f i ® s t m ¤ ñ d @ ¥
PEER-REVIEWED JOURNAL ON THE INTERNET

Toxicity detection sensitive to conversational context

by Alexandros Xenos, John Pavlopoulos, Ion Androutsopoulos, Lucas Dixon, Jeffrey Sorensen, and Léo Laugier

Abstract

User posts whose perceived toxicity depends on conversational context are rare in current toxicity detection datasets. Hence, toxicity detectors trained on existing datasets will also tend to disregard context, making the detection of context-sensitive toxicity harder when it does occur. We construct and publicly release a dataset of 10,000 posts with two kinds of toxicity labels: (i) annotators considered each post with the previous one as context; and (ii) annotators had no additional context. Based on this, we introduce a new task, context sensitivity estimation, which aims to identify posts whose perceived toxicity changes if the context (previous post) is also considered. We then evaluate machine learning systems on this task, showing that classifiers of practical quality can be developed, and we show that data augmentation with knowledge distillation can improve performance further. Such systems could be used to enhance toxicity detection datasets with more context-dependent posts, or to suggest when moderators should consider parent posts, which often may be unnecessary and may otherwise introduce significant additional costs.

Contents

1. Introduction

Online discussion fora can suffer from hateful, insulting, identity-attacking, profane or otherwise abusive posts. Such posts are called toxic (Borkan, *et al.*, 2019), offensive (Sap, *et al.*, 2020) or abusive (Thylstrup and Waseem, 2020), and systems detecting them (Waseem and Hovy, 2016; Pavlopoulos, *et al.*, 2017b; Badjatiya, *et al.*, 2017) are called toxicity (or offensive or abusive language) detection systems. A common characteristic that these systems often have, besides aiming to promote healthy discussions online (Zhang, *et al.*, 2018), is that they disregard much of the conversational context, making the detection of context-sensitive toxicity more difficult.

We consider context to be any information relevant to help understand the meaning and intention of a post; when context is missing, there is more ambiguity in the interpretation of a post. Context is very diverse in nature, because human communication is diverse; people may inhabit any number of roles in their relationships with others. A person on stage in a play about a murder might engage in dialog that would be illegal in other contexts. Far from being inappropriate, people may pay to see this behaviour and applaud it. It is not always clear what social norms, jurisdictional mandates and enforcement regimes apply. A comedian may deliberately engage in provocative language to entertain, inspire or critique society, but a disruptive heckler might still be removed by the venue's bouncers.

In online discourse, context typically includes personal information about the authors (Pavlopoulos, *et al.*, 2017c), interlocutors, metadata about the conversation or subtle references to specific subjects and topics. Within the scope of this work we presume some socially constructed context in the form of common notions about what constitutes appropriate communicative intent in a social media setting — at least enough that persons tasked with evaluating the communicative intent can consensually make

judgements from surface text alone. This concept, of common socially agreed norms, is obviously not a black and white concept, and while certainly worthy of deeper analysis, it is not the focus of our study here. Instead we follow the common practice of having these background social norms manifested through crowd-sourcing platforms and measured at a very abstract level by inter-annotator agreement metrics. Given this approach, we focus on the past conversational context, specifically, the previous post in a discussion. For instance, a post "Keep the hell out" is likely to be considered as toxic by a moderator who has not seen that the previous post was "What was the title of that 'hell out' movie?".

Although toxicity datasets that include conversational context have recently started to appear, in previous work we showed that context-sensitive posts seem to be rare and this makes it hard for models to learn to detect context-dependent toxicity (Pavlopoulos, *et al.*, 2020). To study this problem, we constructed and publicly released a context-aware dataset of 10,000 posts, each of which was annotated by raters who (i) considered the previous (parent) post as context, apart from the post being annotated (the target post), and by raters who (ii) were given only the target post, without any other previous conversational context [1].

We limit the conversational context to the previous post of the thread, as in our previous work (Pavlopoulos, *et al.*, 2020), as a first step towards studying broader context-dependent toxicity. While this is still a very limited form of context, our previous work also highlighted the basic challenges of studying context: it is expensive and time-consuming to consider on crowd-sourcing platforms, because of the challenges of ensuring that a person has in fact considered the context. The more context, and more subtle kinds of context, one attempts to include in a study, the harder it is to ensure annotators have accounted for it. Moreover context-sensitive toxicity in posts is also rare; and thus it is reasonable to wonder if the impact of more indirect and subtle kinds of context is rarer still.

We then use our new dataset to study the nature of context sensitivity in toxicity detection, and we introduce a new task, context sensitivity estimation, which aims to identify posts whose perceived toxicity changes if the context (previous post) is also considered. Using the dataset, we also show that systems of practical quality can be developed for the new task. Such systems could be used to enhance toxicity detection datasets with more context-dependent posts, or to suggest when moderators should consider parent posts, which may not always be necessary and may otherwise increase costs. Finally, we show that data augmentation with teacher-student knowledge distillation can further improve the performance of context sensitivity estimators.

2. Creating the CCC dataset

To build the dataset for this work, we used the publicly available Civil Comments (CC) dataset (Borkan, *et al.*, 2019). Since we extended a previous dataset, we used the same definition for 'toxicity' that was used for the initial dataset, which is: "*toxicity is defined as anything that is rude, disrespectful, or unreasonable that would make someone want to leave a conversation*" (Borkan, *et al.*, 2019). CC was originally annotated by 10 annotators per post, but the parent post (the previous post in the thread) was not shown to annotators. We randomly sampled 10,000 CC posts and gave both the target and the parent post to annotators. We call this new dataset Civil Comments in Context (CCC). Each CCC post was rated either as NON-TOXIC, UNSURE, TOXIC, or VERY TOXIC, as in the original CC dataset [2]. We unified the latter two labels in both CC and CCC to simplify the problem. To obtain the new in-context labels of CCC, we used the Appen [3] platform and five high accuracy annotators per post (annotators from zone 3, allowing adult and warned for explicit content), selected from seven English speaking countries: U.K., Ireland, U.S., Canada, New Zealand, South Africa and Australia [4].

2.1. Inter-annotator agreement

The free-marginal kappa, a measure of inter-annotator agreement (Randolph, 2010), of the CCC annotations was 83.93 percent, while the average (mean pairwise) percentage agreement was 92 percent. In only 71 posts (0.07 percent) an annotator said UNSURE, meaning annotators were confident in their decisions most of the time. We excluded these 71 posts from our study, as there were too few to generalise about.

2.2. Data analysis

The average length of target posts in CCC was only slightly lower than that of parent posts. Figure 1 shows this when counting length in characters, but the same held when counting words (56.5 vs. 68.8 words on average).

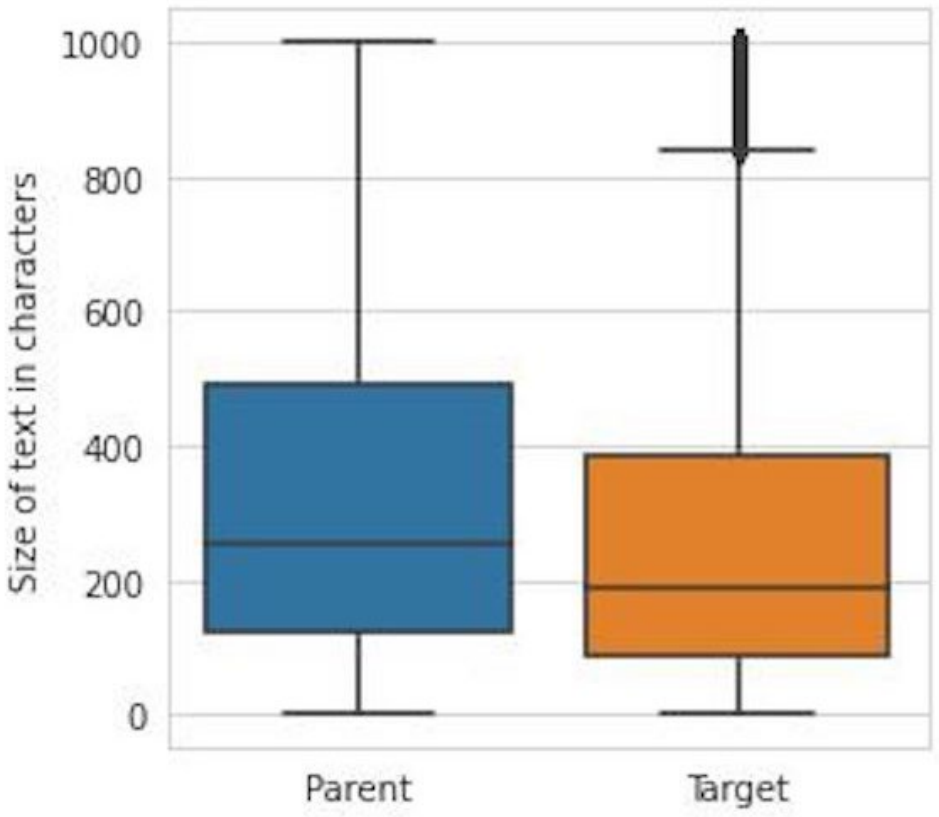

Figure 1: Length of parent/target posts in characters.

To obtain a single toxicity score per post, we calculated the percentage of the annotators who found the post to be insulting, profane, identity-attack, hateful or toxic in another way; all toxicity sub-types provided by the annotators were collapsed to a single toxicity label. This was similar to arrangements in Wulczyn, *et al.* (2017), who also found that training using the empirical distribution (over annotators) of the toxic labels (*i.e.*, a probabilistic gold label per post) led to better toxicity detection performance, compared to using labels reflecting the majority opinion of the raters (a binary gold label per post). See also Fornaciari, *et al.* (2021). With probabilistic gold labels, it was also possible to use regression evaluation measures. Mean Absolute Error (MAE), in particular, which demonstrated the average (over test posts) absolute difference between predicted and gold toxicity scores, was easily interpretable.

