Mitigating Bias in Session-based Cyberbullying Detection: A Non-Compromising Approach

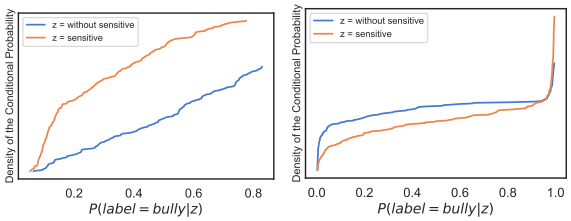
Lu Cheng[1]*, **Ahmadreza Mosallanezhad**[1]*, **Yasin N. Silva**[2], **Deborah L. Hall**[3], and **Huan Liu**[1]
[1] Computer Science and Engineering, Arizona State University
[2] Mathematical and Natural Sciences, Arizona State University
[3] Social and Behavioral Sciences, Arizona State University
{lcheng35, amosalla, ysilva, d.hall, huanliu}@asu.edu

Abstract

The element of repetition in cyberbullying behavior has directed recent computational studies toward detecting cyberbullying based on a *social media session*. In contrast to a single text, a session may consist of an initial post and an associated sequence of comments. Yet, emerging efforts to enhance the performance of session-based cyberbullying detection have largely overlooked unintended social biases in existing cyberbullying datasets. For example, a session containing certain demographic-identity terms (e.g., "gay" or "black") is more likely to be classified as an instance of cyberbullying. In this paper, we first show evidence of such bias in models trained on sessions collected from different social media platforms (e.g., Instagram). We then propose a context-aware and model-agnostic debiasing strategy that leverages a reinforcement learning technique, without requiring any extra resources or annotations apart from a pre-defined set of sensitive triggers commonly used for identifying cyberbullying instances. Empirical evaluations show that the proposed strategy can simultaneously alleviate the impacts of the unintended biases and improve the detection performance.

1 Introduction

Cyberbullying has become a prevalent adverse behavior in online social interactions. Recent findings indicate that over 35% of young people have been victims of cyberbullying and roughly 15% have admitted to cyberbullying others (Hinduja and Patchin, 2020; Kim et al., 2021). The detrimental consequences of cyberbullying have motivated considerable efforts in various fields to combat cyberbullying. For example, in computational studies of cyberbullying detection – which have been largely aimed at classifying text posted on social media platforms with machine learning and natural language processing (NLP) – the primary goal is to improve the overall accuracy and speediness of detection. Partly due to an increased awareness of the repetitive nature of cyberbullying behavior, a number of recent efforts in cyberbullying detection have shifted in focus from classification of a single text to detection in a social media session. A session typically consists of an image/video with a caption, a sequence of comments, and other social content, e.g., number of likes.



(a) HAN. (b) Debiased HAN.

Figure 1: Conditional probability densities of standard HAN and debiased HAN on sessions with and without sensitive triggers z in the Instagram dataset released by (Hosseinmardi et al., 2015).

The promising results, nevertheless, may come from a deeply biased model that captures, uses, and even amplifies the unintended biases embedded in social media data (Zhang et al., 2020). That is, because humans are biased, human-generated language corpora can introduce human social prejudices into model training processes (Caliskan et al., 2017). Evidence of such bias has been found in toxicity detection (Zhang et al., 2020) and hate speech detection (Davidson et al., 2019), revealing that tweets in African-American Vernacular English (AAVE) are more likely to be classified as abusive or offensive. Similarly, a cyberbullying classifier may simply take advantage of sensitive triggers, e.g., demographic-identity information (e.g.,

* Equal contribution

"gay") and offensive terms ("stupid," "ni***r"), to make decisions. Indeed, we find that in the *Instagram* data for benchmarking cyberbullying detection released by (Hosseinmardi et al., 2015), 68.4% of sessions containing the word "gay" are labeled as bullying, 89.4% of sessions containing the word "ni***r," and 64.3% of sessions containing the word "Mexican". In Figure 1, we showcase differences in the performance of a standard hierarchical attention network (HAN) (Yang et al., 2016) – a commonly used model for session-based cyberbullying detection – and a HAN that was debiased using our proposed strategy in sessions with and without sensitive triggers using the benchmark Instagram data. Specifically, the x-axis represents the probability of the classifier predicting a session as bullying, i.e., the decision scores $\mathcal{F} : p(label = bully|Z)$. The y-axis represents the conditional probability densities of the decision scores, i.e., $p(\mathcal{F}|Z)$. Figure 1(a) shows that the densities are dependent on Z and the dependencies are largely reduced by our mitigation strategy, as depicted in Figure 1(b).

