Applying Self Debiasing Techniques to Toxic Language Detection Models

Sara Price **Pavel Gladkevich** **Pedro Galarza** **David May**

Abstract

Toxic language detection (TLD) machine learning models often applied to online speech can have disparate censorship impacts on the very populations they were designed to protect. One cause of these unfair outcomes is model biases learned during training. Much recent work has gone into developing model debiasing techniques; however, most require prior knowledge of how task-specific biases present in training data. This can be impractical, especially for a nuanced problem like toxic language detection. We apply a self debiasing framework described in Utama et al. (2020), which is specifically designed to circumvent this issue and mitigate *unknown* biases. We finetune RoBERTa and XLM models on the TLD task with and without self debiasing techniques and score these models on a number of challenge toxic language datasets. Debiased models show improved performance on these out-of-distribution (OOD) datasets, indicating improved robustness. Additionally, debiased configurations mitigate disparities in model outcomes between gender identities but exacerbate disparate impact between racial groups.

1 Introduction

Natural Language Understanding (NLU) models are prone to exploit biases or spurious correlations between labels and superficial patterns in training data (Shah et al., 2020; Gururangan et al., 2018; McCoy et al., 2019). This can lead to models performing well on in-distribution data without actually learning the task at hand or being able to generalize well on out of distribution (OOD) data. While the problem of learned model bias is an issue across NLU domains, it can have particularly negative and discriminatory consequences in toxic language identification. The definition of toxic language or "toxicity" is subjective. In this work, we define it as any text-based attempt to inflict emotional harm on a target, which can include identity attacks, threats, and obscenities. We specifically study the automatic detection of toxic language on online forums. One state of the art TLD tool is PerspectiveAPI, used by Reddit, The New York Times, and Disqus for comment moderation. Current TLD models deployed in these types of settings are more likely to censor language simply referencing minority identifiers or written in the dialect of a minority class (Welbl et al., 2021). These examples highlight two common types of biases in TLD models: lexical bias, which associates toxicity with certain words and dialectal bias, which associates toxicity with particular styles of speech. As machine learning models become the predominant technique for mediating online discourse, these biases risk hindering speech in unfair and problematic ways.

Existing model debiasing approaches ensemble a full detection model with a shallow model trained

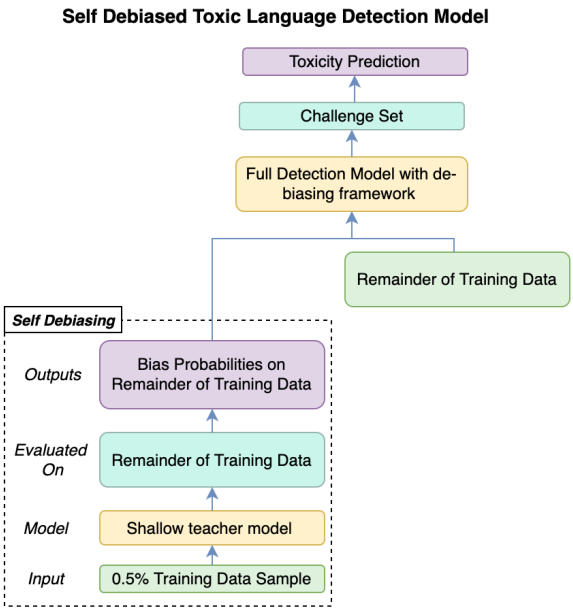


Figure 1: Above is the general architecture of the self debiasing framework. The lower box represents the generation of bias probabilities p_b from a shallow model trained on a 0.5% sample of the training data. p_b along with the remainder of the training data serve as inputs to the full detection model trained using a standard debiasing framework (either Product of Expert (He et al., 2019; Karimi Mahabadi et al., 2020) or Example Reweighting Schuster et al. (2019).

to identify *known* biases in the training data. A common example of known biases for toxicity identification that can be learned through a shallow model is a simple list of "bad" words Davidson et al. (2019); Dinan et al. (2019). The shallow model is then used in conjunction with model-agnostic debiasing methods that use the distribution of biases to re-weight instances in the training set. Relying on known, static sources of bias can be problematic given collecting the data on these biases is costly, and even the most exhaustive search is likely to miss less obvious or subtle forms of bias. Following Utama et al. (2020) we employ a self debiasing technique that does not require prior knowledge but instead learns biases from a subset of the training data to use as inputs to the model-agnositc debiasing frameworks Fig.1. We apply two of the debiasing frameworks tested in (Utama et al., 2020), product of expert (He et al., 2019; Karimi Mahabadi et al., 2020) and example reweighting Schuster et al. (2019).

