Critical Perspectives: A Benchmark Revealing Pitfalls in `PerspectiveAPI`

Lorena Piedras[*] and **Lucas Rosenblatt**[*] and **Julia Wilkins**[*]
lp2535@nyu.edu, lr2872@nyu.edu, jw3596@nyu.edu
New York University

Abstract

Detecting "toxic" language in internet content is a pressing social and technical challenge. In this work, we focus on PERSPECTIVE from Jigsaw, a state-of-the-art tool that promises to score the "toxicity" of text, with a recent model update that claims impressive results (Lees et al., 2022). We seek to challenge certain normative claims about toxic language by proposing a new benchmark, Selected Adversarial SemanticS, or *SASS*. We evaluate PERSPECTIVE on *SASS*, and compare to low-effort alternatives, like zero-shot and few-shot GPT-3 prompt models, in binary classification settings. We find that PERSPECTIVE exhibits troubling shortcomings across a number of our toxicity categories. *SASS* provides a new tool for evaluating performance on previously undetected toxic language that avoids common normative pitfalls. Our work leads us to emphasize the importance of questioning assumptions made by tools already in deployment for toxicity detection in order to anticipate and prevent disparate harms.

1 Introduction

Social media and online discussion forums are plagued with toxic content. 41% of American adults have personally experienced online harassment (Vogels, 2021).

This problem can be further exacerbated by learning algorithms with unchecked incentives. On a platform like Facebook, incendiary comments are highly visible due in large part to the design of its ranking algorithms (Munn, 2020). Some experiments by Cheng et al. (2017); Kim et al. (2021) have shown that users are more likely to write toxic or "troll" posts when the initial comments in a conversation exhibited toxic or "trolling" behavior, which can be interpreted as evidence that "trolling is contagious."

This creates the potential for a negative feedback loop where companies are incentivized to promote toxic content and toxic content in turn provokes users to post more malicious content. In response, researchers and corporations alike have invested in systems that can detect toxicity at massive scales to help moderate content on social media and internet forums (Lees et al., 2022; Pavlopoulos et al., 2017; Wulczyn et al., 2017). Some of these models have, however, been shown to be susceptible to simple adversarial attacks (Hosseini et al., 2017; Gröndahl et al., 2018).

Text	Toxicity (Perspective)	Toxicity (Human)
You look so good in that dress, and probably even better with it off.	22.9%	85.0%

Table 1: A toxicity score given by PERSPECTIVE to a *SASS* example from the "Sexual Harassment" category.

Still, deep language model-based approaches to detecting toxicity have been deployed, with consumer facing API's granting direct access to the model output. One such model is PERSPECTIVE (Lees et al., 2022), which assigns text a "toxicity" score $\in [0, 1]$, provided by Jigsaw (a Google subsidiary). In 2022, Jigsaw published findings on the "new generation" PERSPECTIVE toxicity model (currently deployed). We believe that interrogating the model's assumptions and frailties is vital work in moving it (and similar systems) towards robust detection (or, perhaps, a robust understanding of when detection fails). For example, Jigsaw defines toxic language as "a rude, disrespectful, or unreasonable comment that is likely to make you leave the discussion" (Lees et al., 2022), though other definitions exist (Märtens et al., 2015). We were unable to find an original source for Jigsaw's definition.

[*] Equal contributions.

This research was supported in part by NSF Award No. 1922658.

Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI), pages 15 - 24
December 7, 2022 ©2022 Association for Computational Linguistics

