Introducing CAD: the Contextual Abuse Dataset

Bertie Vidgen[†], **Dong Nguyen**[†⋆], **Helen Margetts**[†‡], **Patricia Rossini**[¶], **Rebekah Tromble**[†§]

[†] The Alan Turing Institute; [⋆] Utrecht University; [‡] University of Oxford;
[¶] University of Liverpool; [§] George Washington University
bvidgen@turing.ac.uk

Abstract

Online abuse can inflict harm on users and communities, making online spaces unsafe and toxic. Progress in automatically detecting and classifying abusive content is often held back by the lack of high quality and detailed datasets. We introduce a new dataset of primarily English Reddit entries which addresses several limitations of prior work. It (1) contains six conceptually distinct primary categories as well as secondary categories, (2) has labels annotated in the context of the conversation thread, (3) contains rationales and (4) uses an expert-driven group-adjudication process for high quality annotations. We report several baseline models to benchmark the work of future researchers. The annotated dataset, annotation guidelines, models and code are freely available.

1 Introduction

Social media platforms have enabled unprecedented connectivity, communication and interaction for their users. However, they often harbour harmful content such as abuse and hate, inflicting myriad harms on online users (Waseem and Hovy, 2016; Schmidt and Wiegand, 2017a; Fortuna and Nunes, 2018; Vidgen et al., 2019c). Automated techniques for detecting and classifying such content increasingly play an important role in moderating online spaces.

Detecting and classifying online abuse is a complex and nuanced task which, despite many advances in the power and availability of computational tools, has proven remarkably difficult (Vidgen et al., 2019a; Wiegand et al., 2019; Schmidt and Wiegand, 2017b; Waseem et al., 2017). As Jurgens et al. (2019) argued in a recent review, research has ‘struggled to move beyond the most obvious tasks in abuse detection.’ One of the biggest barriers to creating higher performing, more robust, nuanced and generalisable classification systems is the lack of clearly annotated, large and detailed training datasets. However, creating such datasets is time-consuming, complicated and expensive, and requires a mix of both social and computational expertise.

We present a new annotated dataset of ∼25,000 Reddit entries. It contains four innovations that address limitations of previous labelled abuse datasets. First, we present a taxonomy with six conceptually distinct primary categories (Identity-directed, Person-directed, Affiliation-directed, Counter Speech, Non-hateful Slurs and Neutral). We also provide salient subcategories, such as whether personal abuse is directed at a person in the conversation thread or to someone outside it. This taxonomy offers greater coverage and granularity of abuse than previous work. Each entry can be assigned to multiple primary and/or secondary categories (Section 3).

Second, we annotate content in context, by which we mean that each entry is annotated in the context of the conversational thread it is part of. Every annotation has a label for whether contextual information was needed to make the annotation. To our knowledge, this is the first work on online abuse to incorporate a deep level of context. Third, annotators provided rationales. For each entry they highlighted the part of the text which contains the abuse (and the relevant parts for Counter Speech and Non-hateful Slurs). Fourth, we provide high quality annotations by using a team of trained annotators and a time-intensive discussion-based process, facilitated by experts, for adjudicating disagreements (Section 4).

This work addresses the need for granular and nuanced abusive content datasets, advancing efforts to create accurate, robust, and generalisable classification systems. We report several baseline models to benchmark the work of future researchers (Section 5). The annotated dataset, annotation

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 2289–2303
June 6–11, 2021. ©2021 Association for Computational Linguistics

codebook and code have been made available.[1] A full description of the dataset is given in our data statement in the Appendix (Bender and Friedman, 2018).

2 Background

Taxonomies of abuse Taxonomies vary in terms of the scope of abusive behaviours they cover. Some offer categories for abuse against both individuals and groups (Zampieri et al., 2020), others cover only abuse against identities (Davidson et al., 2017; Fortuna and Nunes, 2018; Kiela et al., 2020), against only a single identity, such as misogyny (Anzovino et al., 2018) or Islamophobia (Vidgen and Yasseri, 2019), or only abuse against individuals (Wulczyn et al., 2017). Some research distinguishes between content in different languages or taken from different platforms (Kumar et al., 2018).

Waseem et al. (2017) outline two dimensions for characterising online abuse. First, whether it is directed against individuals or groups. Second, whether it is implicit or explicit (also referred to as 'covert' or 'overt' (Kumar et al., 2018) and 'weak' or 'strong' (Vidgen and Yasseri, 2019)). These two dimensions (strength and target) have been further developed in other studies. Zampieri et al. (2019) use a hierarchical three-level approach to annotation, separating (a) offensive from not-offensive tweets, (b) offensive into targeted and untargeted statements and (c) for targeted statements, identification of what is attacked (group, individual or other). Vidgen et al. (2019a) propose a tripartite distinction, also separating 'concept-directed' abuse from group-directed and person-directed abuse. However, this is problematic as concept-directed content may be better understood as legitimate critique.

Many taxonomies include fine-grained labels for complex subcategories of abuse. Palmer et al. (2020) label implicit varieties of hate, including 'adjectival nominalization', 'distancing' and 'Othering' language. Anzovino et al. (2018) label content for six subtypes of misogyny: discrediting, using stereotypes, objectifying, sexually harassing, threatening violence, dominating or derailing. Sanguinetti et al. (2018) provide annotations for which group is targeted and the linguistic action (i.e., dehumanizing, delegitimizing or aiming to inflict harm). They provide flags for aggressiveness, offensiveness, irony and stereotypes. Sap et al. (2020) provide annotations for 'social frames' (i.e., biases and stereotypes) about groups. They provide labels for (a) offence (yes/no), (b) whether a group is targeted, and (c) whether the abuse is intentional. Wulczyn et al. (2017) identify different interpersonal abuse, including toxicity, aggression and attacks.

Some taxonomies explicitly separate abuse from closely-related but non-abusive forms of online expression. This reflects social scientific insights which emphasize the importance, but also difficulty, of making such distinctions (Rossini, 2019, 2020). Vidgen et al. (2020) distinguish hostility against East Asia from criticism of East Asia, as well as counter speech and discussion of prejudice. Procter et al. (2019) distinguish cyber hate from counter speech, as do Qian et al. (2019) and Mathew et al. (2019), amongst others.

