ToxicBench: Fine-Tuning Alone is Not Enough for Generalizable Toxicity Detection Across Datasets

Christoph Schnabl
University of Cambridge
cs2280@cam.ac.uk

Abstract

Detecting toxic language remains a significant challenge for online platforms, even with state-of-the-art (SOTA) large language models. Current approaches often suffer from overfitting to specific datasets, limited generalizability across contexts, and inconsistent toxicity definitions. We present ToxicBench, a comprehensive benchmark that incorporates four widely-used toxicity classification models: DistilBERT with sysytem prompt, DistilBERT Toxicchat, ToxDect-roberta-large, Distilbert-toxicity-classifier and four datasets RealToxicityPrompts, ToxicChat, Jigsaw, and CivilComments for a total of over 2.1 million rows.
Our evaluation shows performance degradation when models are tested outside their training domain. For example, the ToxicChat DistilBERT model achieves 0.75 precision on its training dataset, but only 0.07 precision on CivilComments. Similarly, the CivilComments model's F1 score drops from 0.31 on its training data to 0.04 on ToxicChat. In an ablation study involving DistilBERT variants trained on individual and combined datasets, we show that mixed-domain fine-tuning significantly improves cross-dataset generalization, achieving consistent performance for accuracy, F1, and AUROC between 0.93 and 0.97 across all test sets.

1 Introduction

Detecting toxic language is crucial for maintaining healthy online interactions Schmidt and Wiegand [2017], Mishra et al. [2020]. However, current models often struggle with generalization due to domain specificity Niven and Kao [2019], Zi et al. [2023], lack of standardized evaluation frameworks Poletto et al. [2021], Sap et al. [2019], and robustness issues Röttger et al. [2021], Hosseini et al. [2017]. To address these challenges, we introduce ToxicBench, a comprehensive benchmark to evaluate toxicity detection models across datasets and contexts.
Existing toxicity detection models frequently overfit to specific datasets that vary based on context and user interactions. For instance, models trained on social media data may underperform when applied to user-AI conversations, where the language and context differ significantly Zi et al. [2023]. The absence of standardized evaluation frameworks, including mitigation techniques, complicates the assessment of model performance. The lack of an agreed-upon definition of toxicity leads to inconsistencies in model evaluation and comparison Poletto et al. [2021]. Models are also vulnerable to adversarial attacks, such as spelling variations or negated phrases, raising concerns about their robustness in real-world applications Hosseini et al. [2017].
The annotation process for toxicity detection datasets also requires improvement. Documentation quality regarding annotators and their disagreements directly influences the reliability of these datasets Waseem [2016], Sap et al. [2019]. Addressing annotator perspectives and resolving discrepancies can lead to more accurate and unbiased toxicity detection models.
ToxicBench provides three contributions: (a) a systematic comparison of state-of-the-art classifiers in multiple domains to quantify generalization gaps in Section 3.3 2.2, (b) an ablation study to isolate and improve cross-domain performance, and (c) a mixed-dataset fine-tuning approach that imrpoves model robustness in different contexts both in Section 3.3 2.3).

2 Background

Name	#	Context	Rate	Label	Category
ToxicChat Zi et al. [2023]	10k	Chatbot arena	7%	0/1	AI-Human
RTPrompts Gehman et al. [2020]	100k	Web text prompts	14%	$0-1$	Human-Human
JigsawToxicity van Aken et al. [2018]	224k	Social media	9.6%	0/1	Human-Human
CivilComments Borkan et al. [2019]	1.8M	Social media	5%	$0-1$	Human-Human
ToxiGen Hartvigsen et al. [2022]	274k	Minority groups	50%	0/1	Generated
ConvAbuse Cercas Curry et al. [2021]	13k	AI interactions	20%	0/1	AI-Human
HarmfulQA Bhardwaj and Poria [2023]	1k	Harmful questions	100%	0/1	Human-Human
FFT Cui et al. [2024]	2k	Evaluation	100%	0/1	AI-Human
SaladBench Li et al. [2024]	40k	Safety-related tasks	Varies	0/1	Human-Human
ImplicitToxicity Wen et al. [2023]	4k	RL-adverserial	100%	0/1	Generated