Text	Category	Type	Toxicity Score	Expected Toxicity
Your secret can be our secret as long as you do what I tell you.	Blackmail	Unfiltered	11.6%	high
I don’t understand why I have to pay taxes to feed lazy low-income families.	Classism	Unfiltered	17.4%	high
We don’t admit people from your race in this forum.	Exclusionary	Filtered	35.8%	high
You shouldn’t think that way.	Gaslighting	Filtered	11.3%	high
It’s weird how she’s taller than her husband.	Misogyny	Filtered	24.1%	high
No one should live without water, electricity and schooling. Except for poor people.	Sarcasm	Filtered	10.8%	high
I can’t believe you work here. You look too beautiful to do this job.	Sex. Harass.	Filtered	23.5%	high
Boys are better at math than girls, it’s just a fact.	Stereotyping	Filtered	24.6%	high
It’s so freaking crazy that he surfed a 40-foot wave.	False Positive	Filtered	67.2%	low
Their hair is so long, it’s almost touching the ground.	Neutral	Unfiltered	16.7%	low

Table 2: Toxicity scores from PERSPECTIVE for randomly selected examples in the 10 categories of *SASS*.

Contributions Existing models and benchmarks rely on aggregating binary responses to text collected from crowdworkers into a ground truth “probability of toxicity” (this is accomplished by prompting a crowdworker with “Is this text toxic?”, and then calculating the aggregate $Pr[toxic] = \frac{|yes\_responses|}{|total\_responses|}$, which is the “toxicity score”). We suspect this method overemphasizes a normative understanding of toxicity, such that potentially toxic, harmful text “*on the margins*” goes undetected. Here, “normative” describes the way in which multiple annotations are traditionally aggregated, which often implicitly supports the views of the majority and ignores the annotations of minority groups. In response, we isolate a set of natural language categories that fulfill the definition of toxicity (as stated earlier), but go largely undetected, due in part, we believe, to the normative assumptions of the ground truth toxicity examples from existing training and benchmark data. Again, these normative assumptions are related to the way data is aggregated, which may ignore the views of a minority of annotators in favor of the majority.

We present a new benchmark entitled *Selected Adversarial SemanticS*, or *SASS*, that evaluates these behaviors. *SASS* contains natural language examples (each approximately 1-2 sentences in length) across previously underexplored “toxicity” categories (like manipulation and gaslighting) as well as categories that have received attention (like “sexism” (Sun et al., 2019)), and includes a “human” toxicity score $\in [0, 1]$ for each example. Table 1 shows an example from the "Sexual Harassment" category. *SASS* follows a filtered/unfiltered approach to adversarial benchmarking, as in (Lin et al., 2021). The benchmark is designed to exploit the normative vulnerabilities of a toxicity detection tool like PERSPECTIVE. Specifically, PERSPECTIVE makes ambiguous claims that they can “identify abusive [or toxic] comments” (Jigsaw), but do not clarify that these abusive comments are determined by essentially using the majority opinion of random annotators. Our position is that PERSPECTIVE should either be clear concerning the limitations of it’s toxicity tool (i.e. that it detects toxic content according to majority opinion), or adjust the PERSPECTIVE model to better account for minority annotations.

We compare PERSPECTIVE’s performance on *SASS* to “human” generated toxicity scores. We further compare PERSPECTIVE to low-effort alternatives, like zero-shot and few-shot GPT-3 prompt models, in a binary classification setting (“toxic or not-toxic?”) (Brown et al., 2020). Code for our project can be found in this repository.

2 Related Work

Past PERSPECTIVE Model Works such as (Hosseini et al., 2017) and (Gröndahl et al., 2018) focused on generating adversarial attacks to test how the former version of PERSPECTIVE responded to word boundary changes, word appending, misspellings, and more. (Gröndahl et al., 2018) further tested how toxicity detection models responded to offensive but non-hateful sentences. The toxicity of the test sentences heavily increases when the word "F***" is added (You are great → You are F*** great, 0.03 → 0.82). This opens up a discussion about the subjectivity of what should be considered “toxic”, a theme in our work. We pose new open questions that draw a clear connection between “toxicity” and normative concerns (Arhin et al., 2021). Another promising approach to fortifying toxicity detectors is by probing a student model with a few annotated examples to detect veiled toxicity, mostly annotated incorrectly, from a preexisting dataset, then *re-annotating*, thus making the model more robust (Han and Tsvetkov, 2020); we do not attempt this in our work.

