

ID-XCB: Data-independent Debiasing for Fair and Accurate Transformer-based Cyberbullying Detection

Abstract

The use of swear words is a common proxy to collect datasets with cyberbullying incidents, which increases the chances of collecting such events that are otherwise hard to find. However, datasets collected through this means also have a risk of introducing biases in cyberbullying detection models which can learn spurious associations between swear words and the presence of incidents. In this study, we undertake a pioneering study of measuring and mitigating swearing bias in cyberbullying detection tasks. Initially, we employ word-level bias measures to demonstrate the distinctive features related to swearing biases in transformer-based cyberbullying detection models. Subsequently, we introduce ID-XCB, the first data-independent debiasing technique that combines adversarial training, bias constraints and a debias fine-tuning approach aimed at alleviating model attention to bias-inducing words without impacting overall model performance. Lastly, we explore ID-XCB on two popular session-based cyberbullying detection datasets along with a comprehensive set of ablation studies and model generalisation studies. Our findings show that ID-XCB learns robust cyberbullying detection capabilities while mitigating biases tied to swear word usage. It consistently outperforms state-of-the-art debiasing methods in terms of both performance improvement and bias mitigation. In addition, by combining quantitative and qualitative analyses, we demonstrate the potential for generalisability of our approach when tackling unseen data.

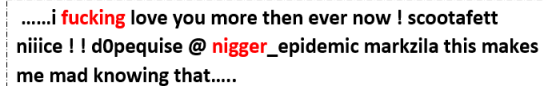
Warning: This paper contains swear words, which do not reflect the views of the authors.

Introduction

Cyberbullying is a form of bullying that takes place online (Smith et al. 2010) and is defined as the repeated, deliberate aggressive behaviour by a group or individual towards a more vulnerable individual (Olweus 2001). Where cyberbullying is characterised by repeated aggression and power imbalance, research (Dadvar et al. 2012, 2013; Menin et al. 2021; Yi and Zubiaga 2022) suggests capturing these characteristics by modelling social media sessions (Yi and Zubiaga 2023a), i.e. a series of conversational exchanges (Cheng et al. 2020), rather than from individual posts.

Cyberbullying detection models often suffer from biases leading to false positive predictions when a swear word is

present (Agrawal and Awekar 2018; Perera and Fernando 2021; Pamungkas, Basile, and Patti 2023), not least because swear words are often used as a proxy for data collection (Van Bruwaene, Huang, and Inkpen 2020). Swear words are however often used in other non-abusive contexts, not necessarily indicating cyberbullying incidents (Stephens and Robertson 2020; Lafreniere, Moore, and Fisher 2022). A disproportionate presence of swear words across both positive and negative samples then leads to model overfitting, exhibiting a ‘swear word bias’ (Hovy and Prabhumoye 2021). Figure 1 shows an example where highly biased profanity is present in a snippet of a full session. A false positive prediction for a model reliant on the presence of swear words.



```
.....i fucking love you more then ever now ! scootafett
niice ! ! d0pequise @ nigger_epidemic markzila this makes
me mad knowing that.....
```

Figure 1: A snippet of false positive sample in Instagram

In the existing body of research debiasing text classification models, and particularly cyberbullying detection models, work has been limited to data-dependent constraints (Gencoglu 2020; Cheng et al. 2021). These methods have proven satisfactory in mitigating biases for models tested on data from the same dataset or with similar characteristics, but risk overfitting on the seen data, limiting their generalisability and preserving both fairness and accurate performance.

To advance research in generalisable debiasing, we propose ID-XCB (Independent Debiasing Cyberbullying Detection), the first data-independent debiasing method that avoids the need to see target data, in turn detaching the link between swear words and cyberbullying incidents (Figure 2 illustrates ID-XCB vs data-dependent strategies). To achieve this, we integrate three strategies: (1) **adversarial training**, using adversarial examples and objective cost function to shift the model’s focus away from profanity; (2) **two-player constraint optimisation** to work in the non-convex setting, as the downstream task is trained on training datasets but fairness constraints are derived from independent validation datasets; and (3) **contextualised embeddings** from transformer models, which support generalisability by transfer-

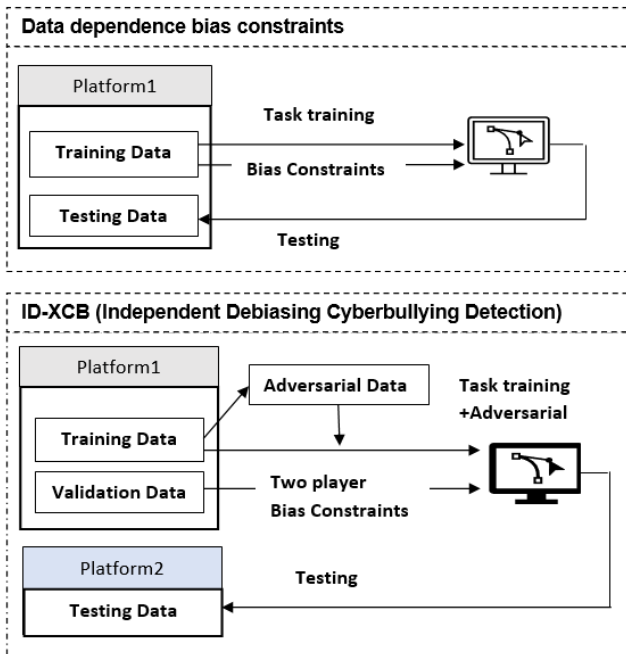


Figure 2: Data-dependent bias constraints vs. ID-XCB.

ring knowledge to the downstream task.¹

Our contributions include:

- We quantify the impact of swear word bias on five transformer models. We find that the same set of low-frequency swear words consistently lead to biases across models, which however vary across datasets.
- We introduce ID-XCB, the first data-independent debiasing method for cyberbullying detection, and experiment it on two datasets showing improved performance and bias reduction over state-of-the-art approaches. It further shows potential for bias mitigation and performance trade-offs in challenging scenarios dealing with unseen data from a different dataset.
- We perform in-depth analyses of specific swear words, model components, ablation experiments, and generalisation highlighting the importance of balancing between performance and fairness in successful debiasing.

Related Work

Cyberbullying detection is generally tackled as a binary text classification task determining if an instance constitutes a case of cyberbullying or not. As cyberbullying generally occurs as a series of social media interactions, a research direction of increasing popularity is to identify cyberbullying incidents observed through multiple user interactions (i.e. sessions). Transformer-based models such as BERT and RoBERTa enable a good understanding of context. Gururangan et al. (2020); Yi and Zubiaga (2023b) demonstrate that transformer-based models can be strong, and compet-

¹The source code is publicly available at https://github.com/Misinformation-emotion/Cyberbullying_debias

itive for session-based cyberbullying detection. However, researchers have shown that transformers trained on unfiltered corpora suffer from degenerated and biased behaviour (Schramowski et al. 2022).

