Capturing Covertly Toxic Speech via Crowdsourcing

Jorge Nario
Google
nario@google.com

Alyssa Lees
Jigsaw
alyssalees@google.com

Ian D. Kivlichan
Jigsaw
kivlichan@google.com

Daniel Borkan
Jigsaw
dborkan@google.com

Nitesh Goyal
Jigsaw
teshg@google.com

Abstract

We study the task of labeling covert or veiled toxicity in online conversations. Prior research has highlighted the difficulty in creating language models that recognize nuanced toxicity such as microaggressions. Our investigations further underscore the difficulty in parsing such labels reliably from raters via crowdsourcing. We introduce an initial dataset, COVERTTOXICITY, which aims to identify and categorize such comments from a refined rater template. Finally, we fine-tune a comment-domain BERT model to classify covertly offensive comments and compare against existing baselines.

1 Introduction

Warning: *this paper contains content that may be offensive or upsetting.*

Tools identifying toxic content are essential for moderation of problematic online conversations (Jigsaw, 2018). Generally, "toxic" is an umbrella label for various types of disrespectful comments, including but not limited to identity attacks, profanity and threats, that could encourage a user to leave a conversation. Besides, ”toxicity”, other terms like ”hate-speech”, and ”online violence” have also been used to refer to similar problems (Chandrasekharan et al., 2017) - yet no single conclusive definition exists. In this work, we chose to use ”toxicity” because our research suggests that ”toxicity” is an easy concept for annotators to understand, meaning we can gather opinions from a diverse range of people, allowing us to capture the inherent subjectivity of the concept. In other words, it was easier for more people to agree on what constituted "toxic" speech—comments likely to make someone leave a conversation—than it was for people to agree on other terms to describe problematic comments.[1]

Previous works have published baseline toxicity detection datasets (Borkan et al., 2019; Wulczyn et al., 2016) along with models for detecting such content (Jigsaw, 2017; Zampieri et al., 2019; Schmidt and Wiegand, 2017; Pavlopoulos et al., 2020; Georgakopoulos et al., 2018).

However, (Han and Tsvetkov, 2020) recently established that covert or veiled toxicity is overwhelmingly misclassified by leading models. Perspective, a free API that uses machine learning to score comments for toxicity, is widely used to aid content moderators on online platforms (Jigsaw, 2017). As example, the deeply offensive comment masked as humor "What's the difference between boy scouts and Jews? Boy scouts come back from camp." is assigned a toxicity probability score of 0.31 by Perspective. Such comments can be as harmful to participants as more overt toxicity (Nadal et al., 2014).

Covert toxicity is an umbrella term which includes types of toxicity that may not be immediately obvious. Covertly toxic comments may use obfuscation, code words, suggestive emojis, dark humor, or sarcasm. It also includes subtle aggression, such as the user comment "'slurp slurp", says Chang', which contains a stereotype regarding Asian individuals. Microaggressions may be unintentional but can have profound psychological repercussions in aggregate (Sue et al., 2007a; Jackson, 2011; Sue et al., 2007b; Stryker and Burke, 2000).

In this work, we seek to close the gap in toxicity models. Our two main contributions include:

- A custom crowd-sourcing instruction template to identify covert toxicity
- A benchmark dataset for identifying covert toxicity.
- An enhanced toxicity model with improved performance in covert/veiled toxicity.

[1] https://jigsaw.google.com/the-current/toxicity/

Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing, pages 14–20
April 20, 2021.

2 Related Work

Ogbonnaya-Ogburu et al. (2020) provides a call for race-conscious efforts within the human-computer interaction community. For example, Dosono and Semaan (2019) argues that Asian American Pacific Islander online community members need support. Further Aroyo et al. (2019) points out that the perceived toxicity of a comment is influenced by a variety of factors not limited to cultural background, rater bias, context and cognitive bias.