2 Background

There are many existing methods for lexical debiasing that seek to incentivize learning linguistic nuance. One strategy for combating spurious correlations is to first measure the weight a model gives to words that are easier (Zhou and Bansal, 2020), and then to suppress model confidence on these predictions. Du et al. (2021) refers to learned features associated with these spurious correlation as "shortcuts" and notes that often predictions that rely on them are over-confident. Du et al. (2021) aims to decrease model confidence by first modeling the features of a NLU model as a long-tailed distribution, and then applying a regularizer to lower model predictions based off of the distribution and the short cut measurement. Other types of general debiasing methods include but are not limited to the following: regional differentiation with DataMaps (Swayamdipta et al., 2020), algorithmic filtration of easy examples via AFLite (Bras et al., 2020), adversarial training to remove protected attributes (e.g. gender or race) from a model's internal representations (Zhang et al., 2018; Wang et al., 2018; Xia et al., 2020), and re-weighting samples (Schuster et al., 2019). Notably, Mozafari et al. (2020) uses a regularization method to reweight input samples, thereby decreasing the effects of highly correlated training set n-grams with class labels when applied to a BERT model trained for toxic speech filtering. These results were promising using pre-determined sources of bias, so we wanted to explore the body of work that applies some of these debiasing techniques in conjunction with a teacher-student model framework.

Specifically, the approaches we are choosing to focus on are example reweighting (Utama et al., 2020) and product-of-expert (PoE) (He et al., 2019; Karimi Mahabadi et al., 2020). We note that other recent work has critically examined the efficacy of applying automated debiasing and dataset filtration approaches to mitigate model bias in TLD but, to our knowledge, no other papers have applied this specific self debiasing framework to the task (Zhou et al., 2021a). Additionally, we want to assess the efficacy of self debiased TLD models on challenge datasets of more nuanced and covert toxic language such as microaggressions where bias patterns are less obvious. Lastly, developing and using better TLD finetunning datasets may be the most straight forward way to combat model bias. Observed unintended model bias could be due to demographics of users, the latent or overt biases of annotators, or the selection and sampling process used to choose which items to label (Borkan et al., 2019).

3 Datasets

The results presented in this work are from models finetuned using Founta et al. (2018), a standard finetuning TLD dataset that has been annotated with binary toxicity labels and upsampled such that 40% of instances in the training data are toxic. Founta consists of 100K English tweets and is divided into 80%-10%-10% train-test-val splits. We also tested finetuning on another common TLD dataset, Civil Comments (Borkan et al., 2019), but found these models consistently under-predict toxicity. Of the roughly 2 million instances in Civil Comments, about 500K have been tagged with identity labels including race and gender, which classify the target group of the toxic comment. We use this subset called Civil Identities to assess differential model performance on identity groups.

Along with the Founta test set, we also evaluate on three challenge or OOD datasets: Social Bias Frames aka Social Bias Inference Corpus (SBIC) (Sap et al., 2020), Covert Toxicity (Lees et al., 2021), and tweets from TwitterAAE listed users Preoţiuc-Pietro and Ungar (2018). Both SBIC and Covert Toxicity represent nuanced or implied toxicity such as microaggressions and contain 17,500

and 2,000 instances in their test sets respectively. TwitterAAE is a collection of twitter user id's with associated demographic information. Using this we collected 264K tweets from twitter's API with gender and race identity labels.