Existing work on debiasing text classification models can be categorised in four main directions: (1) statistically balancing training data, such as data augmentation (Dixon et al. 2018), sample weighting (Zhang et al. 2020), identity term swapping (Badjatiya, Gupta, and Varma 2019) or injecting objective samples (Nozza, Volpetti, and Fersini 2019); (2) mitigating embedding bias, such as fine-tuning pre-trained contextualised word embeddings (Kaneko and Bollegala 2021) or using adversarial learning to reduce the bias (Sweeney and Najafian 2020); (3) proposing a multi-task learning model (Vaidya, Mai, and Ning 2019) with an attention layer that jointly learns to predict the toxicity of a comment as well as the identities present in the comments in order to mitigate bias; and (4) inferring fairness constraints by using Error Rate Equality Difference to restrict the discrimination behaviour of the model (Zafar et al. 2017).

Model debiasing is understudied in cyberbullying detection, where studies to date have focused on feeding fairness constraints. For example, Gencoglu (2020) did so by using sentence-DistilBERT as a base model, adding the Fairness metric as a cost function to constrain bias during training. Cheng et al. (2021) built on a reinforcement learning strategy relying on a pre-defined set of sensitive triggers to constrain a series of hierarchical attention networks. Current works focus on data-dependent fairness constraints, which satisfyingly reduce the false positive rate on the training data. However, if the terms of debiasing are strictly enforced on the training data, this may be beneficial to ensure fairness on similar data, but overfitting will also occur, thereby reducing the fairness of the model on unseen data (Hardt, Price, and Srebro 2016).

Datasets and Lexicon

Datasets We use the two existing and widely-studied session-based cyberbullying datasets from two different social media platforms: Instagram (Hosseinmardi et al. 2015) and Vine (Rafiq et al. 2015). To collect social media sessions likely to contain cyberbullying events, authors of these datasets looked at the presence of toxic words to maximise the chances of collecting positive samples, which were subsequently manually annotated. Table 1 shows statistics of both datasets.

	Instagram(IG)	Vine(VN)
Cyberbullying Ratio	0.29	0.30
# Sessions	2,218	970
# Comments	159,277	70,385
# Users	72,176	25,699
# Avg. length of session	900	698
# Unique Swear words	253	207

Table 1: Dataset statistics.

Lexicon To determine the presence of swear words when

measuring biases, we use the lexicon with 535 words provided in (Van Hee et al. 2018; Google 2010).

Swear Word Bias

We adopt the Oxford Dictionary’s definition of bias as the “inclination or prejudice for or against one person or group, especially in a way considered to be unfair”. By ‘swear word bias’, we refer to the impact of swear words during training on biasing model predictions, or model bias. A prominent swear word bias ultimately leads to the assignment of disproportionately high importance to the presence of swear words in model predictions.

Distribution of swear words

A first look at the distribution of swear words (Table 2) shows that cyberbullying events don’t contain more swear words in cyberbullying detection datasets, likely limiting their utility for the predictions. ($P(S|C) \approx P(S|NC) \& P(C|S) \approx P(C|NC)$). In fact, around 70% of the posts with swear words belong to the negative class and fewer than 3% of all sessions have no swear words in both datasets.

	P(C)	P(NC)	P(S C)	P(S NC)	P(C S)	P(NC S)
Instagram	0.29	0.71	1.0	0.98	0.87	0.87
Vine	0.30	0.69	1.0	0.97	0.86	0.84

Table 2: Distribution of swear words. S: Swear word, C: Cyberbullying, NC: No Cyberbullying.

Measuring swear word bias

To measure the cyberbullying detection model’s bias towards swear words, we adopt the Error Rate Equality Difference approach (Dixon et al. 2018), which relies on the FPR (false positive rate) and FNR (false negative rate) metrics, calculated as follows:

$$FPR = \frac{FP}{FP + TN} \quad ; \quad FNR = \frac{FN}{FN + TP} \quad (1)$$

For each swear word w , this allows to compute the FPRD (FPR difference) and FNRD (FNR difference) as the model bias towards that word:

$$\begin{aligned} FPRD_w &= |FPR - FPR_w| \\ FNRD_w &= |FNR - FNR_w| \end{aligned} \quad (2)$$

Where FPR and FNR are calculated on the entire test set; FPR_w and FNR_w are calculated on the subset of the test set that contains w .

Having those, one can then aggregate the bias towards all swear words W under consideration, i.e. $FPED$ (false positive equality difference), and $FNED$ (false positive equality difference):

$$\begin{aligned} FPED &= \sum_{w \in W} |FPR - FPR_w| \\ FNED &= \sum_{w \in W} |FNR - FNR_w| \end{aligned} \quad (3)$$

Lower values indicate a fairer model.

Quantifying bias with transformers

Prior to moving on to bias mitigation, we first investigate and quantify the impact of swear words $FPRD_w$ on the cyberbullying detection model. Using a vanilla BERT model for initial experiments in both datasets, we show an analysis of the bias measurement of the most frequent swear words as well as most bias-inducing swear words in Table 3. This analysis shows that the most frequent swear words (left) have a low impact on model bias (left). If we look instead at the swear words that cause the highest model bias, we observe that these form a non-intersecting set of swear words that are less frequent (right). We will come back to these top bias-inducing swear words in our study to assess the effectiveness of debiasing strategies.

Experiments involving five transformer models BERT (Devlin et al. 2019), Roberta (Liu et al. 2019), Longformer (Beltagy, Peters, and Cohan 2020), ELECTRA (Clark et al. 2020) and MPNET (Song et al. 2020) reveal different levels of bias, consistently displaying higher bias scores across both datasets (Figure 3).

ID-XCB

In this section, we introduce ID-XCB, our debiasing method and its theoretical implementation in detail. The framework, depicted in Figure 4, is divided into two parts:

1. Transformers Embedding: uses encoders of transformers to generate three kinds of embeddings for training: (i) **clear embeddings** from the original training dataset, (ii) **adversarial embeddings** from the training dataset but with all swear words replaced with a mask, and (iii) **validation embeddings** from validation datasets.

2. Fine-tuning Training: is responsible for integrating three different training processes to break data dependency and improve the generalisability of the debiasing technique and the cyberbullying detection model, which includes adversarial training, tasking training and fairness training. The hidden states selector helps find which layers are best to optimise knowledge.

Training loss functions

The intuition of ID-XCB is that leveraging the combination of the three training loss functions will lead to improved debiasing and performance generalisation on unseen datasets. Details of the three loss functions are provided next.

