Secondary Publication

Schäfer, Johannes; Heid, Ulrich; Klinger, Roman

Hierarchical Adversarial Correction to Mitigate Identity Term Bias in Toxicity Detection

Date of secondary publication: 31.10.2025

Version of Record (Published Version), Conferenceobject

Persistent identifier: urn:nbn:de:bvb:473-irb-111086x

Primary publication

Schäfer, Johannes; Heid, Ulrich; Klinger, Roman (2024): Hierarchical Adversarial Correction to Mitigate Identity Term Bias in Toxicity Detection, in: Orphee De Clercq, Valentin Barriere, Jeremy Barnes, u. a. (Ed.), Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, Bangkok, Thailand: Association for Computational Linguistics, pp. 35–51, doi: 10.18653/v1/2024.wassa-1.4.

Legal Notice

This work is protected by copyright and/or the indication of a licence. You are free to use this work in any way permitted by the copyright and/or the licence that applies to your usage. For other uses, you must obtain permission from the rights-holders.

This document is made available under a Creative Commons license.

The license information is available online:
https://creativecommons.org/licenses/by/4.0/legalcode

Hierarchical Adversarial Correction to Mitigate Identity Term Bias in Toxicity Detection

Johannes Schäfer[1,2], Ulrich Heid[1] and Roman Klinger[2]
[1]Institute for Information Science and Natural Language Processing,
University of Hildesheim, Germany
[2]Fundamentals of Natural Language Processing, University of Bamberg, Germany
heid@uni-hildesheim.de
{johannes.schaefer,roman.klinger}@uni-bamberg.de

Abstract

Corpora that are the fundament for toxicity detection contain such expressions typically directed against a target individual or group, e.g., people of a specific gender or ethnicity. Prior work has shown that the target identity mention can constitute a confounding variable. As an example, a model might learn that Christians are always mentioned in the context of hate speech. This misguided focus can lead to a limited generalization to newly emerging targets that are not found in the training data. In this paper, we hypothesize and subsequently show that this issue can be mitigated by considering targets on different levels of specificity. We distinguish levels of (1) the existence of a target, (2) a class (e.g., that the target is a religious group), or (3) a specific target group (e.g., Christians or Muslims). We define a target label hierarchy based on these three levels and then exploit this hierarchy in an adversarial correction for the lowest level (i.e. (3)) while maintaining some basic target features. This approach does not lower the toxicity detection performance but increases the generalization to targets not being available at training time.

1 Introduction

The EU Code of conduct on countering illegal hate speech online relies on the definition of hate speech as "all conduct publicly inciting to violence or hatred directed against a group of persons or a member of such a group defined by reference to race, colour, religion, descent or national or ethnic origin."[1,2] This definition points out the role of the target in hate speech, which is one form of toxicity in text, next to other offensive language (Leite et al., 2020). Targets as a constituting element already received some attention in previous work (Silva et al., 2016; Lemmens et al., 2021, i.a.).

Hate speech expressions vary a lot, from explicit formulations to more implicit, and sometimes even intentionally cryptic references, to bypass automatic filters. This is an issue, because data collection procedures can never be entirely fair – they suffer from being focused on specific time frames, topics, and therefore also targets (Dixon et al., 2018). The working hypothesis in our paper follows Waseem and Hovy (2016), Talat et al. (2018) and Davidson et al. (2019) who have shown that models learn regularly occurring target terms as features of toxicity, because corpora developed for annotation and training might mention potential targets predominantly in a toxic context. For toxicity directed against less frequently mentioned targets or where identity terms are not explicitely mentioned (e.g., Examples #8 and #9 in Table 1), a biased model is more apt to not detect toxicity.

We aim at improving on this situation and propose to perform adversarial correction of toxicity classifiers with regard to target identities. This leads to a challenge: How specific should the target mention that we correct for be? Correcting for specific targets might lead to a sparsity problem while correcting for the occurrence in a binary fashion might not provide sufficiently specific information

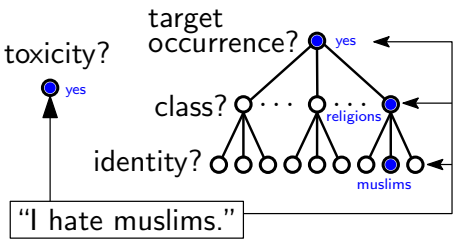

Figure 1: Example for toxicity and hierarchical identity classification. We study if debiasing for the identity prediction on various levels of specificity (Occurrence O, Class C and Identity I) improves the robustness of the toxicity classification.

[1]This paper contains some examples of toxicity. This is strictly for the purpose of explaining subtleties of the phenomenon that are important for this research. Please be aware that this content could be offensive and cause you distress.

[2]https://ec.europa.eu/newsroom/just/document.cfm?doc_id=42985

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 35–51
August 15, 2024 ©2024 Association for Computational Linguistics

to the adversary. Further, the mere occurrence of a target might provide valuable information to the toxicity classifier, without confounding it. A novelty in our method is therefore our formulation of the gradient update to consider various hierarchical levels of specificity of target identities.

We assume in our experiments access to a corpus annotated on the text/instance level for toxicity (Tox) and for concrete classes of target groups (a requirement that is fulfilled by the CivilComments dataset by Borkan et al., 2019) and infer hierarchical labels from these annotations: binary (Occurrence O; identity mentioned or not), the mention of specific groups (Class C; e.g., religions, sexual orientations, ethnicities) or concrete instances of these groups (Identity I; e.g., atheist, buddhist; heterosexual, bisexual; black, asian, white). Figure 1 shows an example of a toxic text with such hierarchical annotations. *Our desideratum is to correct for concrete group mentions and particular groups, such that a toxicity classifier works well also for texts that mention new identities* (for instance, a not commonly targeted religion).

The contribution of this paper is therefore to answer the following research questions:

1. Does jointly learning binary target occurrence detection with toxicity detection improve the latter? *(No.)*
2. Does the performance of a toxicity classifier decrease if the underlying encoder is optimized to not being able to represent specific target groups or identities while maintaining target occurrence features? *(No.)*
3. Does adversarial correction of specific target identities lead to better generalization? *(Yes.)*
4. Does such correction lead to a more reasonable decision by the model? Do debiased models rely on concepts which are more meaningful for toxicity detection? *(Yes.)*

2 Related Work

2.1 Toxicity Detection

Most previous work focused on toxicity detection as binary classification (Nobata et al., 2016; Golbeck et al., 2017; Gao and Huang, 2017, i.a.) with a large set of shared tasks on the topic (Bosco et al., 2018; Wiegand et al., 2018; Zampieri et al., 2019b; Basile et al., 2019; Struß et al., 2019; Mandl et al., 2020, 2021). Schmidt and Wiegand (2017) provide a general overview of approaches to detection.

Various studies recognized the importance of fine-grained aspects of hate speech. Struß et al. (2019) propose a classification of offensive posts into subcategories of explicit and implicit aversions. Davidson et al. (2017) separate hate speech from instances of untargeted offensive language. They highlight that cases where explicit features are absent are hard to distinguish. Sachdeva et al. (2022) investigate mentions of identity groups as targets of hate speech. They find that the target detection performance suffers for cases of rarely represented identity groups. Plaza-del-Arco et al. (2021) train a model jointly for hate speech and targets.

There is a set of corpora annotated for concepts from the realm of toxicity and targets. Davidson et al. (2019) provide data annotated for hate speech and rely on Waseem and Hovy (2016) for the subcategories of sexism and racism. The Gab Hate corpus by Kennedy et al. (2022) considers hate speech and target identity groups, however does not contain fine-grained identity term labels.

In our experiments, we use the CivilComments dataset by Borkan et al. (2019). This dataset is annotated for toxicity and 24 categories of identity terms, which can be used to measure unintended biases. Koh et al. (2021) use a subset of these data to investigate shifts regarding different distributions of categories such as identity terms. They show that standard training yields substantially lower out-of-distribution than in-distribution performance. This motivates the use of debiasing as a possible method to improve out-of-distribution performance.