This paper aims to mitigate the unintended bias in cyberbullying detection in social media sessions. Our task poses multi-faceted challenges that render recent model-agnostic research in fair text classification – especially, data manipulation methods (Dixon et al., 2018; Sun et al., 2019) – inapplicable. First, in contrast to a single text (e.g., a tweet), social media sessions with a sequence of comments contain rich contextual information. Bias mitigation cannot be defined without *context* (Lee et al., 2020). The axiomatic and absolute definitions may render current interventions (e.g., gender-swapping) ineffective and may even misguide cyberbullying classifiers. Second, session-based cyberbullying detection is a sequential decision-making process rather than a one-off operation. Therefore, current decisions made by a cyberbullying classifier can influence its future predictions and debiasing strategies. Third, these data manipulation methods are impractical in our task due to the need for extra data annotation, which is especially time-consuming for sequential social media data with rich context. In addition, these methods consider fairness through a differentiable loss function that may not directly incorporate specific fairness goals or measures.

To address these challenges, we propose a context-aware and model-agnostic debiasing training framework for cyberbullying detection. It does not require additional resources, apart from a predefined set of sensitive triggers. In particular, drawing from recent advances in reinforcement learning (RL), we consider a classifier as an agent that interacts with the environment to accumulate experience in cyberbullying detection and bias mitigation. At each timestep, the agent makes decisions based on all comments observed up to that point in time and is updated by the collected feedback. Empirical evaluations on two real-world datasets show that the proposed debiasing framework can effectively mitigate the unintended biases while improving the performance of cyberbullying detection.

2 Related Work

Cyberbullying Detection. The growing prevalence of social networking sites and convenient access to digital devices and the internet have substantially expedited information-sharing processes. A byproduct of this, however, has been the increased vulnerability of young people, in particular, to one of the most serious online risks – cyberbullying. To help combat cyberbullying, researchers have used various techniques in machine learning and NLP to automate the process of cyberbullying detection. This is also evidenced by a number of recent competitions and workshops for related tasks such as detection of hate speech against immigrants and women (Basile et al., 2019), offensive language identification (Zampieri et al., 2020), and toxic spans detection (Pavlopoulos et al., 2021).

Early works simplified the task as text classification, the input of which are content-based features (e.g., cyberbullying keywords) extracted from a single text (e.g., a tweet) and labels denoting whether the text is relevant to cyberbullying, see, e.g., (Dinakar et al., 2011; Xu et al., 2012). To better leverage the rich information included in social media data, many studies proposed to augment textual features with emotion/sentiment (Dani et al., 2017), social network information such as relational centrality and ego networks (Squicciarini et al., 2015; Huang et al., 2014), and other multi-modal information such as location and time (Cheng et al., 2019b). Extensive experimental results revealed that the improvement of these approaches is significant.

From the data perspective, research in cyberbullying detection has shifted from modeling a single text to multi-modal data and social media sessions. Underpinning these transitions is an increased recognition of two distinct characteris-

tics of cyberbullying behavior – repetitiveness and power imbalance (Smith et al., 2008). To address these characteristics, studies such as (Cheng et al., 2019a, 2021) proposed to model the structure of a session and temporal dynamics among the comments using HAN. Yet, whereas numerous studies have focused on achieving better prediction performance, these approaches tend to carry or reinforce the unintended social biases in the datasets (Gencoglu, 2020). Our work thus complements earlier research by examining and mitigating unintended bias in cyberbullying detection models.

Fairness in NLP. Humans are inherently biased, and many studies have revealed human biases and discrimination in natural language (Garg et al., 2018; Jentzsch et al., 2019). Evidence has, for instance, emerged in biased pre-trained word embeddings and semantics derived from language corpora. However, in the field of NLP, the question of how to alleviate bias and promote fairness has only more recently begun to be addressed. Using text classification tasks as an example, one predominant method to make the classifiers fairer is to balance training data in a statistical sense. In particular, one can augment original data with external labeled data (Dixon et al., 2018). Similar methods include data oversampling/downsampling, sample weighting (Zhang et al., 2020), and identity term swapping (Park et al., 2018). Dixon et al. (Dixon et al., 2018) added non-toxic samples containing identity terms from Wikipedia articles into training data. A similar strategy was used in (Nozza et al., 2019) for misogyny detection. Badjatiya et al. (Badjatiya et al., 2019) proposed to replace sensitive words with neutral words or tokens.

This balancing strategy, while convenient and easy to implement, is not compatible with session-based cyberbullying detection. First, practical considerations impede us from providing additional labeled data with specific sensitive triggers. Data labeling for session-based cyberbullying detection is especially time-consuming and labor-intensive, given that it requires carefully examining a media object and all associated comments in a social media session. Second, because there are potentially many words or tokens sensitive to cyberbullying, identity term swapping is almost impossible. Third, social media sessions contain sequences of comments that provide contextual information important for both cyberbullying detection and bias mitigation. Simple data augmentation can result in the significant loss of such information. Lastly, balancing can introduce additional calibration parameters that can impair classification performance and bias mitigation (Gencoglu, 2020).