Adversarial training The aim is to apply adversarial training against biased latent representations to mitigate unwanted bias. Thus, we utilise the cosine similarity function to generate cosine loss such that the model can’t tell the difference between real training samples or artificially synthesised training samples. The loss function is defined as follows:

$$\text{EmbeddingLoss} = 1 - \cos(x_1, x_2) \quad (4)$$

Where x_1 , and x_2 refer to the clear embeddings and adversarial embeddings. The training goal is to optimise EmbeddingLoss close to 0.

Most frequent swear words			Top bias-inducing swear words			
SW	$FPRD_w$ (IG)	$FPRD_w$ (VN)	SW	$FPRD_w$ (IG)	SW	$FPRD_w$ (VN)
fuck	0.008	0.007	Piece of shit	0.929	cunt	0.900
shit	0.040	0.040	nigger	0.929	cum	0.500
fuckin	0.008	0.010	dickhead	0.929	bitches	0.400
bitch	0.040	0.040	gash	0.929	boob	0.234
hell	0.007	0.001	cunt	0.429	faggot	0.234

Table 3: Bias with most frequent swear words and with top bias-inducing swear words.

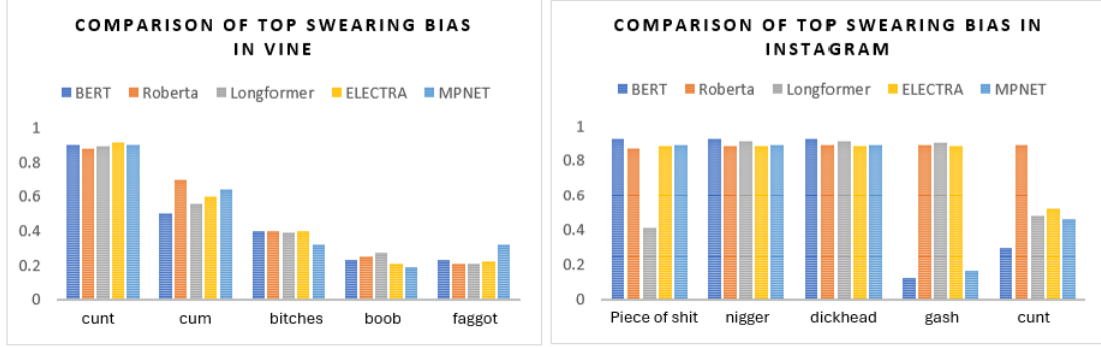


Figure 3: Swear word bias for 5 transformer models on both datasets.

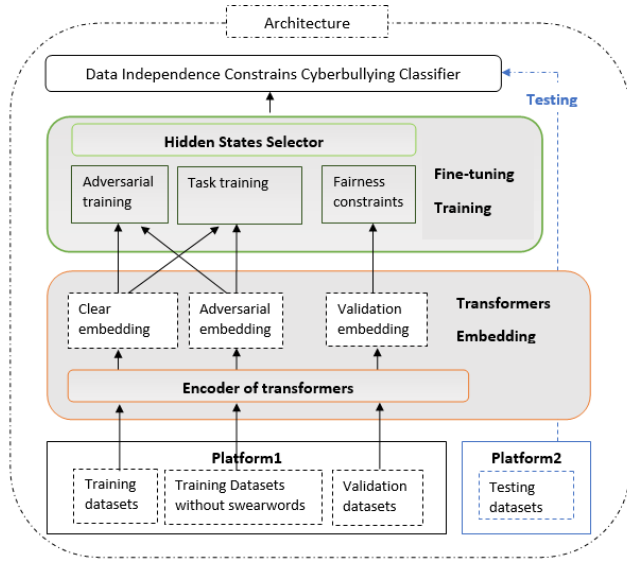


Figure 4: Architecture of ID-XCB.

Task training Binary Cross-Entropy loss is used to train the main task: cyberbullying detection.

$$\begin{aligned}
 BCELoss = & y(\log \frac{(x_1 + x_2)}{2}) \\
 & + (1 - y)(\log(1 - \frac{(x_1 + x_2)}{2})) \quad (5)
 \end{aligned}$$

Where y refers to the training labels. The input is a synthetic embedding which is the average of clear and adversar-

ial embeddings.

Fairness constraints If the terms of debiasing are strictly enforced, this may be beneficial to ensure fairness but may harm model accuracy. In practice, sensitive tuning is performed by using a proportional constraint, which can maintain a more suitable trade-off. We use an independent validation set to derive fairness constraints, which we define as:

$$FC = \beta \left(\sum_{w=1}^n FPRD_w + \sum_{w=1}^n FNRD_w \right) \quad (6)$$

Where β is how tightly the data is bound to adjust the fairness constraints, and w is the swear word under consideration within the validation data.

Hidden states selector

Compared to static embeddings, contextualised embeddings such as BERT, GPT and ELMO are less biased (May et al. 2019), but still show a tendency to adopt biases during training (Zhao et al. 2019; Kurita et al. 2019). These biases are learned in each layer (Bommasani, Davis, and Cardie 2020), thus fine-tuning the orthographic projections in intermediate (hidden) layers (Kaneko and Bollegala 2021) is an efficient method that doesn't depend on the model. To identify which layer is good for fine-tuning, we add a hidden states selector, which iteratively extracts the representation of each layer as input to the classifier for fine-tuning, and then tests the generalisation and debiasing capabilities of each level.

Constraint-based classifier

It comprises two fully connected layers on top. The two-layer feed-forward network is designed with ReLU activa-

tion and 512 hidden sizes for the first layer and Softmax activation for the output layer. Batch normalisation is added to standardise these inputs and reduce the generalisation error, so as to increase the generalisability of the classifier.

Joint training

Following Algorithm 1, we combine the three training losses to fine-tune a classifier. Adversarial training and task training optimise ID-XCB model parameters on a training dataset, and simultaneously enforce the fairness constraints on a validation set to reduce swear word impact. However, as the FC loss is used on an independent validation set V_v , this causes a non-convex combining loss. The non-zero-sum method (Cotter, Jiang, and Sridharan 2019) deals with non-differentiable, even discontinuous constraints. The training goal is not to converge the combined cost function to the lowest point, but to reach a certain trade-off. We set a threshold t to achieve the training target.

Algorithm 1: ID-XCB training. V_s : training embeddings; V_a : adversarial embeddings; V_v : validation embeddings; Y_t : training labels; Y_v : validation labels; Y_a : adversarial labels; n : number of epochs.

```

Require:  $V_s, V_a, V_v, Y_t, Y_v, Y_a, n$ 
0: for epoch in range (n) do
0:   for  $layer(i) = 1, 2, \dots, 12$  do
0:     for step,  $(V_s i, V_a i, V_v i), (Y_t, Y_a)$  do
0:        $l_t = \text{BCELoss}((V_s, V_a), Y_t)$ 
0:        $l_a = \text{EmbeddingLoss}((V_s, V_a), Y_a)$ 
0:        $l_{s_d} = \text{FC}(\text{classifier}, V_v)$ 
0:       if  $(l_t + l_a + l_d) \geq t$  then
0:         classifier.backward( $l_t + l_a + l_d$ )
0:         results[i]=classifier.evaluate( $V_v, Y_v$ )
0:       else
0:         Exit!

