Context Sensitivity Estimation in Toxicity Detection

Alexandros Xenos[♣], **John Pavlopoulos**[♠♣*], **Ion Androutsopoulos**[♣]
[♣]Department of Informatics, Athens University of Economics and Business, Greece
[♠]Department of Computer and Systems Sciences, Stockholm University, Sweden
{a.xenos20,annis,ion}@aueb.gr

Abstract

User posts whose perceived toxicity depends on the conversational context are rare in current toxicity detection datasets. Hence, toxicity detectors trained on current datasets will also disregard context, making the detection of context-sensitive toxicity a lot harder when it occurs. We constructed and publicly release a dataset of 10k posts with two kinds of toxicity labels per post, obtained from annotators who considered (i) both the current post and the previous one as context, or (ii) only the current post. 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 new dataset, we show that systems can be developed for this task. Such systems could be used to enhance toxicity detection datasets with more context-dependent posts, or to suggest when moderators should consider the parent posts, which may not always be necessary and may introduce an additional cost.

1 Introduction

Online fora are used to facilitate discussions, but hateful, insulting, identity-attacking, profane, or otherwise abusive posts may also occur. These posts are called toxic (Borkan et al., 2019) 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 abusive language) detection systems. What most of these systems have in common, besides aiming to promote healthy discussions online (Zhang et al., 2018), is that they disregard the conversational context (e.g., the parent post in the discussion), making the detection of context-sensitive toxicity a lot harder. For instance, the post "Keep the hell out" may be considered as toxic by a moderator, if the previous (parent) post "What was the title of that 'hell out' movie?" is ignored. Although toxicity datasets that include conversational context have recently started to appear, in previous work we showed that context-sensitive posts are still too few in those datasets (Pavlopoulos et al., 2020), which does not allow models to learn to detect context-dependent toxicity. In this work, we focus on this problem. We constructed and publicly release a context-aware dataset of 10k 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 context.[1]

As a first step towards studying context-dependent toxicity, we limit the conversational context to the previous (parent) post of the thread, as in our previous work (Pavlopoulos et al., 2020). We use the 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 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 the parent posts; the latter may not always be necessary and may also introduce additional cost.

2 The dataset

To build the dataset of this work, we used the also publicly available Civil Comments (CC) dataset (Borkan et al., 2019). CC was originally annotated by ten annotators per post, but the parent post

*Corresponding author.

[1]The dataset is released under a CC0 licence. See http://nlp.cs.aueb.gr/publications.html for the link to download it.

Proceedings of the Fifth Workshop on Online Abuse and Harms, pages 140–145

(the previous post in the thread) was not shown to the annotators. We randomly sampled 10,000 CC posts and gave both the target and the parent post to the 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. We unified the latter two labels in both CC and CCC annotations to simplify the problem. To obtain the new in-context labels of CCC, we used the APPEN platform and five high accuracy annotators per post (annotators from zone 3, allowing adult and warned for explicit content), selected from 7 English speaking countries, namely: UK, Ireland, USA, Canada, New Zealand, South Africa, and Australia.[2]

The free-marginal kappa (Randolph, 2010) of the CCC annotations is 83.93%, while the average (mean pairwise) percentage agreement is 92%. In only 71 posts (0.07%) an annotator said UNSURE, i.e., annotators were confident in their decisions most of the time. We exclude these 71 posts from our study, as they are too few. The average length of target posts in CCC is only slightly lower than that of parent posts. Fig. 1 shows this counting the length in characters, but the same holds when counting words (56.5 vs. 68.8 words on average). 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 (i.e., all toxicity sub-types provided by the annotators were collapsed to a single toxicity label). This is similar to arrangements in the work of Wulczyn et al. (2017), who also found that training using the empirical distribution (over annotators) of the toxic labels (a continuous score per post) leads to better toxicity detection performance, compared to using labels reflecting the majority opinion of the raters (a binary label per post). See also Fornaciari et al. (2021).

Combined with the original (out of context) annotations of the 10k posts from CC, the new dataset (CCC) contains 10k posts for which both in-context (IC) and out-of-context (OC) labels are available. Figure 2 shows the number of posts (Y axis) per ground truth toxicity score (X axis). Orange represents the ground truth obtained by annotators who were provided with the parent post when rating (IC), while blue is for annotators who rated the post without context (OC). The vast majority of the

[2]We focused on known English-speaking countries. The most common country of origin was USA.

