Toxicity Detection: Does Context Really Matter?

John Pavlopoulos[†], **Jeffrey Sorensen**[‡]
Lucas Dixon[‡], **Nithum Thain**[‡], **Ion Androutsopoulos**[†]
[†] Department of Informatics, Athens University of Economic and Business, Greece
annis,ion@aueb.gr
[‡] Google
sorenj,ldixon,nthain@google.com

Abstract

Moderation is crucial to promoting healthy online discussions. Although several 'toxicity' detection datasets and models have been published, most of them ignore the context of the posts, implicitly assuming that comments may be judged independently. We investigate this assumption by focusing on two questions: (a) does context affect the human judgement, and (b) does conditioning on context improve performance of toxicity detection systems? We experiment with Wikipedia conversations, limiting the notion of context to the previous post in the thread and the discussion title. We find that context can both amplify or mitigate the perceived toxicity of posts. Moreover, a small but significant subset of manually labeled posts (5% in one of our experiments) end up having the opposite toxicity labels if the annotators are not provided with context. Surprisingly, we also find no evidence that context actually improves the performance of toxicity classifiers, having tried a range of classifiers and mechanisms to make them context aware. This points to the need for larger datasets of comments annotated in context. We make our code and data publicly available.

1 Introduction

Systems that detect abusive language are used to promote healthy conversations online and protect minority voices (Hosseini et al., 2017). Apart from a growing volume of press articles concerning toxicity online,[1] there is increased research interest on detecting abusive and other unwelcome comments labeled 'toxic' by moderators, both for English and other languages.[2] However, the vast majority of current datasets do not include the preceding comments in a conversation and such context was not shown to the annotators who provided the gold toxicity labels. Consequently, systems trained on these datasets ignore the conversational context. For example, a comment like "nope, I don't think so" may not be judged as rude or inflammatory by such a system, but the system's score would probably be higher if the system could also consider the previous (also called *parent*) comment "might it be that I am sincere?". Table 1 shows additional examples of comments that are not easily judged for toxicity without the parent comment. Interestingly, even basic statistics on how often context affects the perceived toxicity of online posts have not been published. Hence, in this paper we focus on the following two foundational research questions:

PARENT	All of his arguements are nail perfect, you're inherently stupid. The lead will be changed.
TARGET	**Great argument!**
PARENT	Really? It's schmucks like you (and Bush) who turn the world into the shithole it is today!
TARGET	**I'd be interested in the reasoning for that comment, personally. (bounties)**
PARENT	Indeed. Hitler was also strongly anti-pornography [...] it sure looks like Hitler is a hot potato that nobody wants to be stuck with.
TARGET	**Well I guess they won't approve the slogan "Hitler hated porn".**
PARENT	?? When did I attack you? I definitely will present this to the arbcom, you should mind WP:CIVIL when participating in discussions in Wikipedia.
TARGET	**I blame you for my alcoholism add that too**

Table 1: Comments that are not easily labeled for toxicity without the 'parent' (previous) comment. The 'target' comment is the one being labeled.

- RQ1: *How often does context affect the toxicity of posts as perceived by humans in online conversations? And how often does context amplify or mitigate the perceived toxicity?*

[1] Following the work of Wulczyn et al. (2017) and Borkan et al. (2019), *toxicity* is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion" (Wulczyn et al., 2017).

[2] For English, see for example TRAC (Kumar et al., 2018), OFFENSEVAL (Zampieri et al., 2019b), or the recent Workshops on Abusive Language Online (https://goo.gl/9HmSzc). For other languages, see for example the German GERMEVAL (https://goo.gl/uZEerk).

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4296–4305
July 5 - 10, 2020.

COMMENT WITH TOXICITY AMPLIFIED IN CONTEXT	
PARENT	But what if the user is a lesbian? Then what?
TARGET	**"Pigs Are People Too". "Avant-garde a clue"**
COMMENT WITH TOXICITY MITIGATED IN CONTEXT	
PARENT	Hmmm. The flame on top of the gay pride emblem can probably be interpreted in a manner that I did not consider. Perhaps one icon on each end using?
TARGET	**Hi Gadget, interpreted in what manner? Flaming gays? Or Burn a gay?**

Table 2: Examples of comments that the annotators labeled differently when the previous (parent) comment was (or not) provided. In the top example, the target comment (the one being annotated) was labeled as toxic only when context was given. In the bottom example, the target comment was considered toxic only without its parent comment.

- RQ2: *Does context actually improve the performance of toxicity classifiers, when they are made context-aware? And how can toxicity classifiers be made context-aware?*

To investigate these questions we created and make publicly available two new toxicity datasets that include context, which are based on discussions in Wikipedia Talk Pages (Hua et al., 2018). The first one is a small dataset of 250 comments, created in an AB test fashion, where two different groups of annotators (crowd-workers) were employed. One group annotated the comments without context, while the other group was given the same comments, this time along with the parent comment and the title of the thread as context. We used this dataset to show that the perceived toxicity of a significant subset of posts (5.2% in our experiment) changes when context is (or is not) provided. We conclude that a small but significant subset of manually labeled posts end up having wrong toxicity labels if the annotators are not provided with context. We also found that context can both amplify (approximately 3.6% of comments in our experiment) and mitigate (approx. 1.6%) the perceived toxicity. Examples of comments that were differently labeled with and without context are shown in Table 2.