Annotation and Data The quality of annotations for abusive datasets has been widely critiqued, and inter-rater agreement scores are often remarkably low. Wulczyn et al. (2017) report an Alpha of 0.45, Sanguinetti et al. (2018) Kappas from k=0.37 for offence to k=0.54 for hate, Gomez et al. (2020) report Kappa of 0.15 in the "MMH150" dataset of hateful memes, and Fortuna and Nunes (2018) report a Kappa of 0.17 for a text-only task. In a classification study of prejudice against East Asia, Vidgen et al. (2020) find that 27% of classification errors are due to annotation mistakes. Low agreement is partly because abuse is inherently ambiguous and subjective, and individuals can perceive the same content very differently (Salminen et al., 2019, 2018).

Many abusive content datasets use crowdsourced annotations (Zampieri et al., 2019; Fortuna and Nunes, 2018; Davidson et al., 2017). They are cheap and scalable but can be low quality and are often ill-suited to complicated tasks (Sabou et al., 2014). Trained experts with clear guidelines are often preferable for ensuring consistency (Vidgen and Derczynski, 2020). Whether experts- or crowdsourced annotators are used, a diverse pool is needed as annotators encode their biases, backgrounds and assumptions into their annotations (Sap et al., 2019; Waseem et al., 2017). Most datasets use a simple majority vote over annotations to determine the final labels. However, majority agreement does not guarantee that content is correctly labelled, especially for complex edge-

[1] https://github.com/dongpng/cad_naacl2021

cases. One option is to use a method that adjusts annotators' impact based on their quality, such as MACE (Hovy et al., 2013). However, this may not work well on the most ambiguous content. Group-decision making processes present a promising way of improving annotation quality. Breitfeller et al. (2019) use a collaborative multi-stage process to label micro-aggression and Card et al. (2015) use a similar process for labelling news articles. This ensures more oversight from experts and reflection by annotators on the difficult content. It also provides a feedback loop for annotators to learn from mistakes and improve.

A well-established problem with abusive content datasets is that each bit of content is marked up individually, without taking into account any content that came before (Gao and Huang, 2017; Mubarak et al., 2017). This can lead to poor quality annotations when content is ambiguous or unclear without knowing the context. Detection systems which do not account for context are likely to be less applicable in the real-world, where nearly all content appears in a certain context (Seaver, 2015). Pavlopoulos et al. (2020) systematically investigate the role of context in a dataset of Wikipedia comments by providing annotators the 'parent' before showing them the 'child' entry. In one experiment at least 5% of the data was affected. In a study of Twitter conversations Procter et al. (2019) label replies to tweets based on whether they 'agree' or 'disagree' with the original message. Notwithstanding these studies, further work is needed to better understand the role of context and how abuse emerges within threads, as well as the challenges of detecting deeply contextual content.

3 Taxonomy

We present a hierarchical taxonomy of abusive content, which comprises six primary categories and additional secondary categories. It builds on critical social scientific research (Marwick and Miller, 2014; Citron and Norton, 2011; Lenhart et al., 2016), and addresses issues in previous taxonomies, including those provided by Zampieri et al. (2020), Waseem et al. (2017), Founta et al. (2018) and Vidgen et al. (2019a). It offers greater coverage by including three conceptually distinct types of abusive content (Identity-directed abuse, Affiliation-directed abuse and Person-directed abuse) as well as three types of non-abusive content (Neutral, Counter Speech and Non-hateful Slurs). The taxonomic structure is shown in Figure 1. Indicative examples are given in Table 1.

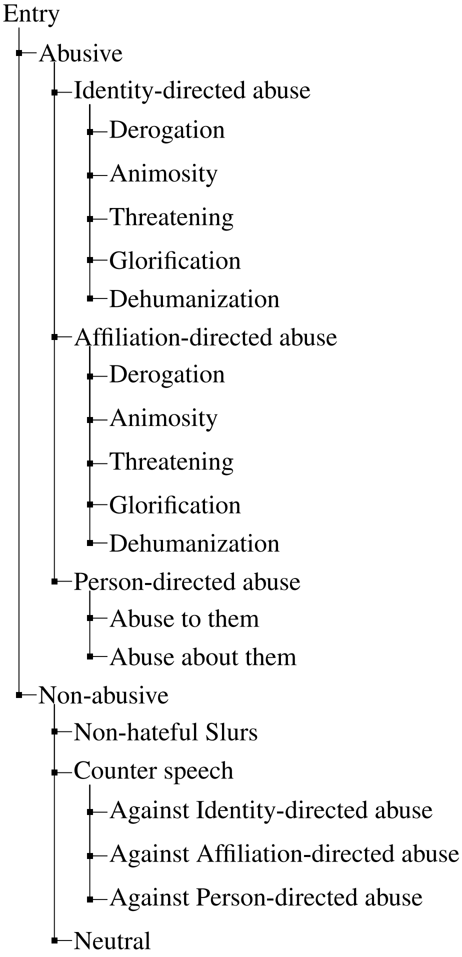

Figure 1: Primary and Secondary categories.

3.1 Identity-directed abuse

Content which contains a negative statement made against an identity. An 'identity' is a social category that relates to a fundamental aspect of individuals' community, socio-demographics, position or self-representation (Jetten et al., 2004). It includes but is not limited to Religion, Race, Ethnicity, Gender, Sexuality, Nationality, Disability/Ableness and Class. The secondary category comprises five subtypes of identity-directed abuse: Derogation, Animosity, Threatening language, Glorification and Dehumanization.

Derogation Language which explicitly attacks, demonizes, demeans or insults a group. Derogation includes representing or describing a group in extremely negative terms and expressing negative emotions about them. Derogation is the basis of most 'explicit' forms of abuse in existing hateful content taxonomies, although it is often referred to

Primary	Secondary	Example
Identity-directed	Derogation	Muslims cant speak English, they're savages
Identity-directed	Animosity	I dont think black people face any discrimination
Identity-directed	Threatening	Gotta kick those immigrants out... now!
Identity-directed	Glorification	Adolf had it right, yknow?
Identity-directed	Dehumanization	Travellers are nothing but trash
Affiliation-directed	Derogation	So sick of these undesirable leftist fools
Person-directed	To them	Fuck off @user
Person-directed	About them	Trump is a massive bellend
Non-hateful Slurs	/	Y'all see me and my n*ggas come in here?
Counter Speech	Identity-directed	Sorry but that's just factually incorrect
Counter Speech	Affiliation-directed	You should be nicer about the mods, they do alot for us
Counter Speech	Person-directed	That's no way to talk to someone!
Neutral	/	I've had a right bloody day of it

Table 1: Indicative examples of the categories.

with different terms. For instance, Davidson et al. (2017) define hate as content that is 'derogatory', Waseem and Hovy (2016) include 'attacks' in their account of hate and Zampieri et al. (2019) 'insults'.