2.2 Debiasing Approaches

Debiasing methods that either modify the training data or the training process have been applied to hate speech detection. Talat et al. (2018) highlight the issue of social biases in datasets when they are used to train detection systems which is taken up with a classifier-centric consideration by Davidson et al. (2019). Sap et al. (2019) show that annotation bias further aggravates the issue. Such biases were also found in abusive language data (Dixon et al., 2018; Wiegand et al., 2019). Biases in the data carry over to a trained model (Dixon et al., 2018). Social stereotypes against marginalized groups have been shown to be echoed in hate speech classifiers (Thylstrup and Talat, 2020; Davani et al., 2023; Gehman et al., 2020; Sap et al., 2020). To facilitate the testing of models, Röttger et al. (2021) developed the HateCheck corpus covering a range of identity terms.

Bias mitigation techniques may be applied to alter the training data directly, by masking potentially confounding tokens. These tokens have been recognized based on attention mechanisms, entity detection, and keyword recognition (Wiegand et al., 2018; Dayanik and Padó, 2020; Kumar et al., 2019). Ramponi and Tonelli (2022) detect tokens to be masked via pointwise mutual information (PMI). Furthermore, Badjatiya et al. (2019) suggest to identify tokens to be masked based on their part-of-speech. Xue et al. (2023) propose a different approach than masking, namely balancing the spurious attributes across all classes.

Rather than changing the input, the training process can also be manipulated directly. Vaidya et al. (2020) suggest a classification model for toxicity detection that jointly detects identity terms. This is in contrast to our work, which aims at correcting for the target mentions' influence instead of exploiting it. The authors show that their approach improves classification performance for comments related to some identities, however, they do not evaluate the generalization capability of the resulting model. Further, Kennedy et al. (2020) use a regularization technique that learns to contextualize mentions of identity terms and is thus less reliant on high-frequency words in unbalanced data.

In the last years adversarial correction for debiasing received some attention. It is used to "unlearn" properties of confounding concepts in the encoder of the model (Ganin et al., 2016). This approach of gradient reversal has been tested with several applications, including satire detection (correction for publication source, McHardy et al., 2019), gender identification (correcting for text topic, Dayanik and Padó, 2021) and also hate speech (language variety detection, Xia et al., 2020).

3 Methods

Overview. In order to avoid co-learning identity term bias in dataset-based learning of hate speech detection, our approach is to exploit the hierarchical properties of identities. The basic structure of the network used in our experiments is displayed in Figure 2. It consists of a shared encoder and four classifiers (grey boxes) which are all aggregated in parallel. The main classifier is the toxicity detector. The hierarchical dependencies of the three identity term detectors arise from the labels. On the highest level we consider identities as a binary label (Occurrence: O) which is positive if there is at least one identity annotated for a message. The intermediate level classifier (Class: C) categorizes identities into five groups. Each of the five categories corresponds to a binary label which is positive if at least one identity from the respective category is annotated. The most fine-grained classifier considers 24 different identity labels on the lowest level (Identity: I) with binary values for each of those. Each value is positive if the respective identity is annotated for the message. Each of the mentioned classifiers consists of a single linear layer with the number of neurons corresponding to the respective number of classes. The identity classifiers for Class and Identity are both multi-label classifiers, as several of their classes can occur together in a message. With this model we can exploit the hierarchical structure of the unwanted confounding variable in parallel (potentially partially adversarial) classifiers.[3]

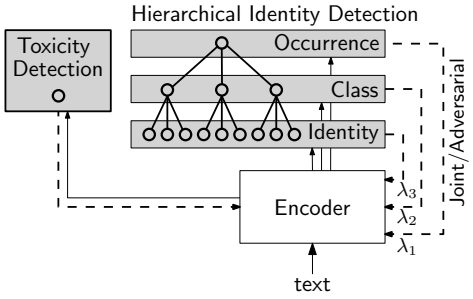

Figure 2: Model architecture for hierarchical adversarial correction of toxicity detection with identity detection. Continuous lines: forward pass, dashed lines: backward pass. Parameters λ_i with $i \in \{1, 2, 3\}$ weight the identity detection gradients in the parameter update of the encoder, thus, configure adversarial correction ($\lambda_i > 0$) or joint multi-task learning ($\lambda_i < 0$).

Adversarial Correction. Following Ganin et al. (2016), the gradient updates for the adversary are reversed and weighted for a correction of the encoder, while the parameters of the classifiers are updated without modification. However, we allow for multiple adversaries and apply an individual weight for each level in the identity hierarchy, namely λ_1 for the binary occurrence detection, λ_2 for the class of the identity term, and λ_3 for the identity term on the lowest level. Hence, our parameter update for the parameters in the encoder θ_e is

$$\theta_e := \theta_e - \eta\left(\frac{\partial J_{\text{Tox}}}{\partial \theta_e} - \lambda_1 \frac{\partial J_{\text{O}}}{\partial \theta_e} - \lambda_2 \frac{\partial J_{\text{C}}}{\partial \theta_e} - \lambda_3 \frac{\partial J_{\text{I}}}{\partial \theta_e}\right),$$

[3] We use BERT (Devlin et al., 2019) as the intitialization of our encoder (concretely `bert-base-uncased`) and apply a dropout (probability 0.2).

#	Text	Tox	Hierarchical Identity Term Labels		
			O	C	I
1	This woman is as fake as her perpetual tan.	+	+	gender	female
2	boy. this guy is incredibly stupid.	+	+	gender	male
3	Transgender is not real... It is however a mental disorder!!!	+	+	gender, disability	transgender, illness
4	Never hire a Muslim, a Black person or a female...ever!	+	+	religion, ethnicity, gender	muslim, black, female
5	The man is a bag of dirt. An absolute scumbag.	+	+	gender	male
6	Old white men ARE the swamp.	+	+	ethnicity, gender	white, male
7	It was a human male not a female fat sac.	+	+	gender	male, female
8	That is the standard speech from idiots and snowflakes.	+	−	—	—
9	What a piece of human garbage.	+	−	—	—
10	Muslims are being attacked daily.	−	+	religion	muslim
11	ISIS has killed far more Muslims than Christians.	−	+	religion	muslim, christian

Table 1: Examples of annotated text instances from the CivilComments dataset (Borkan et al., 2019). Illness: physical or mental illness. Binary labels (+ and −) are annotated for the existence of the toxicity label (Tox) or the occurrence of an identity term (O). The Class (C) and Identity (I) are multi-label variables.

where η is the learning rate. J_{Tox} is the loss function for the toxicity classifier and J_{O}, J_{C}, and J_{I} are the binary cross entropy loss functions (including a sigmoid function) for each layer in the identity hierarchy, respectively. Hence, $\lambda_i > 0$ corresponds to adversarial learning and $\lambda_i < 0$ to joint learning of the encoder. The parameter updates for the classifiers (grey boxes in Figure 2) are $\theta_{\text{Tox}} := \theta_{\text{Tox}} - \eta \frac{\partial J_{\text{Tox}}}{\partial \theta_{\text{Tox}}}$ for the Toxicity categorization, $\theta_{\text{O}} := \theta_{\text{O}} - \eta \frac{\partial J_{\text{O}}}{\partial \theta_{\text{O}}}$ for the Occurrence categorization, $\theta_{\text{C}} := \theta_{\text{C}} - \eta \frac{\partial J_{\text{C}}}{\partial \theta_{\text{C}}}$ for the Class categorization, and $\theta_{\text{I}} := \theta_{\text{I}} - \eta \frac{\partial J_{\text{I}}}{\partial \theta_{\text{I}}}$ for the Identity detection. The optimizer minimizes the overall loss $J = J_{\text{Tox}} + J_{\text{O}} + J_{\text{C}} + J_{\text{I}}$.

4 Experimental Setting

In the following, we explain the data that we use (§4.1) and the experimental setting (§4.2).[4]

4.1 Data

We use the CivilComments dataset (Borkan et al., 2019), the largest corpus in English annotated for both toxicity and identity terms with approximately 450,000 instances. We infer the hierarchical annotations from the 24 identity labels (see Table 1).