3 Preliminaries

Cyberbullying is often characterized as a *repeated* rather than a one-off behavior (Smith et al., 2008). This unique trait has motivated research that focuses on the detection of cyberbullying in entire *social media sessions*. In contrast to a single text, e.g., a Facebook comment or a tweet, a social media session is typically composed of an initial post (e.g., an image with a caption), a sequence of comments from different users, timestamps, spatial location, user profile information, and other social content such as number of likes (Cheng et al., 2020). Session-based cyberbullying detection presents a number of characteristics such as multi-modality and user interaction (Cheng et al., 2020). In this work, because our goal is to mitigate bias in natural language, we focus on text (i.e., a sequence of comments) in a social media session. We formally define session-based cyberbullying detection as follows:

Definition (Cyberbullying Detection in a Social Media Session). We consider a corpus of N social media sessions $\mathcal{C} = \{\int_1, \int_2, ..., \int_N\}$, in which each session consists of a sequence of comments denoted as $\{\mathbf{c}_1, ..., \mathbf{c}_C\}$. A session is labeled as either $y = 1$ denoting a bullying session or $y = 0$ denoting a non-bullying session. Let D be the dimension of extracted textual features (e.g., Bag of Words) $\mathbf{x}_i$ for $\mathbf{c}_i$. Session-based cyberbullying detection aims to learn a binary classifier using a sequence of textual data to identify if a social media session is a cyberbullying instance:

$$\mathcal{F} : \{\mathbf{x}_1, ..., \mathbf{x}_C\} \in \mathbb{R}^D \rightarrow \{0, 1\}. \quad (1)$$

4 Proposed Method

An unbiased model for cyberbullying detection makes decisions based on the semantics in a social media session instead of sensitive triggers potentially related to cyberbullying, such as "gay," "black," or "fat." In the presence of unintended bias, a model may present high performance for sessions with these sensitive triggers without knowing their semantics (Dixon et al., 2018). In this section, we first discuss how to define and assess bias in the context of session-based cyberbullying detection.

We then present the details of our bias mitigation strategy.

4.1 Assessing Bias

Bias in a text classification model can be assessed by the *False Negative Equality Difference* (FNED) and *False Positive Equality Difference* (FPED) metrics, as used in previous studies such as (Zhang et al., 2020; Gencoglu, 2020; Huang et al., 2020). They are a relaxation of *Equalized Odds* (Borkan et al., 2019) and defined as

$$\text{FNED} = \sum_z |\text{FNR}_z - \text{FNR}_{overall}|, \quad (2)$$

$$\text{FPED} = \sum_z |\text{FPR}_z - \text{FPR}_{overall}|, \quad (3)$$

where z denotes cyberbullying-sensitive triggers, such as "gay," "black," and "Mexican." The complete list of sensitive triggers can be found in Appendix A. $\text{FNR}_{overall}$ and $\text{FPR}_{overall}$ denote the False Negative Rate and False Positive Rate over the entire training dataset. Similarly, FNR_z and FPR_z are calculated over the subset of the data containing the sensitive triggers. An unbiased cyberbullying model meets the following condition:

$$P(\hat{Y}|Z) = P(\hat{Y}), \quad (4)$$

where $\hat{Y}$ stands for the predicted label. By Equation 4, we imply that $\hat{Y}$ is independent of the cyberbullying-sensitive triggers Z –that is, a debiased model performs similarly for sessions with and without Z.

Note that the widely-used non-discrimination evaluation sets – Identity Phrase Templates Test Sets (IPTTS) (Dixon et al., 2018) – are not applicable to our task. IPTTS are generated by predefined templates with slots for specific terms, e.g., "I am a boy" and "I am a girl." They only include examples for single text, whereas a social media session includes a sequence of comments. As we will show in subsection 5.1, the average number of comments in the Instagram dataset is 72, which can pose great challenges for generating synthetic social media sessions and the labeling process.

4.2 Mitigating Bias

Essentially, a debiasing session-based cyberbullying detection is a sequential decision-making process where decisions are updated periodically to assure high performance. In this debiasing framework, comments arrive and are observed sequentially. At each timestep, two decisions are made

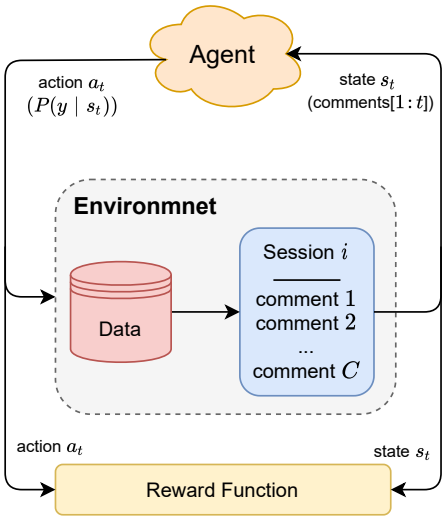


Figure 2: Overview of the proposed model. The agent (a classifier) interacts with the environment to gather experiences M_t that are used to update the agent.