```

Experiment Settings

Our models. BERT and RoBERTa are frequently employed for cyberbullying detection due to their advanced natural language understanding capabilities (Ogunleye and Dharmaraj 2023; Verma et al. 2022; Yi and Zubiaga 2023b). While ID-XCB is flexible and can adopt other transformer models, here we experiment with BERT_base and RoBERTa_base, which we refer to as ID-XCB_{BERT} and ID-XCB_{RoBERTa}. When training these models, we use the training hyper-parameters recommended by Sun et al. (2020); Batch size: 16; Learning rate (Adam): $2e^{-5}$; Number of epochs: 4.

Pre-processing. We follow Ge, Cheng, and Liu (2021) to aggregate and clean session data, and to truncate session lengths to 512 tokens, with the difference that we do not perform oversampling (as our objective is to keep the original data imbalance).

Baseline models. We consider Cheng et al. (2021) and Gencoglu (2020) for standing as highly influential debiasing methods for the cyberbullying detection task. Their approach of bias constraints on training data has consistently

demonstrated state-of-the-art performance in recent experiments. Thus, we compare our method with these two state-of-the-art cyberbullying detection debiasing methods (using BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019) variants of those) as well as vanilla transformers.

- **De-RoBERTa & De-BERT:** applied reinforcement on transformers (Cheng et al. 2021).
- **FC-RoBERTa & FC-BERT:** uses Error Rate Equality Difference to restrict transformers’ discrimination behaviour (Gencoglu 2020).
- **Roberta & BERT.**

Ablated models. Aiming to gain a better understanding of the contribution of each component of ID-XCB, we experiment with ablated models where some of the components are removed. We test a total of three models:

- **ID-XCB_{EB}** Removing fairness constraints.
- **ID-XCB_{BF}** Without Adversarial training.
- **ID-XCB_{EF}** The synthetic embedding is replaced by the original embedding, i.e. ignoring the weakening effect on swear words.

Evaluation. We use five widely-used evaluation metrics for imbalanced datasets and model bias. These include (1) recall, precision and microF1 as performance metrics, and (2) FPED and FNED, as fairness indicators, are cumulative deviation values for each swear word, the scale can be [0, Positive infinity], where 0 indicates no deviation.

Train-test splits. We choose 5 random folds with 80%-20% sessions for training / testing, reporting the average performances across the 5 runs.

Results

We next discuss results of our experiments and delve into numerous aspects of our model.

Overall performance and debiasing

Table 4 shows the results for all models, including our two ID-XCB variants and six baselines.

The average F1 score across five runs for ID-XCB_{BERT} is 0.83 with a standard deviation of 0.01 on both datasets. The best baseline BERT debias model from previous research, De-BERT, achieves an average F1 score of 0.84 with a standard deviation of 0.02 on the IG→IG dataset, and 0.74 with a standard deviation of 0.02 on the VN→VN dataset. A paired t-test shows that the difference in performance between ID-XCB_{BERT} and De-BERT is statistically significant (p-value = 0.031), with a t-statistic of 2.54. This indicates that ID-XCB_{BERT} consistently outperforms De-BERT, based on BERT, in the context of swearing debiasing work. Similarly, we apply the same method to assess the statistical significance between ID-XCB_{RoBERTa} and De-RoBERTa. The average F1 across five runs for ID-XCB_{RoBERTa} is 0.88 (IG→IG) and 0.89 (VN→VN) with a standard deviation of 0.02. In comparison, De-RoBERTa achieves an average F1 of 0.87 with a standard deviation of 0.02 (IG→IG) and 0.76 with a standard deviation of 0.03 (VN→VN). A paired t-test revealed that the difference in performance between the two models is statistically significant (p-value = 0.004), with a t-statistic of 3.714, indi-

Src→Tgt	Model	F1	Rec.	Prec.	FPED	FNED
IG→IG	BERT	0.78 (± 0.02)	0.78 (± 0.02)	0.78 (± 0.02)	8.86 (± 0.36)	27.40 (± 5.98)
	RoBERTa	0.84 (± 0.04)	0.80 (± 0.03)	0.85 (± 0.02)	6.81 (± 3.49)	23.00 (± 3.77)
	De-BERT	0.84 (± 0.02)	0.82 (± 0.03)	0.83 (± 0.02)	14.00 (± 2.79)	20.00 (± 11.68)
	De-RoBERTa	0.87 (± 0.02)	0.85 (± 0.06)	0.88 (± 0.05)	3.20 (± 1.49)	13.00 (± 2.34)
	FC-BERT	0.75 (± 0.03)	0.75 (± 0.03)	0.75 (± 0.03)	5.05 (± 1.24)	16.40 (± 1.83)
	FC-RoBERTa	0.81 (± 0.02)	0.77 (± 0.06)	0.85 (± 0.04)	15.20 (± 4.97)	16.10 (± 3.22)
	ID-XCB _{BERT}	0.83 (± 0.01)	0.84 (± 0.02)	0.83 (± 0.01)	6.70 (± 1.48)	16.29 (± 2.70)
	ID-XCB _{RoBERTa}	0.88 (± 0.02)	0.88 (± 0.02)	0.89 (± 0.02)	3.90 (± 1.00)	15.60 (± 4.56)
VN→VN	BERT	0.77 (± 0.03)	0.75 (± 0.04)	0.80 (± 0.04)	5.73 (± 1.17)	18.00 (± 7.06)
	RoBERTa	0.85 (± 0.03)	0.83 (± 0.07)	0.87 (± 0.02)	4.66 (± 1.67)	16.99 (± 5.62)
	De-BERT	0.74 (± 0.02)	0.73 (± 0.02)	0.86 (± 0.02)	10.40 (± 6.14)	11.40 (± 3.94)
	De-RoBERTa	0.76 (± 0.03)	0.85 (± 0.05)	0.68 (± 0.06)	6.58 (± 3.47)	14.10 (± 9.38)
	FC-BERT	0.66 (± 0.07)	0.67 (± 0.08)	0.66 (± 0.16)	16.00 (± 6.27)	17.00 (± 5.05)
	FC-RoBERTa	0.79 (± 0.02)	0.81 (± 0.02)	0.76 (± 0.03)	8.10 (± 1.51)	13.00 (± 6.34)
	ID-XCB _{BERT}	0.83 (± 0.01)	0.83 (± 0.02)	0.85 (± 0.01)	3.00 (± 0.60)	17.10 (± 2.75)
	ID-XCB _{RoBERTa}	0.89 (± 0.02)	0.86 (± 0.02)	0.93 (± 0.02)	1.51 (± 1.21)	11.73 (± 3.26)

Table 4: Results for ID-XCB and six baseline models.

cating that ID-XCB_{RoBERTa} consistently outperforms De-RoBERTa.”