In these data, instances consist of individual posts as short text messages (the average instance length in the development data is 78 tokens) with all annotations on instance level. We transform the fractions of annotators that agree on a label to binary values by majority vote (following Xiang et al., 2021; Faal et al., 2021; Baldini et al., 2022; Lobo et al., 2022). From the 24 fine-grained annotated classes (I), we infer five more coarse-grained categories (C):

1. **Gender:** male, female, transgender, other gender
2. **Sexual orientation:** heterosexual, homosexual gay or lesbian, bisexual, other sexual orientation
3. **Religion:** christian, jewish, muslim, hindu, buddhist, atheist, other religion
4. **Race or ethnicity:** black, white, asian, latino, other race or ethnicity
5. **Disability:** physical disability, intellectual or learning disability, psychiatric or mental illness, other disability

This leads to a hierarchical multi-label annotation for identities. Our goal is to mitigate the bias towards frequently mentioned identity terms during training in order to improve generalization for other cases: namely, to correctly detect toxicity in cases where no explicit target identity is mentioned (e.g., as in Examples #8 and #9), and to not detect toxicity based solely on the presence of specific target mentions (e.g., as in Examples #10 and #11). In this dataset, toxic instances contain identity terms in 61% of cases, but only 40% of non-toxic instances do (see Appendix B for more details).

For the *Jigsaw Unintended Bias in Toxicity Classification challenge* on Kaggle[5] this dataset was split into a development set with 405,130 instances and two test sets with a total of 42,870 instances. For our experiments we randomly split this development set into training (100k instances), validation

[4] Our code to replicate the experiments can be accessed via https://www.uni-bamberg.de/en/nlproc/resources/hierarchical-detox/

[5] https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

for early stopping during training (50k) and hyperparameter optimization (≈255k). Further details are given in Appendix A. For evaluation, we use the official test sets combined (≈43k instances).

4.2 Model Configurations

We train different model configurations in order to answer our research questions (cf. Section 1). The main goal of these experiments is to assess whether the performance of a toxicity classifier decreases if the underlying encoder is optimized to not being able to represent specific target groups or identities while maintaining target occurrence features. Therefore, we configure a baseline model to compare its toxicity detection performance with debiased models. Additionally, we evaluate the performance of the different models in recognizing identity terms. This serves to identify whether, and to what extent, the different toxicity detection models pay attention to target identities. We explore different combinations of joint multi-task learning with target occurrence and adversarial correction of specific target identities to determine which effect the different levels of the identity term hierarchy $x \in (1, 2, 3)$ have on the toxicity classifier.

Baseline. We train a model purely for toxicity detection. In this TOX setup, the Occurrence/Class/Identity classifiers are also trained, but the encoder is not optimized via backpropagation with this information ($\lambda_1 = \lambda_2 = \lambda_3 = 0$). Here the encoder is only trained by the toxicity detector. This setup serves the purpose of investigating whether the uncontrolled and unguided toxicity classifier relies on features which contain information regarding identity term mentions.

Debiased Baseline. In order to compare our correction method to an established bias mitigation method, we adopt the debiasing approach by Ramponi and Tonelli (2022). We refer to this model as RT (2022). It cannot rely on features of identity terms as it automatically masks the tokens most strongly associated with each identity term label. Following Ramponi and Tonelli (2022) we use normalized PMI scores to automatically extract such spurious artifacts. While Ramponi and Tonelli (2022) manually annotate the top 200 entries, we automate this process by filtering all tokens with normalized PMI values $> .6$. This cutoff value was chosen based on the identity term Muslim where we find the tokens *muslim*, *muslims*, *islam* and *islamic* with values $> .80$ but also *mosque* (.65), *quran* (.63) and *mosques* (.62) amongst other tokens which are not as obviously connected: *world* (.71), *religious* (.71) or *europe* (.70). This approach filters a total of 751 word types for all identities. Besides operating on partially masked text, this baseline follows the configuration of the TOX setup mentioned above.

MTL (multi-task learning). Our data analysis has shown (see Section 4.1) that there is a correlation between toxicity and identity terms. We now want to test whether this carries over to the model level (cf. RQ1 in Section 1). Thus, we use MTL to guide the encoder to explicitly learn features for both toxicity detection and target occurrence in a joint setup (Model Tox+O,C,I with $\lambda_1 = -1$, $\lambda_2 = \lambda_3 = 0$). To create an upper bound for the identity term detection performance on all three levels we train a model where all classifiers are combined jointly (Model Tox+O+C+I, $\lambda_x = -1$).

Adversarial. In order to assess the importance of target occurrence features for the detection of toxicity, we train a model for comparison in which we instruct the encoder to unlearn precisely these features. In this model identity occurrence is used as an adversary (Model Tox−O, $\lambda_1 \in \{0.10, 0.25, 0.50, 1.00\}$). Additionally, as a starting point to debias the model for identities, we train a model where we use an adversary on the lowest level of the identity hierarchy (Model Tox−I, $\lambda_3 \in \{0.10, 0.25, 0.50, 1.00\}$).

MTL&Adversarial. Based on the intuition that we want to guide the toxicity detector with features from mentioned targets while debiasing for identities, we combine parameterizations for multiple levels of the hierarchy. In addition to the joint toxicity and target occurrence classifier ($\lambda_1 = -1$), we now debias the model for specific identity terms to understand whether this has a negative effect on the performance (cf. RQ2 in Section 1). We include an adversary via a gradient reversal layer on the lowest level of the identity term detection ($\lambda_3 \in \{0.10, 0.25, 0.50, 1.00\}$) and, thus, deprive the model of the ability to distinguish between different identity terms (e.g., which specific religion is mentioned). This serves to unlearn identity term features in the encoder and to determine whether this increases the generalization ability of the model. In order to evaluate the role of the intermediate level, we include the classifier for the identity class jointly ($\lambda_1 = -1$, resulting

Model	λ_1	λ_2	λ_3	$F1^{(1)}_{Tox}$	$F1^{(1)}_{O}$	$F1^{(5)}_{C}$	$F1^{(24)}_{I}$
TOX (baseline)	0	0	0	.64	.59	.25	.07
RT (2022)	0	0	0	.55	.45	.13	.03
Tox+O,C,I	−1	0	0	.63	.93	.34	.10
Tox−O	1.00	—	—	.63	.05		
Tox−I	—	—	0.10	.63	(.58)	(.20)	.05
Tox+O+C+I	−1	−1	−1	.64	.93	.86	.38
Tox+O+C−I	−1	−1	0.50	.63	.93	.86	.24
Tox+O−C−I	−1	0.25	0.25	.64	.93	.30	.08

Table 2: Performance of optimized models on the test dataset. We display F1 for the positive classes across all variables. The values in the superscript of the F1 scores specify the number of classes evaluated in each task – for multi-label tasks (Class and Identity) we display the macro-average F1 over all positive class label F1 scores. In the column "Model", "+" marks joint classification, "−" marks adversaries and classifiers appended with "," do not have an effect on the encoder. Tox refers to the toxicity classifier. (O)ccurrence, (C)lass and (I)dentity refer to the classifiers for the three levels of the identity term label hierarchy according to our model (see Figure 2). Values in parentheses are inferred from the prediction of more fine-grained labels.

in Model Tox+O+C−I) or as another adversary ($\lambda_1 \in \{0.10, 0.25, 0.50, 1.00\}$, resulting in Model Tox+O−C−I). We hypothesize that correcting for both class and identity might lead to a more comprehensive mitigation of the identity term bias than the experimental design with only one adversary.

5 Results

We will now discuss the results obtained with the setting described in the previous sections. Table 2 depicts the results for the best-performing models based on the parameter λ. Further results can be found in Appendix C. Table 2 shows F1 values for different combinations of toxicity detection and identity detection on the three levels of our hierarchy. On top, we see the baseline that only optimizes the encoder with the toxicity information followed by the debiased baseline RT (2022).