Animosity Language which expresses abuse against a group in an implicit or subtle manner. The lynchpin of this category is that negativity is directed at the group (i.e., there must be some aspect which is discernibly abusive or demeaning about the group in question) but this is not expressed explicitly. Animosity includes undermining the experiences and treatment of groups, ridiculing them, and accusing them of receiving 'special treatment'. Animosity is similar to the 'implicit' category used in other taxonomies (Waseem et al., 2017; Vidgen and Yasseri, 2019; Kumar et al., 2018).

Threatening language Language which either expresses an intent/desire to inflict harm on a group, or expresses support for, encourages or incites such harm. Harm includes physical violence, emotional abuse, social exclusion and harassment. This is one of the most harmful forms of hateful language (Marwick and Miller, 2014; Citron and Norton, 2011) yet usually it is part of an 'explicit' hate category (Zampieri et al., 2019; Wulczyn et al., 2017; Waseem and Hovy, 2016) and few datasets have treated it as a separate category, see Golbeck et al. (2017), Anzovino et al. (2018), and Hammer (2014) for exceptions.

Dehumanization Language which maliciously describes groups as insects, animals and non-humans (e.g., leeches, cockroaches, insects, germs, rats) or makes explicit comparisons. Dehumanization has been linked with real-world violence and is a particularly important focus for computational work (Leader Maynard and Benesch, 2016; Matsuda et al., 1993), yet is often combined into a broader 'explicit' category (Palmer et al., 2020; Vidgen et al., 2020; Kiela et al., 2020) and has been insufficiently studied on its own, apart from Mendelsohn et al. (2020).

Glorification of hateful entities Language which explicitly glorifies, justifies or supports hateful actions, events, organizations, tropes and individuals (which, collectively, we call 'entities'). It includes denying that identity-based atrocities took place (e.g., Genocide). Glorification is one of the least studied forms of hate computationally, likely because it is more ambiguous, particularly when individuals only express *interest* in the entities (de Gibert et al., 2018).

3.2 Affiliation-directed abuse

Content which express negativity against an affiliation. We define 'affiliation' as a (more or less) voluntary association with a collective. Affiliations include but are not limited to: memberships (e.g. Trade unions), party memberships (e.g. Republicans), political affiliations (e.g. Right-wing people) and occupations (e.g. Doctors). The same secondary categories for Identity-directed abuse apply to Affiliation-directed. In some previous taxonomies, affiliations have been mixed in with identities (Founta et al., 2018; Zampieri et al., 2019), although in general they have been excluded as out of scope (e.g. Waseem and Hovy (2016)).

3.3 Person-directed abuse

Content which directs negativity against an identifiable person, who is either part of the conversation thread or is named. Person-directed abuse includes serious character based attacks, such as accusing the person of lying, as well as aggression, insults and menacing language. Person- and Identity- directed forms of abuse are often addressed in separate taxonomies, although in some studies they have been merged into a more general 'toxic' category (Wulczyn et al., 2017; Golbeck et al., 2017). Recent work have addressed both types of content, recognising that they are conceptually different but often co-occur in the real-world and share syntactical and lexical similarities (Zampieri et al., 2019; Mandl et al., 2019). We provide two secondary categories for person-directed abuse: *Abuse at a person* who is part of the conversation thread and *Abuse about a person* who is not part of the conversation thread. The person must be clearly identified, either by their actual name, username or status (e.g. 'the president of America'). To our knowledge, this distinction has not been used previously.

3.4 Counter Speech

Content which challenges, condemns or calls out the abusive language of others. Counter Speech can take several forms, including directly attacking/condemning abusive language in unambiguous terms, challenging the original content and 'calling out' the speaker for being abusive. We use a similar approach to Qian et al. (2019) and Mathew et al. (2019) who also treat counter speech as a relational act that responds to, and challenges, actual abuse.

3.5 Non-hateful Slurs

A slur is a collective noun, or term closely derived from a collective noun, which is pejorative. Slurs include terms which are explicitly insulting (e.g. 'n*gga' or 'kebabi') as well as terms which implicitly express animosity against a group (e.g. 'Rainy' or 'Chad'). A slur by itself does not indicate identity-directed abuse because in many cases slurs are not used in a derogatory way but, rather, to comment on, counter or undermine genuine prejudice (Jeshion, 2013) — or they have been reclaimed by the targeted group, such as use of 'n*gga' by black communities (Davidson et al., 2017; Davidson and Weber, 2019). In this category we mark up only the non-hateful use of slurs. Hateful uses of slurs would fall under Identity-directed abuse.

3.6 Neutral

Content which does not contain any abuse, Non-hateful Slurs or Counter Speech and as such would not fall into any of the other categories.

4 Data

4.1 Data collection

The low prevalence of online abuse in 'the wild' (likely as little as 0.1% in English language social media (Vidgen et al., 2019b)) means that most training datasets have used some form of purposive (or 'directed') sampling to ensure enough entries are in the positive class (Fortuna et al., 2020). However, this can lead to biases in the dataset (Ousidhoum et al., 2020) which, in turn may impact the performance, robustness and fairness of detection systems trained on them (Sap et al., 2019). Notably, the widely-used practice of keyword sampling can introduce topic and author biases, particularly for datasets with a high proportion of implicit abuse (Wiegand et al., 2019).

Accordingly, like Qian et al. (2019), we use community-based sampling, selecting subreddits which are likely to contain higher-than-average levels of abuse and a diverse range of abuse. This should lead to a more realistic dataset where the abusive and non-abusive content share similarities in terms of topic, grammar and style. We identified 117 subreddits likely to contain abusive content, which we we filtered to just 16, removing subreddits which (1) had a clear political ideology, (2) directed abuse against just one group and (3) did not have recent activity. 187,806 conversation threads were collected over 6 months from 1st February 2019 to 31st July 2019, using the PushShift API (Gaffney and Matias, 2018). We then used stratified sampling to reduce this to 1,394 posts and 23,762 comments (25,156 in total) for annotation. See Data Statement in the Appendix for more information on how the initial 117 subreddits were identified.

4.2 Annotation

All posts and comments were annotated. The titles main body of posts were treated separately, resulting in 1,394 post titles, 1,394 post bodies and 23,762 comments being annotated (26,550 entries in total). All entries were assigned to at least one of the six primary categories. Entries could be assigned to several primary categories and/or several secondary categories. The dataset contains 27,494 distinct labels.