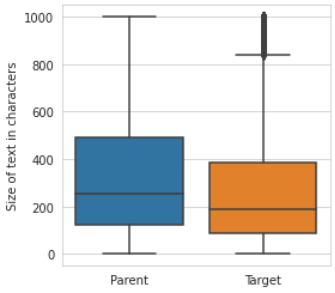

Figure 1: Length of parent/target posts in characters.

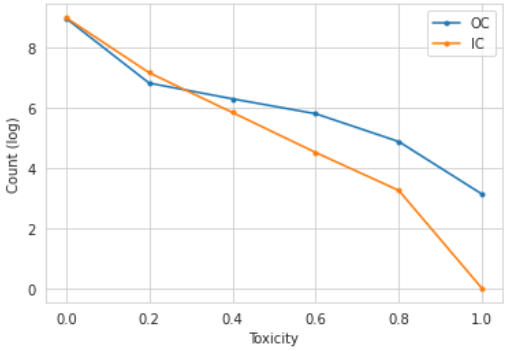

Figure 2: Histogram (converted to curve) of average toxicity according to annotators who were (IC) or were not (OC) given the parent post when annotating.

posts were unanimously perceived 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 is consistent with the findings of our previous work (Pavlopoulos et al., 2020), where we observed that when the parent post is provided, the majority of the annotators perceive fewer posts as toxic, compared to showing no context to the annotators. To study this further, in this work we compared the two scores (IC, OC) per post, as discussed below.

For each post p, we define $s^{ic}(p)$ to be the toxicity (fraction of coders who perceived the post as toxic) derived from the IC coders and $s^{oc}(p)$ to be the toxicity derived from the 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. Fig. 3 shows that δ is most often zero, 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. In numbers, in 66.1% of the posts the toxicity score remained unchanged while out of the remaining 33.9%, in 9.6% it increased

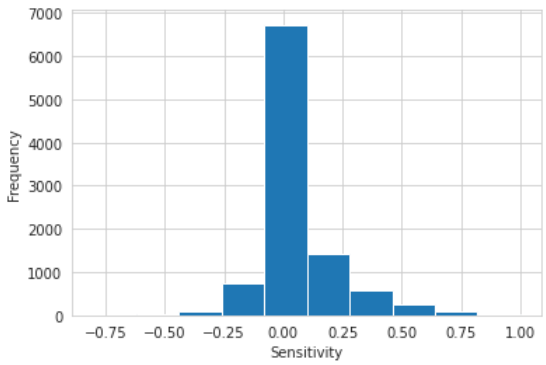

Figure 3: Histogram of context sensitivity. Negative (positive) sensitivity means the toxicity increased (decreased) when context was shown to the annotators.

(960 posts) and in 24.2% it decreased (2,408) when context was provided. If we binarize the ground truth we get a similar trend, but with the toxicity of more posts remaining unchanged (i.e., 94.7%).

When counting the number of posts for which $|\delta|$ exceeds a threshold t, called *context-sensitive posts* in Fig. 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) are rare. Some examples of target posts, along with their parent posts and δ, are shown in Table 1.

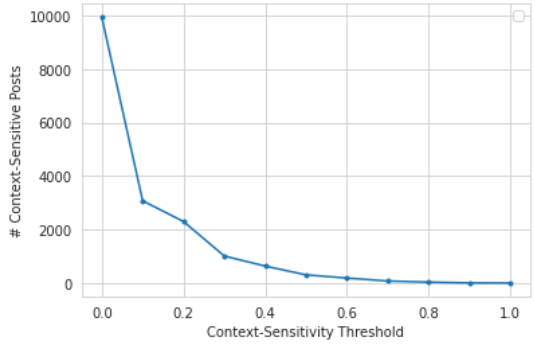

Figure 4: Number of context-sensitive posts ($|\delta| \geq t$), when varying the context-sensitivity threshold t.

3 Experimental Study

Initially, we used our dataset to experiment with existing toxicity detection systems, aiming to investigate if context-sensitive posts are 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 is 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 to classify CCC posts as toxic or not.[3] We either concatenate the parent post to the target one to allow the model to "see" the parent, or not.[4] Figure 5 shows the Mean Absolute Error (MAE) of Perspective, with and without the parent post concatenated, when evaluating on all the CCC posts ($t = 0$) and when evaluating on smaller subsets with increasingly context-sensitive posts ($t > 0$). 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 (Fig. 4).