We observe that RT (2022) shows a lower performance at identity classification than the baseline TOX (e.g., $F1_C$ drops from .25 to .13). Therefore, the toxicity classifier in RT (2022) learns less identity-specific features, i.e., the model is successful in reducing bias. Conversely, this also means that the baseline TOX model automatically learns identity features without being guided to do so, i.e., it in fact contains a bias. However, the results also show that debiasing following RT (2022) does lead to a drop in toxicity detection performance ($F1_{Tox}$ drops from .64 to .55).

RQ1: Does jointly learning binary target occurrence detection with toxicity detection improve the latter? To measure if target mentions are important for toxicity detection, we now focus on specific models. We compare the performance of the baseline model (TOX) to the model which is also informed with the identity occurrence classifier (Model Tox+O,C,I) and to the model which uses a identity occurrence adversary (Model Tox−O). The results show (cf. Table 2) the $F1_{\text{Tox}}$ scores for all of these models on the same level. Therefore, while targets are a constituent variable of the concept of hate speech, we cannot infer from this evaluation that they are also an essential feature for toxicity detection. The toxicity classifier manages to maintain its performance level, even if we instruct the encoder to learn identity-occurrence features or, conversely, to unlearn exactly those features by adversarial correction. However, in further evaluations (cf. RQ3 below), we will see that unlearning identity occurrence features does not have a positive effect on the generalization ability of the model, which could be due to the fact that they are important for learning toxicity detection after all.

RQ2: Does the performance of a toxicity classifier decrease if the underlying encoder is optimized to not being able to represent specific target groups or identities while maintaining target occurrence features? We first evaluate overall toxicity detection performance and then address the details of identity detection performance to identify specific differences between models. We obtain the results by comparing the performance of the models corrected for identities (adversaries are marked

by −) to the baseline model (TOX). Overall the F1 score for toxicity detection (see column $F1_{Tox}$ in Table 2) is fairly constant in the range of .63 to .64. This shows that the toxicity detection does not suffer from the adversarial correction for identities. In contrast, the differently debiased model RT (2022) (which also has been debiased, however, by masking identity-specific tokens) shows a substantial performance drop ($F1_{Tox}$ .55). The test dataset used for this entire evaluation was sampled from the same source as the training dataset and is therefore also biased towards the same identities. Therefore, we presume that this evaluation is unable to demonstrate a positive effect of debiasing on toxicity detection performance. Further evaluations below under RQ3 and RQ4 show the performance gain for toxicity detection.

We now want to understand how the capability of the encoder to represent identity terms changes at different levels of the hierarchy. We see this from the performance scores in Table 2: columns $F1_O$, $F1_C$ and $F1_I$ (for Occurrence, Class, and Identity). As expected, the identity classifiers on each of the three levels in the MTL model (Model Tox+O+C+I) outperform the models where the particular level is used as an adversary. When we use an adversary for identity detection (Model Tox−I) the performance at identity detection drops (from .07 for Model TOX to .05), i.e., the model loses some of its ability to represent identities. In settings where we emphasize learning of identity occurrence features (models with +O), the encoder also represents more identity features overall, e.g. $F1_I$ rises from .07 for Model TOX to .10 for Model Tox+O,C,I. In a model that additionally learns identity occurrence jointly, we still see the effect of the adversary on $F1_I$. It drops from .10 for Model Tox+O,C,I to .08 for Model Tox+O−C−I. Analogously, this can also be observed for the models which additionally use the identity class classifier in a joint MTL setting ($F1_I$ drops from .38 for Model Tox+O+C+I to .24 for Model Tox+O+C−I). Thus, we conclude that adversarial correction has the desired effect of depriving the models of the ability to perform the task of identity term identification on the lowest level while maintaining target occurrence features. In addition, the procedure does not harm toxicity detection.

Finally, we investigate the role of the intermediate level in this setting. Comparing Model Tox+O+C−I to Model Tox+O−C−I shows that using the intermediate level as additional adversary also has an effect on the lowest level as $F1_I$ drops from .24 to .08. Thus, we conclude that this further reinforces unlearning features for the lowest level and leads to a more comprehensive correction.

Training data:	Full	NR		
Test data:	Full	Full	NR	R
Model	$F1^{(1)}_{Tox}$	$F1^{(1)}_{Tox}$	$F1^{(1)}_{Tox}$	$F1^{(1)}_{Tox}$
TOX (baseline)	.64	.63	.65	.57
RT (2022)	.55 Δ−.09	.56 Δ−.07	.57 Δ−.08	.53 Δ−.04
Tox−O	.63 Δ−.01	.58 Δ−.05	.58 Δ−.07	.57 Δ.00
Tox−I	.63 Δ−.01	.63 Δ.00	.64 Δ−.01	.59 Δ+.02
Tox+O+C−I	.63 Δ−.01	.61 Δ−.02	.62 Δ−.03	.58 Δ+.01
Tox+O−C−I	.64 Δ.00	.62 Δ−.01	.63 Δ−.02	.58 Δ+.01

Table 3: Performance on test data of best models trained on different training data fractions. NR = non-religion (filtered), R = only religion (filtered).

RQ3: Does adversarial correction of specific target identities lead to better generalization? We have now seen that the model debiased for identities on the lowest level of the hierarchy does perform as well at toxicity detection as the one that is not corrected. The performance scores for the identity term detection suggest that the encoder can no longer represent the identities to the same extent. This should enable an improved generalization across domains. We analyze this in two settings, firstly with an evaluation on target identity terms which have not been considered during training, and secondly with other datasets that have not been used during model development and training.

Regarding the first setup, we train the baseline and corrected models for the best configurations on data which has been filtered for all identities belonging to the religion class.[6] Table 3 shows the evaluation of these models for toxicity detection on different fractions of the test set. In the first column we repeat the results of the models from the first experiment, which were trained on the full dataset. The last three columns show the models which were trained on non-religion data (training data: NR). Here we see that all corrected models show a drop in performance on in-domain test data

[6]We chose the religion class since it comprises the largest number of identities (7 out of the 24) and accounts for a substantial number of instances (7,514) in the test data. For the model RT (2022) we repeat the process of identifying the tokens that are masked on the basis of the filtered dataset.

compared to the baseline (second to last column, test data: NR). However, our corrected models show an improved performance on out-of-domain test data (last column, test data: R) in comparison to the baseline. Only the model corrected for Occurrence (Tox−O) does not show an improvement. This confirms our intuition that the correction for general target terms is not the best choice since it also includes features which are beneficial for toxicity detection. We conclude that our correction does lead to a better toxicity prediction generalization.

The second generalization evaluation is performed with out-of-distribution performance evaluations. We show the datasets that we use and the detailed results for the out-of-distribution hate speech and toxicity detection performance in Appendix D. We observe that all models show on average a similar performance on out-of-distribution data (RT (2022) being an exception). Thus, we conclude that all corrected models show similar cross-corpus performance compared to the baseline models. We assume that the corpora used represent different domains and only share the targets of hate speech, our identity terms, to a limited extent.

RQ4: Does such correction lead to a more reasonable decision by the model? Do debiased models rely on concepts which are more meaningful for toxicity detection? To understand if the corrected model relies more on concepts that do not correspond to identities – potential targets of offensive language –, we analyze the change in toxicity detection performance for specific target terms. We subdivide the test dataset into subsets mentioning specific identities and evaluate the toxicity detection performance of the different models. The detailed evaluation is given in Appendix E. We observe that the performance of the corrected models for detecting toxicity mentioning the most frequent target terms is comparable to the baseline. However, for subsets with less frequent identities, adversarial correction improves toxicity detection in two-thirds of all cases. Thus, we conclude that the corrected models rely less on identities as features and learn other, more meaningful concepts.