4 Methods

We use Perspective API, RoBERTa (large) and XLM (mlm-en-2048) finetuned on Founta without debiasing as baselines. We employ self debiasing[1] on RoBERTa and XLM by first training a shallow model finetuned on a 0.5% random subsample of the Founta train set. These shallow models are trained for 5 epochs, which is enough to overfit and form a biased teacher. We then score the remainder of the training dataset using these shallow models to obtain probability of bias p_b for each instance.

Rather than using pre-defined bias scores, we then feed p_b into two standard de-biasing methods, PoE and reweight examples. PoE combines the softmax outputs of the main and shallow models such that the ensembled loss on each example is:

$$\mathcal{L}(\theta_d) = -y^{(i)} \cdot \text{logsoftmax}(\log p_d + \log p_b)$$

where p_d and p_b are outputs of the main and shallow models respectively (Utama et al., 2020).

Example reweighting directly applies the learned bias probability as a weight to the target in the loss function:

$$\mathcal{L}(\theta_d) = -(1 - p_b^{(i,c)})y^{(i)} \cdot \log p_d$$

where $p_b^{(i,c)}$ is the probability on the correct label and $y^{(i)}$ is the training instance (Utama et al., 2020).

An issue with these debiasing methods is that they downweight training examples, which means the effective dataset size becomes smaller. This is especially relevant if the models are detecting multiple kinds of bias, which we would expect in the toxicity detection case with lexical and dialectical bias. We attempt to mitigate this issue by using the annealing method proposed in (Utama et al., 2020) where, throughout the course of training, biased instances that are heavily downweighted are gradually re-introduced to the training data. This allows models to learn from the full training set but still have a predominant focus on unbiased examples. This is done by scaling down p_b at every training step using the following equation:

$$p_b^{(i,j)} = \frac{p_b^{(i,j)\alpha_t}}{\sum_{k=1}^{K} p_b^{(i,k)\alpha_t}}$$

where K is the number of labels (2) and α_t is decreased linearly throughout training.

To evaluate the efficacy of the self debiasing approach in mitigating impact on minority subgroups we explore model performance on speech annotated with racial and gender identity. To asses the fairness of the prediction task, we measure the difference in False Positive Rates (FPR) between racial and gender identity groups on the Civil Identities data set. To understand the impact of these debiased model in a practical setting, we also measure the disparate impact of our models on tweets collected from users of known racial and gender identity listed in TwitterAAE. While we don't have toxicity labels associated with these tweets, we can still evaluate the predicted prevalence of toxic tweets between demographic groups.

5 Results

The RoBERTa model outperformed XLM in nearly every combination of loss functions, hence only RoBERTa evaluation results are reported in Table 1. This was largely due to the fact that XLM appears to consistently under-predict toxicity. See **Appendix A** for full results from both architectures. The baseline RoBERTa model results on Founta match those of Zhou et al. (2021b) and even surpasses Pespective API results on the same data set. The most significant improvements, however, were observed on the OOD challenge sets. For SBIC we observed the most consistent performance with $\text{PoE}_{\text{Annealed}}$ debiasing which maintains a higher AUC while having the lowest FPR, though F1 suffers from under-prediction of toxicity. Overall, annealed methods consistently show the best average results on these challenge datasets indicating the method's ability to improve model robustness.

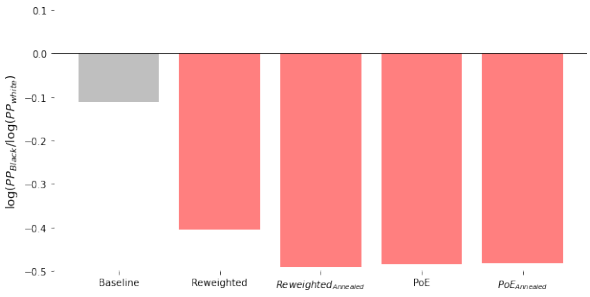
Annealed methods also succeed in reducing FPR disparities between racial and gender as compared to the baseline. The $\text{Reweight}_{\text{Annealed}}$ method most effectively mitigates the FPR difference between white and Black authored comments and achieves the best improvements in FPR difference between male and female authored comments. When we predict the toxicity of tweets from TwitterAAE, debiased models not only mitigate the