In particular, some of these efforts include explorations of microaggressions in social media posts (Breitfeller et al., 2019; Dosono and Semaan, 2019) with disadvantaged groups, surfacing contextual hate speech or words (Taylor et al., 2017) that have alternate meanings or veiled toxicity like codewords, adversarially generated or novel forms of offense (Jain et al., 2018), and exploring humor used as a tool for propagating hate (Billig, 2001).

Small datasets exist (Breitfeller et al., 2019) specifically for exploring microaggressions in online conversations. The Social Bias Inference Corpus (SBIC), a dataset of social media posts with structured annotations of thousands of demographic groups (Sap et al., 2019) is another example. To filter such datasets for veiled toxicity, Han and Tsvetkov (2020) proposed a procedure using probing examples.

Our paper builds on previous work with a strategy for consistently rating covertly toxic content, a shared dataset for training, and a baseline covert toxicity model.

3 Covert Toxicity

In order to formulate a covertly toxic dataset, we first extracted a set of candidate comments from the **CivilCommentsIdentities** dataset (Borkan et al., 2019) as described in Section 4.1. We manually labeled a random sample of these comments, seeking those which could have a negative impact on online conversations, but were frequently missed by raters. From this sample, we defined broad categories of *covert toxicity types*. See Figure 1 for an example.

- **Microaggression** Subtle discrimination towards an identity group
- **Obfuscation** Hidden toxicity via intentional misspellings, coded words, or implied references
- **Emoticons/Emojis** Toxic usage of non-text symbols
- **Sarcasm/Humor** Offensive content in the context of a joke
- **Masked Harm** Implied harm or threats masked by seemingly inoffensive language

3.1 Crowdsourcing Data Collection

We iterated on versions of an instruction templates to assist raters in identifying covertly toxic language with high precision. Previous work by Hube (2020) on biases and crowdworker annotations, illuminated relevant findings for the templates:

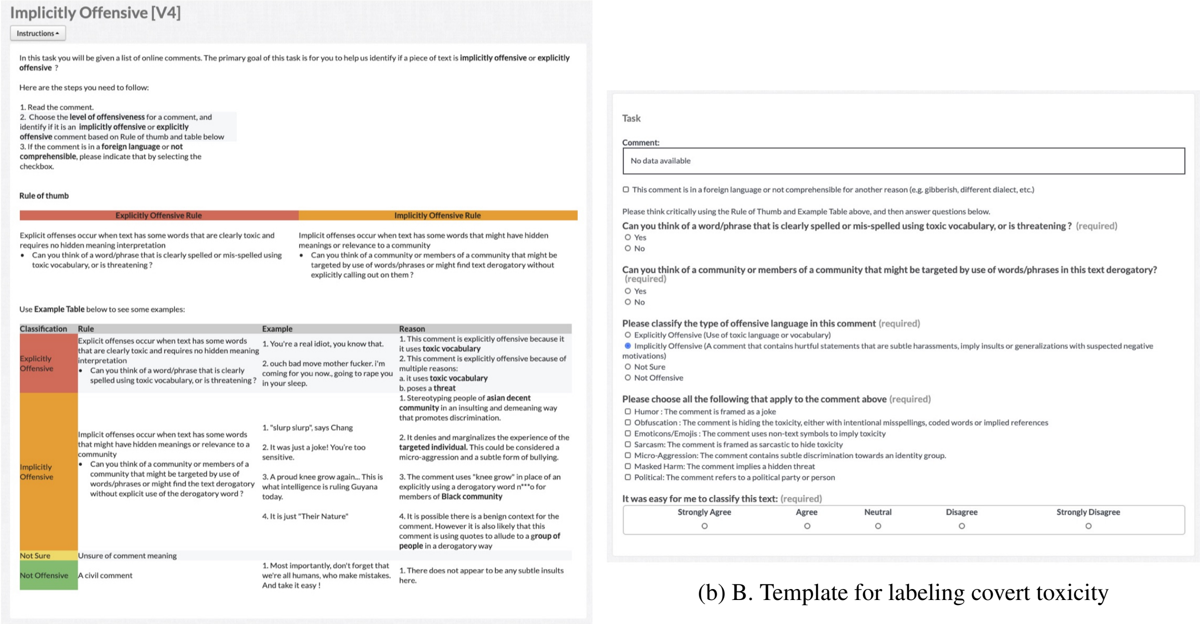
- Ask crowdworkers to reflect and put themselves in the shoes of participant
- Encourage crowdworkers to critically think about the task prior to engagement.