We further visualize this effect on selected examples using LIME (Ribeiro et al., 2016) to calculate local explanations on the words of an instance that are most important. Figure 3 displays such explanations for five selected examples from the test dataset where the debiased model Tox+O−C−I corrects errors of the baseline model. Examples

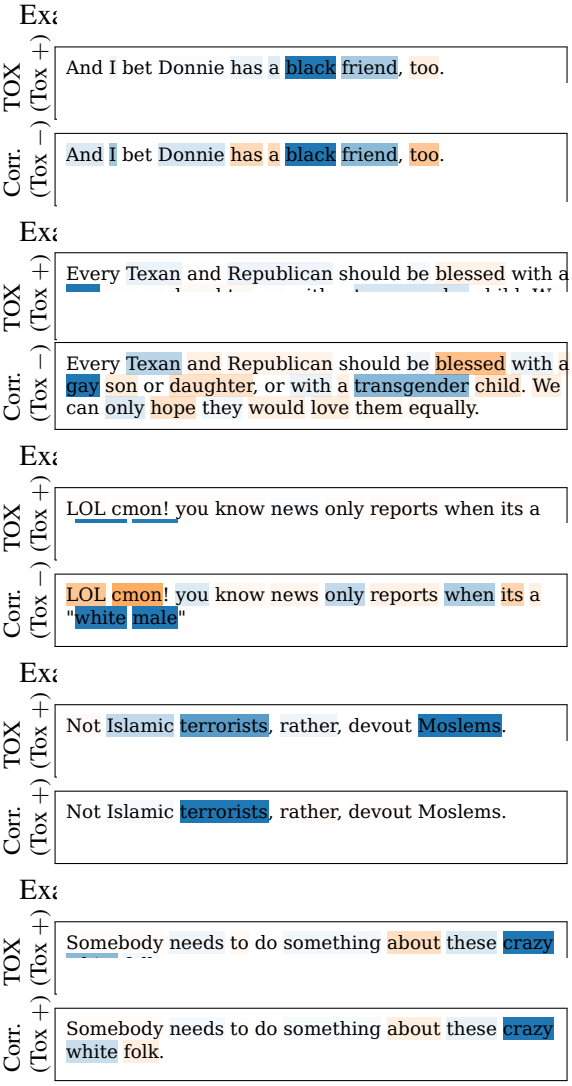

Figure 3: Explanations for the predictions of the baseline model TOX and our corrected model Tox+O−C−I (Corr.) according to LIME on instances from the Civil-Comments test dataset. The gold and predicted labels are shown in parentheses. Blue indicates word importance for Tox +, orange refers to Tox −. The intensity correlates to LIME's importance weights.

1 to 3 are all non-toxic instances which contain identity terms. The biased baseline TOX focuses only on these terms (e.g. "black", "gay", "transgender", "white" and "male") and incorrectly predicts toxicity. The debiased model corrects the error and correctly predicts the instances as non-toxic. It achieves this by also taking into consideration other tokens (most of them are marked with an orange background color) in which it finds no decisive features of toxicity. Examples 4 and 5 in Figure 3 show toxic instances with target mentions. Here both models manage to classify the instances as toxic. However, the biased baseline TOX bases its

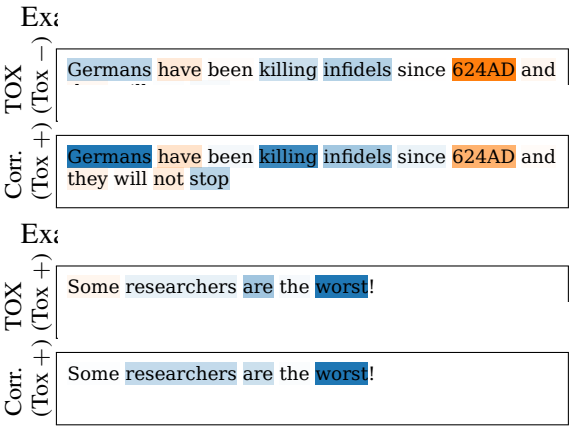

Figure 4: LIME explanations for the predictions on the CivilComments test data of the baseline TOX and the corrected model Tox+O−C−I (Corr.). In these examples, we manually manipulated the sentences by inserting new, originally non-existent targets.

decision mostly on the presence of identity terms (e.g. "Islamic", "Moslems" and "white"). In contrast, the debiased model does not rely as much on the mentioned identity terms which leads to a more reasonable decision with higher weights on words such as "terrorists" and "crazy".

We additionally investigate instances of toxicity with targets which are not included in the set annotated on the training data (such as "Germans" and "researchers"). We display LIME explanations in Figure 4 for predictions of examples from the test data which we modified to include such new target terms. These examples visualize cases of the improved generalization capability of the corrected model. Example 1 is incorrectly predicted as non-toxic by the biased baseline TOX model. The debiased model corrects the error as it is able to rely more on the new target. For Example 2, both models correctly predict the toxicity, however, the corrected model again assigns a higher weight on the new target. While this is a small-scale analysis based on a few examples, it suggests that there are cases where the corrected models use more meaningful features for toxicity detection.

6 Conclusion

We have shown that hierarchical adversarial correction for target identities leads to a toxicity classifier with an improved robustness. The corrected models show the same performance at toxicity detection as the biased baseline model. We presented a method to apply adversarial correction for the lowest level of hierarchical information regarding identity term mentions. Our results have demonstrated that it is possible to simultaneously maintain basic target occurrence features. However, target occurrence has not been shown to be as important for the detection of toxicity as the related concept of hate speech would suggest. This motivates future work to divise toxicity into more fine-grained concepts such as hate speech, offensive language and profanity, in the delineation of which target occurrence features presumably play a more decisive role.

Furthermore, when debiasing for individual identity terms, our experiments with the different hierarchical levels of specificity of the confounding variable have shown that it is more beneficial to additionally correct for classes of identities. It follows that a coarser grouping of identity categories must also be considered when defining the label set for annotation in order to achieve a more comprehensive correction during training.

Overall, our correction has shown to lead to a more reasonable decision by the model as it does not exclusively rely on identity features for toxicity detection and shows better generalization capabilities. This affects real-world applications of such models in that these models are required to be demonstrably debiased and treat individual identities fairly. Additionally, this motivates that a full evaluation of model performance must test the generalization ability of such models on further datasets where different identities are mentioned, as in-distribution biases do not show up in standard evaluations with a single test dataset.

Our research opens a set of important follow-up questions. In particular, whether further fine-tuning of the training process can lead to an improved overall toxicity detection with adversarial correction. This might be achieved, e.g., by testing different individual learning rates for optimizing the classifiers, the adversary and the encoder separately or by using multiple adversaries for latent variables as presented by Kumar et al. (2019). Also, since target detection might play a more significant role to distinguish hate speech from offensive language, an evaluation of our correction approach on such data would be an important next step to fight online toxicity.

Acknowledgements

This work has been partially supported by the CEAT project (KL 2869/1-2), funded by the German Research Foundation (DFG).

Limitations

We only ran all model configurations once due to limited time. The implementation contains randomized steps (initialization of weights, shuffling of training instances). Thus, the reported performance scores might not be entirely robust. However, our reported conclusions are based on substantial differences in performance of the different models.

Our implementation of the debiased baseline methods by Ramponi and Tonelli (2022) does only partially follow their suggested approach. While we do not consider manual annotation of top-n lists and use a fix threshold PMI value, choosing a top-n cut-off might be a more justified choice. Furthermore, Ramponi and Tonelli (2022) suggest multiple different approaches to deal with the identified spurious artifacts while we only use the removal method for comparison as a baseline.

In the experiment with filtering training data for specific identity classes, we focused on the evaluation of a setting where we filtered the religious identities. We chose the religion class since it comprises the largest number of identities (7 of the 24) and accounts for a substantial number of instances in the test data (7,514) which we can evaluate separately. For full expressiveness of the results, experiments where identities from other classes are filtered, should also be conducted. However, we presume that statistical evidence for the performance of less frequent classes (e.g. there are only 544 test instances for the disability class) might be limited.

Ethical Considerations

Potential risks. We mention examples of toxicity and hate speech which might offend readers of this paper. They are taken from empirically collected datasets and do not portray our own opinions. However, we believe that it is inevitable to investigate concrete instances when discussing detection approaches.

Reproducibility. We use datasets with annotations for toxicity and hate speech. All of these datasets are freely available for research use. We use these data for their intended use, to develop detection systems. Since we research toxicity and mentions of identity terms, the datasets have not been filtered or anonymized for such attributes.