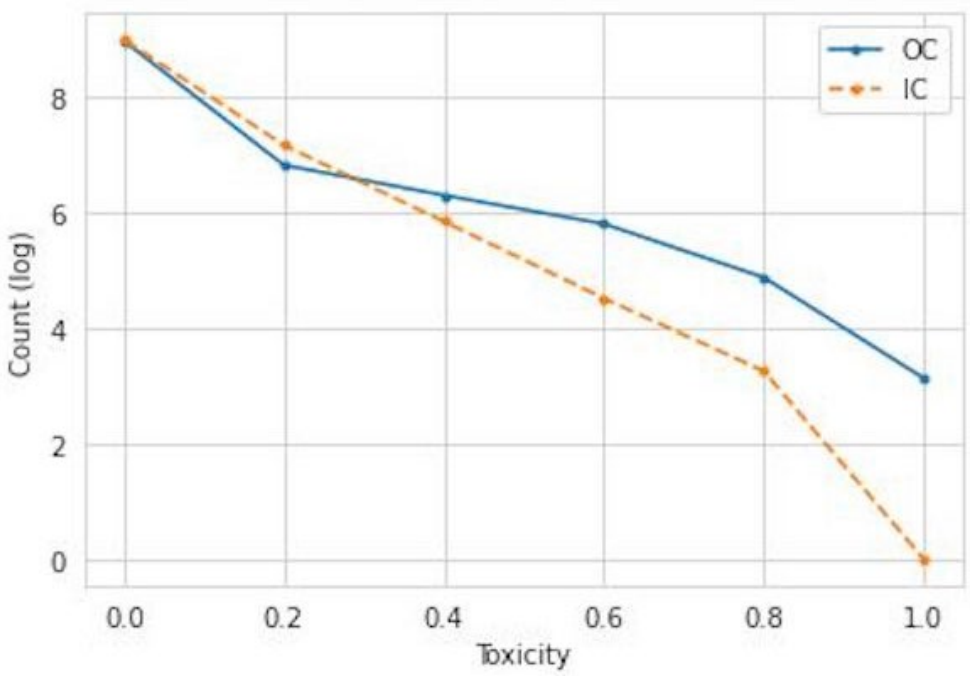

Figure 2: Histogram (converted to curve) showing the distribution of toxicity scores according to annotators who were (IC) or were not (OC) given the parent posts.

Combined with the original (out of context) annotations of the 10,000 posts from CC, the new dataset (CCC) contained 10,000 posts for which both in-context (IC) and out-of-context (OC) labels were available. Figure 2 shows the number of posts (Y axis) per ground truth toxicity score (X axis). Orange (dashed) represents the ground truth obtained by annotators who were provided with the parent post when rating (IC), while blue (solid) is for annotators who rated the post without context (OC). The vast majority of the posts were unanimously labelled as NON-TOXIC (0.0 toxicity), both by the OC and the IC coders. However, IC coders found fewer posts with toxicity greater than 0.2, compared to OC coders. This was consistent with the findings of our previous work (Pavlopoulos, *et al.*, 2020), where we observed that when the parent post was provided, the majority of the annotators perceived fewer posts as toxic, compared to showing no context to annotators. To study this further, in this work we compared the two annotation scores (IC, OC) per post, as discussed below.

2.3. Context-sensitive posts

For each post p, we define $s^{ic}(p)$ to be the toxicity score (fraction of coders who perceived the post as toxic) derived from IC coders, and $s^{oc}(p)$ to be the toxicity derived from OC coders. Then, their difference is $\delta(p) = s^{oc}(p) - s^{ic}(p)$. A positive δ means that raters who were not given the parent post perceived the target post as toxic more often than raters who were given the parent post. A negative δ means the opposite. Figure 3 shows that δ is most often 0, but when the toxicity score changes, δ is most often positive, *i.e.*, showing the context to the annotators reduces the perceived toxicity in most cases when perceived toxicity changes. In 66.1 percent of the posts the toxicity score remained unchanged; out of the remaining 33.9 percent, in 9.6 percent it increased (960 posts) and in 24.3 percent it decreased (2,408) when context was provided. If we binarize the ground truth (both for IC and OC) we reach a similar trend, but with the toxicity of more posts remaining unchanged (*i.e.*, 94.7 percent).

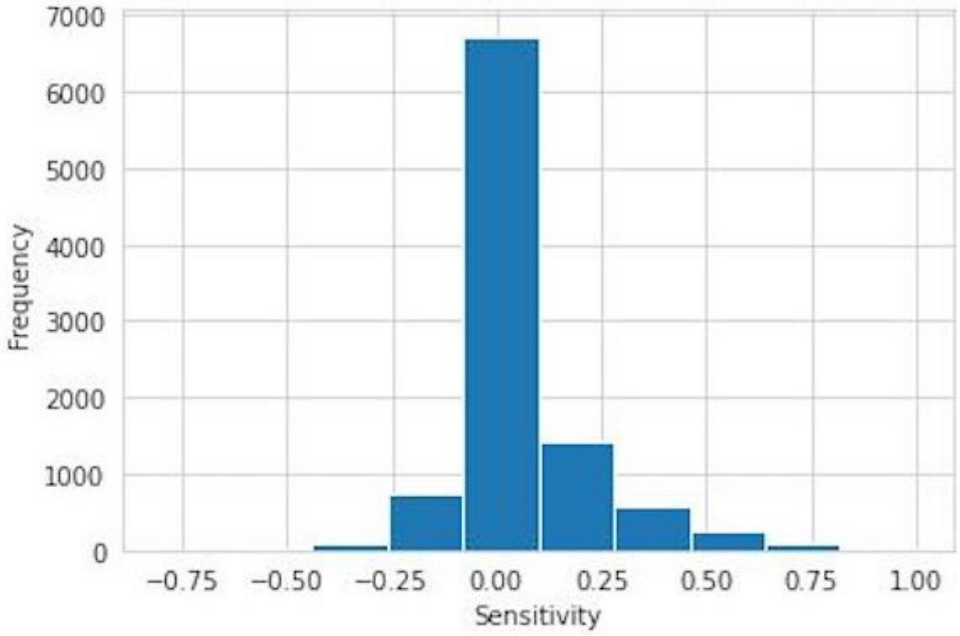

Figure 3: Histogram of context sensitivity. Negative (positive) sensitivity means the toxicity increased (decreased) when context was shown to the annotators.

When counting the number of posts for which $|\delta|$ exceeds a threshold t, called context-sensitive posts in Figure 4, we observe that as t increases, the number of context sensitive posts decreases. This means that clearly context sensitive posts (*e.g.*, in an edge case, ones that all OC coders found as toxic while all IC coders found as non toxic) were rare. Some examples of target posts, along with their parent posts and δ, are shown in Table 1.

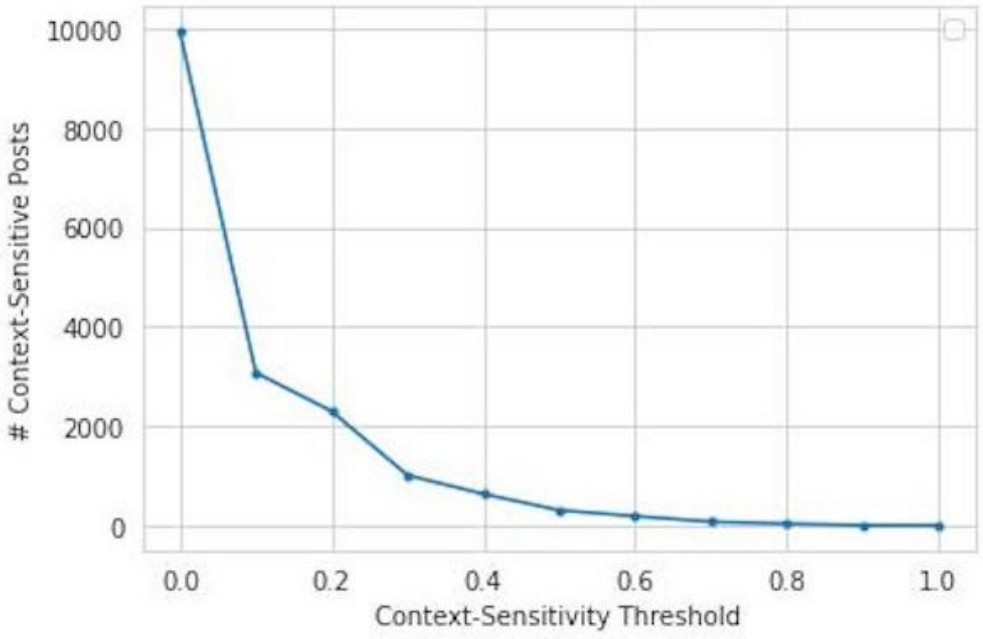

Figure 4: Number of context-sensitive posts ($|\delta| \geq t$), when varying the context-sensitivity threshold t.

2.4. What types of posts tend to be context sensitive?

A manual inspection of posts that had $|\delta| > 0.5$ [5], revealed that the differences between the IC and OC annotations were attributed primarily to four reasons. First, the target post was either *sarcastic* or *ironic*. The first row of Table 1 presents one such example, where the sarcastic tone of the parent text (*i.e.*, "Oh Don ... you are soooo predictable.") was probably what made the IC annotators assign an increased toxicity of the target post, from 36.6 percent to 80 percent. A second type of context sensitive posts comprised target posts that refer to *something toxic mentioned in the parent post*. For example, in the second row of Table 1, the starting phrase of the target post (*i.e.*, "Sucking us all dry") was copied unquoted from the parent post. The third was a supertype of the second type and it comprised target posts that included a *reference to someone or something mentioned in the conversational context*. In the third row of Table 1, for example, the word "ridiculous" referred to the fact that the nearest safe country (which was mentioned in the parent post), was Lebanon. Last, there were posts where the IC annotators perceived differently the toxicity compared to OC annotators, which could be due to a possibly different cultural background between OC and the IC annotators.