All entries were first independently annotated by two annotators. Annotators underwent 4 weeks training and were either native English speakers or fluent. See Data Statement in the Appendix for more information. Annotators worked through entire Reddit conversations, making annotations for each entry with full knowledge of the previous content in the thread. All disagreements were surfaced for adjudication. We used a consensus-based approach in which every disagreement was discussed by the annotators, facilitated by an expert with reference to the annotation codebook. This is a time-consuming process which helps to improve annotators' understanding, and identify areas that guidelines need to be clarified and improved. Once all entries were annotated through group consensus they were then reviewed in one-go by the expert to ensure consistency in how labels were applied. This helped to address any issues that emerged as annotators' experience and the codebook evolved throughout the annotation process. In some cases the labels may appear counter-intuitive. For instance, one entry starts "ITT: Bernie Sanders is imperfect and therefore is a garbage human being." This might appear like an insult, however the remainder of the statement shows that it is intended ironically. Similarly, use of "orange man bad" may appear to be an attack against Donald Trump. However, in reality it is supporting Trump by mocking left-wing people who are opposed to him. Nuances such as these only become apparent after multiple reviews of the dataset and through group-based discussions.

Targets of abuse For Identity-directed, Affiliation-directed and Non-hateful Slurs, annotators inductively identified targets. Initially, 1,500 targets were identified (including spelling variations), which was reduced to 185 through review and cleaning. All important distinctions, including intersectional identities and specific subgroups and outlooks (e.g., 'non-gender dysphoric transgender people') were retained. The identities were then grouped into 8 top level categories. The top level categories for Identity-directed abuse include Gender, Ableness/disability and Race.

Context For every annotation a flag for 'context' was given to capture how the annotation was made. If the primary/secondary label was based on just the entry by itself then 'Current' was selected. If knowledge of the previous content in the conversation thread was required then 'Previous' was selected. Context was primarily relevant in two ways. First, for understanding who a generic pronoun referred to (e.g., 'they'). Second, to express support for another users' abuse (e.g., Person 1 writes 'I want to shoot some X' and person 2 responds 'Go do it!'). If this context is not taken into account then the abuse would be missed. In some cases, only the context of a single previous statement was needed to understand an entry (as with the example just given), whereas in other cases several previous statements were required. For Neutral, no label is given for context. For Non-hateful Slurs, only 'Current' could be selected. Our definition of Counter Speech is relational, and so all Counter Speech require 'Previous' context. For Affiliation-, Identity-, and Person- directed approximately 25-32% of content were labelled with 'Previous' context.

Rationales For all categories other than Neutral, annotators highlighted the part of the entry related to the category. This is important for Reddit data where some comments are very long; the longest entry in our dataset has over 10k characters. As part of the adjudication process, just one rationale was selected for each entry, giving a single 'gold standard'.

Inter annotator agreement Inter annotator agreement for the primary categories was measured using Fleiss' Kappa. It was 'moderate' overall (0.583) (Mchugh, 2012). This compares favourably with other abusive content datasets (Gomez et al., 2020; Fortuna and Nunes, 2018; Wulczyn et al., 2017), especially given that our taxonomy contains six primary categories. Agreement was highest for Non-hateful slurs (0.754). It was consistently 'moderate' for Neutral (0.579), Person (0.513), Affiliation (0.453) and Identity (0.419) but was lower for Counter Speech (0.267). This reflects Counter Speech's low prevalence (meaning annotators were less experienced at identifying it) and the subjective nature of judging whether content counters abuse or is implicitly supportive. One challenge is that if annotators missed a category early on in a thread then they would also miss all subsequent context-dependent entries.

4.3 Prevalence of categories

The prevalence of the primary and secondary categories in the dataset is shown in Table 3. Non-

Category	Fleiss' Kappa
Affiliation directed	0.453
Identity directed	0.498
Person directed	0.513
Counter Speech	0.267
Non-hateful Slurs	0.754
Neutral	0.579
AVERAGE	**0.583**

Table 2: Average Kappa scores for primary categories

hateful Slurs and Neutral entries do not have secondary categories and so only the total is shown. Neutral entries dominate, accounting for 79.8% of the data, followed by Identity-directed abuse which accounts for 9.9%, Affiliation-directed abuse (5.0%), Person-directed abuse (4.0%), Counter Speech (0.8%) and Non-hateful use of slurs (0.5%). Animosity and Derogation are the most frequent secondary categories in Identity-directed and Affiliation-directed abuse, with Threatening language, Dehumanization and Glorification accounting for less than 5% combined. This is unsurprising given the severity of such language. Other training datasets for online abuse generally report similar or slightly higher levels of non-neutral content, e.g., in Gomez et al. (2020) 82% is neutral, in Waseem and Hovy (2016) 68% is not hateful, in both Zampieri et al. (2019) and Vidgen et al. (2020) 67%, and in Founta et al. (2018) 58% is neutral.

5 Experiments

5.1 Experimental setup

Data splits For our classification experiments, we exclude entries that are "[removed]", "[deleted]" or empty because they were either a blank entry associated with a post title or a entry that only contained an image. We also exclude entries written by two prolific bots (SnapshillBot and AutoModerator) and non-English entries, which were identified by langid.py (Lui and Baldwin, 2012) and then manually verified. Entries with an image were included but the image was not used for classification. The dataset used for experiments contains 23,417 entries and is split into a train (13,584; 58%), development (4,526; 19.3%) and test set (5,307; 22.7%). All entries belonging to the same thread are assigned to the same split. A small set of subreddits only occur in either the development or the test set; this allows us to test performance on entries in subreddits that were *not* included in training. Hyperparameters are tuned on the development set.

Primary	Secondary	Number	Percentage
Affiliation-directed	Derogation	629	46.0
	Animosity	676	49.5
	Threatening	30	2.2
	Dehumanization	31	2.3
	Glorification	0	0.0
	Total	1,366	100
Identity-directed	Derogation	1,026	37.8
	Animosity	1,577	58.1
	Threatening	31	1.1
	Dehumanization	29	1.1
	Glorification	49	1.8
	Total	2,712	100
Person-directed	About a person	552	49.6
	To a person	560	50.4
	Total	1,112	100
Counter Speech	Affiliation	53	24.1
	Identity	115	52.3
	Person	52	23.6
	Total	220	100
Non-hateful Slurs	Total	149	100
Neutral	Total	21,935	100
TOTAL		**27,494**	**100**

Table 3: Prevalence of the categories

Classification task We automatically classify the primary categories. Due to the low prevalence of Non-hateful Slurs, these are not used as a separate category in the classification experiments. Instead, for the experiments, we re-assign entries with only a Non-hateful Slur label to Neutral. For entries that have a Non-hateful Slur label and at least one other label, we simply ignore the Non-hateful Slur label[2]. 1.94% of entries in the training set have more than one primary category. When we exclude Neutral entries (because these entries cannot have another category), this increases to 10.5%. The training data has a label cardinality of 1.02 (Tsoumakas and Katakis, 2007). We thus formulate the task as a **multilabel** classification problem. It is challenging given the highly skewed label distributions, the influence of context, and the multilabel setup.