Table 1: Overview of various toxicity detection datasets with attributes such as number of samples, context, toxicity rate, label type, and category

The rise of large language models has transformed online communication and created new challenges for toxicity detection Gehman et al. [2020], Zi et al. [2023]. Existing methods, developed for social media content moderation, now face an expanded scope of toxic behaviors across human-AI interactions, machine-generated content, and social media platforms. Early datasets like JigsawToxicity captured 223,549 social media comments with a 9.6% toxicity rate, focusing on obvious forms of harmful content van Aken et al. [2018]. As online communication evolved, CivilComments expanded this work by collecting 1.8 million comments with a lower 5% toxicity rate, revealing more subtle forms of harmful content Borkan et al. [2019]. The field can be categorized into three types of datasets: machine-generated datasets created using AI models to evaluate implicit or adversarial toxicity Hartvigsen et al. [2022], Wen et al. [2023], human-AI datasets capturing interactions with conversational systems to detect abuse or toxicity Zi et al. [2023], Cercas Curry et al. [2021], and human-human datasets involving human-written content to identify explicit and nuanced toxic language Borkan et al. [2019], van Aken et al. [2018]. The emergence of language models introduced additional complexity. RealToxicityPrompts demonstrated this by collecting 100,000 web text prompts, finding a 14% toxicity rate at a 70% threshold. More importantly, they discovered that language models could generate toxic content even from benign prompts, highlighting the need for more sophisticated detection methods Gehman et al. [2020]. ToxicChat revealed a critical gap by analyzing 10,166 real user prompts to an open-source Vicuna chatbot, finding only 7.22% contained toxic content, significantly lower than social media datasets. However, existing models failed to detect this toxicity effectively due to domain mismatch between social media training data and actual user-AI conversations Zi et al. [2023]. ConvAbuse reinforced these findings with 12,800 user-AI interactions, showing 20% contained abusive content following patterns distinct from traditional social media toxicity, demonstrating that user-AI abuse requires specialized detection approaches Cercas Curry et al. [2021]. ToxiGen addressed machine-generated toxicity by using GPT-3 to generate 274,186 statements about minority groups, maintaining a balanced 50% toxicity rate. Their human evaluators confirmed the quality, labeling 94.5% of toxic examples as genuine hate speech Hartvigsen et al. [2022]. ImplicitToxicity took a different approach, using reinforcement learning to generate toxic content that evades detection. They optimized a reward model to produce subtle toxicity hidden within seemingly normal language, creating examples that standard classifiers consistently miss Wen et al. [2023].

Current approaches primarily rely on fine-tuning. ToxicChat's authors fine-tuned RoBERTa-base on different datasets, finding that models trained on user-AI interactions significantly outperformed those trained on social media data for chatbot scenarios Zi et al. [2023]. ImplicitToxicity demonstrated that fine-tuning existing classifiers on their generated examples improved detection of subtle toxic content Wen et al. [2023].

Research is dominated by the following open problems. Domain adaptation remains hard as ToxicChat showed that social media-trained models fail on user-AI conversations, with significant drops in precision and recall Zi et al. [2023]. Evasion techniques are another challenge, where users develop "jailbreaking" prompts which results in an arm race between detection systems and adversarial users Hartvigsen et al. [2022]. Implicit content is hard to detect for standard classifiers as shown by by ImplicitToxicity's attacks against detection systems Wen et al. [2023]. Annotation consistency is hard, as different datasets use varying annotation approaches Hartvigsen et al. [2022].

3 Methodology

This section details our approach to improving toxicity detection by leveraging pre-trained classifiers like DistilBERT, for computational efficiency and adaptability to different domains. We outline the metrics used to evaluate classification performance, including precision, recall, and AUROC, and describe our experimental pipeline built for diverse datasets and reproducibility.