As such, two priming questions, see Table 3.1, were added in the final template. Note to improve rater comprehension we refer to covert and overt toxicity as implicit and explicit, respectively.See Figure 1 for an example.

Overtly Offensive Rule
Overt offenses occur when text has some words that are clearly toxic and requires no hidden meaning interpretation. **Can you think of a word/phrase that is clearly spelled or mis-spelled using toxic vocabulary, or is threatening?**
Covertly Offensive Rule
Covert offenses occur when text has some words that might have hidden meanings or relevance to a community. **Can you think of a community or members of a community that might be targeted by use of words/phrases or might find text derogatory without explicitly calling out on them?**

The addition of priming questions significantly improved rater performance. In an initial iteration, a test question of the overtly toxic comment “Peter is an idiot”, had a 43% miss rate by raters. The miss rate dropped to 5% with priming questions.

However, raters still frequently missed comments containing stereotypes, identity attacks and veiled threats. As such, we reconfigured the template instructions (see Figure 1) to include a table outlining different types of toxicity with examples. Desired rater accuracy again improved.



Implicitly Offensive [V4]

In this task you will be given a list of online comments. The primary goal of this task is for you to help us identify if a piece of text is **implicitly offensive** or **explicitly offensive** ?

Here are the steps you need to follow:

1. Read the comment.
2. Choose the **level of offensiveness** for a comment, and identify if it is an **implicitly offensive** or **explicitly offensive** comment based on Rule of thumb and table below
3. If the comment is in a **foreign language** or **not comprehensible**, please indicate that by selecting the checkbox.

Rule of thumb

Explicitly Offensive Rule	Implicitly Offensive Rule
Explicit offenses occur when text has some words that are clearly toxic and requires no hidden meaning interpretation • Can you think of a word/phrase that is clearly spelled or mis-spelled using toxic vocabulary, or is threatening ?	Implicit offenses occur when text has some words that might have hidden meanings or relevance to a community • Can you think of a community or members of a community that might be targeted by use of words/phrases or might find text derogatory without explicitly calling out on them ?

Use **Example Table** below to see some examples:

Classification	Rule	Example	Reason
Explicitly Offensive	Explicit offenses occur when text has some words that are clearly toxic and requires no hidden meaning interpretation • Can you think of a word/phrase that is clearly spelled using toxic vocabulary, or is threatening ?	1. You're a real idiot, you know that. 2. ouch bad move mother fucker. i'm coming for you now., going to rape you in your sleep.	1. This comment is explicitly offensive because it it uses **toxic vocabulary** 2. This comment is explicitly offensive because of multiple reasons: a. it uses **toxic vocabulary** b. poses a **threat**
Implicitly Offensive	Implicit offenses occur when text has some words that might have hidden meanings or relevance to a community • Can you think of a community or members of a community that might be targeted by use of words/phrases or might find the text derogatory without explicit use of the derogatory word ?	1. "slurp slurp", says Chang 2. It was just a joke! You're too sensitive. 3. A proud knee grow again... This is what intelligence is ruling Guyana today. 4. It is just "Their Nature"	1. Stereotyping people of **asian decent community** in an insulting and demeaning way that promotes discrimination. 2. It denies and marginalizes the experience of the **targeted individual.** This could be considered a micro-aggression and a subtle form of bullying. 3. The comment uses "knee grow" in place of an explicitly using a derogatory word n***o for members of **Black community** 4. It is possible there is a benign context for the comment. However it is also likely that this comment is using quotes to allude to a **group of people** in a derogatory way
Not Sure	Unsure of comment meaning		
Not Offensive	A civil comment	1. Most importantly, don't forget that we're all humans, who make mistakes. And take it easy !	1. There does not appear to be any subtle insults here.