5.2 Methods

We compare several popular baseline models. We only use the texts of entries as input. The context of entries (e.g., previous entries in a thread) are

[2] This is in-line with our taxonomy, whereby entries assigned to Neutral cannot be assigned to any of the other categories.

not taken into account; integrating context could be explored in future work.

Logistic Regression (LR) We use Logistic Regression with L2 regularization, implemented using scikit-learn (Pedregosa et al., 2011). There are different approaches to multilabel classification (Boutell et al., 2004; Tsoumakas and Katakis, 2007). One common approach is the Label Powerset method, where a new label is created for each unique label combination. However, this approach is not suitable for our data; many label combinations only have a few instances. Furthermore, classifiers would not be able to recognise unseen label combinations. We therefore use a binary relevance setup, where binary classifiers are trained for each label separately. Because the class distribution is heavily skewed, classes are weighted inversely proportional to their frequencies in the training data.

BERT and DistilBERT We finetune the BERT base uncased model (Devlin et al., 2019) with commonly used hyperparameters (see the Appendix). Given BERT's sensitivity to random seeds (Dodge et al., 2020), each setting was run with five different random seeds. Our implementation uses the Hugging Face's Transformers library (Wolf et al., 2019). We use a binary cross entropy loss and encode the labels as multi-hot vectors. Classes are weighted by their ratio of negative over positive examples in the training data. We also finetune DistilBERT (Sanh et al., 2019), a lighter version of BERT trained with knowledge distillation.

5.3 Results

Evaluation metrics The precision, recall and F1 score for each primary category are reported in Table 4. In Table 5, we report micro and macro average F1 scores. Because of the highly skewed class distribution, we favor macro F1 scores. We also report the exact match accuracy (the fraction of entries for which the full set of labels matches).

Classifier comparison BERT performs best and achieves a substantial performance improvement over Logistic Regression (Macro F1 of 0.455 vs. 0.343). The performance of DistilBERT is slightly lower, but very close to BERT's performance. With both BERT and DistilBERT there is still much room for improvement on most categories. Note that a majority class classifier which labels everything as Neutral would achieve a high accuracy (0.818) but a low F1 macro score (0.180). There were no clear performance differences between entries from subreddits that were or were not included in the training data.

Primary categories Performance differs substantially between the different categories (Table 4). All classifiers attain high F1 scores on Neutral entries (LR: 0.859, BERT: 0.902); this is expected as the class distribution is highly skewed towards Neutral. Performance is lowest on Counter Speech (LR: 0.042, BERT: 0.091), possibly due to a combination of factors. First, this category has the lowest number of training instances. Second, inter-annotator agreement was lowest on Counter Speech. And third, all Counter Speech annotations are based on previous content in the thread.

Error analysis Qualitative analysis shows that the BERT model often misclassifies neutral content which mention identities (e.g., non-misogynistic discussions of women) or contains profanities and aggressive language. It tends to classify Affiliation- and Identity-directed abuse which uses less aggressive language and contains fewer abusive keywords as Neutral. Surprisingly, many of the Person-directed entries which are misclassified as Neutral contain clear signals of abuse, such as profanities and overt aggression. No discernible pattern was observed with Counter Speech which was misclassified as a different category. For this category, the low performance may be attributed mostly to its low frequency in the training data.

Context Our benchmark models do not explicitly take into account context for prediction. As expected, all our models are worse at predicting the primary categories of entries where context was required for the annotation. For example, with logistic regression, the recall for Identity-directed abuse is 21.1% for entries where the annotation was based on previous content compared with 46.3% for entries where the annotation is based only on the current content. Similarly, with BERT the recall for Identity-directed abuse increases from 25.3% ('Previous') to 60.1% ('Current').

Secondary categories We compare recall between the secondary categories. For Person-directed abuse, the recall with LR for abuse targeting a person who is not in the thread is substantially lower than for entries that are directed to a person in the thread with (25.2% vs. 35.6%). For BERT and DistilBERT, the performance difference

	LR			DistilBERT			BERT		
	P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
Neutral	0.872	0.845	0.859	0.880	0.917	0.898	0.883	0.922	0.902
Identity-directed	0.281	0.398	0.330	0.414	0.473	0.441	0.411	0.510	0.455
Affiliation-directed	0.229	0.395	0.290	0.368	0.450	0.405	0.368	0.481	0.416
Person-directed	0.145	0.304	0.196	0.359	0.404	0.380	0.356	0.488	0.411
Counter Speech	0.032	0.061	0.042	0.083	0.073	0.076	0.107	0.088	0.091

Table 4: Scores per category on the test set. For DistilBERT and BERT these are the means over 5 runs.

	Accuracy	**F1 (macro)**	**F1 (micro)**
LR	0.634	0.343	0.711
DistilBERT	0.769 (0.005)	0.440 (0.007)	0.797 (0.005)
BERT	0.762 (0.005)	0.455 (0.006)	0.799 (0.005)

Table 5: Results on the test set. For BERT and DistilBERT the standard deviations are also reported.

between these two secondary categories is small (e.g., BERT: 48.6% vs. 49.0%). Furthermore, for Identity-directed abuse the recall for animosity (LR: 36.2%, BERT: 45.3%) tends to be lower than the recall for derogation (LR: 49.0%, BERT: 65.9%), which is expected as animosity expresses abuse in an implicit manner and is often more nuanced. The larger difference for BERT vs. logistic regression shows the promise of more advanced models in distinguishing subcategories. For Affiliation-directed abuse, the differences are smaller. Here, the recall for animosity is (unexpectedly) slightly higher (LR: 43.3%, BERT: 49.5%) than for derogation (LR: 36.1%, BERT: 48.0%).

Label dependence The multilabel setup of this classification task makes this a challenging problem. All models tend to assign too many labels. For example, DistilBERT predicts only too few labels in 1.17% of the cases, the remainder predicting the right number (91.88%) or too many (6.96%). For BERT, the difference is even higher (1.06% too few; 9.21% too many labels).