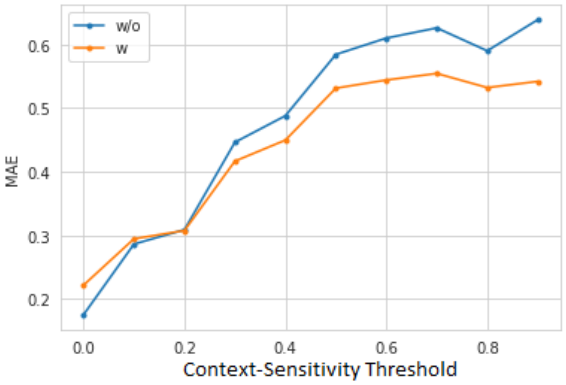

Figure 5: Mean Absolute Error (Y-axis) when *predicting toxicity* for different context-sensitivity thresholds (t; X-axis). We applied Perspective to target posts alone (w/o) or concatenating the parent posts (w).

Figure 5 shows that when we concatenate the parent to the target post (w), MAE is clearly smaller, provided that $t \geq 0.2$. 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 is imposed ($t = 0$) when constructing a dataset, the non-context sensitive posts ($|\delta| = 0$) dominate (Fig. 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 are too rare and, thus, context-aware models do not perform better on existing toxicity datasets.

It is worth observing that the more we move to the right of Fig. 5, the higher the error for both Per-

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

[4] We are investigating better context-aware models.

PARENT OF POST p	POST p	$s^{OC}(p)$	$s^{IC}(p)$	δ
Oh Don..... you are soooo predictable.	oh Chuckie you are such a tattle tale.	36.6%	80%	-43.4%
Oh Why would you wish them well? They've destroyed the environment in their country and now they are coming here to do the same.	"They"? Who is they? Do all Chinese look alike to you? Or are you just revealing your innate bigotry and racism?	70%	0%	70%

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

spective variants (with, without context). This is probably due to the fact that Perspective is trained on posts that have been rated by annotators who were not provided with the parent post (out of context; OC), whereas here we use the in-context (IC) annotations as ground truth. The greater the t in Fig. 5, the larger the difference between the toxicity scores of OC and IC annotators, hence the larger the difference between the (OC) ground truth that Perspective saw and the ground truth that we use here (IC). Experimenting with artificial parent posts (long or short, toxic or not) confirmed that the error increases for context-sensitive posts.

The solution to the problem of increasing error as context sensitivity increases (Fig. 5) would be to train toxicity detectors on datasets that are richer in context-sensitive posts. However, such posts are rare (Fig. 4) and thus, they are hard to collect and annotate. This observation motivated the experiments of the next section, where we train ***context-sensitivity*** detectors, which allow us to collect posts that are likely to be context-sensitive. These posts can 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 (BERTr). The first three regressors use TF-IDF features. In the case of BERTr, we add 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. Preliminary experiments showed that adding simplistic context-mechanisms (e.g., concatenating the parent post) to the context sensitivity regressors does not lead to improvements. This may be due

	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)
RFs	2.2 (0.1)	11.2 (0.2)	21.57 (1.0)	59.67 (0.3)
BERTr	**1.8 (0.1)**	**9.2 (0.3)**	**42.01 (4.3)**	**80.46 (1.3)**

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.

to the fact that it is often possible to decide if a post is *context-sensitive* or not (we do 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 is 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, respectively, and we performed Monte Carlo 3-fold Cross Validation. We used mean square error (MSE) as our loss function and early stopping with patience of 5 epochs. Table 2 presents the MSE and the mean absolute error (MAE) of all the models on the test set. Unsurprisingly, BERTr outperforms the rest of the models in MSE and MAE. Previous work (Wulczyn et al., 2017) reported that training toxicity regressors (based on the empirical distribution of codes) instead of classifiers (based on the majority of the codes) leads to improved classification results too, so 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: $t(p) = \text{SEM}^{oc}(p) + \text{SEM}^{ic}(p)$. By using this binary ground truth, AUPR and AUC ver-

ified that BERTr outperforms the rest of the models, even when the models are used as classifiers.