Table 1: Examples of context-sensitive posts in CCC. Here $s^{oc}(p)$ and $s^{ic}(p)$ are the fractions of out-of-context or in-context annotators, respectively, who found the target post p to be toxic; and $\delta = s^{oc}(p) - s^{ic}(p)$.

Parent of post p	Post p	$s^{oc}(p)$ percent	$s^{ic}(p)$ percent	δ percent
Oh Don ... you are soooo predictable.	oh Chuckie you are such a tattle tale	36.6	80	-43.4
The big three oil corporations are "sucking us all dry" and not the government.	Sucking us all dry by paying for state government for the last 40 years and creating the Permanent fund. Wow. The state pissed all the money away, they could have easily had $100 Billion in there.	80	20	60
Genuine refugees go to the nearest safe country.	ridiculous ... it's Lebanon.	72	20	52

3. Experimental study

Initially, we used our dataset to experiment with state-of-the-art toxicity detectors, aiming to investigate if context-sensitive posts were more difficult to automatically classify correctly as toxic or non-toxic. Then, we trained new systems to solve a different task: that of estimating how sensitive the toxicity score of each post was to its parent post, *i.e.*, to estimate the context sensitivity of a target post.

3.1. Toxicity detection

We employed the Perspective API toxicity detection system, one of the most well known of its kind, as is and with no further fine-

tuning, to classify CCC posts as toxic or not [6]. We also fine-tuned a BERT toxicity classifier in a dataset that comprises 30k randomly sampled posts taken from the CC dataset (excluding the CCC posts). We used Perspective because it performed well and did not require further training, and BERT because it achieves a state-of-the-art performance in many natural language processing (NLP) tasks (Devlin, *et al.*, 2019). In both experiments we either concatenated the parent post to the target to allow the model to consider the parent, or not. The simple concatenation mechanism was simple yet effective in toxicity detection (Pavlopoulos, *et al.*, 2020) and this has been shown also in neural machine translation (Tiedemann and Scherrer, 2017) [7]. Concatenating the parent post was, of course, a rudimentary mechanism to make a toxicity detector context-aware. We plan to investigate more elaborate mechanisms to make toxicity detectors context-aware in future work.

Figure 5 shows the Mean Absolute Error (MAE) of Perspective (blue lines) and BERT (orange lines), with (dashed lines) and without (solid lines) the parent post concatenated, when evaluating on all the CCC posts ($t = 0$) and when evaluating on smaller subsets with increasingly context-sensitive posts ($|\delta| \geq t$, $t > 0$). In all cases, we use the in-context (IC) gold labels as the ground truth. The greater the sensitivity threshold t, the smaller the sample (Figure 4).

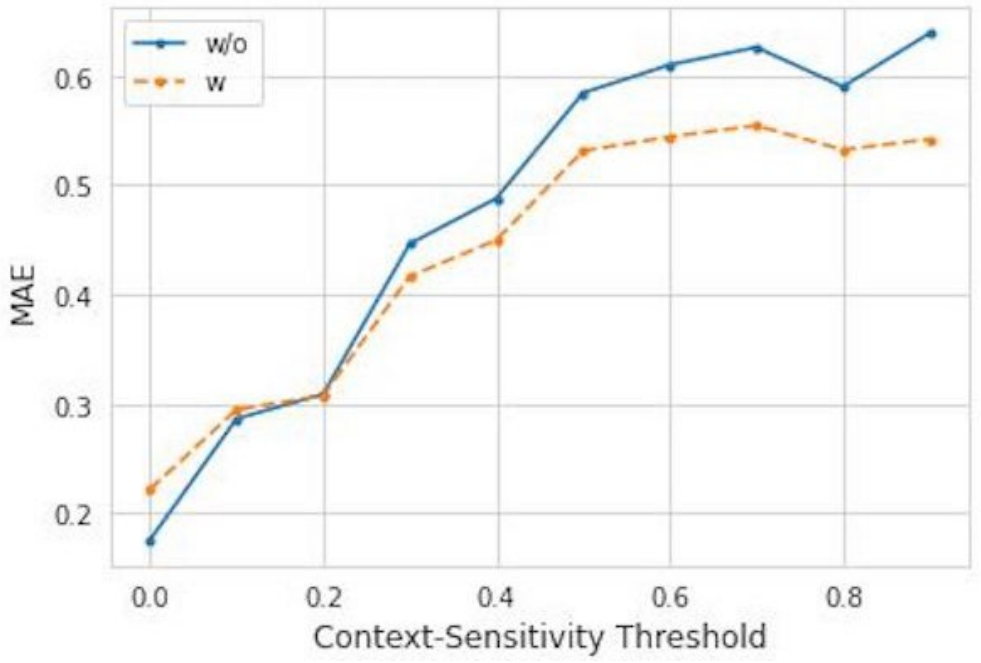

Figure 5: Mean Absolute Error (Y-axis) when predicting toxicity for different context-sensitivity thresholds (t; X-axis). We applied Perspective (blue, higher) and BERT (orange, lower) to target posts alone (solid lines) or concatenating the parent posts (dashed lines).

Figure 5 shows that when we concatenate the parent to the target post (dashed lines), MAE was clearly smaller, provided that $t \geq 0.2$ for both Perspective and BERT. Hence, the benefits of integrating context in toxicity detection systems may be visible only in sufficiently context-sensitive subsets, like the ones we would obtain by evaluating (and training) on posts with $t \geq 0.2$. By contrast, if no context-sensitivity threshold was imposed ($t = 0$) when constructing a dataset, the non-context sensitive posts ($|\delta| = 0$) dominated (Figure 4), hence adding context mechanisms to toxicity detectors has no visible effect in test scores. This explains related observations in our previous work (Pavlopoulos, *et al.*, 2020), where we found that context-sensitive posts were too rare and, thus, context-aware models did not perform better on existing toxicity datasets.

Notice that the more we move to the right of Figure 5, the higher the error for both variants (with, without context) of Perspective and BERT. This is probably due to the fact that both systems were trained on posts that had been rated by annotators who were not provided with the parent post (out of context; OC), whereas here we used the in-context (IC) annotations as ground truth. The greater the value of t in Figure 5, the larger the difference between toxicity scores of OC and IC annotators, hence the larger the difference between the (OC) ground truth that Perspective and BERT saw during their training and the ground truth that we used here (IC). Experimenting with artificial parent posts (long or short, toxic or not) confirmed that the error increased for context-sensitive posts. Finally, BERT had a lower MAE due to its fine-tuning on task specific posts (drawn from the same domain with the CCC posts).

The solution to the problem of increasing error as context sensitivity increases (Figure 5 would be to train toxicity detectors on

datasets that are richer in context-sensitive posts. However, as such posts are rare (Figure 4) they are hard to collect and annotate. This observation motivated the experiments of the next section, where we trained context-sensitivity detectors, which allowed us to collect posts that were likely to be context-sensitive. These posts could then be used to train toxicity detectors on datasets richer in context-sensitive posts.

3.2. Context sensitivity estimation

We trained and assessed four regressors on the new CCC dataset, to predict the context-sensitivity δ. We used Linear Regression, Support Vector Regression, a Random Forest regressor and a BERT-based (Devlin, *et al.*, 2019) regression model ($BERT_r$). The first three regressors used TF-IDF features. In the case of $BERT_r$, we added a feed-forward neural network (FFNN) on top of the top-level embedding of the [CLS] token. The FFNN consists of a dense layer (128 neurons) and a Tanh activation function, followed by another dense layer. The last dense layer has a single output neuron, with no activation function, that produces the context sensitivity score. Two baselines were also included in the experiments, an average and a random. The random baseline always randomly predicts a gold sensitivity score from the training set, while the average baseline always predicts the average of the gold sensitivity scores of all the posts in the training set. Preliminary experiments showed that adding simplistic context-mechanisms (*e.g.*, concatenating the parent post) to the context sensitivity regressors did not lead to improvements. This may be due to the fact that it was often possible to decide if a post was context-sensitive or not (we did not score the toxicity of posts in this section) by considering only the target post without its parent (*e.g.*, in responses like "NO!!"). Future work will investigate this hypothesis further by experimenting with more elaborate context-mechanisms. If the hypothesis can be verified, manually annotating context-sensitivity (not toxicity) may also require only the target post.

We used a train/validation/test split of 80/10/10 percent, respectively, and performed Monte Carlo 3-fold Cross Validation. We used mean square error (MSE) as our loss function and early stopping with patience of five epochs.

Table 2: Mean Squared Error (MSE), Mean Absolute Error (MAE), Area Under Precision-Recall curve (AUPR), and ROC AUC of all context sensitivity estimation models. An average (B1) and a random (B2) baseline have been included. All results averaged over three random splits, standard error of mean in brackets.