based on the feedback from past decisions: (1) predicting whether a session is bullying and (2) gauging the performance differences between sessions with and without sensitive triggers. Our debiasing strategy is built on the recent results of RL (Shi et al., 2018; Zou et al., 2019; Mosallanezhad et al., 2019), particularly, the sequential Markov Decision Process (MDP). In this approach, an agent A interacts with an environment over discrete time steps t: the agent selects action a_t in response to state s_t. a_t causes the environment to change its state from s_t to s_{t+1} and returns a reward r_{t+1}. Therefore, each interaction between the agent and the environment creates an experience tuple $M_t = (s_t, a_t, s_{t+1}, r_{t+1})$. The experience tuple is used to train the agent A through different interactions with the environment. The agent's goal is to excel at a specific task, such as generating text (Shi et al., 2018) or summarizing text (Keneshloo et al., 2019).

In this work, we leverage techniques in RL to alleviate the unintended bias when classifying social media sessions into *bullying* or *non-bullying* based on user comments. In particular, we consider a standard classifier $\mathcal{F}$ (e.g., HAN) as an RL agent and a sequence of comments observed at time $\{1, 2, ..., t\}$ as state s_t. The agent selects an action $a_t \in \{\text{non-bullying}, \text{bullying}\}$ according to a policy function $\pi(s_t)$. $\pi(s_t)$ indicates the probability distribution of actions a in response to state s_t, whereas $\pi(s_t, a_t)$ shows the probability of choosing action a_t in response to state s_t. The action can

be interpreted as the predicted label $\hat{y}$ using the input comments. The reward r_{t+1} is then calculated for the state-action set (s_t, a_t) and the cumulative discounted sum of rewards G_t is used to optimize the policy function $\pi(s_t)$.

Below, we provide details of the (1) environment, (2) states, (3) actions, and (4) the reward function for the proposed debiasing approach.

- *Environment* is a session comments loader. At each episode, the environment chooses a single session and returns its first t comments as state s_t. As such, states are independent from the agent's actions, as they do not affect the next state. When it reaches the maximum number of comments of the selected session C, the process is terminated.
- State s_t is a sequence of comments in a social media session posted by various users from time 1 through time t.
- Action a_t determines a session to be *bullying* or not, given the input comments or state s_t:

$$a_t \in \{\textit{bullying}, \textit{non-bullying}\}. \tag{5}$$

- Reward function R is used to optimize the policy function $\pi(s_t, a_t)$. It is defined based on how successfully the agent predicts the label for the input state s_t and how much bias the classifier currently has. We define the bias of a classifier as the harmonic mean of FPED and FNED characterized by the sensitive triggers in cyberbullying. In a debiased classifier, we expect both FPED and FNED to be close to zero. We define the reward function R as

$$R = -l_{\mathcal{F}} - \beta \times \frac{2 \times \text{FPED} \times \text{FNED}}{\text{FPED} + \text{FNED}}, \tag{6}$$

where l indicates the prediction error of the classifier and β balances between prediction and the debiasing effect of $\mathcal{F}$. The reward function is calculated based on all sessions in the environment, evaluating the performance and bias of the classifier.

4.3 Optimization Algorithm

Given the environment, state, actions, and the reward function, we aim to learn the optimal action selection strategy $\pi(s_t, a_t)$. At each timestep t, the agent classifies a session with t comments and the reward r_{t+1} is calculated using Equation 6, according to the agent's action a_t and state s_t. The goal of the agent is to maximize its reward according to Equation 6. We use the policy gradient algorithm – REINFORCE (Sutton et al., 1999) – to train the agent. As such, the agent has similar properties to a classifier and the classifier's output distribution can be mapped to the agent's policy function $\pi(s_t, a_t)$. We use the following function to update the agent:

$$\Delta\theta = lr \nabla_\theta \mathcal{L}(\theta), \tag{7}$$

where lr denotes the learning rate, θ is the parameter w.r.t. the policy function $\pi_\theta(s_t, a_t)$, and $\mathcal{L}(\theta)$ indicates the policy loss:

$$\mathcal{L}(\theta) = \log(\pi_\theta(s_t, a_t) \cdot G_t), \tag{8}$$

where $G_t = \sum_{i=1}^{t} \gamma^i r_{i+1}$ is the cumulative sum of rewards with discount rate γ. The pseudo-code for the optimization algorithm can be seen in Algorithm 1.