Task

Comment:

No data available

☐ This comment is in a foreign language or not comprehensible for another reason (e.g. gibberish, different dialect, etc.)

Please think critically using the Rule of Thumb and Example Table above, and then answer questions below.

Can you think of a word/phrase that is clearly spelled or mis-spelled using toxic vocabulary, or is threatening ? (required)
○ Yes
○ No

Can you think of a community or members of a community that might be targeted by use of words/phrases in this text derogatory? (required)
○ Yes
○ No

Please classify the type of offensive language in this comment (required)
○ Explicitly Offensive (Use of toxic language or vocabulary)
● Implicitly Offensive (A comment that contains hurtful statements that are subtle harassments, imply insults or generalizations with suspected negative motivations)
○ Not Sure
○ Not Offensive

Please choose all the following that apply to the comment above (required)
☐ Humor : The comment is framed as a joke
☐ Obfuscation : The comment is hiding the toxicity, either with intentional misspellings, coded words or implied references
☐ Emoticons/Emojis : The comment uses non-text symbols to imply toxicity
☐ Sarcasm: The comment is framed as sarcastic to hide toxicity
☐ Micro-Aggression: The comment contains subtle discrimination towards an identity group.
☐ Masked Harm: The comment implies a hidden threat
☐ Political: The comment refers to a political party or person

It was easy for me to classify this text: (required)

Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
○	○	○	○	○

(a) A. Instructions for identifying implicitly (covertly) toxic comments versus explicit toxicity

(b) B. Template for labeling covert toxicity

Figure 1: Rater Instructions and Template

Rater Classification	Count	%
Not Sure	4457	1.4
Overtly Offensive	34094	10.8
Covertly Offensive	120314	38.1
Not Offensive	157259	49.7

Table 1: COVERTTOXICITY rater scores.

4 Datasets

The COVERTTOXICITY[2] dataset consists of *covertly offensive* rater labels using the described template on a subset of CivilCommentsIdentities data (Borkan et al., 2019).

4.1 Training Data

The CivilComments dataset is a publicly available corpus of ∼1.8 million crowd rated comments labeled for toxicity (Borkan et al., 2019). The CivilComments dataset is derived from the Civil Comments platform plugin, deprecated at the end of 2017, for independent news sites. The plugin utilized a peer-review submission process that required commenters to rate the randomly-selected comments before their own was posted for review. The CivilCommentsIdentities dataset is a subset of ∼400K comments, additionally rated for specific identity terms (e.g. gender, religion, or sexual orientation).

A limitation of using this dataset is that the comments are directly targeted towards news related content. As such, this work should not be generalized for other types of online forums, such as 4Chan, which may contain vastly different content and context.

For the COVERTTOXICITY dataset, we applied the methodology of (Han and Tsvetkov, 2020) to the CivilCommentsIdentities set: comments with identity attack annotations and low Perspective API toxicity scores (Jigsaw, 2017) were marked as candidates for covert toxicity.

The CivilCommentsIdentities toxicity label is the fraction of raters who voted for the label. Han and Tsvetkov (2020) noted that comments with veiled toxicity were more likely to have dissent amongst crowd raters and empirically we observed the same. As such, we filtered the dataset using the toxicity label rater fraction, explicitly such that $0 < P(\text{toxicity}) \leq 0.4$. Additionally, we only considered comments with at least one identity label from raters.

The final COVERTTOXICITY training and test subsets consist of ∼48000 and ∼2000 candidate comments, respectively.

[2] https://www.tensorflow.org/datasets/catalog/civil_comments

4.2 Evaluation Datasets

Given the subjective nature of this task, evaluation was performed via two approximate tests sets.