	MSE ↓	MAE ↓	AUPR ↑	AUC ↑
B1	2.3 (0.1)	11.56 (0.2)	12.69 (0.7)	50.00 (0.0)
B2	4.6 (0.0)	13.22 (0.1)	13.39 (0.8)	50.01 (1.6)
LR	2.1 (0.1)	11.0 (0.3)	30.11 (1.2)	71.67 (0.8)
SVR	2.3 (0.1)	12.8 (0.1)	28.66 (1.7)	71.56 (1.0)
RF	2.2 (0.1)	11.2 (0.2)	21.57 (1.0)	59.67 (0.3)
$BERT_r$	**1.8 (0.1)**	**9.2 (0.3)**	**42.01 (4.3)**	**80.46 (1.3)**

Table 2 presents the MSE and MAE of all the models on the test set. Unsurprisingly, $BERT_r$ outperforms the rest of the models in MSE and MAE. As already noted in Section 2.2, Wulczyn, *et al.* (2017) reported that training toxicity regressors (based on the empirical distribution of gold labels per post) instead of classifiers (based on the majority of the gold labels) also leads to improved classification performance. Hence, we also computed classification results. For the latter results, we turned the ground truth probabilities of the test instances to binary labels by setting a threshold t (Section 2) and assigning the label 1 if $\delta > t$ and 0 otherwise. In this experiment, t was set to the sum of the standard error of mean (SEM) of the OC and IC raters for that specific post p, *i.e.*, $t(p) = SEM^{oc}(p) + SEM^{ic}(p)$. The intuition behind this binarisation was that only posts with different mean OC and IC toxicity scores and non-overlapping standard errors were considered positive. By using this binary ground truth, the area under the precision recall curve (AUPR) and the area under the receiver operating characteristic curve (AUC) (Table 2) verified that $BERT_r$ outperforms the other models, even when the models were used as classifiers.

4. Collecting context sensitive posts

In Section 2 we saw that context sensitive posts were very rare in toxicity datasets (Figure 4). In Section 3 we showed that adding

even a rudimentary context mechanism (concatenating the parent post) to an existing toxicity detection system could reduce the system's error on context sensitive posts (Figure 5). However, the error of the toxicity detector remained high for context-sensitive posts. This problem could potentially be addressed by augmenting current datasets with more context-sensitive posts. As shown in Section 3, a regressor trained to predict the context sensitivity of a post can achieve low error (Table 2). Hence, we assessed the scenario where a context sensitivity regressor was employed to obtain a dataset richer in context-sensitive posts.

We used our best context-sensitivity regressor ($BERT_r$) to select the 250 most likely context-sensitive posts from the 2M CC posts, excluding the 10,000 CCC posts. We then crowd-annotated the 250 posts in context (IC) as with CCC posts, keeping also the original out-of-context (OC) annotations [8]. Table 3 shows examples of the 250 target posts obtained, along with their parent posts and δ.

Table 3: Examples of context-sensitive posts in the sampled dataset. Here $s^{oc}(p)$ and $s^{ic}(p)$ are the fractions of out-of-context or in-context annotators, respectively, who found the target post p to be toxic; and $\delta = s^{oc}(p) - s^{ic}(p)$.

Parent of post *p*	Post *p*	$s^{oc}(p)$ percent	$s^{ic}(p)$ percent	δ percent
And since Thomas Aquinas never observed animals having gay sex in the wild homosexuality never made it into the annals of natural law theory	Animals having "gay" sex? You mean there are "gay" animals. So, when they're not "doing it" they do other things like go to Madonna concerts?	60	0	-60
Making a cake is MUCH different then selling gasoline or renting hotel rooms. Making a cake is a form of artistry and requires the cake maker to artistically express him/herself which means the cake maker is actively participating. Owning a gas station where random people pump their own gas does not require active participation.	Oh, ok. So the if the gas guy had to pump gas for that gay man, he should be able to refuse that, right?	83.3	20	63.3
And SCOTUS will slap Watson & Chinp down yet again ... these Odummy Sock-Puppets never	Is the post implying that the judge is gay? I don't understand the comment, please	83.3	20	63.3

learn. That threesome they shared back in the day must have been amazing	explain? Are gays involved in this and not Muslims and their relatives?			
The appeal courts have one thing to do, ask is it legal or not, thats it, that is what appeals judges do, and they didnt, they coward away cause they knew then could not rule it illegal. sorry for your ignorance	The case has not yet been adjudicated on its merits (whether the Executive Order is illegal or not). Both the trial decision and the appeal decision were about staying the EO *until the trial on its merits* — ie, an injuction. I'd think about finding out some facts before calling someone else ignorant, Rex.	80	20	60
"...."marriage," by definition, meant one man, one woman ..." Actually no. The definition restricting it to one man one woman unions was only introduced into USA law 2004/5/6 across numerous states in a frantic attempt to avoid courts making similar findings to those of the Massachusetts Supreme court ruling. Prior to that it had always been expressly defined as between "two people", which is what triggered the Massachusets challenge. The fact that only opposite sex	The definitive dictionary of the English language, the OED, does not contain a single instance in which "modern" civilized society has included gay marriage. It does mention instances of "group" marriage in small, primitive societies, where all the men in a village are married to all the women. But those, as you know, are by far the exception. Actually, your argument bolsters my point. It was so universally	0	60	-60

marriages were performed in the past, does not mean marriage was defined as only between opposite sex couples, it simply illustrates that couples who were not opposite sex were being denied a fundamental right. The evidence of the existence of discrimination is not proof that the discrimination was justified or justifiable.	understood at the founding of the Nation that marriage meant man-woman that marriage did not need to be defined. Indeed, in most States, marriage could not have been defined so as to allow gay marriage, because until 1961, ALL 50 STATES outlawed sahdemy. Do you begin to get at least part of the point?			
May be Trudeau should do a double apology just to one up Harper and then apologize for Papa Trudeau and no himself ruining the Canadian economy.	What has this got to do with the rape and abuse of boys and girls in residential schools?	30	60	-30

We then repeated the same experiment, this time using 250 randomly selected posts from the 2M CC posts, excluding the 10,000 CCC posts and the 250 posts that were selected using $BERT_r$; the 250 random posts represented the complement of the posts selected by $BERT_r$. Figure 6 is the same as Figure 4, but we now considered the 250 randomly selected posts (dashed line) and the 250 posts that were selected using $BERT_r$ (solid line). As in Figure 4, we vary the context-sensitivity threshold t on the horizontal axis. The 250 posts that were sampled using $BERT_r$ clearly included more context-sensitive posts than the 250 random ones, with the threshold (t) in the range $0.1 \leq t < 0.7$, indicating that $BERT_r$ could be successfully used to obtain datasets richer in context-sensitive posts. As in Figure 4, there were very few context-sensitive posts for $t \geq 0.7$.

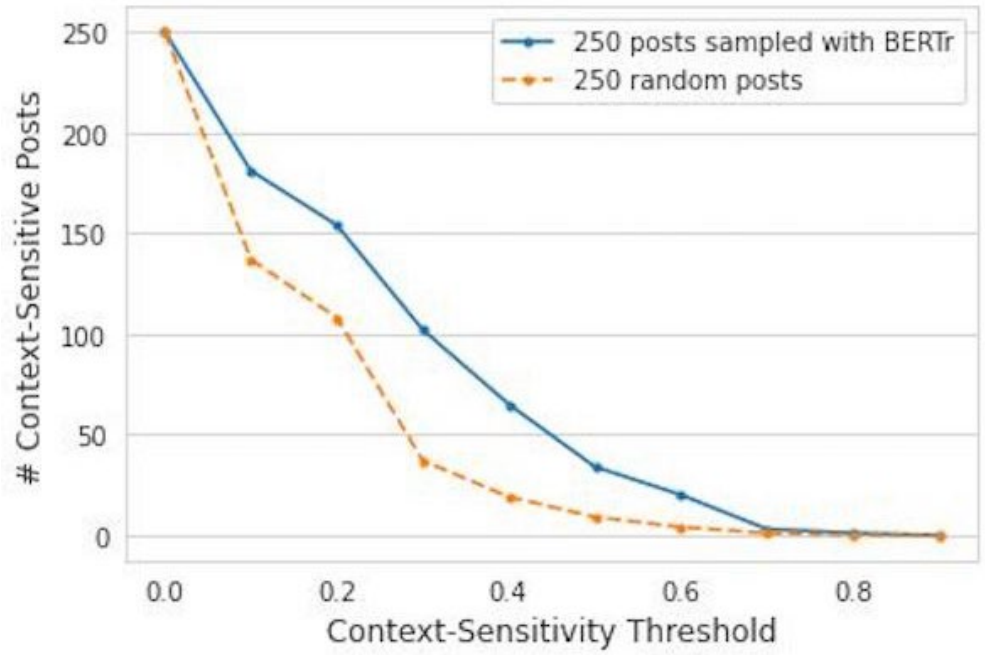

Figure 6: Number of context-sensitive posts ($|\delta| \geq t$), for different context-sensitivity thresholds (t), using 250 likely context-sensitive posts sampled with $BERT_r$ (solid) or 250 randomly selected posts (dashed line).

In this experiment, we also asked the crowd-annotators to indicate whether the parent post was helpful or not, when assessing the toxicity of each target post. Figure 7 shows how many of the 250 target posts (sampled using $BERT_r$ or random) the majority of the annotators responded that the parent post was useful. We vary the sensitivity threshold (t) on the horizontal axis up to $t = 0.7$, since no posts were context-sensitive for $t > 0.7$ (Figure 6). The perceived utility of the parent posts was clearly higher for the 250 posts sampled with $BERT_r$, compared to the 250 random ones, for all sensitivity thresholds. This again indicated that $BERT_r$ could be used to obtain datasets richer in context-sensitive posts.