Algorithm 1 The Optimization Algorithm

Require: The dataset $\{\mathbf{x}, z, y\}$, initialized $\pi_\theta(s_0, a_0)$, discount rate γ, balancing weight β, learning rate lr, number of episode E.

while Episode $e < E$ **do**
  Initialize s_t, M
  for $t \in \{0, 1, ..., C\}$ **do**
    A selects action a_t according to distribution $\pi(s_t)$
    $M \leftarrow M + (s_t, a_t, r_{t+1}, s_{t+1})$
    $s_t \leftarrow s_{t+1}$
    for each timestep t, reward in M_t **do**
      $G_t \leftarrow \sum_{i=1}^{t} \gamma^i r_{i+1}$
    end for
    Calculate mean policy loss for all timesteps according to Equation 8.
    Update the policy according to Equation 7.
  end for
end while

5 Evaluation

In this section, we conduct both quantitative and qualitative evaluations to examine the efficacy of our debiasing strategy.[1] In particular, we show that our method can effectively mitigate the impacts of unintended data biases without impairing the model's prediction performance by answering:

[1]The source code is publicly available at https://github.com/GitHubLuCheng/MitigateBiasSessionCB

Table 1: Statistics of the Instagram and Vine datasets.

Datasets	# Sessions	# Bullying	# Non-bullying	# Comments
Instagram	2,218	678	1,540	155,260
Vine	970	304	666	78,250

(1) Can we mitigate the unintended bias of machine learning models for detecting cyberbullying sessions by leveraging techniques in RL?
(2) If so, will this debiasing strategy impair the cyberbullying detection performance? and
(3) If 'no' to (2), what is the source of gain?

5.1 Data.

Two benchmark datasets for cyberbullying detection – *Instagram* (Hosseinmardi et al., 2015) and *Vine* (Rafiq et al., 2015) – are used for empirical evaluation. The number of sessions in *Instagram* and *Vine* is 2,218 and 970, respectively. Both datasets were crawled using a snowball sampling method and manually annotated via the crowdsourcing platform CrowdFlower.[2] Sessions containing less than 15 comments were removed to ensure data annotation quality. Annotators were asked to examine the image/video, associated caption, and all of the comments in a session before making the final decisions.

Instagram: Instagram[3] is a social networking site ranked as one of the top five networks with the highest percentage of users reporting experiences of cyberbullying (the Label Anti Bullying Charity, 2013). Each social media session consists of image content, a corresponding caption, and a sequence of comments in temporal order. In total, this dataset is composed of 2,218 sessions, with an average number of 72 comments in each session.

Vine: Vine[4] was a mobile application that allowed users to upload and comment on six-second looping videos. Each social media session consists of video content, the corresponding caption, and a sequence of comments in temporal order. This dataset contains 970 sessions and each session contains, on average, 81 comments.

5.2 Experimental Setup

For social media sessions, standard fairness methods, such as identity swapping and data supplementation, are not applicable. We compare our approach with commonly used machine learning models for classification with sequential text data, including HAN, Convolutional Neural Network (CNN), and Gated Recurrent Unit (GRU), as well as a recent model proposed for session-based cyberbullying detection – HANCD (Cheng et al., 2019a). HANCD leverages multi-task learning to jointly model the hierarchical structure of a social media session and the temporal dynamics of its sequence of comments to improve the performance of cyberbullying detection.

We also include the state-of-the-art model *Constrained* (Gencoglu, 2020) that imposes two fairness constraints on cyberbullying detection to mitigate biases. In our implementation, we use the HANCD classifier as the cyberbullying model in *Constrained* for a fair comparison. The parameter w.r.t. the fairness constraints is set to 0.005, as suggested. Both HAN and HANCD use GRU to extract the context of the input data. We use 1-layer GRUs with a hidden size of 100 and 200 neurons for word and comment attention networks, respectively. As our approach is model-agnostic, for each standard machine learning model, there is a corresponding debiased counterpart.

For the proposed method, $l_{\mathcal{F}}$ in the reward function (Equation 6) is computed as the cross entropy loss between the true label y and the predicted probability p:

$$l_{\mathcal{F}} = -\frac{1}{2}\sum_{i=1}^{2} y_i \log(p_i) + (1-y_i)\log(1-p_i). \quad (9)$$

In Algorithm 1, the classifier $\mathcal{F}$ is pre-trained for 5 iterations using loss function $l_{\mathcal{F}}$, learning rate $3e-3$, and the Adam optimizer (Kingma and Ba, 2014). $\mathcal{F}$ is then placed in the RL setting discussed in subsection 4.2. We apply the REINFORCE method with $E = 500$ episodes, learning rate $1e-5$, $\beta = 1.0$, and $\gamma = 0.5$ using the Adam optimizer to further update the classifier.

Evaluations focus on both the prediction accuracy and the debiasing effect of a model. For prediction performance, we adopt standard metrics for binary classification, including Precision, Recall, F1, and AUC scores. Following (Zhang et al., 2020; Gencoglu, 2020), we use FPED, FNED, and total bias (FPED+FNED) to evaluate how biased a model is w.r.t. sessions with and without sensitive triggers. Lower scores indicate less bias. For all models, pre-trained GloVe word embeddings (Pennington et al., 2014) and 10-fold cross validation with 80/20 split are used for fair comparison.

[2] https://www.figure-eight.com/
[3] https://www.instagram.com/
[4] https://vine.co/. It was shut down in 2017.