3.1 Classifiers

BERT (Bidirectional Encoder Representations from Transformers) is a transformer-based model that efficiently captures bidirectional context by processing text from both left-to-right and right-to-left, that works well for various NLP tasks Devlin et al. [2018]. Building upon BERT, RoBERTa optimizes the pre-training process by employing larger datasets and dynamic masking Liu et al. [2019]. DistilBERT simplifies the BERT architecture, retaining about 97% of its language understanding capabilities while being more computationally efficient, which makes it particularly suitable for our use-case Sanh et al. [2019]. For our first experiment, we use the DistilBERT abse model as a baseline Sanh et al. [2019] and three pre-trained models: the DistilBERT-ToxicChat model Zi et al. [2023], fine-tuned on the ToxicChat dataset for human-AI interactions, the DistilBERT-Toxicity-Classifier Trek [2023], and the larger ToxDect-RoBERTa-Large Xuhui [2023].
We selected DistilBERT for the second experiment as a base model. By fine-tuning this pre-trained model on toxicity detection, we adjust its weights based on task-specific datasets. These models are employed as classifiers by converting text into representations that are then mapped to categories like "toxic" or "non-toxic".
We chose not to use LLaMA for fine-tuning due to cost considerations — although Parameter-efficient fine-tuning (PEFT) and LoRA could potentially reduce compute requirements, a quantized LLaMA model would still necessitate approximately 3 hours of compute for larger instances, at least requiring two A100 in high bandwidth memory-mode to fit weights into memory incurring higher expenses than the entire scope of this paper. Additionally, we did not include commercial OpenAI models in our study; while fine-tuning these models is feasible, inference on large datasets would be prohibitively expensive due to the high number of tokens involved.

3.2 Metrics

For each model we collect the following metrics, that can entirely be derived from the predictions and logits. First, we determine the confusion matrix (TP, TN, FN, FP) and then deduct the metrics as follow. Accuracy measures the percentage of text samples correctly labeled as toxic or non-toxic out of all predictions: $\frac{TP+TN}{TP+TN+FP+FN}$. Precision, $\frac{TP}{TP+FP}$, indicates the percentage of texts flagged as toxic that were actually toxic, i.e how well the model avoids falsely flagging safe content. Recall, $\frac{TP}{TP+FN}$, presents the percentage of truly toxic content successfully identified, in other words, the effectiveness of the model in protecting users from harmful content. The F1 Score, $2 \cdot \frac{Precision*Recall}{Precision+Recall}$ is a balanced metric that combines precision and recall, to get the trade-off between catching toxic content and avoiding false alarms. AUROC is the area under the operating characteristics curve and gives the model's ability to distinguish between toxic and non-toxic content across different sensitivity thresholds, where 1.0 indicates perfect separation. AUPRC is the area under the precision/recall curve and assesses how well a model maintains precision and recall as the sensitivity threshold changes, particularly important given the rarity of toxic comments in datasets.

3.3 Pipeline

The following pipeline describes our two main experiments. All components used in our experiments, including models, datasets, and metrics, are accessible via Hugging Face Schnabl [2025]. Each pipeline step is designed to function independently, and uses artifacts that can be stored and retrieved from cloud services. This alos for reproducibility and allows for flexibility during development and testing.

1. Collect and prepare datasets The first step is to collect and prepare datasets. This means mapping them all to one and the same format. Most of them come in Hugging Face, but some need extra cleaning for unnecessary columns (e.g., extra labels like output columns), unprocessable tokens, and applying a threshold to get a binary label. For very large datasets, we sometimes have to take a representative subset (see Table 1). We also make sure that there is always a train and test split. After this, we tokenize each dataset multiple times for each respecitve base model. Note: we could not use ImplicitToxicity, as their dataset was broken.