To investigate the second question, concerning the effect of context on the performance of toxicity classifiers, we created a larger dataset of 20k comments; 10k comments were annotated out of context, 10k in context. This time we did not require the *same* comments to be annotated with and without context, which allowed us to crowd-source the collection of a larger set of annotations. These two new subsets were used to train several toxicity detection classifiers, both context-aware and context-unaware, which were evaluated on held out comments that we always annotated in context (based on the assumption that in-context labels are more reliable). Surprisingly, we found no evidence that context actually improves the performance of toxicity classifiers. We tried a range of classifiers and mechanisms to make them context aware, and having also considered the effect of using gold labels obtained out of context or by showing context to the annotators. This finding is likely related to the small number of context-sensitive comments. In turn this suggests that an important direction for further research is how to efficiently annotate larger corpora of comments in context. We make our code and data publicly available.[3]

2 Related Work

Toxicity detection has attracted a lot of attention in recent years (Nobata et al., 2016; Pavlopoulos et al., 2017b; Park and Fung, 2017; Wulczyn et al., 2017). Here we use the term 'toxic' as an umbrella term, but we note that the literature uses several terms for different kinds of toxic language or related phenomena: 'offensive' (Zampieri et al., 2019a), 'abusive' (Pavlopoulos et al., 2017a), 'hateful' (Djuric et al., 2015; Malmasi and Zampieri, 2017; ElSherief et al., 2018; Gambäck and Sikdar, 2017; Zhang et al., 2018), etc. There are also taxonomies for these phenomena based on their directness (e.g., whether the abuse was unambiguously implied/denoted or not), and their target (e.g., whether it was a general comment or targeting an individual/group) (Waseem et al., 2017). Other hierarchical taxonomies have also been defined (Zampieri et al., 2019a). While most previous work does not address toxicity in general, instead addressing particular subtypes, toxicity and its subtypes are strongly related, with systems trained to detect toxicity being effective also at subtypes, such as hateful language (van Aken et al., 2018). As is customary in natural language processing, we focus on aggregate results when hoping to answer our research questions, and leave largely unanswered the related epistemological questions when this does not preclude using classifiers in real-world applications.

Table 3 lists all currently available public datasets for the various forms of toxic language that we are aware of. The two last columns show that

[3] https://github.com/ipavlopoulos/context_toxicity

Dataset Name	Source	Size	Type	Lang.	C_a	C_t
CCTK	Civil Comments Toxicity Kaggle	2M	Toxicity sub-types	EN	✗	-
CWTK	Wikipedia Toxicity Kaggle	223,549	Toxicity sub-types	EN	✗	-
Davidson et al. (2017)	Twitter	24,783	Hate/Offense	EN	✗	-
Zampieri et al. (2019a)	Twitter	14,100	Offense	EN	✗	-
Waseem and Hovy (2016)	Twitter	1,607	Sexism/Racism	EN	✗	-
Gao and Huang (2017)	Fox News	1,528	Hate	EN	✓	Title
Wiegand et al. (2018)	Twitter	8541	Insult/Abuse/Profanity	DE	✗	-
Ross et al. (2016)	Twitter	470	Hate	DE	✗	-
Pavlopoulos et al. (2017a)	Gazzetta.gr	1,6M	Rejection	EL	✓	-
Mubarak et al. (2017)	Aljazeera.net	31,633	Obscene/Offense	AR	✓	Title

Table 3: Publicly available datasets for toxicity detection. The Size column shows the number of comments. Column C_a shows if annotation was context-aware or not. Column C_t shows the type of context provided. Pavlopoulos et al. (2017a) used professional moderator decisions, which were context-aware, but context is not included in their dataset. The datasets of Gao and Huang (2017) and Mubarak et al. (2017) include context-aware labels, but provide only the titles of the news articles being discussed.

no existing English dataset provides both context (e.g., parent comment) and context-aware annotations (annotations provided by humans who also considered the parent comment).

Both small and large toxicity datasets have been developed, but approximately half of them contain tweets, which makes reusing the data difficult, because abusive tweets are often removed by the platform. Moreover, the textual content is not available under a license that allows its storage outside the platform. The hateful language detection dataset of Waseem and Hovy (2016), for example, contains 1,607 sexism and racism annotations for IDs of English tweets. A larger dataset was published by Davidson et al. (2017), containing approx. 25k annotations for tweet-IDs, collected using a lexicon of hateful terms. Research on forms of abusive language detection is mainly focused on English (6 out of 10 datasets), but datasets in other languages also exist, such as Greek (Pavlopoulos et al., 2017a), Arabic (Mubarak et al., 2017), and German (Ross et al., 2016; Wiegand et al., 2018).