Dependencies between labels are sometimes violated. In our taxonomy, entries which are Neutral cannot have another label, but our models violate this constraint in many cases. With DistilBERT 3.8% of the entries are classified as Neutral and at least one other class, this is even more so for BERT (5.4%) and (LR: 10.7%). Future work could therefore explore modeling relationships between labels.

6 Discussion and Conclusion

We have presented a detailed dataset for training abusive content classification systems. It incorporates relevant social scientific concepts, providing a more nuanced and robust way of characterising — and therefore detecting — abuse. We have also presented benchmark experiments, which show much room for improvement.

Our analyses indicate numerous areas to explore further, including creating systems which explicitly model the conversation threads to account for context. Predictive methods could be applied to understand and forecast when a conversation is turning toxic, potentially enabling real-time moderation interventions. More powerful models could also be applied to better distinguish the primary categories and to begin classification of the secondary categories. This could be achieved by also using the images to classify the content, which we did not do. Finally, we would also expect the rationales to be of considerable use in future experiments, both for classification and to understand the annotation process.

The current work has several limitations. First, the class distribution is heavily skewed towards the Neutral class and some abusive categories have low frequencies. This better reflects real-world prevalence of abuse but can limit the signals available for classification. Second, inter-annotator agreement was in-line with other research in this domain but could still be improved further, especially with ‘edge case’ content.

7 Ethical considerations

We follow the ACM’s Code of Ethics and Professional conduct[3], as well as academic guidelines for ethically researching activity on social media (Townsend and Wallace, 2017; Williams, 2019). Online abuse poses substantial risk of harm to online users and their communities, and there is a

[3] https://www.acm.org/code-of-ethics

strong social justification for conducting this work.

Dataset collection We used the Pushshift API to collect data from Reddit[4], which we accessed through the data dumps on Google's BigQuery using R[5]. The Pushshift API is a wrapper which allows large quantities of Reddit data to be accessed reliably and easily (Baumgartner et al., 2020; Gaffney and Matias, 2018). Our collection is consistent with Reddit's Terms of Service.

Ethical approval This project was given ethical approval on 18th March 2019, before any research had started, by The Alan Turing Institute (submission C1903-053). Reddit can be considered a public space in that discussion are open and posts are aimed at a large audience. In this way, it differs from a one-to-one or 'private' messaging service. When users sign up to Reddit, they consent to have their data made available to third parties, such as academics. Many users are aware of this and choose to use non-identifiable pseudonyms. Existing ethical guidance indicates that in this situation explicit consent is not required from each user (which is often infeasible), provided that harm to users is minimized at all times (Williams, 2019) and no 'real' quotes are attributed to them in the paper. We follow this guidance and do not provide any direct quotes. The examples given in Table 1 are synthetic. We also minimized how many entries we collected from each user so that each one comprises only a small part of the total dataset. At no point did any of the research team contact any Reddit users, minimizing the risk that any harm could be caused to them. Further, we decided not to review any profile information about the users, substantially minimizing the risk that any personally identifiable information is included in the dataset.

Treatment of annotators We used trained annotators that were carefully recruited through the host institution (in line with their HR procedures). Crowdsourced workers were not used. Annotators were carefully supervised with weekly meetings and regular one-to-one discussions. We followed the guidelines provided by Vidgen et al. (2019a) for ensuring annotator welfare during the work. We provided annotators with access to support services throughout the project, including counselling support, although they were not used. Annotators were paid substantially above the living wage. They were paid holiday and all meetings and training time was paid.

Research team wellbeing To protect the wellbeing of the research team, we had regular catchup discussions, and made sure that the lead researchers were not exposed excessively to harmful content. We did not post anything about the project whilst it was conducted (to minimize the risk of attracting the attention of malicious online actors) and did not engage with any of the Reddit users or communities being studied.

Dataset information and quality We provide a Data Statement in the Appendix, following Bender and Friedman (2018), with full information about the dataset.

Baseline models We present baseline classification models in the paper. We have carefully considered how these models could be deployed and believe that this is highly unlikely given their performance. There is a risk of bias in any dataset, and associated models, and we have sought to provide as much information as possible in our dataset, documentation and other artefacts to enable future researchers to investigate these issues. We do not use demographic or identity characteristics in the formation of the dataset. We also do not provide information about individual annotators, only giving the overall profile of the annotation team. The computational time/power involved in creating the baselines was minimal.

Acknowledgments

This work was supported by Wave 1 of The UKRI Strategic Priorities Fund under the EPSRC Grant EP/T001569/1, particularly the "Criminal Justice System" theme within that grant, and The Alan Turing Institute.

The authors would like to thank Zeerak Waseem and Paul Röttger for their helpful advice, as well as feedback from attendees at the Hertie School's 2020 workshop on online hate regulation.

References

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on Twitter. In *NLDB*, pages 57–64.

[4] https://pushshift.io/api-parameters/

[5] https://pushshift.io/using-bigquery-with-reddit-data/

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. In *Proceedings of the 14th International Conference on Web and Social Media*, pages 830–839.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Matthew R. Boutell, Jiebo Luo, Xipeng Shen, and Christopher M. Brown. 2004. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757 – 1771.

Luke M. Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedingsof the 9th EMNLP-IJCNLP*, pages 1664–1674.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 438–444.

Danielle Citron and Helen Norton. 2011. Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age. *Boston University Law Review*, 91(16):1435–2131.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *Proceedings of the 11th ICWSM*, pages 1–4.

Thomas Davidson and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *3rd Workshop on Abusive Language Online (ACL)*, pages 1–11.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the Second Workshop on Abusive Language Online (ACL)*, pages 11–20.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jesse Dodge, Gabriel Ilharco, Roy Schwartz, Ali Farhadi, Hannaneh Hajishirzi, and Noah Smith. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Comput. Surv.*, 51(4).

Paula Fortuna, Juan Soler-company, and Leo Wanner. 2020. Toxic, Hateful, Offensive or Abusive? What Are We Really Classifying? An Empirical Analysis of Hate Speech Datasets. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC)*, pages 6786–6794.

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 *ICWSM*, pages 1–11.

Devin Gaffney and J Nathan Matias. 2018. Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *Plos ONE*, 13(7):1–13.

Lei Gao and Ruihong Huang. 2017. Detecting Online Hate Speech Using Context Aware Models. In *Proceedings of Recent Advances in Natural Langugae Processing*, pages 260–266.