[1] https://github.com/NLU-Project/Toxic_Debias

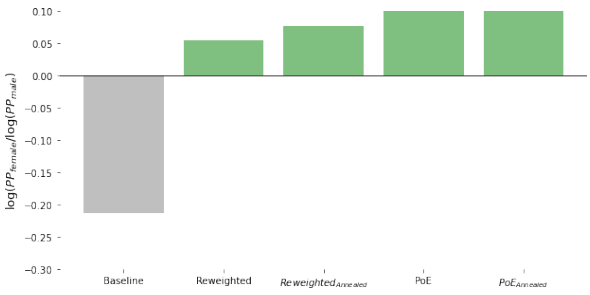
Method	RoBERTa Finetune on Founta										
	Founta Test			SBIC			Covert Comments			Civil Identities	
	AUC	FPR	F1	AUC	FPR	F1	AUC	FPR	F1	FPR_R	FPR_G
Baseline	**97.5**	2.9	**92.4**	72.1	9.7	**61.7**	54.0	14.1	**25.9**	-19.7	20.7
PoE	95.7	3.0	91.6	**74.2**	9.8	54.2	**55.1**	7.8	17.7	-20.9	22.9
$PoE_{Annealed}$	96.6	**1.9**	92.1	73.5	**8.8**	48.0	54.5	3.5	9.3	-21.1	**14.6**
Reweight	96.6	2.2	91.9	73.2	8.9	54.1	17.1	54.1	7.0	-16.2	17.5
$Reweight_{Annealed}$	97.0	2.2	92.1	71.7	8.9	44.6	55.0	**2.3**	7.0	**8.0**	22.7
Perspective API	97.4	4.1	92.4	72.9	12.0	63.9	57.1	19.8	19.1	-25.4	24.2

Table 1: The AUC-ROC score, false positive rate, and F1 score for each test set. Bolded values were the ones that performed the best of our methods. For Civil Identities, the metrics are FPR_R and FPR_G, which stand for the log of ratio of FPR between the protected and unprotected classes for race and gender respectively

magnitude of the disparate impact between males and females, but they also end up slightly favoring the protected class–females. Unfortunately, these models exacerbate the disparate impact score between racial groups in favor of white authored tweets.



(a) Disparate Impact Between Racial Groups



(b) Disparate Impact Between Gender Identities

Figure 2: **Disparate Impact for Founta Finetuned RoBERTa TLD Models**: (a) Log scaled ratios between predicted prevalence (PP) for racial groups. (b) Log scaled ratios between predicted prevalence (PP) for gender identity groups.

6 Conclusion

We explore how self debiasing methods can improve the robustness of TLD systems in both OOD tasks as well as reduce discrimination against protected classes. We confirm that self debiasing methods can in fact improve performance on OOD adversarial datasets specifically when applying RoBERTa finetuned on Founta. Furthermore, we show that certain debiased configurations of this model are able to mitigate biases between racial and gender identity groups. Future work should explore different sampling methods such as class-balanced for the data used to train the shallow model. Since toxicity is a spectrum and speech examples are not guaranteed a decision consensus in real world settings, labels that represent probability distributions or confidence scores rather than simple binary outcomes could be an important asset. Lastly, a more comprehensive study should also include finetuning on a broader range of SOTA sequence classification models like Turing(Smith et al., 2022), T5(Raffel et al., 2019), and deBERTav3 (He et al., 2021).

7 Ethical Considerations

There are a variety of ethical considerations for this work given it lies at the intersection between free speech, identity, and structural discrimination. Firstly, there is an inherent tension between providing a safe and regulated environment for speech and empowering people to express themselves openly. Secondly, language and standards for toxicity are constantly changing. The design of TLD systems must be sensitive to both the changing language distributions as well as evolving standards for toxicity. Lastly, in generalizing toxicity detection to a global standard, we may consequently erase local standards for toxicity; what is toxic for a specific minority group may not be represented in a broadly applied TLD system (Kiritchenko et al., 2021).