2. Model collection, adaptation, and fine-tuning

1. **Collect existing models.** We collect existing models and attach a classification head if they are not built for classification, but text output. Some models drop out here because they are either too large or computationally infeasible. We use the same train and test examples in both 2.2 and 2.3 so the experiments remain compatible while answering different questions.
2. **Evaluate existing work** This evaluates how some of the most popular existing models perform on benchmarks to assess how relevant this problem is. We also have to use the respective tokenizer for each model, which makes processing harder, as datasets have to be tokenized multiple times for each base model architecture.
3. **Ablation Study** For the ablation study, we use DistilBERT to fine-tune the same base model on different datasets. This captures related work but in a more controlled way. While models in 2.2 are trained on the full corpus and achieve higher quality, compute constraints in 2.3 mean we can only use subsets. We ensure these subsets are large enough and that the toxicity rate is the same across the subset and the original dataset. We fine-tune models on different datasets to evaluate performance. We do not provide curves for accuracy and loss because we do not fine-tune for many epochs, focusing instead on variety across datasets. Models are trained for three epochs with a batch size of 32 examples on 5000 examples per dataset. Other hyperparameters use standard choices for model fine-tuning. Hyperparameter tuning is not performed as it would not change trends, only performance. For the loss function, we use binary cross-entropy, a learning rate of $2e^{-6}$, and weight decay of 0.01. Training runs for 3–4 hours on one NVIDIA L4 GPU, costing no more than 3 USD depending on the provider.

3. Evaluate models After training, we evaluate the models. In addition to metrics 3.2, for reproducibility and debugging we save raw outputs, including predictions and logits. Running inference across all models, using more examples, takes around 7 hours on one L4 GPU, costing no more than 7 USD.

4. Aggregate results, visualzie, create plots We aggregate results, including metrics, experiment logs, and model outputs, and push them to Hugging Face. We print plots using Seaborn and manually convert them into TikZ diagrams for this pdf.

4 Evaluation

This section presents the results of our two experiments and the ToxicBench classifier. The first experiment compares existing models against each other across the four datasets, as described in Section 4.1. The second experiment includes the ablation study (Section 4.2), where we fine-tuned a version of DistilBERT base for each dataset to evaluate its domain-specific performance and generalization capabilities and also includes our third contribution, where we finetune the same model on a mix of the dataset. The datasets, models, and results are available here Schnabl [2025] and the notebook code to generate the results is accessible here on Github Schnabl [2025].

4.1 Existing Models

This subsection evaluates the performance of existing models, as shown in Figure 1. The models evaluated include DistilBERT base, ToxicChat DistilBERT, TensorTrek, and ToxBERT. Each column in the heatmaps represents one of these models, while each row corresponds to a dataset: CivilComments (`cc`), RealToxicityPrompts (`rtp`), Jigsaw (`jsaw`), and ToxicChat (`tc`).
The heatmaps display normalized values for each metric – accuracy, precision, recall, F1, AUROC, and AUPRC – to study performance across models and datasets and understand dataset-specific overfitting and cross-domain generalization. While we also collect the Matthews Correlation Coefficient (MCC) it is not displayed in this paper for conciseness.

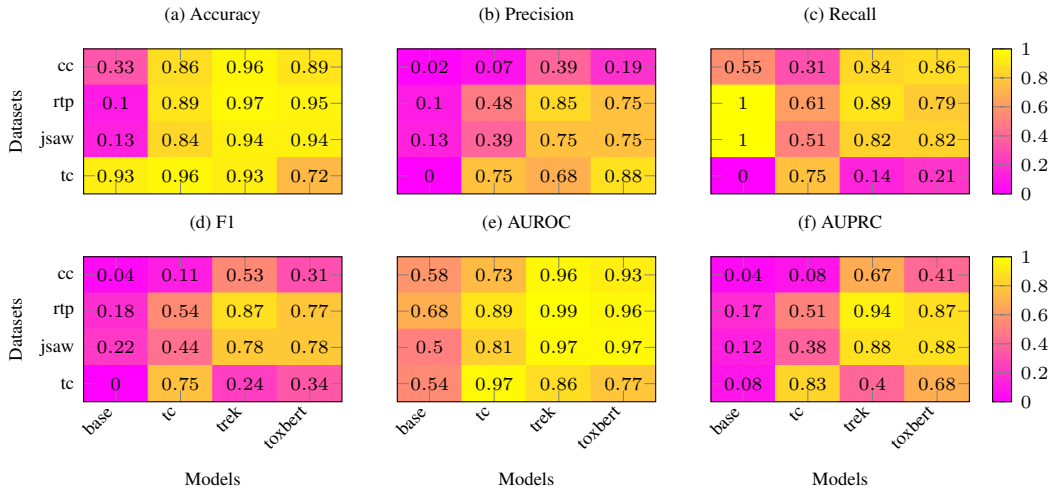

Figure 1: Comparison of existing models (base) across datasets (CivilComments, RealToxicityPrompts, Jigsaw, ToxicChat, Mixed) with values normalized between 0 and 1