These results demonstrate the general superiority of our ID-XCB model, of which ID-XCB_{RoBERTa} shows superior performance. We observe that in the in-dataset experiments it is capable of obtaining higher performance scores. It also achieves superior bias mitigation than most of the models, showing the best FPED score across all models for the Vine dataset. It is also only behind De-BERT for FNED in the Vine dataset and only behind De-RoBERTa in the Instagram dataset.

Interestingly, we observe that ID-XCB achieves competitive debiasing while leading to improved overall performance. Debiasing can often sacrifice performance, especially when applying data rebalancing, regularisation, and adversarial learning debiasing methods (Chen et al. 2023). In our study, our goal was not to achieve the lowest possible debiasing result but to find a balance between performance and bias reduction. We observe that all our models mitigate the bias of the original base model, but we did not further constrain the bias, as doing so would reduce model performance. In this challenging scenario, ID-XCB achieves a trade-off between debias and performance.

Ablation study

We conduct ablation tests without adversarial training, fairness constraints and synthetic embedding in ID-XCB_{RoBERTa}. The results in Table 5 show that all components contribute to a noticeable improvement to both performance and debiasing. It is particularly worth mentioning that when the synthetic embedding (ID-XCB_{EF}) is replaced by the original embedding, the bias increases sharply. This aligns with our expectations, as the component was designed to mitigate the model’s fixation on specific swear words, but still preserving their contribution to the task.

Revisiting bias-inducing swear words

In Table 3 we have identified the top 5 most bias-inducing swear words for both datasets when using a vanilla BERT model, which we revisit here to assess the extent to which ID-XCB helps mitigate the bias induced by these words. Table 6 shows the FPRD scores for these top 5 swear words in Instagram and Vine, respectively. We observe a significant decrease in bias score (FPRD) in both cases for all the words individually as well as on aggregate when we average their scores. In our analysis, we observe a more significant decrease in bias scores for the Instagram dataset compared to the Vine dataset when focusing on the most bias-inducing words. This difference could be attributed to different data collection strategies and different data structures on different platforms. Yi and Zubiaga (2023a) observe remarkable differences between the two datasets. Where the majority of cyberbullying incidents occur at the beginning in the Instagram dataset, these are more uniformly spread throughout the entire session for the Vine dataset, which makes model debiasing more challenging. Additionally, the Instagram dataset may contain more diverse or expressive language patterns, which can amplify the impact of bias-inducing words.

Generalisation

To assess the generalisability of ID-XCB, we next look at experiments crossing datasets, i.e. using a dataset for training and the other for testing. Looking again at the highest bias-inducing words, Figure 5 demonstrates that ID-XCB smoothly shrinks the highest bias at the word level without stimulating new bias, proving its effectiveness on mitigating bias for unseen data.

Table 7 shows performance results for cross-platform experiments, highlighting in the first instance the overall superior performance of ID-XCB_{RoBERTa}. Despite performance improvement, the ID-XCB model can outperform

Src→Tgt	Model	F1	FPED	FNED
IG→IG	ID-XCB	0.89 (± 0.01)	3.90 (± 1.00)	15.60 (± 4.56)
	ID-XCB _{EB}	0.85 (± 0.03)	23.00 (± 4.65)	16.00 (± 4.97)
	ID-XCB _{BF}	0.86 (± 0.02)	10.00 (± 1.42)	21.00 (± 3.78)
	ID-XCB _{EF}	0.84 (± 0.02)	15.00 (± 3.26)	20.00 (± 3.74)
VN→VN	ID-XCB	0.89 (± 0.02)	1.51 (± 1.21)	11.73 (± 3.26)
	ID-XCB _{EB}	0.87 (± 0.01)	6.46 (± 1.76)	16.38 (± 0.70)
	ID-XCB _{BF}	0.88 (± 0.02)	4.98 (± 1.19)	16.88 (± 1.24)
	ID-XCB _{EF}	0.86 (± 0.01)	6.71 (± 1.51)	16.90 (± 0.04)

Table 5: Performance comparison of ID-XCB_{RoBERTa} and ablated variants.

SW	RoBERTa	ID-XCB _{RoBERTa}
Instagram		
piece of shit	0.929	0.449
nigger	0.929	0.135
dickhead	0.929	0.706
gash	0.929	0.049
cunt	0.429	0.363
Average	0.829	0.340
Vine		
cunt	0.900	0.242
cum	0.500	0.026
bitches	0.400	0.338
boob	0.234	0.075
faggot	0.234	0.214
Average	0.454	0.179

Table 6: Bias scores (FPRD).

some of the baseline models in terms of debiasing, but it doesn’t achieve the top result in the debiasing metric for this study case. Because the debiasing metric or performance metric might not adequately reflect the trade-offs between bias reduction and maintaining high accuracy or other essential performance measures. The context in which the model is deployed requires a balanced measure of overall model effectiveness in this case. This calls for further consideration of the trade-off between debiasing and performance. Delving into the significance of this, we next look into the trade-off between the two metrics.

Assessing the trade-off between performance and debiasing

While assessing model generalisability, we need to consider the performance vs debiasing trade-off. This is important as we observe that, for example, lowest performing models, FC-RoBERTa (19% and 40% below our model in the two datasets), achieve some of the best debiasing scores. Given the lack of a joint metric, we analyse the interplay between them. Therefore, we design constraint weights and layer selectors to achieve a balance between debiasing and main-

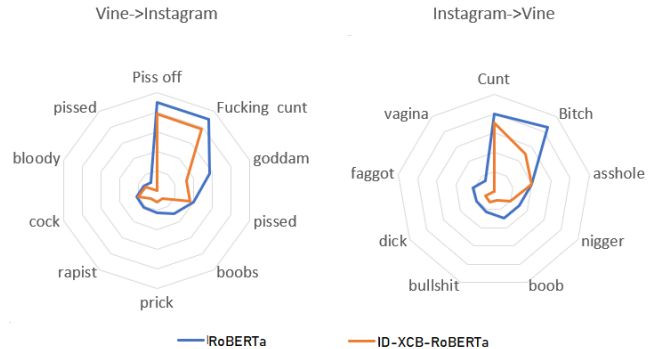


Figure 5: Swear word bias in RoBERTa vs ID-XCB_{RoBERTa} for cross-platform experiments.

taining performance. Figure 5 shows our constraints didn’t encourage new bias generated during cross-platform tasks. Models smoothly shrink the highest bias in word level without stimulating new bias.