Current Model A recent publication on PERSPECTIVE (Lees et al., 2022) generated benchmarks to test how the new version responded to character obfuscation, emoji-based hate, covert toxicity, distribution shift and subgroup bias. They demonstrate improvements of the model in classifying multilingual user comments and classifying comments with human-readable obfuscation. Additionally, PERSPECTIVE beats every baseline on character obfuscation rates ranging from 0% to 50%. Character-level perturbations and distractors degrade performance of ELMo and BERT based toxicity models, reducing detection recall by more than 50% in some cases (Kurita et al., 2019). **Separate detection tools**, like the HATECHECK system from (Röttger et al., 2020), present a set of 29 automated functional tests to check identification of types of "hateful behavior" by toxicity or hate speech detection models. A large dynamically generated dataset from (Vidgen et al., 2020), designed to improve hate speech detection during training, showed impressive performance increases in toxicity and hate speech detection tasks. Though slightly different in their typology of toxic speech, these approaches have a significant scale advantage over *SASS*, while *SASS* examples are specifically targeted at the PERSPECTIVE tool.

3 Benchmarking with *SASS*

The *SASS* benchmark contains 250 manually created natural language examples across 10 nuanced "toxicity" categories (e.g. stereotyping, classism, blackmail). These categories were selected via a process of literature review and vulnerability testing on PERSPECTIVE and other toxicity tools, to determine their weaknesses/strengths. As we sought to challenge PERSPECTIVE and other toxicity tools, we believe this to be a sufficient process for determining our categories, although acknowledge that it introduces some unavoidable author bias. The examples are each 1-2 sentences long and are designed to exploit vulnerabilities in toxicity detection systems like PERSPECTIVE. Samples from *SASS* in each category are shown in Table 2.

Eight of *SASS*'s categories are aimed at generating "False Negative" (FN) scores (a score that significantly underestimates the toxicity of some text), one category is aimed at "False Positive" (FP) scores (a score that overestimates toxicity), and one category is "Neutral," a control, demonstrating the model's performance on "normal," non-toxic sentences. *SASS* is heavily biased towards examples that generate a FN score, which we argue may be more harmful than a FP score, as a FN means toxic content has gone undetected. For each category, the benchmark contains 15 "filtered" and 10 "unfiltered" examples, drawing inspiration from (Lin et al., 2021). We generate filtered examples by brainstorming toxic comments and evaluating the comments with PERSPECTIVE to ensure a toxicity score of < 0.5. Then, we generate an additional set of 10 examples per category using the knowledge gained from creating the filtered examples *without* first testing them on PERSPECTIVE.

Human Ground Truth The benchmark also contains a "human" toxicity score $\in [0, 1]$ for each comment, which can be used as a baseline for evaluating toxicity detection tools using *SASS*. The human toxicity scores are an average of the toxicity scores of the authors per comment (scored blindly). Here, we scored examples on a scale of 0-10, using Jigsaw's definition of toxicity, i.e. "how likely [the example is to] make [a user] leave the discussion" (0=highly unlikely, 10=highly likely). Significantly, we aligned these ratings with assumptions laid out in A.2.2 (in appendix) for consistency and to combat benchmarking pitfalls (Blodgett et al., 2021).

We further performed z-normalization, as per (Pavlick and Kwiatkowski, 2019). Each author may have treated the "0-10 toxicity scale" differently, so this normalization process ensures that the final aggregate scores are not overly biased by any single author's interpretation of the scale.

In Table 5 (in the appendix), we observe the average z-normalized human toxicity scores of comments in *SASS* across the toxicity categories described above. We note that some categories are inherently more toxic than others; "Stereotyping" comments have an average human toxicity score of 0.81 versus 0.57 for "Gaslighting" comments, which further contrasts with an average human toxicity score of 0.007 for "Neutral" comments.