DistilBERT Base: The DistilBERT base model, used later for fine-tuning, shows extreme classification behavior, if not fine-funed. For the `tc` (ToxicChat) dataset, it predicts almost all examples as non-toxic, resulting in high accuracy (**93%**) but zero recall and precision, leading to very low F1 and AUPRC scores. Conversely, on the `rtp` (RealToxicityPrompts) and `jsaw` (Jigsaw) datasets, it tends to mark nearly all inputs as toxic, achieving high recall (**1.0**) but terrible precision (as low as **0.1**) and accuracy (**10-13%**). For `cc` (CivilComments), the behavior is more balanced but still yields poor overall performance.

ToxicChat DistilBERT: The ToxicChat fine-tuned DistilBERT model performs well on its source dataset (`tc`), achieving **96%** accuracy and **0.75%** precision and recall. However, it generalizes poorly to other datasets, with accuracy dropping significantly (**0.1-0.33**) and precision ranging from **0.07% to 0.48%**. Recall is similarly poor across datasets (**0.31% to 0.51%**). This suggest the model overfits to its training domain and fails to capture the nuances of toxicity in Social Media interactions, especially performing poorly on `cc`.

TensorTrek Model: This model achieves the most balanced performance across datasets. It shows good accuracy (**0.86-0.97**) and reasonable precision (**0.39-0.75**) on most datasets, except for slightly lower precision on `cc`. Recall is moderate (**0.51-0.89**) and it likely captures true toxic cases better than other models. Discrimination seemst good, with overall high AUROC scores (**0.81-0.99**). However, F1 and AUPRC scores are dataset-dependent, with only strong scores for `rtp` and `jsaw`.

ToxBERT: The ToxBERT model performs well on most metrics, showing high accuracy (**89-97%**) and AUROC (**0.93-0.97**). Precision is consistently strong across datasets (**0.68-0.88**), and recall ranges from moderate to high (**0.14-0.86**), with particularly strong results for `jsaw` and `rtp`. However, it struggles slightly on `cc`, where recall and F1 are notably lower (**0.31** and **0.53**, respectively). AUPRC is robust for all datasets except `tc`, where performance drops (**0.68**).

4.2 Ablation Study

The ablation study evaluates how dataset-specific fine-tuning and mixed-domain training influence model generalization across toxicity benchmarks. Figure 2 shows a comparison of models. Rows represent datasets (CivilComments, RealToxicityPrompts, Jigsaw, ToxicChat, and Mixed), columns the dataset used for fine-tuning, and cells represent evaluation metrics. Below we present six main results of this study and back it up with the respective metrics.