It is difficult to find a simple linear relationship between task performance and debiasing effects. We use FPRD and FNED as the metrics to measure bias associated with specific words. As pointed out by Borkan et al. (2019), these are metrics of fairness rather than measurements of performance. Given the different objectives of performance and bias metrics, we analyse their interplay. In Table 7, we observe that the model (FC-BERT) with the worst cross-platform performance (F1 score) also obtain the lowest bias value. This is because FPR and FNR appear to differ minimally if they are both very large across the dataset and bias triggers.

Impact of constraint weighting The weight is an important parameter to adjust how many constraints are applied to the bias. We assess 10 values ranging from [0.1-1] at intervals of 0.1. Figure 6 illustrates the correlation between bias constraint weights and task performance. Both cross-dataset settings show the same pattern: gradually debiasing on the training dataset enhances the model’s performance on unseen datasets. The model achieves optimal performance when parameters are set to 0.6 or 0.7. However, tightening the constraints beyond this point leads to a significant drop in model performance. This demonstrates that excessive de-

Source→Target	Model	F1	Rec.	Prec.	FPED	FNED
IG→VN	BERT	0.54 (± 0.02)	0.52 (± 0.03)	0.61 (± 0.01)	4.00 (± 1.31)	18.00 (± 2.58)
	RoBERTa	0.70 (± 0.02)	0.71 (± 0.02)	0.70 (± 0.01)	28.30 (± 3.58)	31.90 (± 8.43)
	De-BERT	0.64 (± 0.01)	0.68 (± 0.02)	0.63 (± 0.01)	14.00 (± 3.49)	30.00 (± 4.58)
	De-RoBERTa	0.62 (± 0.01)	0.71 (± 0.03)	0.60 (± 0.02)	32.00 (± 2.05)	27.00 (± 1.65)
	FC-BERT	0.41 (± 0.02)	0.50 (± 0.01)	0.34 (± 0.02)	3.00 (± 1.12)	0.00
	FC-RoBERTa	0.57 (± 0.01)	0.65 (± 0.02)	0.54 (± 0.02)	7.20 (± 1.32)	10.90 (± 2.62)
	ID-XCB _{BERT}	0.73 (± 0.02)	0.72 (± 0.02)	0.74 (± 0.02)	23.00 (± 3.37)	31.00 (± 4.56)
	ID-XCB _{RoBERTa}	0.76 (± 0.01)	0.78 (± 0.01)	0.75 (± 0.01)	20.30 (± 3.56)	17.40 (± 6.02)
VN→IG	BERT	0.70 (± 0.04)	0.66 (± 0.01)	0.79 (± 0.06)	24.00 (± 2.32)	37.00 (± 3.06)
	RoBERTa	0.71 (± 0.02)	0.69 (± 0.04)	0.79 (± 0.02)	32.40 (± 4.54)	35.40 (± 6.72)
	De-BERT	0.64 (± 0.03)	0.68 (± 0.02)	0.63 (± 0.01)	31.00 (± 4.94)	30.00 (± 2.70)
	De-RoBERTa	0.61 (± 0.01)	0.59 (± 0.02)	0.69 (± 0.01)	12.00 (± 1.18)	15.00 (± 4.32)
	FC-BERT	0.31 (± 0.04)	0.53 (± 0.02)	0.63 (± 0.03)	13.60 (± 2.27)	3.50 (± 1.00)
	FC-RoBERTa	0.41 (± 0.05)	0.50 (± 0.03)	0.34 (± 0.06)	5.00 (± 2.32)	16.00 (± 3.42)
	ID-XCB _{BERT}	0.74 (± 0.01)	0.73 (± 0.01)	0.75 (± 0.01)	20.00 (± 3.21)	34.00 (± 4.01)
	ID-XCB _{RoBERTa}	0.81 (± 0.02)	0.80 (± 0.01)	0.81 (± 0.01)	23.00 (± 5.31)	35.00 (± 3.48)

Table 7: Cross-dataset results for ID-XCB and six baseline models.

biasing during training may not generalise well.

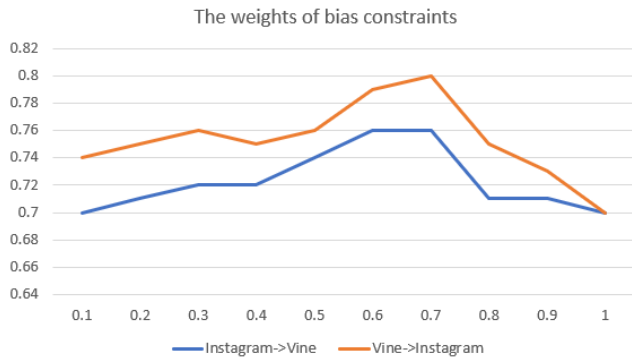


Figure 6: Impact of constraint weighting. X axis for constraint weights (β) and Y axis for F1 score.

The bias through different layers The hidden state selector is used to select the most transferable layer in transformers, so as to improve the transferability of the overall model. We are interested in quantifying its impact across layers, particularly as the last layer of the model is not necessarily the one leading to the best performance. To gain a deeper understanding into how the performance of various layers impacts debiasing and cyberbullying detection transferability, we measure the F1 score and FPED in each layer, fixing the last layer as the reference. Figure 7 shows that layer 10 on IG→VN leads to the best performance-debiasing trade-off, and layer 8 on VN→IG. We conclude two interesting phenomena: 1) A non-linear relationship exists between debiasing and performance. For instance, the eighth and fifth layers demonstrate the smallest FPRD values (IG→VN), yet their performance reaches two extremes; 2) The first few layers focus on learning general features, rendering debiasing

less impactful on overall model performance. However, in deeper layers, the network pays more attention to specific features, making debiasing efforts more effective in enhancing performance.

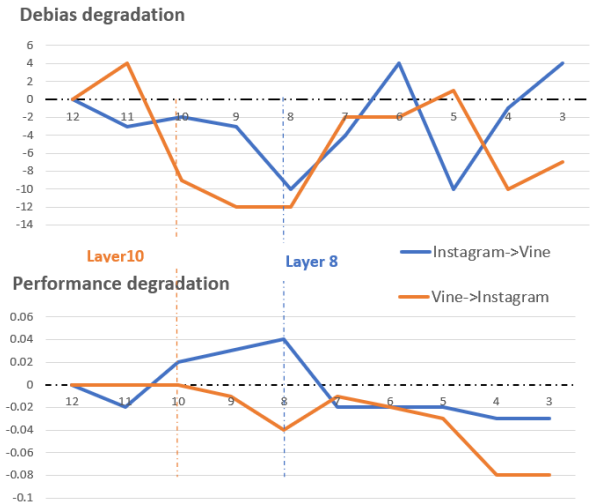


Figure 7: Bias across layers: X axis refers to 12 transformer layers. Y axis refers to relative performance and debiasing compared to layer 12.