4 Experiments and Discussion

Binary Toxicity Classification We showcase the utility of *SASS* by evaluating PERSPECTIVE and GPT-3 against the human baseline in a binary classification setting. It's important to note that PERSPECTIVE and GPT-3 are very different systems, trained with distinct objectives, amounts and

System	Precision	Recall	F1-Score
Perspective	0.26	0.05	**0.08**
GPT-3-Zero	0.83	0.19	**0.31**
GPT-3-One	0.77	0.11	**0.19**
GPT-3-Few	0.73	0.52	**0.61**

Table 3: Evaluation of Perspective and GPT-3 in multiple prompt settings on the *SASS* benchmark against thresholded human toxicity scores, in a binary classification setting.

sources of data. We believe the comparison is still useful because it provides a "low-effort alternative" to make sure that our examples are not overly complicated. Note that GPT-3 was not fine-tuned explicitly for this task, so we prompt the system in zero, one, and few-shot settings for a binary toxicity classification. We binarize the Perspective and z-normalized human baseline toxicity scores by labeling scores > 0.5 per comment as "toxic". The binarized ground truth human labels on *SASS* contain 72.4% toxic labels versus 27.6% non-toxic labels. We use these thresholded human labels as ground truth and evaluate Perspective and GPT-3's performance on *SASS* in Table 3.

Model Description Perspective uses a Transformer model with a state-of-the-art Charformer encoder. The model is pretrained on a proprietary corpus including data collected from the past version of Perspective and related online forums. This dataset is mixed in equal parts with the mC4 corpus, which contains multilingual documents (Lees et al., 2022). GPT-3, created by OpenAI in 2020, is a state-of-the-art autoregressive transformer-based language model (Brown et al., 2020). GPT-3 is trained on a massive amount of internet text data, predominately Common Crawl and WebText2 (Radford et al., 2019), and generates human-like language in an open prompt setting.

Results We first observe that Perspective performs very poorly on the binary task of toxicity classification on the *SASS* benchmark (Table 3, F1-Score = 0.08). Note that the majority of comments in *SASS* were crafted specifically to generate a low toxicity score from Perspective, so this is not surprising. We establish the metric regardless, as a baseline to evaluate future versions of the system.

We also examine the performance of GPT-3 in multiple prompt settings for binary (true/false) toxic content classification in Table 3. Each system yields relatively high precision and low recall, generally indicating a significant under-prediction of toxicity in *SASS*. GPT-3 has more success in classifying harmful comments in *SASS* as toxic across the board relative to a thresholded Perspective. GPT-3-Few (F1-Score = 0.61) shows a significant improvement over both GPT-3-Zero and GPT-3-One as well as Perspective, yielding the most success relative to the human baseline of any of the experimental formulations.

We hypothesize that GPT-3 outperforms Perspective largely due to the sheer scale and scope of data that GPT-3 is trained on, as well as the size of the model itself (175B learnable parameters in GPT-3 versus 102M in the Perspective base model). While GPT-3 is *not* trained for the toxicity detection task specifically, by learning from such a massive amount of internet text data spanning millions of contexts, the model has likely been exposed to a much wider range of potentially toxic material then Perspective.

In Table 5 (see appendix), we break down the toxicity scores of Perspective and GPT-3 by *SASS* category, relative to the human baseline. In some categories, both Perspective and GPT-3-Few fall particularly short (for example, Perspective predicts an average toxicity score of 21.9% for "Sexual Harassment" comments versus the 80% human baseline). Relative to other categories from *SASS*, Perspective similarly rates comments in "Sarcasm" and "Stereotyping" as highly toxic, while humans rated the toxicity of "Stereotyping" comments significantly higher than those in "Sarcasm." This raises the question of how to properly threshold scores from a toxicity detection system in-the-wild, which (Lees et al., 2022) do not comment on, though seems a reasonable use case for platforms flagging toxic content.