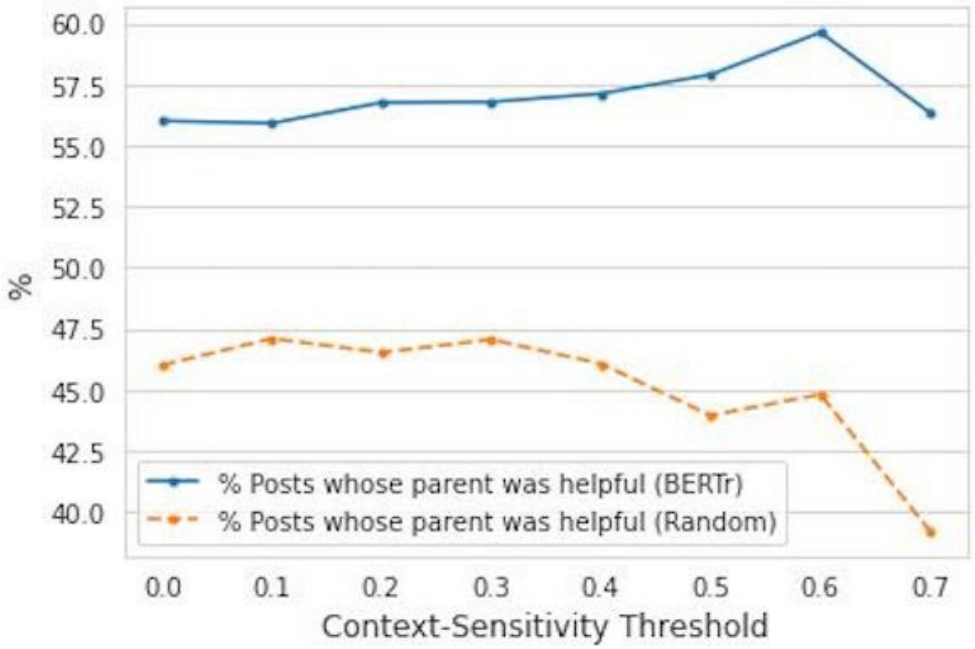

Figure 7: Percentage of the 250 target posts, sampled with $BERT_r$ (solid) or random (dashed line), for which the majority of annotators found the parent post useful when assessing the toxicity of the target post.

4.1. Verifying the statistical significance of the findings

To verify the statistical significance that the annotators found the parent post useful more frequently in posts sampled with $BERT_r$ than in random posts, we performed a paired bootstrap resampling, following the experimental setting of Koehn (2004). We sampled 100 posts from the 250 random posts, and 100 posts from the 250 posts obtained by using $BERT_r$, and we computed the percentage of posts where the majority of annotators found the parent post helpful, for random posts and $BERT_r$ posts. By resampling 1,000 times, we discovered that this percentage was greater for $BERT_r$ posts than for random posts, with a p-value of 0.05.

Finally, by turning the ground truth toxicity probabilities (for IC and OC annotation) into binary labels as in Section 3, we estimated a context sensitivity class ratio (fraction of context-sensitive posts out of all 250 posts), for the $BERT_r$-sampled and the randomly sampled posts. By using these binary labels, we found that 99 out of the 250 $BERT_r$-sampled posts (39.6 percent) were context sensitive, while only 43 out of the 250 randomly sampled posts (17.2 percent) were context-sensitive (22 percent points lower; *i.e.*, a 57 percent decrease). We verified the statistical significance of this finding (lower fraction) by using bootstrapping with a p-value of 0.05, as noted earlier. We conclude that sampling with $BERT_r$ led to a higher context-sensitivity class ratio than random sampling.

5. Improving the context-sensitivity regressor with data augmentation

We showed, in the previous section, that by employing a context sensitivity regressor ($BERT_r$ was our best one) one can sample new sets of posts (*e.g.*, from the 2M CC posts) that are richer in context-sensitive posts (by 22 percent points in our previous experiments) compared to random samples. By adding such richer (in context-sensitive posts) sets to an existing context-sensitivity dataset (*e.g.*, our CCC dataset), one can gradually increase the ratio of context-sensitive posts (which is low in CCC, see Figure 4). A natural question then is whether one could improve the context-sensitivity regressor by re-training it on the augmented dataset, which is less dominated by context-insensitive posts (more balanced in terms of context-sensitivity). Ideally the newly sampled (and overall more context-sensitive) posts would be crowd-annotated for context-sensitivity (by IC and OC raters) to obtain ground truth (gold context sensitivity scores). To avoid this additional annotation cost, however, in this section we explore a teacher-student approach (Hinton, *et al.*, 2015). The teacher is the initial $BERT_r$ context-sensitivity regressor (Section 3.2), which provides silver context sensitivity scores for the newly sampled posts. The student is another $BERT_r$ instance, which is trained on the augmented dataset (the data with gold sensitivity scores the teacher was trained on, plus the newly sampled posts with silver sensitivity scores).

These steps can be repeated in cycles, by using the student as the new teacher to sample and silver-score additional posts in each cycle. Similar teacher-student approaches have recently been used in several NLP and computer vision tasks (Xie, *et al.*, 2020a; Yu, *et al.*, 2018; Xie, *et al.*, 2020b), often using teacher and student models with different capacities. In our case, the teacher and student are the same model, but the student is trained on additional data silver-scored by the teacher, which is very similar to classical semi-supervised learning with Expectation Maximisation (Bishop, 2006).

Following this teacher-student approach, we experimented with data augmentation to improve the context-sensitivity estimator, using two different settings. In both settings, the teacher silver-scores the newly sampled additional training posts. In the setting discussed first, the teacher is also used to sample the new training posts. By contrast, in the second setting the new posts were randomly sampled, and the teacher was only used to silver-score them.

Teacher-student with teacher sampling: In this setting, we randomly sampled 20,000 posts from the Civil Comments (CC) dataset and used them as a pool to select (and silver-score) new training instances from, as follows:

1. Train a $BERT_r$ teacher on the gold-scored (by crowd-annotators) training instances of our CCC dataset (Section 2).
2. Use the $BERT_r$ teacher to silver-score for context-sensitivity all the posts of the pool (initially 20,000).
3. Select from the pool the 1,000 posts with the highest silver sensitivity scores, remove them from the pool, and add them (with their silver sensitivity scores) to the training set.
4. Train a BERT student on the new training set (augmented by 1,000 silver-scored posts).

r

5. Evaluate the student using exactly the same splits as in Section 3.2.
6. (Optional) Go back to step 2, using the student as a new teacher in a new cycle.

We repeated this process for five cycles and ended up with a training set augmented by 5,000 likely context-sensitive posts. Experimental results (Figure 8, blue solid line) show performance gains in MSE even from the first cycle. We also compared against using a single cycle with 5,000 new posts added at once (blue dashed line), instead of adding only 1,000 posts per cycle and re-training the teacher. Performing cycles and re-training the teacher clearly led to lower MSE, but with diminishing returns after cycle 4.

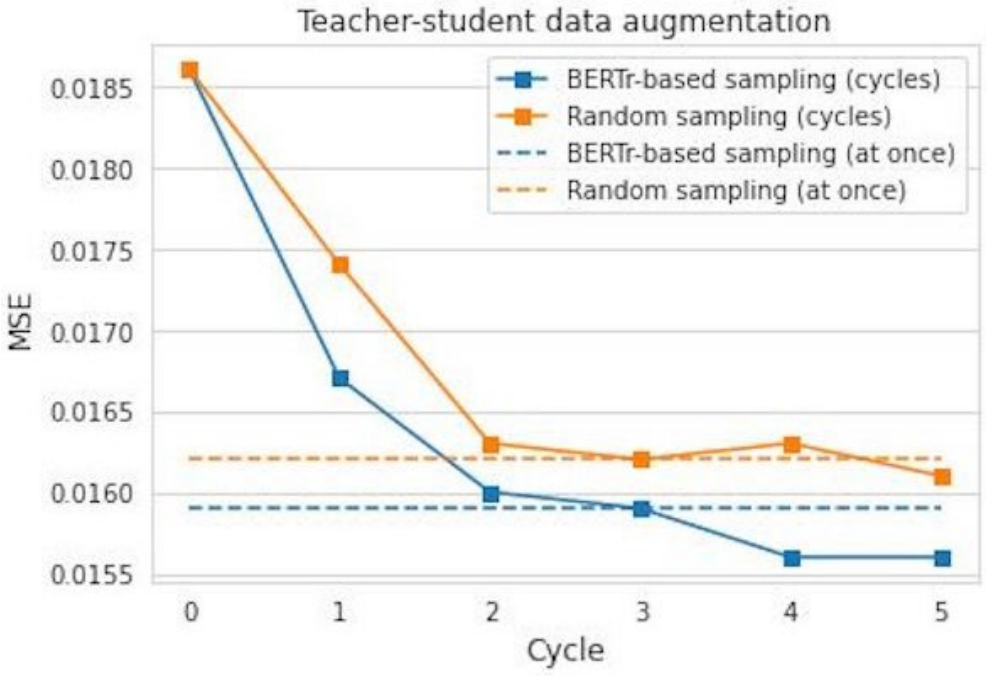

Figure 8: Data augmentation with knowledge distillation to improve $BERT_r$ context-sensitivity regressor. lue solid line: the teacher model is used both to silverscore the new training instances and to sample them. Orange solid line: the teacher model is used only to silver-score the new training instances, which are randomly selected. Dashed lines: same as the solid ones, but only one cycle is performed, which adds 5,000 silver-scored new training instances at once.

Teacher-student with random sampling: This setting is the same as the previous one, but in step 3 we randomly select 1,000 posts from the pool, instead of selecting the 1,000 posts with the highest silver sensitivity scores. Again, we used five cycles (Figure 8, orange solid line) and we also compared to a single cycle that adds 5,000 silver-scored training instances at once (orange dashed line). Sampling with the teacher's scores (blue solid line) was clearly better than random sampling (orange lines).

6. Related work

Toxicity detection

Abusive language detection is a difficult task due to its subjective nature. Cyberbullies attack victims on different topics such as race, religion and gender across multiple social media platforms (Agrawal and Awekar, 2018). Thus, the vocabulary used and the perceived meaning of words may vary when abusive language occurs in a different context. Several approaches have been examined in order to tackle the problem of abusive language detection. Researchers initially experimented with machine learning techniques using hand-crafted features, such as lexical and syntactic features (Davidson, *et al.*, 2017; Waseem and Hovy, 2016;

Djuric, *et al.*, 2015). Then, deep learning methods were employed, operating on word embeddings (Park and Fung, 2017; Pavlopoulos, *et al.*, 2017b, 2017c; Chakrabarty, *et al.*, 2019; Badjatiya, *et al.*, 2017; Haddad, *et al.*, 2020). More recently, fine-tuned large pre-trained Transformers were used in order to tackle toxicity detection (Mozafari, *et al.*, 2019; Pavlopoulos, *et al.*, 2019; D'Sa, *et al.*, 2020a; Ozler, *et al.*, 2020). As in many NLP tasks, deep learning approaches seemed to work better for toxicity detection than older machine learning methods based on hand-crafted features (Badjatiya, *et al.*, 2017).