Conclusion

In this study, we introduce ID-XCB, the first data-independent debiasing approach for cyberbullying detection. Our study looks at both measurement and mitigation of biases in cyberbullying datasets and detection. We first quantify the impact and complexity of the bias produced by

swear words in transformer-based models. While tackling the problem, we propose a novel bias mitigation method, ID-XCB, which encapsulates adversarial training, constraint optimisation and layer fine-tuning. The method pays attention to the damage to the module due to excessive constraints and considers the trade-off between fairness and accuracy in the selection of algorithms. Our experiments on two cyberbullying detection datasets show the effectiveness of ID-XCB, achieving competitive bias constraints enabling data-independent debiasing and training. This improvement is generally consistent in in-dataset experiments, whereas the improvement is particularly on performance for cross-dataset settings, with a good balance on debiasing, but which shows room for improvement in these challenging settings.

Limitations

Our work is however not without limitations. Most importantly, the dearth of available datasets leads to inevitable limitations in further studying generalisability across a more diverse set of datasets and across other social media platforms beyond Instagram and Vine. While there has been a more substantial body of work in related tasks within the umbrella of online abuse detection, such as hate speech detection, research on cyberbullying detection, and particularly on session-based cyberbullying detection, is much more limited to date and would greatly benefit from access to a broader collection of available datasets.

Our proposed ID-XCB model demonstrates state-of-the-art performance on the cyberbullying detection task, enabling some generalization across different datasets and platforms. This improvement comes with a competitive trade-off on performance and debiasing, however the model is not always consistently best across both metrics, which shows an area for further improvement.

Ethics Statement

The aim of our research is to contribute to society and to human well-being by curbing incidents of cyberbullying online and particularly on social media. Our approach to mitigating biases in keyword presence across cyberbullying and non-cyberbullying incidents is performed without leveraging any identity information to support the debiasing, and hence our approach is designed to support all individuals equally and with no intended discrimination or potential for harm towards any vulnerable groups.

There is an inevitable risk, as adversaries, who actually engage in cyberbullying incidents, may use this kind of research for malicious purposes such as to learn how to circumvent detection. This is however not the intended use of our research.

References

Agrawal, S.; and Awekar, A. 2018. Deep learning for detecting cyberbullying across multiple social media platforms. In *Advances in Information Retrieval: 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings*, 141–153. Springer.

Badjatiya, P.; Gupta, M.; and Varma, V. 2019. Stereotypical Bias Removal for Hate Speech Detection Task using Knowledge-based Generalizations. In Liu, L.; White, R. W.; Mantrach, A.; Silvestri, F.; McAuley, J. J.; Baeza-Yates, R.; and Zia, L., eds., *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, 49–59. ACM.

Beltagy, I.; Peters, M. E.; and Cohan, A. 2020. Longformer: The long-document transformer. *ArXiv preprint*, abs/2004.05150.

Bommasani, R.; Davis, K.; and Cardie, C. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4758–4781. Online: Association for Computational Linguistics.

Borkan, D.; Dixon, L.; Sorensen, J.; Thain, N.; and Vasserman, L. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, 491–500.

Chen, J.; Dong, H.; Wang, X.; Feng, F.; Wang, M.; and He, X. 2023. Bias and debias in recommender system: A survey and future directions. *ACM Transactions on Information Systems*, 41(3): 1–39.

Cheng, L.; Mosallanezhad, A.; Silva, Y.; Hall, D.; and Liu, H. 2021. Mitigating Bias in Session-based Cyberbullying Detection: A Non-Compromising Approach. 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)*, 2158–2168. Online: Association for Computational Linguistics.

Cheng, L.; Silva, Y. N.; Hall, D.; and Liu, H. 2020. Session-based cyberbullying detection: Problems and challenges. *IEEE Internet Computing*, 25(2): 66–72.

Clark, K.; Luong, M.; Le, Q. V.; and Manning, C. D. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Cotter, A.; Jiang, H.; and Sridharan, K. 2019. Two-player games for efficient non-convex constrained optimization. In *Algorithmic Learning Theory*, 300–332. PMLR.

Dadvar, M.; Jong, F. d.; Ordelman, R.; and Trieschnigg, D. 2012. Improved cyberbullying detection using gender information. In *Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012)*. University of Ghent.

Dadvar, M.; Trieschnigg, D.; Ordelman, R.; and Jong, F. d. 2013. Improving cyberbullying detection with user context. In *European Conference on Information Retrieval*, 693–696. Springer.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.
- Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 67–73.
- Ge, S.; Cheng, L.; and Liu, H. 2021. Improving cyberbullying detection with user interaction. In *Proceedings of the Web Conference 2021*, 496–506.
- Gencoglu, O. 2020. Cyberbullying detection with fairness constraints. *IEEE Internet Computing*, 25(1): 20–29.
- Google, I. 2010. google badwords, <https://code.google.com/archive/p/badwordslst/downloads> [Accessed: (Use the date of access)].
- Gururangan, S.; Marasović, A.; Swayamdipta, S.; Lo, K.; Beltagy, I.; Downey, D.; and Smith, N. A. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.
- Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. In Lee, D. D.; Sugiyama, M.; von Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 3315–3323.
- Hosseinmardi, H.; Mattson, S. A.; Ibn Rafiq, R.; Han, R.; Lv, Q.; and Mishra, S. 2015. Analyzing labeled cyberbullying incidents on the instagram social network. In *International conference on social informatics*, 49–66. Springer.
- Hovy, D.; and Prabhumoye, S. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8): e12432.
- Kaneko, M.; and Bollegala, D. 2021. Debiasing Pre-trained Contextualised Embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 1256–1266. Online: Association for Computational Linguistics.
- Kurita, K.; Vyas, N.; Pareek, A.; Black, A. W.; and Tsvetkov, Y. 2019. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, 166–172. Florence, Italy: Association for Computational Linguistics.
- Lafreniere, K. C.; Moore, S. G.; and Fisher, R. J. 2022. The power of profanity: The meaning and impact of swear words in word of mouth. *Journal of Marketing Research*, 59(5): 908–925.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.
- May, C.; Wang, A.; Bordia, S.; Bowman, S. R.; and Rudinger, R. 2019. On Measuring Social Biases in Sentence Encoders. 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)*, 622–628. Minneapolis, Minnesota: Association for Computational Linguistics.
- Menin, D.; Guarini, A.; Mameli, C.; Skrzypiec, G.; and Brighi, A. 2021. Was that (cyber) bullying? Investigating the operational definitions of bullying and cyberbullying from adolescents' perspective. *International journal of clinical and health psychology*, 21(2): 100221.
- Nozza, D.; Volpetti, C.; and Fersini, E. 2019. Unintended bias in misogyny detection. In *Ieee/wic/acm international conference on web intelligence*, 149–155.
- Ogunleye, B.; and Dharmaraj, B. 2023. The use of a large language model for cyberbullying detection. *Analytics*, 2(3): 694–707.
- Olweus, D. 2001. Bullying at school: Tackling the problem. *OECD observer*, 24–24.
- Pamungkas, E. W.; Basile, V.; and Patti, V. 2023. Investigating the role of swear words in abusive language detection tasks. *Language Resources and Evaluation*, 57(1): 155–188.
- Perera, A.; and Fernando, P. 2021. Accurate cyberbullying detection and prevention on social media. *Procedia Computer Science*, 181: 605–611.
- Rafiq, R. I.; Hosseinmardi, H.; Han, R.; Lv, Q.; Mishra, S.; and Mattson, S. A. 2015. Careful what you share in six seconds: Detecting cyberbullying instances in Vine. In Pei, J.; Silvestri, F.; and Tang, J., eds., *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2015, Paris, France, August 25 - 28, 2015*, 617–622. ACM.
- Schramowski, P.; Turan, C.; Andersen, N.; Rothkopf, C. A.; and Kersting, K. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3): 258–268.
- Smith, P. K.; Mahdavi, J.; Carvalho, M.; Fisher, S.; Russell, S.; and Tippett, N. 2010. Cyberbullying: Its nature and impact in secondary school pupils. *ArXiv preprint*, abs/10.1111.
- Song, K.; Tan, X.; Qin, T.; Lu, J.; and Liu, T.-Y. 2020. Mpnnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33: 16857–16867.
- Stephens, R.; and Robertson, O. 2020. Swearing as a response to pain: Assessing hypoalgesic effects of novel “swear” words. *Frontiers in psychology*, 11: 723.
- Sun, C.; Qiu, X.; Xu, Y.; and Huang, X. 2020. How to Fine-Tune BERT for Text Classification? In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11856 LNAI, 194–206. Springer. ISBN 9783030323806.
- Sweeney, C.; and Najafian, M. 2020. Reducing sentiment polarity for demographic attributes in word embeddings using adversarial learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 359–368.
- Vaidya, A.; Mai, F.; and Ning, Y. 2019. Empirical analysis of multi-task learning for reducing model bias in toxic comment detection. *ArXiv preprint*, abs/1909.09758.