4 Related Work

Following the work of Borkan et al. (2019), this work uses toxicity as an umbrella term for hateful, identity-attack, insulting, profane or posts that are toxic in another way. Toxicity detection is a popular task that has been addressed by machine learning approaches (Davidson et al., 2017; Waseem and Hovy, 2016; Djuric et al., 2015), including deep learning approaches (Park and Fung, 2017; Pavlopoulos et al., 2017b,c; Chakrabarty et al., 2019; Badjatiya et al., 2017; Haddad et al., 2020; Ozler et al., 2020). Despite the plethora of computational approaches, what most of these have in common is that they disregard context, such as the parent post in discussions. The reason for this weakness is that datasets are developed while annotators ignore the context (Nobata et al., 2016; Wulczyn et al., 2017; Waseem and Hovy, 2016). Most of the datasets in the field are in English, but datasets in other languages have the same weakness (Pavlopoulos et al., 2017a; Mubarak et al., 2017; Chiril et al., 2020; Ibrohim and Budi, 2018; Ross et al., 2016; Wiegand et al., 2018). We started to investigate context-sensitivity in toxicity detection in our previous work (Pavlopoulos et al., 2020) using existing toxicity detection datasets and a much smaller dataset (250 posts) we constructed with both IC and OC labels. Comparing to our previous work, here we constructed and released a much larger dataset (10k posts) with IC and OC labels, we introduced the new task of context-sensitivity estimation, and we reported experimental results indicating that the new task is feasible.

5 Conclusions and Future Work

We introduced the task of estimating the context-sensitivity of posts in toxicity detection, i.e., estimating the extent to which the perceived toxicity of a post depends on the conversational context. We constructed, presented, and release a new dataset that can be used to train and evaluate systems for the new task, where context is the previous post. 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. 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.

Acknowledgement

We thank L. Dixon and J. Sorensen for their continuous assistance and advice. This research was funded in part by a Google Research Award.

References

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *WWW*, pages 491–500, San Francisco, USA.

Tuhin Chakrabarty, Kilol Gupta, and Smaranda Muresan. 2019. Pay "attention" to your context when classifying abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 70–79, Florence, Italy. Association for Computational Linguistics.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. An annotated corpus for sexism detection in French tweets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1397–1403, Marseille, France. European Language Resources Association.

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*, 11(1).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th International*

Conference on World Wide Web, WWW '15 Companion, page 29–30, New York, NY, USA. Association for Computing Machinery.

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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

Bushr Haddad, Zoher Orabe, Anas Al-Abood, and Nada Ghneim. 2020. Arabic offensive language detection with attention-based deep neural networks. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 76–81, Marseille, France. European Language Resource Association.

Muhammad Okky Ibrohim and Indra Budi. 2018. A dataset and preliminaries study for abusive language detection in indonesian social media. *Procedia Computer Science*, 135:222–229. The 3rd International Conference on Computer Science and Computational Intelligence (ICCSCI 2018) : Empowering Smart Technology in Digital Era for a Better Life.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, Vancouver, BC, Canada. Association for Computational Linguistics.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, page 145–153, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

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. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 28–33, Online. Association for Computational Linguistics.

Ji Ho Park and Pascale Fung. 2017. One-step and two-step classification for abusive language detection on Twitter. In *Proceedings of the First Workshop on Abusive Language Online*, pages 41–45, Vancouver, BC, Canada. Association for Computational Linguistics.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017a. Deep learning for user comment moderation. In *Proceedings of the First Workshop on Abusive Language Online*, pages 25–35, Vancouver, BC, Canada. Association for Computational Linguistics.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017b. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1125–1135, Copenhagen, Denmark. Association for Computational Linguistics.

John Pavlopoulos, Prodromos Malakasiotis, Juli Bakagianni, and Ion Androutsopoulos. 2017c. Improved abusive comment moderation with user embeddings. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 51–55, Copenhagen, Denmark. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter?

Justus Randolph. 2010. Free-marginal multirater kappa (multirater $\varkappa$free): An alternative to fleiss fixed-marginal multirater kappa. volume 4.

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. In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, pages 6–9.

Nanna Thylstrup and Zeerak Waseem. 2020. Detecting 'dirt' and 'toxicity': Rethinking content moderation as pollution behaviour.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018), Vienna, Austria – September 21, 2018*.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1391–1399, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

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. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.