A common characteristic of most of the datasets listed in Table 3 is that, during annotation, the human workers were not provided with, nor instructed to review, the context of the target text. Context such as the preceding comments in the thread, or the title of the article being discussed, or the discussion topic. A notable exception is the work of Gao and Huang (2017), who annotated hateful comments under Fox News articles by also considering the title of the news article and the preceding comments. However, this dataset has three major shortcomings. First, the dataset is very small, comprising approximately 1.5k posts retrieved from the discussion threads of only 10 news articles. Second, the authors did not release sufficient information to reconstruct the threads and allow systems to consider the parent comments. Third, only a single annotator was used for most of the comments, which makes the annotations less reliable.

Two other datasets, both non English, also include context-aware annotations. Mubarak et al. (2017) provided the title of the respective news article to the annotators, but ignored parent comments. This is problematic when new comments change the topic of the discussion and when replies require the previous posts to be judged. Pavlopoulos et al. (2017a) used professional moderators, who were monitoring entire threads and were thus able to use the context of the thread to judge for the toxicity of the comments. However, the plain text of the comments for this dataset is not available, which makes further analysis difficult. Moreover, crucially for this study, the context of the comments was not released in any form.

In summary, of the datasets we know of (Table 3), only two include context (Gao and Huang, 2017; Mubarak et al., 2017), and this context is limited to the title of the news article the comment was about. As discussed above, Gao and Huang (2017) include the parent comments in their dataset, but without sufficient information to link the target comments to the parent ones. Hence *no toxicity dataset includes the raw text of both target and parent comments with sufficient links between the two*. This means that toxicity detection methods cannot exploit the conversational context when being trained on existing datasets.

Using previous comments of a conversation or preceding sentences of a document is not uncommon in text classification and language modeling. Mikolov and Zweig (2012), for example, used LDA to encode the preceding sentences and pass the en-

Dataset Statistics	CAT-SMALL	CAT-LARGE
#comments (N/C)	250	10k/10k
avg. length (N/C)	100	161/161
#toxic (GN/GC)	11/16	59/151

Table 4: Dataset statistics. CAT-SMALL contains 250 comments. CAT-LARGE contains 10k comments without (N) and 10k comments with context (C). Average length in characters. GN is the group of annotators with no access to context, and GC the group with context. For each comment and group of annotators, the toxicity scores of the annotators were averaged and rounded to the nearest binary decision (toxic, non-toxic) to compute the number of toxic comments (#toxic).

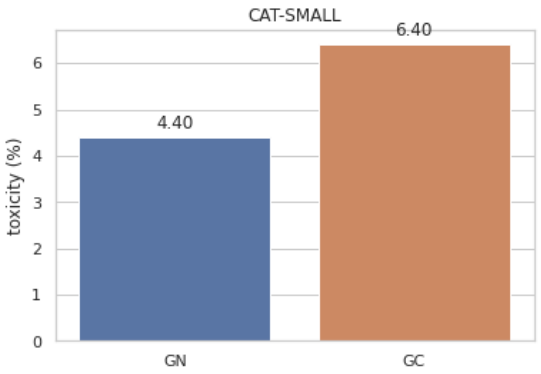

Figure 1: Toxicity ratio (%) of the comments of CAT-SMALL when using the toxicity labels of GN (annotators with no context) or GC (annotators with context). The difference is statistically significant ($P < .01$).

coded sentence history to an RNN language model (Blei et al., 2003). Their approach achieved state of the art language modeling results and was used as an alternative solution (e.g., to LSTMs) for the problem of vanishing gradients. Sordoni et al. (2015) experimented with concatenating consecutive utterances (or their representations) before passing them to an RNN to generate conversational responses. They reported gains up to 11% in BLEU (Papineni et al., 2002). Ren et al. (2016) reported significant gains in Twitter sentiment classification, when adding contextual features.

3 Experiments

3.1 Experiments with CAT-SMALL for RQ1

To investigate how often context affects the perceived toxicity of posts, we created CAT-SMALL, a small Context-Aware Toxicity dataset of 250 randomly selected comments from the Wikipedia Talk Pages (Table 4). We gave these comments to two groups of crowd-workers to judge their toxicity. The first group (GC, Group with Context) was also given access to the parent comment and the discussion title, while the second group (GN, Group with No context) was provided with no context. No annotator could belong to both groups, to exclude the case of an annotator having seen the context of a post and then being asked to label the same post without its context. We used the Figure Eight crowd-sourcing platform, which provided us with these mutually exclusive groups of annotators.[4] We collected three judgments per comment, per group. All comments were between 10 and 400 characters long. Their depth in their threads was from 2 (direct reply) to 5.

We used the parent comment and discussion title only, instead of a larger context (e.g., the entire thread), to speed up our machine learning experiments, and also because reading only the previous comment and the discussion title made the manual annotation easier. In preliminary experiments, we observed that including more preceding comments had the side effect of workers tending to ignore the context completely.[5] We addressed this problem by asking the annotators an extra question: "Was the parent comment less, more, or equally toxic?"