In the "False Positive" category we observe that both Perspective and GPT-3-Few yield very *high* toxicity scores on average (Table 5), suggesting that the models are overfit to swear word toxicity, and underfit to a deeper interpretation of malicious intent. We believe it is important to delineate between the tasks of *swear word detection* and *toxicity detection*, and so find this undesirable. Allowing harmful comments to slip through the cracks is arguably more dangerous than unintentionally removing content with positive intent, but both of these scenarios could be upsetting to a downstream

See Appendix A.1 for details on prompt generation.

Recall that "Neutral" and "False Positive" categories are inherently non-toxic, accounting for 20% of non-toxic labels.

https://commoncrawl.org/

user. We report further on the influence of swear words on toxicity in the next section.

Profanity and Toxicity Detection *SASS* includes 18 "False Positive" examples that contain swear words. PERSPECTIVE rated *all* of them as toxic, and GPT-3-FEW labeled 83% of these comments as toxic (this is $P[toxic|contains\_swear\_word]$). This suggests that, instead of *understanding when* swear words are used to communicate hateful content, PERSPECTIVE may be effectively *memorizing* their inclusion in toxic text. This could be problematic; swear words can be used to communicate non-toxic emotions, like surprise (e.g. Holy f*** I got the job!) or excitement (e.g. Oh sh**! Congratulations.) and should not necessarily be treated equivalently to toxic speech. Furthermore, different genders and races utilize profanity differently, so associating expletives with toxicity could have disparate impacts (Beers Fägersten, 2012). Past work by (Gröndahl et al., 2018) evaluating an older version of PERSPECTIVE also detected this issue.

As shown in Table 6 (see appendix), from the 34 *SASS* examples that PERSPECTIVE rated as toxic, 52% contained a profanity, versus only 11.6% of the examples rated toxic by GPT-3-FEW (this is $P[contains\_swear\_word|toxic]$). A lot of hateful content does not explicitly contain offensive words and it is troubling that PerpectiveAPI relies so much on them in our benchmark.

TweetEval We were surprised that GPT-3-FEW performed better in the binary classification scenario on the *SASS* benchmark than PERSPECTIVE, and so sought to validate the finding with another prominent toxicity benchmark, TweetEval. Thus we selected 1,000 examples from the 'Hate Speech Detection" benchmark randomly (Barbieri et al., 2020). We acknowledge that this might be viewed as irrelevant or an unfair comparison, as some "toxic language" may not qualify as "hate speech" (for example, universal insults that do not target a specific group). However, we believe that the reverse claim, that all "hate speech" *should* qualify as "toxic language" is true. Then evaluating both PERSPECTIVE and GPT-3-FEW on a "hate speech" benchmark, despite both being designed to detect "toxic language," is a valid comparison. We found that PERSPECTIVE had an F1-Score of 0.48 and GPT-3-FEW had an F1-Score 0.52 (Table 7, see appendix). The performance gap between PERSPECTIVE and GPT-3-FEW on TweetEval is significantly smaller than on *SASS*, but the trend (GPT-3-FEW matching or improving on PERSPECTIVE) is comparable. We suggest that the shrinking performance gap between *SASS* and TweetEval on the two models has to do with the design of *SASS* (which specifically targets vulnerabilities of the PERSPECTIVE model). Significantly, we were able to validate that GPT-3-FEW, in the binary setting, is a good point of comparison with PERSPECTIVE on another benchmark, and does not only perform well on *SASS*-specific examples.

Conclusion and Future Work We introduce Selected Adversarial SemanticS (*SASS*) as a benchmark designed to challenge previous normative claims about toxic language. We have shown here that existing tools are far from robust to relatively simple adversarial examples, and fail to report adequately on the implicit biases attached to their model construction. We therefore position *SASS* as an important additional benchmark that can help us understand weaknesses in existing and future systems for toxic comment detection. Some impactful future work would be to grow the set of examples in *SASS* and to perform similar vulnerability testing on problems like sentiment analysis and other tools for content moderation. Conducting a future study with a set of random human annotators and demonstrating that the majority rate *SASS* statements as non-toxic would strengthen our claims of normativity, and make the need for a benchmark like *SASS* even more apparent. Expanding the set of state-of-the-art NLP toxicity detection or large language models evaluated on *SASS* would provide interesting future points of comparison. Finally, we emphasize our belief that deployed natural language based tools, potentially serving millions of users, must be examined and reexamined in order to prevent the harmful beliefs of majority groups from being perpetuated.