Table 2: Bias comparisons of different models. Lower FPED and FNED indicate lower bias in the model.

Model	*Instagram*			*Vine*		
	FPED	FNED	Total	FPED	FNED	Total
Constrained	0.061	**0.073**	0.134	**0.018**	0.065	0.083
HAN	0.134	0.180	0.314	0.070	0.031	0.101
CNN	0.243	0.180	0.424	0.115	0.098	0.214
GRU	0.211	0.169	0.380	0.092	0.076	0.168
HANCD	0.125	0.167	0.293	0.063	0.042	0.105
De-HAN	0.057	0.078	0.135	0.020	**0.030**	**0.050**
De-CNN	0.198	0.178	0.376	0.099	0.081	0.180
De-GRU	0.116	0.156	0.272	0.072	0.035	0.107
De-HANCD	**0.050**	0.081	**0.131**	0.019	0.041	0.060

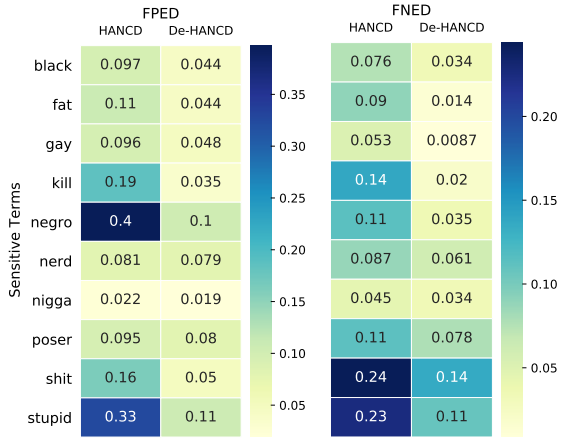


Figure 3: Comparison for fairness measures of *HANCD* and *De-HANCD* on the Instagram dataset, in which FNED $= \sum_z |\text{FNR}_z - \text{FNR}_{overall}|$ and FPED $= \sum_z |\text{FPR}_z - \text{FPR}_{overall}|$. Values closer to 0 indicate better equity. Best viewed in color.

Furthermore, we perform McNemar's test to examine whether a statistically significant difference between baseline and debiased models exists in terms of cyberbullying classification accuracy and equity. The best results are highlighted in bold font.

5.3 Can we mitigate unintended bias?

In this section, we show experimental results to answer the first question: "Can the proposed framework mitigate unintended bias?" As expected, the proposed RL framework can effectively mitigate the impact of the unintended bias embedded in the datasets for cyberbullying detection. We report results for both *Instagram* and *Vine* in Table 2. "De-" denotes a debiased model, e.g., De-HAN is a HAN debiased by the proposed RL framework. "Total" stands for the total bias (FPED+FNED). All McNemar's tests resulted in statistical significance with p-values < 0.05.

We observe the following: (1) Compared to the standard classifiers, the debiased counterparts significantly improve FNED and FPED scores, indicating that our proposed debiasing strategy can mitigate the unintended bias in data used for predicting cyberbullying sessions, regardless of the dataset or machine learning model. For example, when tested on *Instagram* with the HAN model, our debiasing method can decrease FPED, FNED, and total bias by 95.7%, 56.7%, and 57.0%, respectively. For *Vine*, the improvement with HAN is 71.4%, 3.3%, and 50.5%, respectively. (2) Total biases of standard classifiers come from both the FPRs and FNRs for the *Instagram* experiments, while the main contributor of biases is the FPRs for the *Vine* experiments. Our approach mitigates total bias in both scenarios. (3) Our debiasing strategy based on RL techniques is also more effective than the fairness constraints proposed in (Gencoglu, 2020), as indicated by the decreased total biases for both *Instagram* and *Vine*. By comparing HANCD, Constrained, and De-HANCD, we see that Constrained decreases FPED by sacrificing FNED, while De-HANCD can decrease both.

In addition to the quantitative results, we provide qualitative analyses by visualizing FPED and FNED of both the standard and debiased HANCD models. In an experiment with *Instagram* for sessions containing ten sensitive triggers, as illustrated in Figure 3, we can observe that compared to De-HANCD, HANCD is more biased toward some sensitive triggers, such as "fat" and "stupid." Demographic-identity related bias is also detected in HANCD. For example, sessions containing identity terms including "ne**o," "gay," and "ni**a" are more likely to be falsely identified as "bullying," as indicated by FPED. By contrast, De-HANCD mitigates various types of unintended biases and has more consistent performance across all of the sensitive triggers.

5.4 Is there a trade-off between accuracy and bias mitigation?

A dilemma often faced by researchers studying bias and fairness in machine learning is the trade-off between fairness and efficiency (Bertsimas et al., 2012). Under this trade-off theory, forcing cyberbullying classifiers to follow the proposed debiasing strategy would invariably decrease the accuracy. This section shows that, somewhat counterintuitively, our approach can outperform biased models w.r.t. overall cyberbullying detection ac-

Table 3: Performance comparisons of different models on the *Instagram* dataset. Higher AUC, precision (PREC), recall (REC), and F1 scores indicate better performance. p-value < 0.05 for all McNemar's tests.