- **COVERTTOXICITY** COVERTTOXICITY test set of ∼2000 comments with continuous rater fractions as covertly toxic label.
- **SBIC Microaggressions Dataset** The Social Bias Inference Corpus (SBIC) consists of over 150k social media posts annotated for toxicity and implied offense over thousands of demographic groups (Sap et al., 2019). We again follow the the protocol used by Han and Tsvetkov (2020) for creating a synthetically binary label test set. The test set is the subset of SBIC that scored $P(\text{toxicity}) < 0.5$. A positive **covertly toxic** is applied if the comment includes at least one type of identity attack in the annotations and negative otherwise. This yielded an evaluation set with ∼3100 marked covertly offensive and ∼9000 marked not covert (and presumed not offensive given the low toxicity scores).

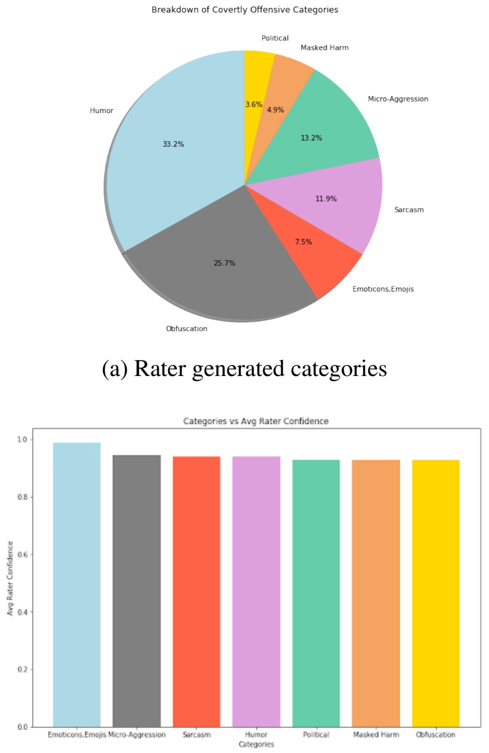
5 Results

Table 1 includes the breakdown of rater scores for COVERTTOXICITY, with 38% labeled as covertly offensive. The table results align with our expectations. The original comments were filtered for low (explicit) toxicity scores and as such a sizeable portion were confirmed as not offensive.

Raters marked humor (33.2%) and obfuscation (25.7%) as the predominant categories of covertly offensive comments. Humor and sarcasm are difficult for language models to detect in general. However, obfuscation, where offensive words are hidden with different spellings, symbols or other means is an easier target for improvement.

Comments with higher covertly offensive rating fractions (i.e. rater agreement), were more likely to be categorized as containing obfuscation, and less likely to be categorized as containing humor or emoticons/emoji. Additionally, obfuscation and comment length are anti-correlated in our dataset.

For example, the New Moon with Face emoji in 'But that's just "Their Nature". 🌚 ' is likely added to further target people of color. The comment should absolutely be labeled as toxic. However, such symbols remain challenging for raters to recognize and as such are not flagged. We include plots summarizing raters reported categories and self-reported confidence in figure 2.



(a) Rater generated categories

(b) Average rater confidence per category

Figure 2: Rater Covert Toxicity Categories for the CivilCommentsIdentities set.

5.1 Model Evaluation

We evaluate the efficacy of modeling with the COVERTTOXICITY dataset using a fine-tuned BERT model (Devlin et al., 2018). We used a pre-trained Comment-Domain BERT model (Lees et al., 2020) as checkpoint and jointly fine-tuned the **covertly toxic** and **toxic** labels. The entire CivilCommentsIdentities dataset was used for training toxicity with zero-weights assigned for missing covertly toxic labels. We employed cross-entropy loss with the two labels weighted by prevalence. The model is labeled as **Covert-BERT** in experiments below. A baseline model fine-tuned solely on the 'toxicity' label in CivilCommentsIdentities is marked as **Toxic-BERT**.