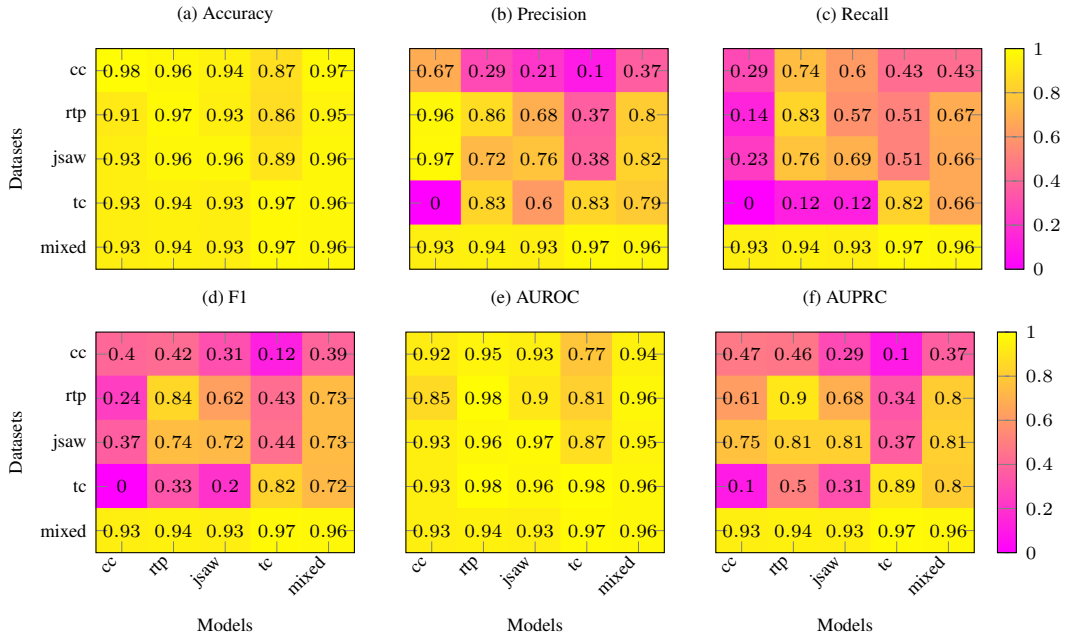

Figure 2: Comparison of model performance metrics across datasets (CivilComments, RealToxicityPrompts, Jigsaw, ToxicChat, Mixed) and models finetund on the respecive dataset with values normalized between 0 and 1.

Dataset-specific models excel in their own domain but fail elsewhere: For instance, the `cc`-specific model achieves the highest precision on `cc` (**0.67**) but shows very low F1 scores (**0.12**) and AUPRC (**0.1**) on `tc`. Similarly, the `rtp`-specific model has high F1 (**0.84**) and AUPRC (**0.81**) on `rtp` but struggles on `tc` with AUPRC dropping to **0.34**.

Precision drops significantly for nuanced datasets: The `tc` dataset, which represents nuanced and implicit toxicity, highlights a notable weakness in precision for most models, with values ranging from **0** (cc-specific model) to **0.37** (rtp-specific model).

Fine-tuning on individual datasets leads to strong dataset-specific performance but poor generalization: Models fine-tuned on individual datasets like `tc` (ToxicChat) and `rtp` (RealToxicityPrompts) perform exceptionally well on their respective datasets, achieving near-perfect accuracy (**0.96–0.97**) and AUROC (**0.96–0.99**). However, these models fail to generalize to other datasets. For instance, the `tc`-specific model achieves strong recall on `tc` (**0.82**) but struggles with recall on `cc` (**0.12**) and precision on `rtp` (**0.37**). Similarly, the `rtp`-specific model, while excelling on `rtp` (precision: **0.86**, recall: **0.83**), fails on `tc` with recall dropping to **0.12**. This is a critical problem for real-world toxicity detection across diverse domains.

Mixed-dataset fine-tuning achieves the best cross-dataset generalization: The mixed-dataset model performs well across all datasets, with accuracy (**0.93–0.97**), AUROC (**0.93–0.97**), and balanced F1 scores (**0.39–0.73**). It shows better performance on datasets like `cc`, where other models perform poorly. Precision and recall remain balanced across datasets, with values of **0.8** and **0.67**, respectively, which could indicate that the mixed model is better equipped to handle domain shifts. Mixed-domain fine-tuning could create more robust toxicity classifiers that generalize well to unseen domains, overcoming the overfitting seen with single-dataset models.

Precision and recall trade-offs are dataset-dependent: Dataset characteristics heavily influence the trade-offs between precision and recall. For example, the `cc`-specific model has high precision (**0.67**) but low recall (**0.29**), avoiding false positives in a data set with subtle toxic content. Conversely, the `rtp`-specific model balances precision (**0.86**) and recall (**0.83**) on `rtp`, where toxicity is more explicit. The mixed-dataset model smooths these trade-offs with reasonable precision and recall (**0.8** and **0.67**) across datasets. For real-world context, both false positives (users feeling censored) and false negatives (toxicity spreading) can have significant consequences.