Jennifer Golbeck, Alicia A. Geller, Jayanee Thanki, Shalmali Naik, Kelly M. Hoffman, Derek Michael Wu, Alexandra Berlinger, Priyanka Vengataraman, Shivika Khare, Zahra Ashktorab, Marianna J. Martindale, Gaurav Shahane, Paul Cheakalos, Jenny Hottle, Siddharth Bhagwan, Raja Rajan Gunasekaran, Rajesh Kumar Gnanasekaran, Rashad O. Banjo, Piyush Ramachandran, Lisa Rogers, Kristine M. Rogers, Quint Gergory, Heather L. Nixon, Meghna Sardana Sarin, Zijian Wan, Cody Buntain, Ryan Lau, and Vichita Jienjitlert. 2017. A Large Labeled Corpus for Online Harassment Research. In *Proceedings of the ACM Conference on Web Science*, pages 229–233.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring Hate Speech Detection in Multimodal Publications. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 1–8.

Hugo Lewi Hammer. 2014. Detecting threats of violence in online discussions using bigrams of important words. In *Proceedings - 2014 IEEE Joint Intelligence and Security Informatics Conference, JISIC 2014*, page 319.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning Whom to Trust with MACE. In *Proceedings of NAACL-HLT*, pages 1120–1130.

Robin Jeshion. 2013. Slurs and Stereotypes. *Analytic Philosophy*, 54(3):314–329.

Jolanda Jetten, Russell Spears, and Tom Postmes. 2004. Intergroup Distinctiveness and Differentiation: A Meta-Analytic Integration. *Journal of personality and social psychology*, 86(6):862–879.

David Jurgens, Eshwar Chandrasekharan, and Libby Hemphill. 2019. A Just and Comprehensive Strategy for Using NLP to Address Online Abuse. In *Proceedings of the 57th Annual Meeting of the ACL*, pages 3658–3666.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. *arXiv:2005.04790v2*, pages 1–17.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking Aggression Identification in Social Media. In *Proceedings ofthe First Workshop on Trolling, Aggression and Cyberbullying,*, 1, pages 1–11.

Jonathan Leader Maynard and Susan Benesch. 2016. Dangerous Speech and Dangerous Ideology: An Integrated Model for Monitoring and Prevention. *Genocide Studies and Prevention*, 9(3):70–95.

Amanda Lenhart, Michelle Ybarra, Kathryn Zickuhr, and Myeshia Price-Feeney. 2016. *Online harrassment, digital abuse, and cyberstalking in America*. Data & Society, New York.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the HASOC track at FIRE 2019: Hate speech and offensive content identification in Indo-European languages. In *FIRE '19: Proceedings of the 11th Forum for Information Retrieval Evaluation*, pages 14–17.

Alice E Marwick and Ross Miller. 2014. *Online Harassment, Defamation, and Hateful Speech: A Primer of the Legal Landscape Recommended Citation*. Center on Law and Information Policy at Fordham Law School, New York.

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

Mary Matsuda, Charles Lawrence, Richard Delgado, and Kimberlé Crenshaw. 1993. *Words that Wound: Critical race theory, assaultive speech and the First Amendment*. Routledge, New York.

Mary L Mchugh. 2012. Interrater reliability: the Kappa statistic. *Biochemia Medica*, 22(3):276–282.

Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. 2020. A Framework for the Computational Linguistic Analysis of Dehumanization. *Frontiers in Artificial Intelligence*, 3(August):1–24.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive Language Detection on Arabic Social Media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56.

Nedjma Ousidhoum, Yangqiu Song, and Dit-yan Yeung. 2020. Comparative Evaluation of Label-Agnostic Selection Bias in Multilingual Hate Speech Datasets. In *Proceedings of EMNLP*. Association for Computational Linguistics.

Alexis Palmer, Christine Carr, Melissa Robinson, and Jordan Sanders. 2020. COLD: Annotation scheme and evaluation data set for complex offensive language in English. *The Journal for Language Technology and Computational Linguistics*, 34(1):1–28.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity Detection: Does Context Really Matter? *arXiv:2006.00998*, pages 1–10.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rob Procter, Helena Webb, Pete Burnap, William Housley, Adam Edwards, Matthew Williams, and Marina Jirotka. 2019. A Study of Cyber Hate on Twitter with Implications for Social Media Governance Strategies. In *Proceedings of Truth and Trust 2020*.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang 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*, pages 4757–4766.

Patrícia Rossini. 2019. Toxic for whom? Examining the targets of uncivil and intolerant discourse in online political talk. In *Voices: Exploring the shifting contours of communication*, pages 221–242. Peter Lang, New York.

Patricia Rossini. 2020. Beyond Incivility: Understanding Patterns of Uncivil and Intolerant Discourse in Online Political Talk. *Communication Research*, 47(8).

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus Annotation through Crowdsourcing: Towards Best Practice Guidelines. In *LREC 2018, Eleventh International Conference on Language Resources and Evaluation*, pages 859–866.

Joni Salminen, Hind Almerekhi, Ahmed Mohamed Kamel, Soon Gyo Jung, and Bernard J. Jansen. 2019. Online hate ratings vary by extremes: A statistical analysis. In *CHIIR 2019 - Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 213–217.

Joni Salminen, Fabio Veronesi, Hind Almerekhi, Soongyo Jung, and Bernard J. Jansen. 2018. Online Hate Interpretation Varies by Country, But More by Individual. In *Proceedings of SNAMS*, pages 1–7.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter Corpus of Hate Speech against Immigrants. In *LREC*, pages 2798–2805.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. In *5th Workshop on Energy Efficient Machine Learning and Cognitive Computing*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, Noah A Smith, and Paul G Allen. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the ACL*, pages 1668–1678.

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

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

Anna Schmidt and Michael Wiegand. 2017b. A Survey on Hate Speech Detection using Natural Language Processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10. Association for Computational Linguistics.

Nick Seaver. 2015. The nice thing about context is that everyone has it. *Media, Culture and Society*, 37(7):1101–1109.

Leanne Townsend and Prof Claire Wallace. 2017. *Social Media Research : A Guide to Ethics*. The University of Aberdeen, Aberdeen.

Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13.

Bertie Vidgen and Leon Derczynski. 2020. Directions in Abusive Language Training Data: Garbage In, Garbage Out. *arXiv:2004.01670*, pages 1–26.

Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020. Detecting East Asian prejudice on social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172, Online. Association for Computational Linguistics.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019a. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online (ACL)*, pages 80–93.