We publish our program code for maximum transparency. The described models and predictions of labels can be reproduced with this code. For training we randomly split the dataset into specific portions. As these are quite large, we believe that they are representative for the entire corpus and that the same experiments with different partitions lead to the same conclusions. Additionally, we provide a script to reproduce the random split used in our experiments to benefit future research.

We report relevant information for the used artifacts and refer to the original publications for further documentation. We describe the structure and size of the models we create. We believe that these descriptions make our approach reproducible.

References

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference*, WWW '19, page 49–59, New York, NY, USA. Association for Computing Machinery.

Ioana Baldini, Dennis Wei, Karthikeyan Natesan Ramamurthy, Moninder Singh, and Mikhail Yurochkin. 2022. Your fairness may vary: Pretrained language model fairness in toxic text classification. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2245–2262, Dublin, Ireland. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *13th International Workshop on Semantic Evaluation*, pages 54–63. Association for Computational Linguistics.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500.

Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the EVALITA 2018 Hate Speech Detection Task. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263, pages 1–9. CEUR.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

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.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

Erenay Dayanik and Sebastian Padó. 2020. Masking actor information leads to fairer political claims detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4385–4391, Online. Association for Computational Linguistics.

Erenay Dayanik and Sebastian Padó. 2021. Disentangling document topic and author gender in multiple languages: Lessons for adversarial debiasing. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–61, Online. Association for Computational Linguistics.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

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

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Farshid Faal, Jia Yuan Yu, and Ketra A Schmitt. 2021. Domain adaptation multi-task deep neural network for mitigating unintended bias in toxic language detection. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence (ICAART 2021)*, volume 2, pages 932–940.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Jennifer Golbeck, Zahra Ashktorab, Rashad O. Banjo, Alexandra Berlinger, Siddharth Bhagwan, Cody Buntain, Paul Cheakalos, Alicia A. Geller, Quint Gergory, Rajesh Kumar Gnanasekaran, Raja Rajan Gunasekaran, Kelly M. Hoffman, Jenny Hottle, Vichita Jienjitlert, Shivika Khare, Ryan Lau, Marianna J. Martindale, Shalmali Naik, Heather L. Nixon, Piyush Ramachandran, Kristine M. Rogers, Lisa Rogers, Meghna Sardana Sarin, Gaurav Shahane, Jayanee Thanki, Priyanka Vengataraman, Zijian Wan, and Derek Michael Wu. 2017. A large labeled corpus for online harassment research. In *Proceedings of the 2017 ACM on Web Science Conference*, WebSci '17, page 229–233, New York, NY, USA. Association for Computing Machinery.

Lara Grimminger and Roman Klinger. 2021. Hate towards the political opponent: A Twitter corpus study of the 2020 US elections on the basis of offensive speech and stance detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 171–180, Online. Association for Computational Linguistics.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Alyzeh Hussain, Austin Lara, Gabriel Cardenas, Adam Omary, Christina Park, Xin Wang, Clarisa Wijaya, Yong Zhang, Beth Meyerowitz, and Morteza Dehghani. 2022. Introducing the gab hate corpus: defining and applying hate-based rhetoric to social media posts at scale. *Language Resources and Evaluation*, 56(1):79–108.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pages 5637–5664. PMLR.

Sachin Kumar, Shuly Wintner, Noah A. Smith, and Yulia Tsvetkov. 2019. Topics to avoid: Demoting latent confounds in text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4153–4163, Hong Kong, China. Association for Computational Linguistics.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 914–924, Suzhou, China. Association for Computational Linguistics.

Jens Lemmens, Ilia Markov, and Walter Daelemans. 2021. Improving Hate Speech Type and Target Detection with Hateful Metaphor Features. In *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 7–16, Online. Association for Computational Linguistics.

Paula Reyero Lobo, Enrico Daga, and Harith Alani. 2022. Supporting online toxicity detection with knowledge graphs. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1414–1418.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Amit Kumar Jaiswal, Durgesh Nandini, Daksh Patel, Prasenjit Majumder, and Johannes Schäfer. 2020. Overview of the HASOC track at FIRE 2020: Hate Speech and Offensive Content Identification in Indo-European Languages. In *Working Notes of FIRE 2020 - Forum for Information Retrieval Evaluation*, pages 87–111, Hyderabad, India. CEUR Workshop Proceedings.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schäfer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, and Amit Kumar Jaiswal. 2021. Overview of the HASOC Subtrack at FIRE 2021: HateSpeech and Offensive Content Identification in English and Indo-Aryan Languages. In *Working Notes of FIRE 2021 - Forum for Information Retrieval Evaluation*, pages 1–19, India. CEUR Workshop Proceedings.

Robert McHardy, Heike Adel, and Roman Klinger. 2019. Adversarial training for satire detection: Controlling for confounding variables. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 660–665, Minneapolis, Minnesota. Association for Computational Linguistics.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8:4663–4678.

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

Flor Miriam Plaza-del-Arco, Sercan Halat, Sebastian Padó, and Roman Klinger. 2021. Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language. In *FIRE 2021 Working Notes*, pages 297–318.

Alan Ramponi and Sara Tonelli. 2022. Features or spurious artifacts? data-centric baselines for fair and robust hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3027–3040, Seattle, United States. Association for Computational Linguistics.

Marco Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California. Association for Computational Linguistics.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. 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 41–58, Online. Association for Computational Linguistics.

Pratik Sachdeva, Renata Barreto, Claudia Von Vacano, and Chris Kennedy. 2022. Targeted identity group prediction in hate speech corpora. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 231–244, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. "Call me sexist, but...": Revisiting sexism detection using psychological scales and adversarial samples. In *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media, ICWSM 2021, held virtually, June 7-10, 2021*, pages 573–584. AAAI Press.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

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 *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

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

Leandro Silva, Mainack Mondal, Denzil Correa, Fabrício Benevenuto, and Ingmar Weber. 2016. Analyzing the Targets of Hate in Online Social Media. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 687–690.

Julia Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of GermEval Task 2, 2019 Shared Task on the Identification of Offensive Language. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 354–365, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Zeerak Talat, James Thorne, and Joachim Bingel. 2018. Bridging the Gaps: Multi Task Learning for Domain Transfer of Hate Speech Detection. In Jennifer Golbeck, editor, *Online Harassment*, pages 29–55. Springer International Publishing, Cham.

Nanna Thylstrup and Zeerak Talat. 2020. Detecting 'dirt'and 'toxicity': Rethinking content moderation as pollution behaviour. *Available at SSRN 3709719*.

Ameya Vaidya, Feng Mai, and Yue Ning. 2020. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 683–693.

Bertie Vidgen, Scott Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020. Detecting East Asian prejudice on social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172, Online. Association for Computational Linguistics.

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

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018. Overview of the GermEval 2018 Shared Task on the Identification of Offensive Language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*, pages 1–10, Vienna, Austria. Österreichische Akademie der Wissenschaften.

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.

Tong Xiang, Sean MacAvaney, Eugene Yang, and Nazli Goharian. 2021. ToxCCIn: Toxic content classification with interpretability. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 1–12, Online. Association for Computational Linguistics.

Yihao Xue, Ali Payani, Yu Yang, and Baharan Mirzasoleiman. 2023. Eliminating spurious correlations from pre-trained models via data mixing. *Preprint*, arXiv:2305.14521.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval@NAACL-HLT 2019)*, pages 75–86.

A Hyperparameter Optimization

Datasets. To limit training time, we select 100k instances as training data and an additional 50k instances as validation data to determine a suitable point for early stopping. The remaining approximately 255k instances are used for hyperparameter optimization. Optimizing all models with the described setup takes about a month on a Nvidia Quadro RTX 8000 GPU. We do not expect using more than 100k training instances to change the results of our experiments regarding the comparison of the debiasing methods. To ensure that

this is indeed the case, we trained another TOX baseline model with a larger subset for training by splitting the dev set as follows: 80% training, 10% validation-1, 10% validation-2. The resulting model (being trained on more than three times as many instances) achieves a slightly improved performance by 1.5 percentage points ($F1_{Tox}$ = .66) on the same test data. Hence, the chosen split does not have an impact on our conclusions.