For each comment and group of annotators, the toxicity scores of the annotators were first averaged and rounded to the nearest binary decision, as in Table 4. Figure 1 shows that the toxicity ratio (toxic comments over total) of CAT-SMALL is higher when annotators are given context (GC), compared to when no context is provided (GN). A one-sided Wilcoxon-Mann-Whitney test shows this is a statistically significant increase. This is a first indication that providing context to annotators affects their decisions. The toxicity ratio increases by 2 percentage points (4.4% to 6.4%) when context is provided, but this is an aggregated result, possibly hiding the true size of the effect of context. The perceived toxicity of some comments may be increasing when context is provided, but for other comments it may be decreasing, and these effects may be partially cancelling each other when measuring the change in toxicity ratio.

To get a more accurate picture of the effect of

[4] See https://www.figure-eight.com/. The annotators were high-performing workers from previous jobs. The demographics and backgrounds of the crowdworkers are detailed in Posch et al. (2018).

[5] We experimented with providing the GC annotators with all the parent comments in the discussion. We also experimented with preselection strategies, such as employing the score from a pre-trained toxicity classifier for a stratified selection and using a list of terms related to minority groups.

context, we measured the number of comments of CAT-SMALL for which the (averaged and rounded) toxicity label was different between the two groups (GN, GC). We found that the toxicity of 4 comments out of 250 (1.6%) decreased with context, while the toxicity of 9 comments (3.6%) increased. Hence, perceived toxicity was affected for 13 comments (5.2% of comments). While the small size of CAT-SMALL does not allow us to produce accurate estimates of the frequency of posts whose perceived toxicity changes with context, the experiments on CAT-SMALL indicate that context has a statistically significant effect on the perceived toxicity, and that context can both amplify or mitigate the perceived toxicity, thus making a first step to addressing our first research question (RQ1). Nevertheless, larger annotated datasets need to be developed to estimate more accurately the frequency of context-sensitive posts in online conversations, and how often context amplifies or mitigates toxicity.

3.2 Experiments with CAT-LARGE for RQ2

To investigate whether adding context can benefit toxicity detection classifiers, we could not use CAT-SMALL, because its 250 comments are too few to effectively train a classifier. Thus, we proceeded with the development of a larger dataset. Although the best approach would be to extend CAT-SMALL, which had two mutually exclusive groups of annotators labeling each comment, we found that the annotation process was very slow in that case, largely because of the small size of annotator groups we had access to in Figure Eight (19 and 23 for GC and GN respectively).[6] By contrast, when we did not request mutually exclusive annotator groups, we could get many more workers (196 and 286 for GC and GN respectively) and thus annotation became significantly faster.

For this larger dataset, dubbed CAT-LARGE, we annotated 20k randomly selected comments from Wikipedia Talk Pages. 10k comments were annotated by human workers who only had access to the comment in question (group with no context, GN). The other 10k comments were annotated by providing the annotators also with the parent comment and the title of the discussion (group with context, GC). Each comment was annotated by three workers. We selected comments of length from 10 and 400 characters, with depth in thread from 2 (direct reply) to 5. Inter-annotator agreement was computed with Krippendorff's alpha on 123 texts, and it was found to be 0.72% for GN and 0.70% for GC.

[6]Figure Eight provided us with the two mutually exclusive annotator groups, which could not grow in size.

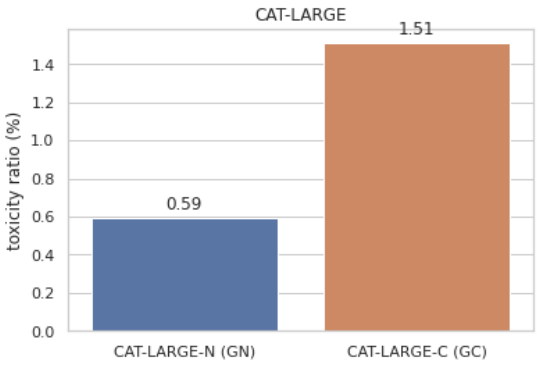

Figure 2: Toxicity ratio (%) of the comments of CAT-LARGE-N (10k comments annotated with no context, left) and CAT-LARGE-C (10k other comments annotated with context, right). For each comment, the toxicity scores of the annotators were first averaged and rounded to the nearest binary decision, as in Table 4. The difference is statistically significant ($P < .001$).

Figure 2 shows that the toxicity ratio increased (from 0.6% to 1.5%) when context was given to the annotators. A one-sided Wilcoxon-Mann-Whitney test shows this is a statistically significant increase ($P < .001$). Again, the change of toxicity ratio is an indication that context does affect the perceived toxicity, but it does not accurately show how many comments are affected by context, since the perceived toxicity may increase for some comments when context is given, and decrease for others. Unlike CAT-SMALL, in CAT-LARGE we cannot count for how many comments the perceived toxicity increased or decreased with context, because the two groups of annotators (GN, GC) did not annotate the same comments. The toxicity ratios of CAT-LARGE (Fig. 2) are lower than in CAT-SMALL (Fig. 1), though they both show a trend of increased toxicity ratio when context is provided. The toxicity ratios of CAT-LARGE are more reliable estimates of toxicity in online conversations, since they are based on a much larger dataset.