Bertie Vidgen, Helen Margetts, and Alex Harris. 2019b. *How much online abuse is there? A systematic review of evidence for the UK*. The Alan Turing Institute, London.

Bertie Vidgen, Rebekah Tromble, Alex Harris, Scott Hale, Dong Nguyen, and Helen Margetts. 2019c. Challenges and frontiers in abusive content detection. In *3rd Workshop on Abusive Language Online*.

Bertie Vidgen and Taha Yasseri. 2019. Detecting weak and strong Islamophobic hate speech on social media. *Journal of Information Technology & Politics*, 17(1):66–78.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding Abuse: A Typology of Abusive Language Detection Subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *NAACL-HLT*, pages 88–93.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *NAACL-HLT*, pages 602–608, Minneapolis. ACL.

Matthew Williams. 2019. *Hatred behind the scenes: a report on the rise of online hate speech*. Mishcon de Reya, London.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the International World Wide Web Conference*, pages 1391–1399.

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 NAACL HLT 2019*, volume 1, pages 1415–1420.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

A Data Statement

A. CURATION RATIONALE In order to study the classification of abusive content online, and the role of context, we collected data from Reddit, using the PushShift API. We sampled data by collecting content from specific cubreddits, rather than by using keywords.

To identify subreddits which contain substantial levels of abuse (and as such are suitable for inclusion in this dataset), we reviewed four sources, which returned 117 unique subreddits. Note that many prolific abusive subreddits had been banned prior to our study, such as r/AgainstGayMarriage, r/BlackPeopleHate, r/coontown and r/physical_removal.

1. Curated list of 21 hateful subreddits, hosted on r/AgainstHateSubreddits.[6]
2. Curated list of 30 ‘reactionary/fascist’ subreddits, hosted on r/GenderCynicial.[7]
3. List of 13 hateful subreddits from a news article in Vice.[8]
4. Every subreddit mentioned on r/AgainstHateSubreddits in the 12 months from January to December 2018 (93 in total).

Initially, the number of conversation threads that we sampled from each subreddit was proportionate to the total number of conversation threads they hosted during the period (such that more active subreddits featured more heavily in the dataset). Due to large differences in how active users were across subreddits, we then boosted the number of conversation threads from subreddits with fewer posts. We also stratified by date and limited the number of entries from each author to maximize diversity. The number of conversation threads from each subreddit is shown in Table 6.

Subreddit	# threads
r/Drama	148
r/conspiracy	145
r/bakchodi	123
r/TrueOffMyChest	123
r/subredditcancer	106
r/ImGoingToHellForThis	104
r/ShitPoliticsSays	100
r/TumblrInAction	98
r/SubredditDrama	96
r/4chan	77
r/WatchRedditDie	77
r/CCJ2	44
r/Negareddit	43
r/HateCrimeHoaxes	41
r/smuggies	38
r/imgoingtohellforthis2	31
TOTAL	**1,394**

Table 6: Number of conversation threads from each subreddit.

B. LANGUAGE VARIETY Most of the content was in English. Annotators were instructed to mark up non-English where possible, in order to retain the conversation structure in the final datastet. We checked language by first applying langid.py[9] and then manually checking all of the entries which were flagged as non-English. 1,407 entries were flagged as non-English, of which 353 were identified as genuinely non-English by human review. We excluded non-English entries from the dataset for experiments.

C. SPEAKER DEMOGRAPHICS The creators of the Reddit entries (‘speakers’) were not directly approached and thus we could not ask them for demographic information. Further, Reddit users provide relatively little information about themselves in their profiles, and we opted not to collect the little information that was available due to ethical concerns.

Outside of two moderator bot accounts (‘Snapshillbot’ and ‘Automoderator’), the most common users appeared over 60 times in the dataset. 11,122

[6] https://www.reddit.com/r/AgainstHateSubreddits

[7] https://www.reddit.com/r/GenderCynical/comments/bdrtvq/a_comprehensive_list_of_transphobic_subreddits/

[8] https://www.vice.com/en_uk/article/8xxymb/here-are-reddits-whiniest-most/low-key-toxic-subreddits

[9] https://github.com/saffsd/langid.py

users appear in total.

D. ANNOTATOR DEMOGRAPHICS The dataset includes annotations from 12 trained analysts. They were recruited through a competitive process. They underwent 4 weeks of training, including numerous one-to-one sessions. Work was conducted over 12 weeks, with each annotator working between 10 and 20 hours each week. Of the 12 annotators who contributed to the final dataset, 11 consented to provide information about their demographics. Age: 7 annotators were 18–29, 3 were 30–39 and 1 was 40–49. Gender: 4 were female and 7 were male. Ethnicity: 8 were white, 1 Latino, 1 of Middle Eastern ethnic origin and 1 was mixed. National identity: 7 were British, 1 American, 1 Ecuadorean, 1 Jordanian and 1 Polish. Social media use: 9 used social media more than once per day, and 2 use it once per day. Exposure to online abuse: All annotators had witnessed online abuse in the previous year, with 10 stating they had witnessed it more than 3 times and 1 stating they had witness it 2–3 times. Disagreements were adjudicated through group discussion with an expert in abusive online content. They are a post-doctoral researcher with extensive experience.

E. SPEECH SITUATION All Reddit comments and posts were made between 1st February 2019 and 31st July 2019. The intended audience is unknown but was most likely the other members of the subreddit.

F. TEXT CHARACTERISTICS The composition of the dataset, including the distribution of the Primary and Secondary categories, is described in the paper.

B Data and Model fitting

B.1 Data

The application of the Context flag for the primary categories is shown in Table 7.

Primary category	**Context**	**Number**	**Percent**
Affiliation	Current	1,033	75.6%
	Previous	333	24.4%
Identity	Current	2,046	75.4%
	Previous	666	24.6%
Person	Current	765	68.8%
	Previous	347	31.2%
Counter Speech	Previous	220	100.0%
Non-hateful Slurs	Current	149	100.0%
Neutral	None	21,935	100.0%

Table 7: Use of context flag, split by primary category.

B.2 Model fitting

Training details We fine-tune for 3 epochs, a training batch size of 16, and 100 warm up steps. We experiment with learning rates of {2e-5,3e-5, 4e-5, 5e-5} and a weight decay of {0, 0.01, 0.03}. Experiments were run with a single NVIDIA Quadro RTX 6000 GPU. Finetuning one BERT model took around 24 minutes. Finetuning one DistilBERT model took around 12 minutes.