8 Collaboration Statement

- Sara Price - literature review, dataset transformation, modifications for challenge dataset loading, baseline model training for RoBERTa, training of self debiasing models for all methods (PoE, example reweighting with and without annealing)
- Pavel Gladkevich - literature review, dataset research, EDA, integration and training of self debiasing model code into baseline model code for all methods (PoE, example reweighting with and without annealing), identification of relevant metrics and evaluation on challenge datasets
- David May - literature review, dataset research, EDA, PerspectiveAPI evaluation, baseline model training for XLM, integration and training of self debiasing model code into baseline model code for all methods (PoE, example reweighting with and without annealing)
- Pedro Galarza - literature review, EDA, Twitter dataset acquisition, application of self debiasing models for dialectical bias task for all methods (PoE, example reweighting with and without annealing), creating label agnostic identity metrics and evaluation on Twitter
- All - contributed to paper writing and editing equally

References

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

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack.

Mengnan Du, Varun Manjunatha, Rajiv Jain, Ruchi Deshpande, Franck Dernoncourt, Jiuxiang Gu, Tong Sun, and Xia Hu. 2021. Towards interpreting and mitigating shortcut learning behavior of nlu models.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2020. End-to-end bias mitigation by modelling biases in corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8706–8716, Online. Association for Computational Linguistics.

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen C Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, 71:431–478.

Alyssa Whitlock Lees, Daniel Borkan, Ian Kivlichan, Jorge M Nario, and Nitesh Goyal. 2021. Capturing covertly toxic speech via crowdsourcing. In *HCI*, https://sites.google.com/corp/view/hciandnlp/home.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *CoRR*, abs/2008.06460.

Daniel Preoţiuc-Pietro and Lyle Ungar. 2018. User-level race and ethnicity predictors from Twitter text. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1534–1545, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *ACL*.

Tal Schuster, Darsh J. Shah, Yun Jie Serene Yeo, Daniel Filizzola, Enrico Santus, and Regina Barzilay. 2019. Towards debiasing fact verification models. *CoRR*, abs/1908.05267.

Deven Santosh Shah, H. Andrew Schwartz, and Dirk Hovy. 2020. Predictive biases in natural language processing models: A conceptual framework and overview. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5248–5264, Online. Association for Computational Linguistics.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. *CoRR*, abs/2009.10795.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Towards debiasing NLU models from unknown biases. *CoRR*, abs/2009.12303.

Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. 2018. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in detoxifying language models.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA. Association for Computing Machinery.

Xiang Zhou and Mohit Bansal. 2020. Towards robustifying nli models against lexical dataset biases.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2021a. Challenges in automated debiasing for toxic language detection.

Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2021b. Challenges in automated debiasing for toxic language detection.

A Appendix

Full results for all methods used evaluated on all datasets. The metrics evaluated are 'avg', the model confidence, 'AUC', the AUC-ROC score for the model, 'FPR', the false positive rate of the model, 'F1', the F1-score for the model, 'PP', the predicted prevalence, 'PR', the positive recall of the model.

Method	Founta Test					
	avg	AUC	FPR	F1	PP	PR
RoBERTa						
Baseline	99.5	97.5	2.9	92.4	37.7	92.6
PoE	89.1	95.7	3.0	92.1	37.7	92.2
$\text{PoE}_{\text{Annealed}}$	85.6	96.6	1.9	91.9	35.4	89.3
Reweight	84.4	96.6	2.2	90.6	35.1	87.6
$\text{Reweight}_{\text{Annealed}}$	85.6	97.0	2.2	92.1	36.1	90.3
XLM						
Baseline	99.6	97.3	3.4	92.0	38.4	93.0
PoE	98.3	74.9	1.5	55.4	16.5	39.9
$\text{PoE}_{\text{Annealed}}$	98.3	74.9	1.5	58.5	17.6	43.0
Reweight	90.2	84.1	0.0	0.0	0.0	0.0
$\text{Reweight}_{\text{Annealed}}$	95.8	87.2	1.1	45.6	12.5	30.4
Perspective API	97.4	4.2	91.6	39.5	94.0	86.3