We used CAT-LARGE to experiment with both context-insensitive and context-sensitive toxicity classifiers. The former only consider the post being rated (the target comment), whereas the latter also consider the context (parent comment).

Context Insensitive Toxicity Classifiers

BILSTM Our first context-insensitive classifier is a bidirectional LSTM (Hochreiter and Schmidhuber, 1997). On top of the concatenated last states (from the two directions) of the BILSTM, we add

a feed-forward neural network (FFNN), consisting of a hidden dense layer with 128 neurons and tanh activations, then a dense layer leading to a single output neuron with a sigmoid that produces the toxicity probability. We fix the bias term of the single output neuron to $\log \frac{T}{N}$, where T and N are the numbers of toxic and non-toxic training comments, respectively, to counter-bias against the majority (non-toxic) class.[7] This BILSTM-based model could, of course, be made more complex (e.g., by stacking more BILSTM layers, and including self-attention), but it is used here mainly to measure how much a relatively simple (by today's standards) classifier benefits when a context mechanism is added (see below).

BERT At the other end of complexity, our second context-insensitive classifier is BERT (Devlin et al., 2019), fine-tuned on the training subset of each experiment, with a task-specific classifier on top, fed with BERT's top-level embedding of the [CLS] token. We use BERT-BASE pre-trained on cased data, with 12 layers and 768 hidden units. We only unfreeze the top three layers during fine-tuning, with a small learning rate (2e-05) to avoid catastrophic forgetting. The task-specific classifier is the same FFNN as in the BILSTM classifier.

BERT-CCTK We also experimented with a BERT model that is the same as the previous one, but fine-tuned on a sample (first 100k comments) of the CCTK dataset (Table 3). We used the general toxicity labels of that dataset, and fine-tuned for a single epoch. The only difference of this model, compared to the previous one, is that it is fine-tuned on a much larger training set, which is available, however, only without context (no parent comments). The annotators of the dataset were also not provided with context (Table 3).

PERSPECTIVE The third context-insensitive classifier is a CNN-based model for toxicity detection, trained on millions of user comments from online publishers. It is publicly available through the PERSPECTIVE API.[8] The publicly available form of this model cannot be retrained, fine-tuned, or modified to include a context-awareness component. Like BERT-CCTK, this model uses an external (but now much larger) labeled training set. This training set is not publicly available, it does not include context, and was labeled by annotators who were not provided with context.

[7] See an example in http://tiny.cc/m572gz.

[8] https://www.perspectiveapi.com/

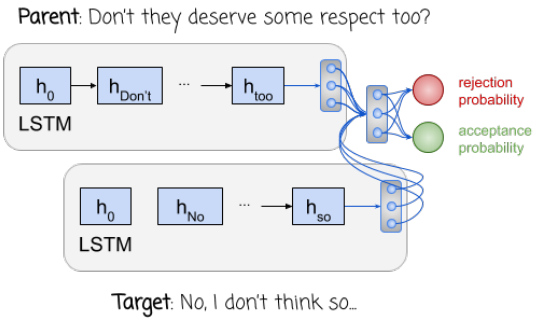

Figure 3: Illustration of CA-BILSTM-BILSTM. Two BILSTMs, shown unidirectional for simplicity, encode the parent and target comment. The concatenation of the vector representations of the two comments is then passed to a FFNN.

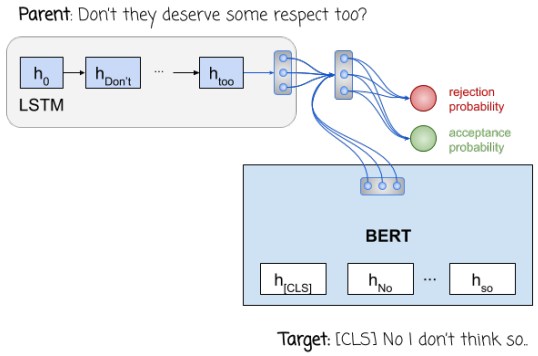

Figure 4: Illustration of CA-BILSTM-BERT. BERT encodes the target comment. BILSTM (shown unidirectional for simplicity) encodes the parent comment. The vector representations of the two comments are concatenated and passed to a FFNN.

Context Sensitive Toxicity Classifiers

CA-BILSTM-BILSTM In a context-aware extension of the context-insensitive BILSTM classifier, dubbed CA-BILSTM-BILSTM, we added a second BILSTM to encode the parent comment (Fig. 3). The vector representations of the two comments (last states from the two directions of both BILSTMs) are concatenated and passed to a FFNN, which is otherwise identical to the FFNN of the context-insensitive BILSTM.