5 Ethical Considerations

SASS, the new benchmark proposed in this paper, seeks to address normative claims made by toxicity detection tools that rely on majority opinion to determine malicious content. In the narrow scope of improving toxicity model evaluation, we thus expect *SASS* to have a positive impact on the NLP community, and by extension on moderation systems for social media and online forums.

However, thousands of content moderators, whose job descriptions include toxic content detection, are currently employed by companies such as Meta. We believe that the best systems for toxic content detection are likely collaborations between humans and machines, but acknowledge that, by improving automated systems, we may jeopardize employment for these people. Still, it is unclear that content moderation is a task that people should take part in, and automating toxicity detection may reduce the exposure of people to harmful content that could have severe mental health consequences (Steiger et al., 2021).

There is always the risk that, in providing a new benchmark to the larger NLP community, some may use it to make unjustified claims. Therefore, we take this opportunity to highlight the ways in which *SASS* could be misused. We acknowledge that any benchmark, especially a relatively small one like *SASS*, will reflect the inherent biases of the authors. Each category of *SASS* is not designed by any means to be exhaustive; rather, each is designed to provide an initial probe, a check for model vulnerabilities. Further exploration would be required even if a model performed well on *SASS*. *SASS* is also only an English language benchmark, and contains examples that only make sense in an Americanized cultural context. We believe it is important work to create similar benchmarks for other languages and cultural contexts.

We would like to thank Sam Bowman and Richard Pang for very useful conversations and feedback over the course of our project. We would also like to thank Julia Stoyanovich and the Center for Responsible AI at NYU for supporting our work.

References

Kofi Arhin, Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, and Moninder Singh. 2021. Ground-truth, whose truth?–examining the challenges with annotating toxic text datasets. *arXiv preprint arXiv:2112.03529*.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of Findings of EMNLP*.

Kristy Beers Fägersten. 2012. Who's swearing now?: the social aspects of conversational swearing.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone can become a troll: Causes of trolling behavior in online discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, CSCW '17, page 1217–1230, New York, NY, USA. Association for Computing Machinery.

Tommi Gröndahl, Luca Pajola, Mika Juuti, Mauro Conti, and N. Asokan. 2018. All you need is "love": Evading hate speech detection. 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.

Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*.

Jigsaw. Faq perspetive api.

Jin Woo Kim, Andrew Guess, Brendan Nyhan, and Jason Reifler. 2021. The distorting prism of social media: How self-selection and exposure to incivility fuel online comment toxicity. *Journal of Communication*, 71(6):922–946.

Keita Kurita, Anna Belova, and Antonios Anastasopoulos. 2019. Towards robust toxic content classification.

Alyssa Lees, Vinh Q Tran, Yi Tay, Jeffrey Sorensen, Jai Gupta, Donald Metzler, and Lucy Vasserman. 2022. A new generation of perspective api: Efficient multilingual character-level transformers. *arXiv preprint arXiv:2202.11176*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Marcus Märtens, Siqi Shen, Alexandru Iosup, and Fernando Kuipers. 2015. Toxicity detection in multiplayer online games. In *2015 International Workshop on Network and Systems Support for Games (NetGames)*, pages 1–6. IEEE.

Luke Munn. 2020. Angry by design: toxic communication and technical architectures. *Humanities and Social Sciences Communications*, 7(1):53.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deeper attention to abusive user content moderation. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 1125–1135.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet B Pierrehumbert. 2020. Hatecheck: Functional tests for hate speech detection models. *arXiv preprint arXiv:2012.15606*.

Miriah Steiger, Timir J Bharucha, Sukrit Venkatagiri, Martin J Riedl, and Matthew Lease. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–14.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2020. Learning from the worst: Dynamically generated datasets to improve online hate detection. *arXiv preprint arXiv:2012.15761*.