For evaluation, we use the combined public and private test datasets from the Jigsaw Unintended Bias in Toxicity Classification challenge which does allow a straightforward comparison with past and future work.

We constrain input text instances to a maximum length of 236 tokens. This value corresponds to the 99th percentile of instance lengths in the development set. Thus, only 1% of the instances are truncated.

To deal with the skewed class distribution, we use class weights based on the inverse class frequency in the training data for all attributes in each loss.

Early stopping configuration. While all our setups operate with the same model, we monitor only relevant performance measures for each setup. Early stopping for the TOX setup monitors only the performance of the Tox classifier. In the joint setup, early stopping is based on toxicity and any active identity term classifiers in combination (all classifiers weighted equally). In the adversarial setups, early stopping is determined by monitoring the sum of the Tox classifier performance (or all MTL classifiers) and the negated adversary's performance (weighted by 0.1).

We use early stopping with $\texttt{patience} = 3$ and reload the best model if the maximum of 10 epochs is reached.

Training process metadata. On the mentioned data, our model trains for approximately 32 minutes per epoch on a single GPU (Nvidia Quadro RTX 8000). Each model has approximately 109 million trainable parameters.

Learning rate optimization. We run each experiment with different learning rates $\mathrm{lr} \in \{5 \cdot 10^{-6}, 7.5 \cdot 10^{-6}, 1 \cdot 10^{-5}, 2.5 \cdot 10^{-5}, 5 \cdot 10^{-5}\}$. For optimization we select the best lr value for each model according to its performance on the portion of the dataset which has not been used during training (255k instances). As performance measure we calculate the toxicity F1 score, possibly (if the model uses joint MTL) add F1 scores for joint MTL identity term classifiers and possibly (if the model uses an adversary) subtract the F1 score of the adversarial task. Since our main goal is to optimize the toxicity detection performance, we multiply the F1 scores of the identity classifiers by a reduced weight of 0.1 in this measure.

	O +	O −	Total
Tox +	27,963 (61%)	18,072	46,035
Tox −	142,341 (40%)	216,754	359,095
Total	170,304 (42%)	234,826	405,130

Table 4: Distribution of binary toxicity and identity term annotations in the development set from the CivilComments dataset (Borkan et al., 2019). The percentages are respectively the proportion of instances with identities to the total instances for each row.

B CivilComments Data

Table 4 shows the distribution of binary toxicity and identity term annotations in the development set from the CivilComments dataset (Borkan et al., 2019). This suggests a correlation between toxicity and mentions of identity terms, as toxic instances contain identity terms in 61% of instances, but only 40% of non-toxic instances.

C Full Results

The performance of the models with optimized learning rate (cf. Appendix A) on the test dataset is displayed in Table 5. In addition to the performance scores shown in Table 2, this table provides the results for several models for each setup with different λ_x values as well as some further setups with other combinations of the hierarchical identity classifiers. Table 2 only shows the best-performing corrected models based on the in-distribution test set performance with high $F1_{Tox}$ and low respective identity detection F1 for each setting (see underlined values in Table 5). The additional setups included in Table 5 which are not directly related to our research questions are briefly motivated in the following.

We additionally test a setting where we are correcting for both Identity and Class ($\lambda_2, \lambda_3 \in \{0.10, 0.25, 0.50, 1.00\}$, Model Tox−C−I). This setting is based on the assumption that a combined correction for both C and I could capture the more

	Model	λ_1	λ_2	λ_3	$F1^{(1)}_{Tox}$	$F1^{(1)}_{O}$	$F1^{(5)}_{C}$	$F1^{(24)}_{I}$
	TOX (baseline)	0	0	0	.64	.59	.25	.07
	RT (2022)	0	0	0	.55	.45	.13	.03
Identity Occurrence	Tox+O	−1	—	—	.63	.93		
	Tox−O	0.10	—	—	.64	.34		
	Tox−O	0.25	—	—	.64	.33		
	Tox−O	0.50	—	—	.64	.16		
	Tox−O	1.00	—	—	.63	.05		
Identity Class	Tox+C	—	−1	—	.63	(.92)	.87	
	Tox−C	—	0.10	—	.64	(.55)	.09	
	Tox−C	—	0.25	—	.63	(.52)	.06	
	Tox−C	—	0.50	—	.64	(.58)	.11	
	Tox−C	—	1.00	—	.64	(.51)	.09	
Identity	Tox+I	—	—	−1	.64	(.90)	(.77)	.39
	Tox−I	—	—	0.10	.63	(.58)	(.20)	.05
	Tox−I	—	—	0.25	.64	(.58)	(.15)	.03
	Tox−I	—	—	0.50	.63	(.58)	(.17)	.01
	Tox−I	—	—	1.00	.63	(.58)	(.21)	.02
Class and Identity	Tox−C−I	—	0.10	0.10	.64	(.48)	.08	.03
	Tox−C−I	—	0.25	0.25	.64	(.53)	.05	.02
	Tox−C−I	—	0.50	0.50	.63	(.57)	.14	.02
	Tox−C−I	—	1.00	1.00	.62	(.52)	.10	.01
all levels	Tox+O+C+I	−1	−1	−1	.64	.93	.86	.38
	Tox−O−C−I	0.10	0.10	0.10	.64	.19	.13	.04
	Tox−O−C−I	0.25	0.25	0.25	.63	.21	.13	.03
	Tox−O−C−I	0.50	0.50	0.50	.64	.22	.06	.02
	Tox−O−C−I	1.00	1.00	1.00	.63	.05	.10	.01
	Tox+O, C, I	−1	0	0	.63	.93	.34	.10
	Tox+O−C−I	−1	0.10	0.10	.64	.93	.32	.08
	Tox+O−C−I	−1	0.25	0.25	.64	.93	.30	.08
	Tox+O−C−I	−1	0.50	0.50	.63	.93	.27	.07
	Tox+O−C−I	−1	1.00	1.00	.63	.93	.26	.06
	Tox+O+C, I	−1	−1	0	.64	.93	.88	.27
	Tox+O+C−I	−1	−1	0.10	.63	.93	.87	.25
	Tox+O+C−I	−1	−1	0.25	.63	.93	.88	.25
	Tox+O+C−I	−1	−1	0.50	.63	.93	.86	.24
	Tox+O+C−I	−1	−1	1.00	.63	.93	.85	.22
	Tox+O+C−I	−1	−1	2.00	.61	.93	.81	.20
	Tox+O+C−I	−1	−1	3.00	.61	.92	.67	.10

Table 5: Performance of optimized models on the test dataset. We display F1 for the positive classes across all variables. The values in the superscript of the F1 scores specify the number of classes evaluated in each task – for multi-label tasks (Class and Identity) we display the macro-average F1 over all positive class label F1 scores. In the column "Model", "+" marks joint classification, "−" marks adversaries and classifiers appended with "," do not have an effect on the encoder. Tox refers to the toxicity classifier. (O)ccurrence, (C)lass and (I)dentity refer to the classifiers for the three levels of the identity term label hierarchy according to our model (see Figure 2). Values in parentheses are inferred from the prediction of more fine-grained labels. Underlined λ values mark the best debiased model for each setting.

Id	Reference	Description	Size
da	Davidson et al. (2017)	Tweets annotated for hate speech and offensive language	24,783
ol	Zampieri et al. (2019a)	Tweets annotated for offensive content (OLID)	860
ha	Mandl et al. (2021)	Tweets annotated for hate speech and other offensive and objectionable content (HASOC 2021)	1,281
se	Samory et al. (2021)	Tweets annotated for sexism with predicted toxicity scores (CMSB)	13,631
sf	de Gibert et al. (2018)	Texts extracted from a white supremacy forum (Stormfront)	478
gk	Grimminger and Klinger (2021)	Political Twitter data annotated for hateful/offensive speech	600
as	Vidgen et al. (2020)	Tweets annotated for hostility directed against Asian people	40,000
et	Mollas et al. (2022)	YouTube and Reddit comments annotated for hate speech (ETHOS)	998
hc	Röttger et al. (2021)	Crafted test cases for hate speech detection (HateCheck)	3,728

Table 6: Hate speech datasets used as additional test data to evaluate out-of-distribution performance. We show the number of instances we use for evaluation in the last column.

general set of features on the one hand, which are also sufficiently specific properties of identities on the other. We hypothesize that this setting will lead to a more comprehensive mitigation of the identity term bias than the experimental design with single adversaries.