Toxic-Bert scored an average model probability for toxicity of 0.21 on comments with a majority of raters voting covertly toxic in the COVERTTOXICITY test set. In contrast, on these same comments, Covert-BERT achieved a 0.44 average model probability of toxicity. Figure 3 illustrates the shift in model probabilities for covert comments in Covert-BERT, suggesting the model is more adept at identifying covertly toxic comments.

However, the eval ROC-AUC on the held-out

Comment	Toxicity	Perspective	Toxic-BERT	Covert-BERT
You're an engineering major? Wow, totally wouldn't have seen you as an engineer... I always picture them as men.	covert	0.282	0.185	**0.705**
If minorities in this country were truly oppressed you wouldn't have so many people desperately pretending to be one.	covert	0.160	0.205	**0.680**
Why do Ethiopians like paralyzed children? Because their food can't run now.	covert	0.191	0.261	**0.657**
I wonder who will be in the WWE hall of fame in 2018.	not toxic	0.081	0.016	0.285
you are an idiot	overt	**0.98**	0.928	0.478

Table 2: Scored comments from micro-aggressions SBIC subset

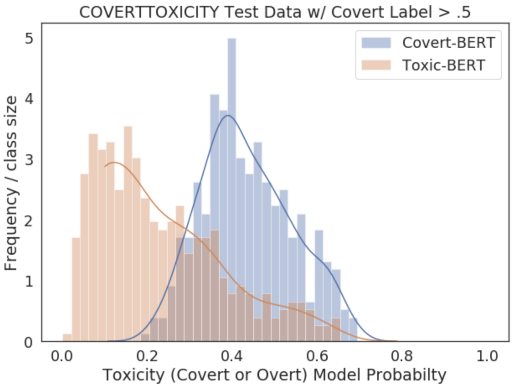


Figure 3: Shift in Covert-BERT model probabilities on covert label in COVERTTOXICITY dataset.

COVERTTOXICITY test set for the covert label established with majority rater consensus of covert toxicitiy (≥ 0.5) remained poor at 0.59. This is in contrast to Toxic-BERT with ROC-AUC of 0.52 for the same covertly toxic labels.

The Covert-BERT model shows more promise on the synthetic covert label subset of microaggressions from the SBIC dataset. The model showed substantially improvements in average model probability of toxicity for covert labels, ROC-AUC for covert toxicity, and recall as shown in Table 3. Similarly, Figure 4 illustrates the distribution of Covert-BERT model probabilities across covert and non covert synthetic labels. The Covert-BERT model appears better suited for extracting covert-toxicity among microaggression specific data, as demonstrated with sample rated comments in Table 2.

Model	Avg Cov	AUC	R@0.3	R@0.5
T-BERT	0.231	0.673	0.29	0.04
C-BERT	**0.381**	**0.781**	**0.62**	**0.18**

Table 3: Model evaluation on subset of SBIC. Avg Cov and Recall are the average covert model probability and recall across positive covert labels, respectively.

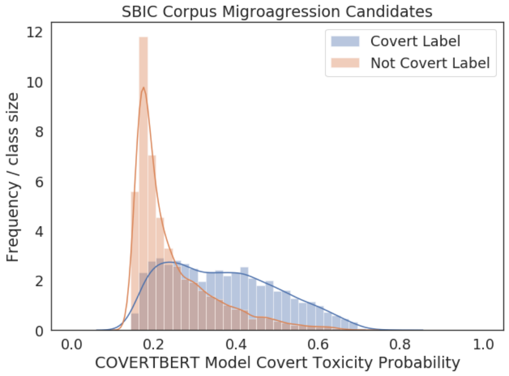


Figure 4: Distribution of COVERT-BERT model probabilities for SBIC microaggressions set.

6 Conclusion

We iterate on rater feedback to create an initial baseline dataset, COVERTTOXICITY, that encapsulates a variety of often mislabeled online toxicity. While progress is still needed in extracting coherent rater signals and modelling, our initial work demonstrates the possibility of capturing veiled toxic language with machine learning models.