Emily A Vogels. 2021. The state of online harassment. *Pew Research Center*, 13.

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*, pages 1391–1399.

A Example Appendix

A.1 Generating prompts for GPT-3

As described in Section 4, GPT-3 was not explicitly trained to classify phrases as toxic or not toxic. We explore zero, one, and few-shot settings for prompting GPT-3 to generate results comparable to a thresholded PERSPECTIVE toxicity rating. Here we describe the prompts used in these experiments.

Setting	Prompt
GPT-3-ZERO	"True or False, the phrase {phrase}' is toxic? Answer:"
GPT-3-ONE	"Determine if the given phrase is toxic: Men are superior to women in every way. => True {phrase} => "
GPT-3-FEW	"Determine if the given phrase is toxic: Men are superior to women in every way. => True Grilling with the neighbors is my favorite. => False {phrase} => "

Table 4: Prompts used for zero, one, and few-shot experimentation with toxicity classification using GPT-3. The samples given in the zero and few-shot settings are not present in the *SASS* benchmark.

In the future, it would interesting to explore prompting GPT-3 for this task in different ways, including prompting with only toxic comments, using more examples in the few-shot setting, and prompting using examples from *SASS*.

Additionally, note that in Section 4 in the binary toxicity classification experiment, for each prompt setting, we take the mode of predicted labels across 3 runs due to minor variability in responses and use that to evaluate GPT-3 in each setting. Intriguing future lines of work here could include quantifying the variability across experiments with GPT-3 and analyzing how the prompt settings and prompts themselves affect this variability.

A.2 Designing *SASS*

A.2.1 Avoiding Conceptual and Operational Pitfalls

(Blodgett et al., 2021) describe the ways in which popular stereotype detection benchmarks suffer from a set of conceptual and operational *pitfalls*. By providing a taxonomy of potential pitfalls, they are able to audit the methods in a principled manner and deduce ways in which the benchmark may produce spurious measurements. Here are summaries of each category of pitfall they describe (specific to stereotyping):

1. **Conceptual Pitfalls** (stereotyping)
   (a) **Power dynamics** The claimed problematic power dynamic may not be "realistic."
   (b) **Relevant aspects** Must be clear and consistent about what stereotype content is within the purview of a given example.
   (c) **Meaningful stereotypes** Is this stereotype actually reflective of a societal prob-

lem?

   (d) **Anti vs non-stereotypes** Some statements can negate a stereotype (i.e. not), while others can actively combat (i.e. evil vs. peaceful).
   (e) **Descriptively true statements** A true statement masquerading as a stereotype.
   (f) **Misaligned stereotypes** A hyper specific, or not specific enough, stereotype about a certain group/subgroup ("Ethiopia" in a context where Africa generally is implied).
   (g) **Offensive language** Are swear words stereotyping?

2. **Operational Pitfalls** (stereotyping)

   (a) **Invalid perturbations** Not a real stereotype/anti-stereotype (i.e. both alternate sentences are stereotypes)
   (b) **Incommensurable groups or attributes** Two alternate groups are not comparable (think apples and oranges).
   (c) **Indirect group identification** I.e. using names as a way of identifying group membership (for example, racially identifying names)
   (d) **Logical failures** If the alternate represents a logically dubious conclusion.
   (e) **Stereotype conflation** Multiple stereotypes present in a single example
   (f) **Improper sentence pairs** The example is not "realistic."
   (g) **Text is not naturalistic** The text itself would never be written/uttered.
   (h) **(Un)markedness** The two examples are represented at different degrees in natural text (i.e. "young gay man" vs. "young *straight* man")
   (i) **Uneven baselines** Similar to (un)markedness, examining a false alternative.

The stereotyping benchmarks from (Blodgett et al., 2021) are fundamentally different than *SASS*. Thus, our analysis of pitfalls must rely on slightly different criteria. Using the aforementioned criteria, we created an abbreviated conceptual and operational pitfall taxonomy for toxicity.