CA-BILSTM-BERT We also used a BILSTM to encode the parent in a context-aware extension of the BERT-based classifier, called CA-BILSTM-BERT (Fig. 4). Now BERT encodes the target comment, whereas a BILSTM (the same as in CA-BILSTM-BILSTM) encodes the parent. (We could not use two BERT instances to encode both the parent and the target comment, because the resulting model did not fit in our GPU.) The concatenated representations of the two comments are passed to a FFNN,

which is otherwise the same as as in previous models. BERT is fine-tuned on the training subset, as before, and the BILSTM encoder of the parent is jointly trained (with a larger learning rate).

CA-SEP-BERT We also experimented with another context-aware version of the BERT-based classifier, dubbed CA-SEP-BERT. This model concatenates the text of the parent and target comments, separated by BERT's [SEP] token, as in BERT's next sentence prediction pre-training task (Fig. 5). Unlike CA-BILSTM-BERT, it does not use a separate encoder for the parent comment. The model is again fine-tuned on the training subset.

CA-CONC-BERT-CCTK, CA-CONC-PERSPECTIVE These are exactly the same as BERT-CCTK and PERSPECTIVE, respectively, trained on the same data as before (no context), but at test time they are fed with the concatenation of the text of the parent and target comment, as a naive context-awareness mechanism.

Context Sensitive vs. Insensitive Classifiers

Table 5 reports ROC AUC scores, averaged over a 5-fold Monte Carlo (MC) cross-validation, i.e., using 5 different random training/development/test splits (Gorman and Bedrick, 2019); we also report the standard error of mean over the folds. The models are trained on the training subset(s) of CAT-LARGE-N (@N models) or CAT-LARGE-C (@C models), i.e., they are trained on comments with gold labels obtained *without* or *with* context shown to the annotators, respectively. All models are always evaluated (in each fold) on the test subset(s) of CAT-LARGE-C, i.e., with gold labels obtained *with* context shown to annotators, assuming that those labels are more reliable (the annotators had a broader view of the discussion). In each fold (split) of the MC cross-validation, the training, development, and test subsets are 60%, 20%, and 20% of the data, respectively, preserving in each subset the toxicity ratio of the entire dataset. We always use the test (and development) subsets of CAT-LARGE-C, as always noted. We report ROC AUC, because both datasets are heavily unbalanced, with toxic comments being rare (Fig. 2). [9]

A first observation from Table 5 is that the best results are those of PERSPECTIVE, BERT-CCTK, and their context-aware variants (last four rows).

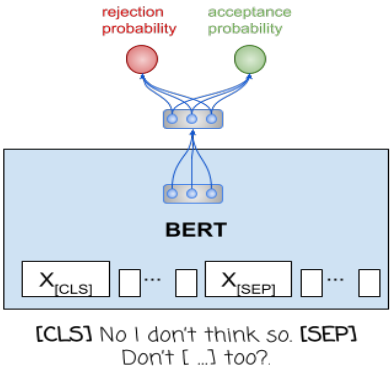

Figure 5: Illustration of CA-SEP-BERT. A single BERT instance encodes the parent and target comments, separated by [SEP]. The top-level representation of the [CLS] token is passed to a FFNN.

This is not surprising, since these systems were trained (fine-tuned in the case of BERT-CCTK) on much larger toxicity datasets than the other systems (upper two zones of Table 5), and BERT-CCTK was also pre-trained on even larger corpora.

What is more surprising is that *any kind of information about the context does not lead to any consistent (or large) improvement in system performance*. PERSPECTIVE and BERT-CCTK seem to improve slightly with the naive context-awareness mechanism of concatenating the parent and target text during testing, but the improvement is very small and we did not detect a statistically significant difference.[10] Training with gold labels obtained from annotators that had access to context (@C models) also leads to no consistent (or large) gain, compared to training with gold labels obtained out of context (@N models). This is probably due to the fact that context-sensitive comments are few (5.2% in the experiments on CAT-SMALL) and, hence, any noise introduced by using gold labels obtained out of context does not significantly affect the performance of the models.

There was also no consistent (or large) improvement when encoding the parent comments with a BILSTM (CA-BILSTM-BILSTM, CA-BILSTM-BERT) or directly as in BERT's next sentence prediction pre-training task (CA-SEP-BERT). This is again probably a consequence of the fact that context-sensitive comments are few. The small number of context-sensitive comments does not allow the BILSTM- and BERT-based classifiers to learn how to use the context encodings to cope with

[9] Recall that we also fix the bias term of the output neuron of each model (apart from PERSPECTIVE) to $-\log \frac{T}{N}$, to bias against the majority class. We also tried under-sampling to address class imbalance, but this technique worked best.