Model	AUC	PREC	REC	F1
Constrained	0.9042	0.8099	0.9101	0.8570
HAN	0.9032	**0.8434**	0.8879	0.8651
CNN	0.7120	0.6872	0.7380	0.7117
GRU	0.7352	0.7003	0.7265	0.7132
HANCD	0.9087	0.8218	**0.9206**	0.8684
De-HAN	0.9057	0.8292	0.9115	0.8684
De-CNN	0.7068	0.7011	0.6940	0.6975
De-GRU	0.7565	0.7355	0.7498	0.7426
De-HANCD	**0.9089**	0.8357	0.9102	**0.8714**

Table 4: Performance comparisons of different models on the *Vine* dataset. Higher AUC, precision (PREC), recall (REC), and F1 scores indicate better performance. p-value < 0.05 for all McNemar's tests.

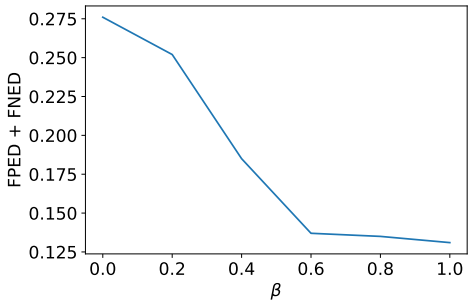
Model	AUC	PREC	REC	F1
Constrained	0.8077	0.7644	0.8113	0.7871
HAN	0.8527	0.5203	0.8127	0.6344
CNN	0.6245	0.4603	0.7119	0.5591
GRU	0.6759	0.4801	0.7651	0.5900
HANCD	0.9223	0.6841	0.8590	0.7616
De-HAN	**0.9365**	0.8924	0.9079	0.9001
De-CNN	0.6288	0.4306	0.6532	0.5190
De-GRU	0.6890	0.5237	0.7568	0.6190
De-HANCD	0.9350	**0.9015**	**0.9156**	**0.9085**

curacy, while also decreasing unintended biases in the data.

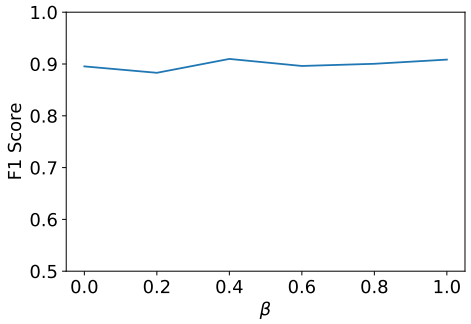
Results are presented in Tables 3-4. We see that the proposed debiasing strategy can both alleviate the bias and retain high prediction accuracy. For instance, for *Instagram*, our approach achieves the highest AUC and F1 score of all evaluated models. For *Vine*, the improvement of De-HAN over HAN is 9.8% and 41.9% for AUC and F1 score, respectively. The improvement over Constrained is 15.8% and 15.4%, respectively. Biased models present much lower Precision than Recall for *Vine*. This result is in line with the findings in Table 2, where we observe that the larger bias component is associated with FPRs in *Vine*. This indicates that when the sample size is small, these models overfit to sensitive triggers for detecting bullying instances. The debiasing strategy effectively reduces models' reliance on those terms and utilizes contextual information for prediction.

5.5 What is the source of gain?

What is the ingredient that enables our approach to achieve both the lowest bias and highest accuracy? This non-compromising approach may be attributed to the proposed RL framework that effectively captures contextual information. In this section, we examine the impact of parameter β in Equation 6 by varying $\beta \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. We show performance w.r.t. bias mitigation (total bias) and cyberbullying detection (F1 score) in Figure 4.



(a) Performance w.r.t. bias mitigation.



(b) Performance w.r.t. cyberbullying detection.

Figure 4: Total bias and F1 score of *De-HANCD* using different values of β in Equation 6. The total bias is calculated as the sum of FPED and FNED.

The results clearly show the efficacy of the proposed RL framework for bias mitigation. In particular, as we increase β, the RL agent puts more effort toward alleviating biases by minimizing both FPED and FNED simultaneously. Moreover, by interacting with the environment, the RL agent also leverages contextual information in order to minimize the prediction error and receive a larger reward. As a result, the RL agent largely reduces biases while improving the prediction accuracy, as shown by the slight increase in detection performance of the classifier in Figure 4b.