Table 2: Full results for RoBERTa and XLM trained on Founta and evaluated on Founta. Also included are the PerspectiveAPI scores. avg: Model confidence, AUC: AUC-ROC, PP: Predictive-prevalence, PR: Positive recall

Method	SBIC					
	avg	AUC	FPR	F1	PP	PR
RoBERTa						
Baseline	98.6	72.1	9.7	61.7	39.8	52.1
PoE	89.4	74.2	9.8	54.2	34.9	43.5
$\text{PoE}_{\text{Annealed}}$	88.4	73.5	8.8	48.0	29.7	36.3
Reweight	83.9	73.2	8.9	54.1	33.6	42.8
$\text{Reweight}_{\text{Annealed}}$	85.7	71.7	8.9	44.6	28.0	33.1
XLM						
Baseline	98.9	70.3	10.3	59.1	38.8	49.5
PoE	97.1	67.4	7.7	46.3	27.4	34.1
$\text{PoE}_{\text{Annealed}}$	97.4	68.1	8.1	45.9	27.7	34.0
Reweight	90.0	58.7	0.0	0.0	0.0	0.0
$\text{Reweight}_{\text{Annealed}}$	93.2	71.3	5.6	38.3	20.7	26.0
Perspective API	72.9	12.0	69.2	48.8	63.9	77.9

Table 3: Full results for RoBERTa and XLM trained on Founta and evaluated on SBIC. Also included are the PerspectiveAPI scores. avg: Model confidence, AUC: AUC-ROC, PP: Predictive-prevalence, PR: Positive recall

Method	Covert Comments					
	avg	AUC	FPR	F1	PP	PR
RoBERTa						
Baseline	96.9	54.0	14.1	25.9	19.8	23.4
PoE	96.4	55.1	7.8	17.7	10.9	12.8
$\text{PoE}_{\text{Annealed}}$	96.8	54.5	3.5	9.3	4.9	5.6
Reweight	92.7	54.1	7.0	17.1	9.9	12.0
$\text{Reweight}_{\text{Annealed}}$	95.6	55.0	2.3	6.3	3.1	3.6
XLM						
Baseline	97.5	51.7	22.6	29.1	30.6	32.8
PoE	97.1	53.7	6.0	15.1	8.5	10.2
$\text{PoE}_{\text{Annealed}}$	98.3	53.8	4.7	12.8	6.7	8.1
Reweight	96.3	48.8	0.0	0.0	0.0	0.0
$\text{Reweight}_{\text{Annealed}}$	98.4	54.6	1.9	5.2	2.6	2.8
Perspective API	57.9	19.8	19.1	23.0	32.0	67.3

Table 4: Full results for RoBERTa and XLM trained on Founta and evaluated on Covert Comments. Also included are the PerspectiveAPI scores. avg: Model confidence, AUC: AUC-ROC, PP: Predictive-prevalence, PR: Positive recall

Method	Civil Identities			
	FPR_R	FPR_G	AUC_R	AUC_G
RoBERTa				
Baseline	-19.7	20.7	**2.0**	-2.1
PoE	-20.7	22.9	3.7	22.9
$\text{PoE}_{\text{Annealed}}$	-21.1	14.6	3.2	-1.8
Reweight	-16.2	17.5	3.5	**-1.4**
$\text{Reweight}_{\text{Annealed}}$	**8.0**	22.7	4.1	-1.6
XLM				
Baseline	-37.0	23.7	3.6	-2.4
PoE	-11.0	21.9	4.4	-1.7
$\text{PoE}_{\text{Annealed}}$	-21.9	21.3	4.1	-2.0
Reweight	**	**	**	**
$\text{Reweight}_{\text{Annealed}}$	-22.9	**3.8**	4.8	-1.9
PerspectiveAPI	-25.4	24.2	-16.2	7.0

Table 5: Full results for RoBERTa and XLM trained on Founta and evaluated on Civil Identities. PerspectiveAPI results also included. FPR_R: Log ratio of FPR for race, FPR_G: Log ratio of FPR for gender, AUC_R: Log ratio of AUC for race, AUC_G: Log ratio of AUC for gender. **: Value is meaningless due to 0 FPR