References

Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. *Crowdsourcing Subjective Tasks: The Case Study of Understanding Toxicity in Online Discussions*, page 1100–1105. Association for Computing Machinery, New York, NY, USA.

Michael Billig. 2001. Humour and hatred: The racist jokes of the ku klux klan. *Discourse & Society – DISCOURSE SOCIETY*, 12:267–289.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. *CoRR*, abs/1903.04561.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Eshwar Chandrasekharan, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert. 2017. You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Bryan Dosono and Bryan Semaan. 2019. Moderation practices as emotional labor in sustaining online communities: The case of aapi identity work on reddit. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, page 1–13, New York, NY, USA. Association for Computing Machinery.

Spiros V. Georgakopoulos, Sotiris K. Tasoulis, Aristidis G. Vrahatis, and Vassilis P. Plagianakos. 2018. Convolutional neural networks for toxic comment classification. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, SETN '18, New York, NY, USA. Association for Computing Machinery.

Xiaochuang Han and Yulia Tsvetkov. 2020. Fortifying toxic speech detectors against veiled toxicity. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7732–7739, Online. Association for Computational Linguistics.

Christoph Hube. 2020. *Methods for detecting and mitigating linguistic bias in text corpora*. Ph.D. thesis, Hannover: Institutionelles Repositorium der Leibniz Universität Hannover.

Kelly Jackson. 2011. Derald wing sue, microaggressions in everyday life: Race, gender, and sexual orientation. *Social Service Review*, 85:519–521.

E. Jain, S. Brown, J. Chen, E. Neaton, M. Baidas, Z. Dong, H. Gu, and N. S. Artan. 2018. Adversarial text generation for google's perspective api. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1136–1141.

Google Jigsaw. 2017. Perspective api. `https://www.perspectiveapi.com/`. Accessed: 2021-02-02.

Google Jigsaw. 2018. Nyt moderators. `https://blog.google/technology/ai/new-york-times-using-ai-host-better-conversations/`. Accessed: 2021-02-02.

Alyssa Whitlock Lees, Ian Kivlichan, and Jeffrey Scott Sorensen. 2020. Jigsaw @ ami and haspeede2: Finetuning a pre-trainedcomment-domain bert model. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online.

Kevin Nadal, Katie Griffin, Yinglee Wong, Sahran Hamit, and Morgan Rasmus. 2014. The impact of racial microaggressions on mental health counseling implications for clients of color. *Journal of counseling and development: JCD*, 92:57–66.

Ihudiya Finda Ogbonnaya-Ogburu, Angela D.R. Smith, Alexandra To, and Kentaro Toyama. 2020. Critical race theory for hci. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–16, New York, NY, USA. Association for Computing Machinery.

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

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2019. Social bias frames: Reasoning about social and power implications of language. *CoRR*, abs/1911.03891.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain. Association for Computational Linguistics.

Sheldon Stryker and Peter Burke. 2000. The past, present, and future of an identity theory. *Social Psychology Quarterly*, 63:284.

D. W. Sue, Christina M. Capodilupo, Gina C. Torino, Jennifer M Bucceri, Aisha M. B. Holder, Kevin L Nadal, and Marta Esquilin. 2007a. Racial microaggressions in everyday life: implications for clinical practice. *The American psychologist*, 62 4:271–86.

Derald Sue, Jennifer Bucceri, Annie Lin, Kevin Nadal, and Gina Torino. 2007b. Racial microaggressions and the asian american experience. *Cultural diversity & ethnic minority psychology*, 13:72–81.

Jherez Taylor, Melvyn Peignon, and Yi-Shin Chen. 2017. Surfacing contextual hate speech words within social media. *CoRR*, abs/1711.10093.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2016. Ex machina: Personal attacks seen at scale. *CoRR*, abs/1610.08914.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.