[10] We used single-tailed stratified shuffling (Dror et al., 2018; Smucker et al., 2007), $P < 0.01$, 10,000 repetitions, 50% swaps in each repetition.

model @training	ROC AUC @C
BILSTM @N	56.48±1.42
BILSTM @C	56.38±1.51
CA-BILSTM-BILSTM @N	56.13±1.27
CA-BILSTM-BILSTM @C	58.00±2.70
BERT @N	75.94±2.73
BERT @C	73.49±1.49
CA-BILSTM-BERT @N	74.60 ±3.08
CA-BILSTM-BERT @C	74.46±1.84
CA-SEP-BERT @N	73.29±3.89
CA-SEP-BERT @C	73.54±3.36
PERSPECTIVE	79.27±2.87
CA-CONC-PERSPECTIVE	81.89 ± 2.79
BERT-CCTK	78.08±1.50
CA-CONC-BERT-CCTK	81.69±2.22

Table 5: ROC AUC scores (%) averaged over five-fold MC cross-validation (and standard error of mean) for models trained on CAT-LARGE-N (@N models, gold labels obtained *without* showing context) or CAT-LARGE-C (@C models, gold labels obtained *with* context). All models evaluated on the test subset of CAT-LARGE-C (AUC @C, gold labels obtained *with* context). PERSPECTIVE and BERT-CCTK were trained on larger external training sets with no context, but are tested on the same test subset (in each fold) as the other models.

context-sensitive comments, and failing to cope with context-sensitive comments does not matter much during testing, again since context-sensitive comments are so few.

We conclude for our second research question (RQ2) that we found no evidence that context actually improves the performance of toxicity classifiers, having tried both simple (BILSTM) and more powerful classifiers (BERT), having experimented with several methods to make the classifiers context aware, and having also considered the effect of gold labels obtained out of context vs. gold labels obtained by showing context to annotators.

4 Conclusions and Future Work

We investigated the role of context in detecting toxicity in online comments. We collected and share two datasets for investigating our research questions around the effect of context on the annotation of toxic comments (RQ1) and its detection by automated systems (RQ2). We showed that context does have a statistically significant effect on toxicity annotation, but this effect is seen in only a narrow slice (5.2%) of the (first) dataset. We also found no evidence that context actually improves the performance of toxicity classifiers, having tried both simple and more powerful classifiers, having experimented with several methods to make the classifiers context aware, and having also considered the effect of gold labels obtained out of context vs. gold labels obtained by showing context to the annotators. The lack of improvement in system performance seems to be related to the fact that context-sensitive comments are infrequent, at least in the data we collected.

A limitation of our work is that we considered a narrow contextual context, comprising only the previous comment and the discussion title.[11] It would be interesting to investigate in future work ways to improve the annotation quality when more comments in the discussion thread are provided, and also if our findings hold when broader context is considered (e.g., all previous comments in the thread, or the topic of the thread as represented by a topic model). Another limitation of our work is that we used randomly sampled comments. The effect of context may be more significant in conversations about particular topics, or for particular conversational tones (e.g. sarcasm), or when they reference communities that are frequently the target of online abuse. Our experiments and datasets provide an initial foundation to investigate these important directions.

Acknowledgments

We thank the anonymous reviewers for their comments. This research was funded in part by Google.

References

B. van Aken, J. Risch, R. Krestel, and A. Löser. 2018. Challenges for toxic comment classification: An in-depth error analysis. In *2nd Workshop on Abusive Language Online*, pages 33–42, Brussels, Belgium.

D. M. Blei, A. Y Ng, and M. I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500. Association for Computing Machinery.

[11] The discussion title was used only by the human annotators that examined context, not by context-aware systems, which considered only the parent comment.

T. Davidson, D. Warmsley, M. Macy, and I. Weber. 2017. Automated hate speech detection and the problem of offensive language. In *ICWSM*, pages 512–515, Montreal, Canada.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, page 4171–4186, Minneapolis,MN,USA.

N. Djuric, J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. 2015. Hate speech detection with comment embeddings. In *ICWWW*, pages 29–30.

R. Dror, G. Baumer, S. Shlomov, and R. Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *ACL*, Melbourne, Australia.

M. ElSherief, V. Kulkarni, D. Nguyen, W. Y. Wang, and E. Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. *arXiv preprint*.

B. Gambäck and U. K. Sikdar. 2017. Using convolutional neural networks to classify hate-speech. In *1st Workshop on Abusive Language Online*, pages 85–90, Vancouver, Canada.

L. Gao and R. Huang. 2017. Detecting online hate speech using context aware models. In *RANLP*, pages 260–266.

Kyle Gorman and Steven Bedrick. 2019. We need to talk about standard splits. In *ACL*, pages 2786–2791.

S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran. 2017. Deceiving Google's perspective api built for detecting toxic comments. *arXiv preprint*.

Y. Hua, C. Danescu-Niculescu-Mizil, D. Taraborelli, N. Thain, J. Sorensen, and L. Dixon. 2018. Wikiconv: A corpus of the complete conversational history of a large online collaborative community. *arXiv preprint*.

R. Kumar, A. K. Ojha, S. Malmasi, and M. Zampieri. 2018. Benchmarking aggression identification in social media. In *TRAC*, Santa Fe, USA.

S. Malmasi and M. Zampieri. 2017. Detecting hate speech in social media. In *RANLP*, pages 467–472.

T. Mikolov and G. Zweig. 2012. Context dependent recurrent neural network language model. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 234–239. IEEE.