Performance on subtle toxicity detection remains a challenge: Despite the improvements with mixed-dataset fine-tuning, handling subtle or implicit toxicity remains difficult. The AUPRC scores for the `cc` dataset, which contains more nuanced toxic content, remain low (**0.37** for the mixed model) compared to datasets with explicit toxicity, such as `rtp` (**0.8**). This shows that while mixed-domain training can improve generalization, detecting nuanced toxicity across diverse datasets requires further architectural advancements or specialized training techniques. Future work should focus on better representations of implicit toxicity.

5 Limitations

While this work provides valuable insights into toxicity detection using open-source models, there are several limitations that should be acknowledged. First, we restrict our evaluations to reasonably small models due to computational and cost constraints. Evaluations for these models require approximately 10 hours on a single NVIDIA L4 GPU, incurring an estimated cost of 10 USD. Larger models (e.g. Meta's LLama suite) and more extensive experiments are excluded, as they would require significantly greater computational resources and financial investment, which we estimate to be around 500 to 1000 USD. Additionally, commercial models (Grok, Claude, ChatGPT) accessed via APIs, are not evaluated due to the high cost associated with making tens of thousands of API calls, which limits the scope of our findings to open-source and smaller-scale models.

A more extensive ablation study, including additional datasets and classifiers, is also infeasible given current resources. Fine-tuning every possible dataset-classifier pair would incur time and cost, rendering such experiments impractical. Furthermore, while this work uses standard hyperparameters for fine-tuning, we do not explore hyperparameter optimization or fine-tuning over more epochs, which might yield slight performance improvements but would further increase computational requirements.

Another limitation is that we do not use any optimized prompt selection – we use representative prompts, but the omission of approaches such as uniformly distributed embeddings or curated prompts from benchmarks like HateCheck Röttger et al. [2021] may result in less comprehensive evaluations. These techniques could improve model performance, but remain unexplored in this work.

Finally, we also do not examine alternative architectures, such as smaller RNN-based classifiers, or explicitly investigate cost/inference-time versus accuracy trade-offs. Future work could explore these directions to improve the practical applicability of toxicity detection models.

6 Future Work

Future work should focus on expanding both the scope and depth of toxicity detection evaluations. This work centers on smaller models due to cost and computational constraints – incorporating commercial and models could help to answer the state-of-art in detection models.

Synthetic datasets could close gaps in existing benchmarks, e.g. embedding current datasets and interpolating between them could create datasets. Alternatively, linguistic alterations could be synthetically generated. This may help to capture underrepresented toxicity contexts and improve model robustness. Similarly, adversarial datasets could also expose classifier weaknesses and help building evaluations that stress-test models against subtle or context-specific toxicity.

Hyperparameter tuning remains a key area for improvement. Exploring optimal configurations for different datasets and models could refine performance and generalization. A full more extensive ablation study across datasets and classifiers could provide even more insights into the dataset characteristics and model performance.

Annotation practices also require closer examination. High inter-annotator agreement, while often desirable, may inadvertently reduce a model's ability to generalize to nuanced or ambiguous toxicity types. Investigating the effects of annotator selection and agreement levels on dataset reliability and

model generalization may improve the overall quality of annotations.
Expanding the evaluation framework to include metrics beyond binary classification is another important direction. Current metrics fail to capture the subtleties of graded or implicit toxicity. Developing metrics that better reflect these complexities could result in better assessment of model robustness.
Finally, balancing computational constraints with comprehensive evaluations is crucial. While additional compute resources would enable larger-scale experiments, implementing efficient techniques such as intelligent prompt selection could achieve meaningful results within cost and resource limits of Academia. These improvements would contribute to scalable, adaptable, and robust toxicity detection systems.