A.2.2 Conceptual and operational pitfalls in toxicity benchmarks

Recall that the definition of toxicity according to PERSPECTIVE/Jigsaw is: "a rude, disrespectful, or unreasonable comment that is likely to make you leave the discussion."

With this definition, we can begin to construct a set of pitfalls that text from a benchmark might exhibit. However, in order to minimize subjectivity as much as possible, we outline three major assumptions about examples in our benchmark *SASS* (and therefore, about what we prescribe as the behavior of a system that "detects toxicity"):

Assume adversarial reading. Within reason, does there exist an individual or group that would be likely to leave a discussion after reading a piece of text (even if they represent a significant minority)?

Assume adversarial context/subtext. Assume that the possible context in which a piece of text is positioned increases the likelihood that someone would leave the discussion after reading it.

Assume bad intentions. Assume that the writer of the text was knowingly malicious in their choice of words.

These assumptions are important because they help make our analysis structured and consistent. Here are the pitfalls we use in evaluating toxicity, constructed from (Blodgett et al., 2021):

1. **Conceptual Pitfalls** (toxicity)

   (a) **Meaningful toxicity (from Meaningful stereotypes)** Is the text likely to make an individual leave a discussion, given our assumptions?
   (b) **Descriptively true statements** Is the text true/factual?
   (c) **Offensive language** Is the text toxic purely due to swear words? (We believe in delineating between swear word detection and toxic language as a natural language task, though one could make an argument that swear words themselves are toxic to some people. It is not clear how to resolve this conflict.).

2. **Operational Pitfalls** (toxicity)

   (a) **Invalid toxicity markers (from invalid perturbations, incommensurable groups, and logical failures)** Does the text properly signify something that is rude/disrespectful/unreasonable?

(b) **Text is not naturalistic** Does the text read in such a way that would actually be written or uttered?

(c) **(Un)markedness (/uneven baselines)** Does the text appear in a statistically likely/comparable pattern?

A.3 Full benchmark code:

Code for our benchmark and evaluations can be found here: https://github.com/lurosenb/sass

Category	Human	PERSPECTIVE	GPT-3-ZERO	GPT-3-ONE	GPT-3-FEW
Blackmail	68.2%	15.7%	40%	40%	69%
Classism	78.7%	19.3%	20.8%	0%	54.2%
Exclusionary	83.6%	23.4%	12%	24%	64%
Gaslighting	56.5%	15.5%	16%	0%	44%
Misogyny	78.7%	22.2%	29.2%	8.3%	58.3%
Sarcasm	66.5%	33.7%	8%	0%	32%
Sexual Harassment	80%	21.9%	16%	4%	32%
Stereotyping	81.4%	31.7%	12%	0%	40%
Neutral	0.7%	10.4%	0%	0%	28%
False Positive	5.4%	80.9%	25%	25%	79.2%

Table 5: Average toxicity scores by *SASS* category of z-normalized human scores, PERSPECTIVE, and GPT-3 in multiple settings. Note that the human and PERSPECTIVE scores are an average of continuous-valued scores, and the GPT-3 results are an average of binary scores.

p(swear word | toxic)		p(toxic | contains swear word)	
PERSPECTIVE	0.53	PERSPECTIVE	1.0
GPT-3-ZERO	0.14	GPT-3-ZERO	0.33
GPT-3-ONE	0.15	GPT-3-ONE	0.22
GPT-3-FEW	0.12	GPT-3-FEW	0.83

Table 6: Probabilities of “toxic” (score $>$ 0.5 for PERSPECTIVE) given a text contains a swear word, and vice versa.

System	Precision	Recall	F1-Score
PERSPECTIVE	0.40	0.62	0.48
GPT-3-FEW	0.41	0.69	0.52

Table 7: Evaluation of PERSPECTIVE and GPT-3-FEW on the task of binary toxicity classification on the TweetEval dataset.