To facilitate research in this field, researchers have published several datasets containing different types of toxicity. Nobata, *et al.*, (2016) developed a corpus of user comments posted on Yahoo Finance and News annotated for abusive language, the first of its kind. Wulczyn, *et al.* (2017) created and experimented with three new datasets; the 'Personal Attack' dataset where 115k comments from Wikipedia Talk pages were annotated as containing personal attack or not, the 'Aggression' dataset where the same comments were annotated as being aggressive or not, and the 'Toxicity' dataset that includes 159k comments again from Wikipedia Talk pages that were annotated as being toxic or not. Waseem and Hovy (2016) experimented on hate speech detection using a corpus of more than 16k tweets containing sexist, racist and non-toxic posts that they annotated by themselves. Most of the published toxicity datasets contain posts in English, but datasets in other languages also exist, such as Greek (Pavlopoulos, *et al.*, 2017a), Arabic (Mubarak, *et al.*, 2017), French (Chiril, *et al.*, 2020), Indonesian (Ibrohim and Budi, 2018) and German (Roß, *et al.*, 2016; Wiegand, *et al.*, 2018).

Context-aware NLP

Incorporating context into human language technology has been successfully applied to various NLP applications and domains. In text/word representation, context has a central role (Mikolov, *et al.*, 2013; Pennington, *et al.*, 2014; Melamud, *et al.*, 2016; Peters, *et al.*, 2018; Devlin, *et al.*, 2019). Integrating context is crucial in the sentiment analysis task too, where the semantic orientation of a word changes according to the domain or the context in which that word is being used (Agarwal, *et al.*, 2015). Vanzo, *et al.* (2014) explored the role of contextual information in supervised sentiment analysis over Twitter. They proposed two different types of contexts, a conversation-based context and a topic-based context, which includes several tweets in the history stream that contain overlapping hashtags. They modelled each tweet and its context as a sequence of tweets and used a sequence labelling model, SVM^{HMM}, to predict their sentiment labels jointly. They found that the kind of context they considered led to specific consistent benefits in sentiment classification. Ren, *et al.* (2016) proposed a context-based neural network model for Twitter sentiment analysis, incorporating contextual features from relevant tweets into the model in the form of word embedding vectors. They experimented with three types of context, a conversation-based context, an author-based context and a topic-based context. They found that integrating contextual information about the target tweet in their neural model offers improved performance compared with the state-of-the-art discrete and continuous word representation models. They also reported that topic-based context features were most effective for this task.

Despite the wide use of context in other NLP tasks, such as dialogue systems (Lowe, *et al.*, 2015; Dusek and Jurčček, 2016) and informational bias detection (van den Berg and Markert, 2020), very few researchers have focused on context-aware toxic language detection. Gao and Huang (2017) provided a corpus of speech labelled by annotators as hateful, obtained from full threads of online discussion posts under Fox News articles. They proposed two types of hate speech detection models that incorporate context information, a logistic regression model with context features and a neural network model with learning components for context. They reported performance gains in F1 score when incorporating context and that combining the two types of models they considered further improved performance by another seven percent in F1 score. Mubarak, *et al.* (2017) provided the title of the respective news article to the annotators during the annotation process, but they ignored parent comments since they did not have the entire thread. As Pavlopoulos, *et al.* (2020) already noticed, this presents the following problem: new comments may change the topic of the conversation and replies may require previous posts to be assessed correctly. Pavlopoulos, *et al.* (2017a) provided the annotators with the whole conversation thread for each target comment as context during the annotation process. However, the plain text of the comments was not available, which makes further analysis difficult.

In later work Pavlopoulos, *et al.* (2020) published two new toxicity datasets containing posts from the Wikipedia Talk pages, where during the annotation process, annotators were provided with the previous post in the thread and the discussion title. Pavlopoulos, *et al.* found that providing annotators with context could result both in amplification and mitigation of the perceived toxicity of posts. Moreover, they found no evidence that context actually improved the performance of toxicity classifiers. In a similar work, Menini, *et al.* (2021) investigated the role of textual context in abusive language detection on Twitter. They first re-annotated tweets in the dataset of Founta, *et al.* (2018) in two settings, with and without context. During the annotation process, they provided annotators with the whole conversational thread of each post that was being annotated. After comparing the two datasets (with and without context-aware annotations) they found that context was sometimes necessary to understand the real intent of the user, and that it was more likely to mitigate the abusiveness of a tweet even if it contained profanity, a finding consistent with our work. Finally, they experimented with several classifiers, using both context-aware and context-unaware architectures. Regarding the context they experimented with different context lengths, ranging from one (*i.e.*, only the preceding tweet, similar to our context), to all. They observed that context size did not show a consistent impact and different context lengths did not have a statistically significant effect, which they measured for each classification algorithm using an approximate randomization test. Their experimental results showed that when classifiers were given context and were evaluated on context-aware datasets, their performance dropped dramatically compared to a setting where classifiers were not given context and were evaluated on context-unaware datasets. However, Menini, *et al.* noted that additional work was needed to find better approaches for effectively including context in the classification. Vidgen, *et al.* (2021) introduced a new annotated dataset for abusive

language of approximately 25k Reddit entries, which were annotated in context, using the conversational thread of each entry as context. They provided high-quality annotations by using a team of trained annotators and a time-intensive discussion-based process, facilitated by experts, for adjudicating disagreements. For every annotation they provided a label for whether contextual information was needed to make the annotation which was very similar to our helpful codes (asking the annotators to indicate whether the parent post was helpful or not, when assessing the toxicity of each target post). Although they incorporated a deep level of context, their dataset did not include OC labels which made the analysis of context sensitivity impossible. Research conducted by Pavlopoulos, *et al.* (2020), Menini, *et al.* (2021) and Vidgen, *et al.* (2021) were very similar to ours in terms of studying context-aware toxicity detection, however, none studied context-sensitivity of context-dependent posts.

Regression as classification in NLP

Approaching a text classification problem as a regression-based problem has been tested by researchers in various NLP tasks, such as sentiment analysis (Wang, *et al.*, 2016), emotional analysis (Buechel and Hahn, 2016), metaphor detection (Parde and Nielsen, 2018) and toxicity detection. Wulczyn, *et al.'s* (2017) efforts were similar to ours in that they noticed that estimating the likelihood of a post to be personal attack (or not), using the empirical distribution of human-ratings, rather than the majority vote, produced a better classifier, even in terms of the ROC AUC metric. D'Sa, *et al.* (2020b) experimented on the English Wikipedia Detox corpus by designing both binary classification and regression-based approaches aiming to predict whether a comment was toxic or not. They compared different unsupervised word representations and different deep learning based classifiers. In most of their experiments, the regression-based approach showed slightly better performance than the classification setting, which was consistent with the findings of Wulczyn, *et al.* (2017). Fornaciari, *et al.* (2021) proposed a new method for leveraging instance ambiguity, as expressed by probability distribution over label annotations. They created Multi Task Learning (MTL) models to predict this label distribution as an auxiliary task in addition to the standard classification task. They found that this auxiliary task reduced the penalty for errors on ambiguous entities and thereby mitigated overfitting. Finally, they demonstrated that incorporating this auxiliary task in training could result in significantly improved performance across two tasks (Part-of-speech tagging and Morphological stemming) beyond the standard approach and prior work.

7. Limitations and considerations

We limited our study to the parent post of the conversational context, but we noted that more posts or even the entire thread could be used. Also, other possible sources of context exist and could be examined along with the thread's posts. For instance, the discussion title could provide annotators with more information when they annotate each post of the discussion. We consider this study as the first of a series of steps that need to be taken in order to investigate the relation of context in toxicity detection.

Online discussions are currently moderated by human raters and machine learning models. Both may carry bias introduced by annotators (*e.g.*, if all of the annotators originate from the same cultural background). The same limitation applies to this study, for which we employed crowd annotators, but without trying to control for possibly different social norms.

We used crowd-sourcing to develop our dataset, but crowd-sourcing has weaknesses and has been criticised (Akhtar, *et al.*, 2021, 2020; Al Kuwatly, *et al.*, 2020; Waseem, 2016). We consider the high inter-annotator agreement that we observed in this work (kappa = 83.93 percent) satisfying. However, future work may further investigate our hypotheses by using alternative annotation means.

OC and IC labels were obtained at different times, though we used the same platform, same guidelines and we requested annotators of the same type. However, it was impossible to determine if toxicity change was due to the inclusion of context or due to other reasons. For example, annotators were not identical, demographic features of annotator groups were not guaranteed to be the same, and time had elapsed between the two annotations, which may have changed perceptions of toxicity. Other variables may be interfering.

We focused on posts in English and we employed English-speaking annotators. The English-centric nature of the Internet is a widely acknowledged problem. The ways that the communicative intent is mixed into culture, and the notions of what is appropriate in a given context are problems that are sometimes simpler for a language like English that does not mark gender and has lost its formal and casual distinctions. This is also related to the difficulty of doing work with social norms, which is challenging in well resourced languages and virtually impossible in languages that lack modelling resources.