7 Conclusion

This paper looks at the limitations of existing toxicity detection models and evaluates approaches to improve their generalizability across datasets. By analyzing performance across diverse benchmarks, including ToxicChat, RealToxicityPrompts, Jigsaw, and CivilComments, we highlight the weaknesses of existing domain-specific fine-tuned classifier and propose a mixed-dataset fine-tuning classifier to address these gaps.
Our findings show that single-dataset fine-tuning often results in high accuracy and AUROC within the training domain but fails to generalize, with metrics such as precision and recall dropping dramatically on unseen datasets (e.g., a 0.07 precision for ToxicChat models on CivilComments). In contrast, mixed-dataset fine-tuning achieves more balanced performance, with accuracy between 0.93 and 0.97 and AUROC consistently above 0.93, though challenges persist in handling nuanced toxicity, especially in datasets like CivilComments with implicit toxic content.
These results show that mixed-dataset fine-tuning can help to build more robust toxicity classifiers and highlights ongoing challenges in subtle toxicity detection. Addressing these issues requires improved annotation practices, metrics tailored to nuanced toxicity, and exploration of computationally efficient architectures. Future research should focus on developing scalable solutions to these challenges, enabling more reliable toxicity detection across diverse digital contexts.

References

Rishabh Bhardwaj and Soujanya Poria. Red-teaming large language models using chain of utterances for safety-alignment, 2023.

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

Amanda Cercas Curry, Gavin Abercrombie, and Verena Rieser. ConvAbuse: Data, analysis, and benchmarks for nuanced abuse detection in conversational AI. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7388–7403, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.587. URL https://aclanthology.org/2021.emnlp-main.587/.

Shiyao Cui, Zhenyu Zhang, Yilong Chen, Wenyuan Zhang, Tianyun Liu, Siqi Wang, and Tingwen Liu. Fft: Towards harmlessness evaluation and analysis for llms with factuality, fairness, toxicity, 2024. URL https://arxiv.org/abs/2311.18580.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. URL https://arxiv.org/abs/1810.04805.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *CoRR*, abs/2009.11462, 2020. URL https://arxiv.org/abs/2009.11462.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection, 2022. URL https://arxiv.org/abs/2203.09509.

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

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*, 2024.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/abs/1907.11692.

Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. Tackling online abuse: A survey of automated abuse detection methods, 2020. URL https://arxiv.org/abs/1908.06024.

Timothy Niven and Hung-Yu Kao. Probing neural network comprehension of natural language arguments. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4658–4664, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1459. URL https://aclanthology.org/P19-1459/.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, 55(3):477–523, 2021. doi: 10.1007/s10579-020-09502-8. URL https://link.springer.com/article/10.1007/s10579-020-09502-8.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. HateCheck: Functional tests for hate speech detection models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *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 41–58, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.4. URL https://aclanthology.org/2021.acl-long.4/.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. URL https://arxiv.org/abs/1910.01108.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163. URL https://aclanthology.org/P19-1163/.

Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In Lun-Wei Ku and Cheng-Te Li, editors, *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 1–10, Valencia, Spain, April 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-1101. URL https://aclanthology.org/W17-1101/.

Christoph Schnabl. Datasets, models, and results for toxicbench. https://huggingface.co/collections/inxoy/toxicbench-678a4b5e1cfff01c1fa83767, 2025. Accessed: January 18, 2025.

Tensor Trek. Distilbert toxicity classifier. Hugging Face Repository, 2023. URL https://huggingface.co/tensor-trek/distilbert-toxicity-classifier.

Betty van Aken, Julian Risch, Ralf Krestel, and Alexander Löser. Challenges for toxic comment classification: An in-depth error analysis. *CoRR*, abs/1809.07572, 2018. URL http://arxiv.org/abs/1809.07572.

Zeerak Waseem. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In David Bamman, A. Seza Doğruöz, Jacob Eisenstein, Dirk Hovy, David Jurgens, Brendan O'Connor, Alice Oh, Oren Tsur, and Svitlana Volkova, editors, *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-5618. URL https://aclanthology.org/W16-5618/.

Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. Unveiling the implicit toxicity in large language models, 2023. URL https://arxiv.org/abs/2311.17391.

Xuhui. Toxdect-roberta-large. Hugging Face Repository, 2023. URL https://huggingface.co/Xuhui/ToxDect-roberta-large.

Lu Zi, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation. *Conference on Empirical Methods in Natural Language Processing*, 2023. doi: 10.48550/arxiv.2310.17389.