H. Mubarak, K. Darwish, and W. Magdy. 2017. Abusive language detection on arabic social media. In *1st Abusive Language Workshop*, pages 52–56, Vancouver, Canada.

C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang. 2016. Abusive language detection in online user content. In *ICWWW*, pages 145–153.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

J. H. Park and P. Fung. 2017. One-step and two-step classification for abusive language detection on twitter. In *1st Workshop on Abusive Language Online*, pages 41–45.

J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos. 2017a. Deep learning for user comment moderation. In *1st Workshop on Abusive Language Online*, pages 25–35.

J. Pavlopoulos, P. Malakasiotis, and I. Androutsopoulos. 2017b. Deeper attention to abusive user content moderation. In *EMNLP*, pages 1125–1135, Copenghagen, Denmark.

L. Posch, A. Bleier, F. Flöck, and M. Strohmaier. 2018. Characterizing the global crowd workforce: A cross-country comparison of crowdworker demographics. *arXiv preprint*.

Y. Ren, Y. Zhang, M. Zhang, and D. Ji. 2016. Context-sensitive twitter sentiment classification using neural network. In *30th AAAI Conference on Artificial Intelligence*.

B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, and M. Wojatzki. 2016. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. In *NLP4CMC*, Bochum, Germany.

M. D. Smucker, J. Allan, and B. Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *CIKM*, pages 623–632. ACM.

A. Sordoni, M. Galley, M. Auli, C. Brockett, Y. Ji, M. Mitchell, J.-Y. Nie, J. Gao, and B. Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint*.

Z. Waseem, T. Davidson, D. Warmsley, and I. Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *1st Workshop on Abusive Langauge Online*, Vancouver, Canada.

Z. Waseem and D. Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *NAACL SRW*, pages 88–93, San Diego, California.

M. Wiegand, M. Siegel, and J. Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proceedings of GermEval*.

E. Wulczyn, N. Thain, and L. Dixon. 2017. Ex machina: Personal attacks seen at scale. In *ICWWW*, pages 1391–1399.

M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. 2019a. Predicting the Type and Target of Offensive Posts in Social Media. In *NAACL*.

M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, and R. Kumar. 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *SemEval*.

Z. Zhang, D. Robinson, and J. Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *Lecture Notes in Computer Science*. Springer Verlag.

A Data Annotation

Annotators were asked to judge the toxicity of each comment, given the following definitions:

- VERY TOXIC: A very hateful, aggressive, disrespectful comment or otherwise very likely to make a user leave a discussion or give up on sharing their perspective.
- TOXIC: A rude, disrespectful, unreasonable comment or otherwise somewhat likely to make a user leave a discussion or give up on sharing their perspective.
- UNSURE: Due to polysemy, lack of context or other reasons.
- NOT TOXIC: Not containing any toxicity.

For annotation, we used the 'Figure Eight' platform and we invested 5 cents per row.[12] For the CAT-SMALL we employed high accuracy annotators (i.e., from zone 3), selected from 7 English speaking countries (i.e., UK, Ireland, USA, Canada, New Zealand, South Africa, Australia), and only ones allowing explicit content (we also warned about the explicit content in the title). 62 quiz questions were used. For the CAT-LARGE, we invested the same amount of money but all the annotators were able to participate (again, they ware warned for the explicit content). Inter annotator agreement was measured on the quiz questions with Krippendorff's alpha and was found to be 70% and 72% for the C and N sets.

GC annotators had one more question, which was asking them to compare the toxicity of the target comment to that of the parent comment. The main scope of that question was to make it less easy for annotators to ignore the parent comment.

[12] https://www.figure-eight.com/

B Hyper parameters

All systems were trained for 100 epochs with patience of 3 epochs. We performed early stopping by monitoring the validation ROC AUC.

BILSTM

The hidden size of the LSTM cells had size 128. We used batch size 128, max length 512, and we concatenated the forward and backward last hidden states before the FFNN. We used binary cross entropy for loss and Adam optimizer was used with default parameters (learning rate 1e-03).

CA-BILSTM-BILSTM

We used the same hyper-parameters with BILSTM but included one more bidirectional LSTM to encode the parent text. The parent biLSTM had 64 hidden nodes and we concatenated the forward and backward last hidden states. The parent and the target embeddings (the ones generated by the two biLSTMS) were concatenated before being passed to the FFNN.

BERT

We used a learning rate of 2e-05 for BERT and only unfroze the top three layers during training to our data. On top of the BERT [CLS] representation, we added a FFNN of 128 hidden nodes and a sigmoid to yield the toxicity probability. 128 tokens were used as maximum sequence length.

CA-SEP-BERT

A [SEP] token separated the two texts and the [CLS] token was used as with BERT. Same parameters with BERT were used.

CA-BILSTM-BERT

We used a bidirectional LSTM to encode the parent comment, similarly to CA-BILSTM-BILSTM. The biLSTM representation was concatenated with the [CLS] representation before the FFNN. All other parameters were set to the same values as BERT.