8. Conclusions and future work

We introduced the task of estimating context sensitivity of posts in toxicity detection, *i.e.*, estimating the extent in which the perceived toxicity of a post depends on a conversational context. We constructed, presented and released a new dataset that can be

used to train and evaluate systems for the new task, where context is the previous post. We also showed that context-sensitivity estimation systems can be used to collect larger samples of context-sensitive posts, which is a prerequisite to train toxicity detectors to better handle context-sensitive posts. Furthermore, we showed that the performance or our best context sensitivity estimator was further improved by augmenting the training dataset with teacher-student knowledge distillation. Context-sensitivity estimators can also be used to suggest when moderators should consider the context of a post, which is more costly and may not always be necessary. In future work, we hope to incorporate context mechanisms in toxicity detectors and train (and evaluate) them on datasets sufficiently rich in context-sensitive posts.

About the authors

Alexandros Xenos holds a B.Sc. and M.Sc. degree in computer science from the AUEB and is a member of the NLP Group AUEB.
E-mail: a [dot] xenos20 [at] aueb [dot] gr

John Pavlopoulos is a visiting scholar at Ca'Foscari University of Venice, Italy and an affiliated researcher at the Athens University of Economics and Business, Greece, the Stockholm University, Sweden, and the Ritsumeikan University, Japan. His research is focused on machine learning, especially deep learning, for natural language processing and digital humanities.
Send comments to: annis [at] aueb [dot] gr

Ion Androutsopoulos is Professor and head of the NLP Group in the Dept. of Informatics, Athens University of Economics & Business. His interests include biomedical question answering, natural language generation from medical images, text classification, including filtering abusive content, information extraction and opinion mining, including legal text analytics and sentiment analysis.
E-mail: ion [at] aueb [dot] gr

Jeffrey Sorensen was part of the original team at Google Jigsaw that launched the Perspective API. He joined Google in 2010 to work with the speech team, developed compact language models for use in the on-device recognizer for mobile devices, and lead a team responsible for data collection and annotation. Before Google he worked for IBM on speech recognition and translation.
E-mail: ldixon [at] google [dot] com

Lucas Dixon was Chief Scientist and founder of engineering efforts at Jigsaw and tech-lead of the Perspective API. He is now a research scientist in the PAIR team in Google. Before Google, he was a research scientist at the University of Edinburgh. His research interests include NLP, deep learning, automated reasoning, symbolic computation and quantum computing.
E-mail: sorenj [at] google [dot] com

Leo Laugier is a Ph.D. candidate in the Department of Computer Sciences and Networks (INFRES) at Institut Polytechnique de Paris (IP Paris), funded by a Google doctoral fellowship and working on conversation AI research. He is broadly interested about leveraging AI models and theory in social science computing, and specifically in applying techniques of NLP for social good.
E-mail: leo [dot] laugier [at] telecom-paris [dot] fr

Acknowledgments

We thank *First Monday's* reviewers for their comments and suggestions. This research was funded in part by an unrestricted gift from Google.

Notes

1. The dataset is released under a CC0 licence. It can be downloaded from https://github.com/ ipavlopoulos/context_toxicity/tree/master/data.

2. See Appendix for more details on the annotation guidelines.

3. https://appen.com.

4. We chose populous majority English-speaking countries. The most common country of origin was the U.S.

5. One hundred posts with $|\delta| > 0.5$ and fourteen with $\delta < 0.5$ were studied.

6. https://www.perspectiveapi.com

7. We concatenated the parent post at inference time, instead of training with the concatenated text, because in preliminary experiments the former performed best.

8. The MAE computed between the 250 crowd-annotated and the 250 $BERT_r$-generated context sensitivity scores, with the latter expected to be high, is 19.33 percent.

References

Sweta Agrawal and Amit Awekar, 2018. "Deep learning for detecting cyberbullying across multiple social media platforms," *arXiv*:1801.06482 (19 January).
doi: https://doi.org/10.48550/arXiv.1801.06482, accessed 10 August 2022.

Basant Agarwal, Namita Mittal, Pooja Bansal and Sonal Garg, 2015. "Sentiment analysis using common-sense and context information," *Computational Intelligence and Neuroscience*, volume 2015, article ID 715730.
doi: https://doi.org/10.1155/2015/715730, accessed 10 August 2022.

Sohail Akhtar, Valerio Basile and Viviana Patti, 2021. "Whose opinions matter? Perspective-aware models to identify opinions of hate speech victims in abusive language detection," *arXiv*:2106.15896 (30 June).
doi: https://doi.org/10.48550/arXiv.2106.15896, accessed 10 August 2022.

Sohail Akhtar, Valerio Basile and Viviana Patti, 2020. "Modeling annotator perspective and polarized opinions to improve hate speech detection," *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, number 1, pp. 151–154, and at https://ojs.aaai.org/index.php/HCOMP/article/view/7473, accessed 10 August 2022.

Hala Al Kuwatly, Maximilian Wich and Georg Groh, 2020. "Identifying and measuring annotator bias based on annotators demographic characteristics," *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 184–190, and at https://aclanthology.org/2020.alw-1.21/, accessed 10 August 2022.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta and Vasudeva Varma, 2017. "Deep learning for hate speech detection in tweets," *WWW '17 Companion: Proceedings of the 26th International Conference on World Wide Web Companion* pp. 759–760.
doi: https://doi.org/10.1145/3041021.3054223, accessed 10 August 2022.

Christopher M. Bishop, 2006. *Pattern recognition and machine learning*. New York: Springer.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain and Lucy Vasserman, 2019. "Nuanced metrics for measuring unintended bias with real data for text classification," *WWW '19: Companion Proceedings of the 2019 World Wide Web Conference*, pp. 491–500.
doi: https://doi.org/10.1145/3308560.3317593, accessed 10 August 2022.

Sven Buechel and Udo Hahn, 2016. "Emotion analysis as a regression problem — dimensional models and their implications on emotion representation and metrical evaluation," *ECAI'16: Proceedings of the Twenty-second European Conference on Artificial Intelligence*, pp. 1,114–1,122.
doi: https://doi.org/10.3233/978-1-61499-672-9-1114, accessed 10 August 2022.

Tuhin Chakrabarty, Kilol Gupta and Smaranda Muresan, 2019. "Pay 'attention' to your context when classifying abusive language," *Proceedings of the Third Workshop on Abusive Language Online*, pp. 70–79, and at https://aclanthology.org/W19-3508/, accessed 10 August 2022.

Patricia Chiril, Veronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi and Marlene Coulomb-Gully, 2020. "An annotated corpus for sexism detection in French tweets," *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 1,397–1,403, and at https://aclanthology.org/2020.lrec-1.175/, accessed 10 August 2022.

Thomas Davidson, Dana Warmsley, Michael Macy and Ingmar Weber, 2017. "Automated hate speech detection and the problem of offensive language," *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11, number 1, at https://ojs.aaai.org/index.php/ICWSM/article/view/14955, accessed 10 August 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova, 2019. "BERT: Pre-training of deep bidirectional transformers for language understanding," *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. 4,171–4,186, and at https://aclanthology.org/N19-1423/, accessed 10 August 2022.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic and Narayan Bhamidipati, 2015. "Hate speech detection with comment embeddings," *WWW '15 Companion: Proceedings of the 24th International Conference on World Wide*

Web, pp. 29–30.
doi: https://doi.org/10.1145/2740908.2742760, accessed 10 August 2022.

Ashwin Geet D'Sa, Irina Illina and Dominique Fohr, 2020a. "BERT and fastText embeddings for automatic detection of toxic speech," *2020 International Multi-Conference on: "Organization of Knowledge and Advanced Technologies" (OCTA)*.
doi: https://doi.org/10.1109/OCTA49274.2020.9151853, accessed 10 August 2022.

Ashwin Geet D'Sa, Irina Illina and Dominique Fohr. 2020b. "Towards non-toxic landscapes: Automatic toxic comment detection using DNN," *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pp. 21–25, and at https://aclanthology.org/2020.trac-1.4/, accessed 10 August 2022.

Ondrej Dusek and Filip Jurčček, 2016. "A context-aware natural language generator for dialogue systems," *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 185–190, and at https://aclanthology.org/W16-3622/, accessed 10 August 2022.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy and Massimo Poesio, 2021. "Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning," *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2,591–2,597, and at https://aclanthology.org/2021.naacl-main.204/, accessed 10 August 2022.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos and Nicolas Kourtellis, 2018. "Large scale crowdsourcing and characterization of Twitter abusive behavior," *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*, volume 12, number 1, at https://ojs.aaai.org/index.php/ICWSM/article/view/14991, accessed 10 August 2022.

Lei Gao and Ruihong Huang, 2017. "Detecting online hate speech using context aware models," *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pp. 260–266, and at https://aclanthology.org/R17-1036/, accessed 10 August 2022.

Bushr Haddad, Zoher Orabe, Anas Al-Abood and Nada Ghneim, 2020. "Arabic offensive language detection with attention-based deep neural networks," *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pp 76–81, and at https://aclanthology.org/2020.osact-1.12/, accessed 10 August 2022.

Geoffrey Hinton, Oriol Vinyals and Jeff Dean, 2015. "Distilling the knowledge in a neural network," *arXiv*:1503.02531 (9 March).
doi: https://doi.org/10.48550/arXiv.1503.02531, accessed 10 August 2022.

Muhammad Okky Ibrohim and Indra Budi 2018. "A dataset and preliminaries study for abusive language detection in indonesian social media," *Procedia Computer Science*, volume 135, pp. 222–229.
doi: https://doi.org/10.1016/j.procs.2018.08.169, accessed 10 August 2022.

Philipp Koehn, 2004. "Statistical significance tests for machine translation evaluation," *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 388–395, and at https://aclanthology.org/W04-3250/, accessed 10 August 2022.

R. Lowe, Nissan Pow, Laurent Charlin, Joelle Pineau and Iulian V. Serban, 2015. "Incorporating unstructured textual knowledge sources into neural dialogue systems," at http://blueanalysis.com/julianserban/Files/IncorporatingExternalKnowledge.pdf, accessed 10 August 2022.