Van Bruwaene, D.; Huang, Q.; and Inkpen, D. 2020. A multi-platform dataset for detecting cyberbullying in social media. *Language Resources and Evaluation*, 54: 851–874.

Van Hee, C.; Jacobs, G.; Emmerly, C.; Desmet, B.; Lefever, E.; Verhoeven, B.; De Pauw, G.; Daelemans, W.; and Hoste, V. 2018. Automatic detection of cyberbullying in social media text. *PLoS one*, 13(10): e0203794.

Verma, K.; Milosevic, T.; Cortis, K.; and Davis, B. 2022. Benchmarking language models for cyberbullying identification and classification from social-media texts. In *Proceedings of the First Workshop on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference*, 26–31.

Yi, P.; and Zubiaga, A. 2022. Cyberbullying detection across social media platforms via platform-aware adversarial encoding. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, 1430–1434.

Yi, P.; and Zubiaga, A. 2023a. Session-based cyberbullying detection in social media: A survey. *Online Social Networks and Media*, 36: 100250.

Yi, P.; and Zubiaga, A. 2023b. Session-based cyberbullying detection in social media: A survey. *Online Social Networks and Media*, 36: 100250.

Zafar, M. B.; Valera, I.; Gomez-Rodriguez, M.; and Gum-madi, K. P. 2017. Fairness Constraints: Mechanisms for Fair Classification. In Singh, A.; and Zhu, X. J., eds., *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20–22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, 962–970. PMLR.

Zhang, H.; Lu, A. X.; Abdalla, M.; McDermott, M.; and Ghassemi, M. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *proceedings of the ACM Conference on Health, Inference, and Learning*, 110–120.

Zhao, J.; Wang, T.; Yatskar, M.; Cotterell, R.; Ordonez, V.; and Chang, K.-W. 2019. Gender Bias in Contextualized Word Embeddings. 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)*, 629–634. Minneapolis, Minnesota: Association for Computational Linguistics.

Ethics Checklist

For most authors...

1. Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? [Yes, this research question advances science without violating social contracts.](#)
2. Do your main claims in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes, see the Abstract and the Introduction.](#)

3. Do you clarify how the proposed methodological approach is appropriate for the claims made? [Yes, see the ID-XCB.](#)

4. Do you clarify what are possible artifacts in the data used, given population-specific distributions? [Yes, see Datasets and Lexicon.](#)

5. Did you describe the limitations of your work? [Yes, see Limitations.](#)

6. Did you discuss any potential negative societal impacts of your work? [Yes, see Ethics Statement.](#)

7. Did you discuss any potential misuse of your work? [Yes, see Ethics Statement.](#)

8. Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? [NA](#)

9. Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes, we read and ensure our paper conforms to them.](#)

Additionally, if your study involves hypotheses testing...

1. Did you clearly state the assumptions underlying all theoretical results? [Yes, see Introduction.](#)

2. Have you provided justifications for all theoretical results? [Yes, see Ablation study.](#)

3. Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? [Yes, see Assessing the trade-off between performance and debiasing](#)

4. Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? [Yes, see Experiment Settings](#)

5. Did you address potential biases or limitations in your theoretical framework? [Yes, see Quantifying bias with transformers](#)

6. Have you related your theoretical results to the existing literature in social science? [Yes, see Related Work](#)

7. Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? [Yes, see Conclusion and limitations](#)

Additionally, if you are including theoretical proofs...

1. Did you state the full set of assumptions of all theoretical results? [NA](#)

2. Did you include complete proofs of all theoretical results? [NA](#)

Additionally, if you ran machine learning experiments...

1. Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes, we included source code and related URL.](#)

2. Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes, see Experiment settings](#)

3. Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes, We report the standard deviation \(STD\) values in “Table 4” and “Table 5”.](#)
4. Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [No, because our work is not compute intensive](#)
5. Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? [Yes, see Experiment Settings and Results](#)
6. Do you discuss what is “the cost“ of misclassification and fault (in)tolerance? [NA](#)

Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

1. If your work uses existing assets, did you cite the creators? [Yes, we cited them all.](#)
2. Did you mention the license of the assets? [No, because it’s not stated in the original sources, but we cited.](#)
3. Did you include any new assets in the supplemental material or as a URL? [Yes, we included source code](#)
4. Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [NA](#)
5. Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [NA](#)
6. If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see [?](#))? [NA](#)
7. If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see [?](#))? [NA](#)

Additionally, if you used crowdsourcing or conducted research with human subjects...

1. Did you include the full text of instructions given to participants and screenshots? [NA](#)
2. Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? [NA](#)
3. Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [NA](#)
4. Did you discuss how data is stored, shared, and deidentified? [NA](#)