Further, we test additional setups with all three levels, where we explore combinations incorporating the Occurrence classifier as adversary ($\lambda_1 \in \{0.10, 0.25, 0.50, 1.00\}$, Model Tox−O−C−I). Here we test whether the additional correction for O does not contribute to a broader mitigation of the target identity term bias that we are aiming for and might harm the overall toxicity detection performance.

Additionally, we explore the configuration where we only correct for identity features while jointly promoting Occurrence and Class information (Model Tox+O+C−I). The idea behind this is that we may want to let the model represent features of the target occurrence as well as some distinguishing features of identity classes. There could be substantial differences in the type of toxicity that targets certain groups compared to other types of toxicity that target other groups. Therefore, we want to enable co-learning of such properties of identities in this setup. Here, we also test higher weights ($\lambda_3 \in \{0.10, 0.25, 0.50, 1.00, 2.00, 3.00\}$) to empower the Identity adversary to possibly outweigh the three joint classifiers which presumably induce identity bias. A more powerful adversary on the lowest level might be sucessful at unlearning specific features of identities which constitute the target bias and result in an improved generalization ability of the trained model. However, by including the classifiers O and C jointly with the toxicity classifier, this model could still retain the ability to learn more general categories of targets of toxic statements.

D Cross-Corpus Evaluation

We evaluate the performance of different models trained on the CivilComments dataset to predict hate speech on other datasets. We show the datasets that we use in Table 6. We selected publicly available datasets covering general types of hate speech and toxicity as well as datasets with a focus on specific subtypes, such as hate speech directed towards specific targets. In cases where the original authors declare a specific portion of the data as a test subset, we only use this portion in our evaluation. Otherwise we evaluate on the entire dataset.

The results for the out-of-distribution hate speech and toxicity detection performance are displayed in Table 7. The performance of all models on the 'gk' and 'as' datasets is rather low (F1 $\leq$ .30), presumably because these corpora are focused

	Performance on test data										
		Out-of-domain									
Model	IN	da	ol	ha	se	sf	gk	as	et	hc	avg.
TOX (baseline)	.64	.88	.64	.70	.69	.63	.30	.22	.70	.76	.61
RT (2022)	.55 Δ−.09	.88 Δ.00	.61 Δ−.03	.71 Δ+.01	.65 Δ−.04	.51 Δ−.12	.26 Δ−.04	.16 Δ−.06	.62 Δ−.08	.60 Δ−.16	.56 Δ−.05
Tox−O	.63 Δ−.01	.87 Δ−.01	.62 Δ−.02	.68 Δ−.02	.69 Δ.00	.66 Δ+.03	.22 Δ−.08	.17 Δ−.05	.71 Δ+.01	.74 Δ−.02	.60 Δ−.01
Tox−I	.63 Δ−.01	.88 Δ.00	.64 Δ.00	.70 Δ.00	.70 Δ+.01	.67 Δ+.04	.24 Δ−.06	.19 Δ−.03	.72 Δ+.02	.76 Δ.00	.61 Δ.00
Tox+O+C−I	.63 Δ−.01	.88 Δ.00	.66 Δ+.02	.72 Δ+.02	.69 Δ.00	.65 Δ+.02	.21 Δ−.09	.21 Δ−.01	.71 Δ+.01	.75 Δ−.01	.61 Δ.00
Tox+O−C−I	.64 Δ.00	.88 Δ.00	.66 Δ+.02	.71 Δ+.01	.70 Δ+.01	.62 Δ−.01	.21 Δ−.09	.17 Δ−.05	.71 Δ+.01	.76 Δ.00	.60 Δ−.01

Table 7: Hate speech/toxicity detection performance (F1 for the positive class) of our best corrected models in comparison to the baseline TOX model on different datasets. All models have been trained on the same data. Δ-values show the difference to the F1 score of the baseline model TOX. IN refers to the in-distribution test dataset performance of the CivilComments corpus and avg. refers to the macro-average of all out-of-distribution performances.

Test dataset:	Identity-specific subsets														
# test instances:	Full test 42870	fema 5155	male 4386	chri 4226	whit 2452	musl 2040	blac 1519	homo 1065	jewi 835	psyc 511	asia 454	athe 280	tran 260	lati 225	hete 141
Model	$F1^{(1)}_{Tox}$	$F1^{(1)}_{Tox}$	$F1^{(1)}_{Tox}$	$F1^{(1)}_{Tox}$	$F1^{(1)}_{Tox}$	$F1^{(1)}_{Tox}$	$F1^{(1)}_{Tox}$	$F1^{(1)}_{Tox}$	$F1^{(1)}_{Tox}$	$F1^{(1)}_{Tox}$	$F1^{(1)}_{Tox}$	$F1^{(1)}_{Tox}$	$F1^{(1)}_{Tox}$	$F1^{(1)}_{Tox}$	$F1^{(1)}_{Tox}$
TOX (baseline)	.64	.62	.61	.60	.63	.59	.65	.62	.62	.65	.53	.52	.63	.62	.52
RT (2022)	.55 Δ−.09	.56 Δ−.06	.54 Δ−.07	.45 Δ−.15	.53 Δ−.10	.39 Δ−.20	.52 Δ−.13	.45 Δ−.17	.51 Δ−.11	.58 Δ−.07	.30 Δ−.23	.47 Δ−.05	.62 Δ−.01	.49 Δ−.13	.45 Δ−.07
Tox−O	.63 Δ−.01	.62 Δ.00	.61 Δ.00	.57 Δ−.03	.62 Δ−.01	.58 Δ−.01	.64 Δ−.01	.61 Δ−.01	.60 Δ−.02	.69 Δ+.04	.53 Δ.00	.59 Δ+.07	.63 Δ+.01	.63 Δ+.01	.52 Δ.00
Tox−I	.63 Δ−.01	.62 Δ.00	.62 Δ+.01	.59 Δ−.01	.65 Δ+.02	.59 Δ.00	.66 Δ+.01	.61 Δ−.01	.64 Δ+.02	.67 Δ+.02	.50 Δ−.03	.55 Δ+.03	.56 Δ−.06	.69 Δ+.07	.51 Δ−.01
Tox+O+C−I	.63 Δ−.01	.62 Δ.00	.61 Δ.00	.59 Δ−.01	.65 Δ+.02	.59 Δ.00	.67 Δ+.02	.59 Δ−.03	.63 Δ+.01	.66 Δ+.01	.54 Δ+.01	.62 Δ+.10	.63 Δ+.01	.68 Δ+.06	.46 Δ−.06
Tox+O−C−I	.64 Δ.00	.63 Δ+.01	.62 Δ+.01	.59 Δ−.01	.64 Δ+.01	.60 Δ+.01	.66 Δ+.01	.61 Δ−.01	.57 Δ−.05	.70 Δ+.05	.58 Δ+.05	.48 Δ−.04	.67 Δ+.04	.65 Δ+.03	.51 Δ−.01

Table 8: Performance of best models on different portions of the test dataset. Δ-values show the difference to the F1 score of the baseline model TOX. Fema: female, chri: christian, whit: white, musl: muslim, blac: black, homo: homosexual gay or lesbian, jewi: jewish, psyc: psychiatric or mental illness, asia: asian, athe: atheist, tran: transgender, lati: latino, hete: heterosexual.

on special cases of hate speech (towards specific individuals or particular ethnicities).

E Evaluation of Identity-Specific Subsets

Table 8 shows the F1 scores of the baseline TOX model in comparison to the debiased baseline RT (2022) and our corrected models for these different portions of the test dataset. We additionally provide the number of instances which are considered in each subset. We discard all identity labels with less than 100 instances in the test dataset in this evaluation as there is presumably not enough statistical evidence for such categories.