Oren Melamud, Jacob Goldberger and Ido Dagan, 2016. "context2vec: Learning Generic Context Embedding with Bidirectional LSTM," *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 51–61, and at https://aclanthology.org/K16-1006/, accessed 10 August 2022.

Stefano Menini, Alessio Palmero Aprosio and Sara Tonelli, 2021. "Abuse is contextual, what about NLP? The role of context in abusive language annotation and detection," *arXiv*:2103.14916 (27 March).
doi: https://doi.org/10.48550/arXiv.2103.14916, accessed 10 August 2022.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean, 2013. "Distributed representations of words and phrases and their compositionality," *NIPS'13: Proceedings of the 26th International Conference on Neural Information Processing Systems*, volume 2, pp. 3,111–3,119.

Marzieh Mozafari, Reza Farahbakhsh and Noël Crespi, 2019. "A BERT-based transfer learning approach for hate speech detection in online social media," In: Cherifi, H., Gaito, S., Mendes, J., Moro, E., Rocha, L. (editors). *Complex networks and their applications VIII. Complex networks 2019. Studies in Computational Intelligence*, volume 881. Cham, Switzerland: Springer, pp. 928–940.
doi: https://doi.org/10.1007/978-3-030-36687-2_77, accessed 10 August 2022.

Hamdy Mubarak, Kareem Darwish and Walid Magdy, 2017. “Abusive language detection on Arabic social media,” *Proceedings of the First Workshop on Abusive Language Online*, pp. 52–56, and at https://aclanthology.org/W17-3008/, accessed 10 August 2022.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad and Yi Chang, 2016. “Abusive language detection in online user content,” *WWW ’16: Proceedings of the 25th International Conference on World Wide Web*, pp. 145–153. doi: https://doi.org/10.1145/2872427.2883062, accessed 10 August 2022.

Kadir Bulut Ozler, Kate Kenski, Steve Rains, Yotam Shmargad, Kevin Coe and Steven Bethard, 2020. “Fine-tuning for multi-domain and multi-label uncivil language detection,” *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pp. 28–33, and at https://aclanthology.org/2020.alw-1.4/, accessed 10 August 2022.

Natalie Parde and Rodney Nielsen, 2018. “Exploring the terrain of metaphor novelty: A regression-based approach for automatically scoring metaphors,” *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, number 1, article number 658, pp. 5,366–5,373. doi: https://doi.org/10.1609/aaai.v32i1.11940, accessed 10 August 2022.

Ji Ho Park and Pascale Fung, 2017. “One-step and two-step classification for abusive language detection on Twitter,” *Proceedings of the First Workshop on Abusive Language Online*, pp. 41–45, and at https://aclanthology.org/W17-3006/, accessed 10 August 2022.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain and Ion Androutsopoulos, 2020. “Toxicity detection: Does context really matter?” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4,296–4,305, and at https://aclanthology.org/2020.acl-main.396.pdf, accessed 10 August 2022.

John Pavlopoulos, Nithum Thain, Lucas Dixon and Ion Androutsopoulos, 2019. “ConvAI at SemEval2019 Task 6: Offensive language identification and categorization with perspective and BERT,” *Proceedings of the 13th International Workshop on Semantic Evaluation*, pp. 571–576, and at https://aclanthology.org/S19-2102, accessed 10 August 2022.

John Pavlopoulos, Prodromos Malakasiotis and Ion Androutsopoulos, 2017a. “Deep learning for user comment moderation,” *Proceedings of the First Workshop on Abusive Language Online*, pp. 25–35, and at https://aclanthology.org/W17-3004/, accessed 10 August 2022.

John Pavlopoulos, Prodromos Malakasiotis and Ion Androutsopoulos, 2017b. “Deeper attention to abusive user content moderation,” *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1,125–1,135, and at https://aclanthology.org/D17-1117, accessed 10 August 2022.

John Pavlopoulos, Prodromos Malakasiotis, Juli Bakagianni and Ion Androutsopoulos, 2017c. “Improved abusive comment moderation with user embeddings,” *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing Meets Journalism*, pp. 51–55, and at https://aclanthology.org/W17-4209/, accessed 10 August 2022.

Jeffrey Pennington, Richard Socher and Christopher Manning, 2014. “GloVe: Global vectors for word representation,” *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1,532–1,543, and at https://aclanthology.org/D14-1162/, accessed 10 August 2022.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee and Luke Zettlemoyer, 2018. “Deep contextualized word representations,” *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1 (long papers), pp. 2227–2237, and at https://aclanthology.org/N18-1202/, accessed 10 August 2022.

Justus Randolph, 2010. “Free-Marginal Multirater Lappa (multirater κ_{free}): An alternative to Fleiss’ Fixed-Marginal Multirater Kappa,” at https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.59.8776&rep=rep1&type=pdf, accessed 10 August 2022.

Yafeng Ren, Yue Zhang, Meishan Zhang and Donghong Ji, 2016. “Context-sensitive Twitter sentiment classification using neural network,” *AAAI’16: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 215–221. doi: https://doi.org/10.1609/aaai.v30i1.9974, accessed 10 August 2022.

Björn Roß, Michael Rist, Guillermo Carbonell, Ben Cabrera, Nils Kurowsky and Michael Wojatzki, 2016. “Measuring the reliability of hate speech annotations: The case of the European Refugee Crisis,” *Proceedings of NLP4CMC III: Third Workshop on Natural Language Processing for ComputerMediated Communication*, pp. 6–9. doi: https://doi.org/10.17185/duepublico/42132, accessed 10 August 2022.

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,” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5,477–5,490, and at https://aclanthology.org/2020.acl-main.486.pdf, accessed 10 August 2022.

Nanna Thylstrup and Zeerak Waseem. 2020. “Detecting ‘dirt’ and ‘toxicity’: Rethinking content moderation as pollution behaviour,” *SSRN Electronic Journal* (22 December).
doi: https://doi.org/10.2139/ssrn.3709719, accessed 10 August 2022.

Jörg Tiedemann and Yves Scherrer, 2017. “Neural machine translation with extended context,” *Proceedings of the Third Workshop on Discourse in Machine Translation*, pp. 82–92, and at https://aclanthology.org/W17-4811/, accessed 10 August 2022.

Esther van den Berg and Katja Markert, 2020. “Context in informational bias detection,” *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6,315–6,326, and at https://aclanthology.org/2020.coling-main.556/, accessed 10 August 2022.

Andrea Vanzo, Danilo Croce and Roberto Basili, 2014. “A context-based model for sentiment analysis in Twitter,” *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 2,345–2,354, and at https://aclanthology.org/C14-1221/, accessed 10 August 2022.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini and Rebekah Tromble, 2021. “Introducing CAD: The contextual abuse dataset,” *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2,289–2,303, and at https://aclanthology.org/2021.naacl-main.182/, accessed 10 August 2022.

Jin Wang, Liang-Chih Yu, K. Robert Lai and Xuejie Zhang, 2016. “Dimensional sentiment analysis using a regional CNN-LSTM model,” *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, short papers, pp. 225–230, and at https://aclanthology.org/P16-2037/, accessed 10 August 2022.

Zeerak Waseem. 2016. “Are you a racist or am i seeing things? Annotator influence on hate speech detection on Twitter,” *Proceedings of the First Workshop on NLP and Computational Social Science*, pp. 138–142, and at https://aclanthology.org/P16-2037/, accessed 10 August 2022.

Zeerak Waseem and Dirk Hovy, 2016. “Hateful symbols or hateful people? predictive features for hate speech detection on Twitter,” *Proceedings of the NAACL Student Research Workshop*, pp. 88–93 and at https://aclanthology.org/N16-2013, accessed 10 August 2022.

Michael Wiegand, Melanie Siegel and Josef Ruppenhofer, 2018. “Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language,” *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*, and at https://epub.oeaw.ac.at/0xc1aa5576_0x003a10d2.pdf, accessed 10 August 2022.

Ellery Wulczyn, Nithum Thain and Lucas Dixon, 2017. “Ex machina: Personal attacks seen at scale,” *WWW ’17: Proceedings of the 26th International Conference on World Wide Web*, pp. 1,391–1,399.
doi: https://doi.org/10.1145/3038912.3052591, accessed 10 August 2022.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong and Quoc V. Le, 2020a. “Unsupervised data augmentation for consistency training,” *NIPS’20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, article number 525, pp. 6,256–6,268.

Qizhe Xie, Minh-Thang Luong, Eduard Hovy and Quoc V. Le, 2020b. “Self-training with noisy student improves ImageNet classification,” *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
doi: https://doi.org/10.1109/CVPR42600.2020.01070, accessed 10 August 2022.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi and Quoc V. Le, 2018. “Fast and accurate reading comprehension by combining self-attention and convolution,” *International Conference on Learning Representations*, at https://openreview.net/references/pdf?id=H1-o1a0DG, accessed 10 August 2022.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli and Nithum Thain, 2018. “Conversations gone awry: Detecting early signs of conversational failure,” *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1: Long papers, pp. 1,350–1,361, and at https://aclanthology.org/P18-1125/, accessed 10 August 2022.

Appendix

Annotation guidelines

We used the same annotation guidelines that were used in Borkan, *et al.* (2019).

The guidelines are available at
https://github.com/conversationai/conversationai.github.io/blob/main/crowdsourcing_annotation_schemes/toxicity_with_subattributes.md.

Editorial history

Received 3 September 2021; revised 18 April 2022; accepted 10 August 2022.

Toxicity detection sensitive to conversational context
by Alexandros Xenos, John Pavlopoulos, Ion Androutsopoulos, Lucas Dixon, Jeffrey Sorensen, and Léo Laugier.
First Monday, volume 27, number 9 (September 2022).
doi: https://dx.doi.org/10.5210/fm.v27i9.12285