6 Conclusion and Future Work

In this work, we examined unintended biases in datasets for session-based cyberbullying detection. In contrast to conventional data for bias mitigation in text classification, social media sessions consist of a sequence of comments with rich contextual

information. To alleviate these unintended biases, we propose an effective debiasing strategy by leveraging techniques in RL. Our approach is context-aware, model-agnostic, and does not require additional resources or annotations aside from a predefined set of potentially sensitive triggers related to cyberbullying. Empirical evaluations demonstrated that our approach can mitigate unintended bias in the data without impairing a model's prediction accuracy.

Other types of decisions in sequential decision-making processes can impact the underlying user population, thereby influencing future comments generated by users. Future research can be directed towards studying the long-term impact of the debiasing strategy, as well as investigating different types of biases in session-based cyberbullying detection, such as gender bias, racial bias, and language bias. Our approach can also benefit from integrating previous studies that use data augmentation or swapping methods to counteract bias. Due to the challenges of data collection and labeling, validating our approach on datasets across different social media platforms is also an important avenue for future work.

Ethics Statement

This work seeks to advance collaborative research efforts aimed at mitigating bias in session-based cyberbullying detection, a topic that has yet to be studied extensively. Here, we provide preliminary solutions, but more work is needed to elucidate ways to build debiased and effective models. While all data used in this study are publicly available, we are committed to securing the privacy of the individuals in our datasets. To this end, we automatically replaced user names with ordered indices in our analysis. The insulting or offensive terms and the figures used in this paper are for illustrative purposes only and do not represent the views or ethical attitudes of the authors.

Acknowledgements

This material is based upon work supported by the National Science Foundation (NSF) Grants 1719722 and 2036127.

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*, pages 49–59.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Nozza Debora, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, Manuela Sanguinetti, et al. 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.

Dimitris Bertsimas, Vivek F Farias, and Nikolaos Trichakis. 2012. On the efficiency-fairness trade-off. *Management Science*, 58(12):2234–2250.

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.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Lu Cheng, Ruocheng Guo, Yasin Silva, Deborah Hall, and Huan Liu. 2019a. Hierarchical attention networks for cyberbullying detection on the instagram social network. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 235–243. SIAM.

Lu Cheng, Ruocheng Guo, Yasin N Silva, Deborah Hall, and Huan Liu. 2021. Modeling temporal patterns of cyberbullying detection with hierarchical attention networks. *ACM/IMS Transactions on Data Science*, 2(2):1–23.

Lu Cheng, Jundong Li, Yasin N Silva, Deborah L Hall, and Huan Liu. 2019b. Xbully: Cyberbullying detection within a multi-modal context. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 339–347.

Lu Cheng, Yasin N Silva, Deborah Hall, and Huan Liu. 2020. Session-based cyberbullying detection: Problems and challenges. *IEEE Internet Computing*.

Harsh Dani, Jundong Li, and Huan Liu. 2017. Sentiment informed cyberbullying detection in social media. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 52–67. Springer.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. *arXiv preprint arXiv:1905.12516*.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 5.

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.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Oguzhan Gencoglu. 2020. Cyberbullying detection with fairness constraints. *IEEE Internet Computing*.

Sameer Hinduja and Justin W Patchin. 2020. Cyberbullying fact sheet: Identification, prevention, and response. *Cyberbullying Research Center. Retrieved January*, 30:2011.

Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, and Shivakant Mishra. 2015. Detection of cyberbullying incidents on the instagram social network. *arXiv preprint arXiv:1503.03909*.

Qianjia Huang, Vivek Kumar Singh, and Pradeep Kumar Atrey. 2014. Cyber bullying detection using social and textual analysis. In *Proceedings of the 3rd International Workshop on Socially-Aware Multimedia*, pages 3–6.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J Paul. 2020. Multilingual twitter corpus and baselines for evaluating demographic bias in hate speech recognition. *arXiv preprint arXiv:2002.10361*.

Sophie Jentzsch, Patrick Schramowski, Constantin Rothkopf, and Kristian Kersting. 2019. Semantics derived automatically from language corpora contain human-like moral choices. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 37–44.

Yaser Keneshloo, Naren Ramakrishnan, and Chandan K Reddy. 2019. Deep transfer reinforcement learning for text summarization. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 675–683. SIAM.

Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela Wisniewski, and Munmun De Choudhury. 2021. You Don't Know How I Feel:Insider-Outsider Perspective Gaps in Cyberbullying Risk Detection. page 13.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ditch the Label Anti Bullying Charity. 2013. Ditch the label anti bullying charity: The annual cyberbullying survey 2013. https://www.ditchthelabel.org/wp-content/uploads/2016/07/cyberbullying2013.pdf. Accessed: 2020-09-18.

Michelle Seng Ah Lee, Luciano Floridi, and Jatinder Singh. 2020. From fairness metrics to key ethics indicators (keis): a context-aware approach to algorithmic ethics in an unequal society. *Available at SSRN*.

Ahmadreza Mosallanezhad, Ghazaleh Beigi, and Huan Liu. 2019. Deep reinforcement learning-based text anonymization against private-attribute inference. 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 2360–2369.

Debora Nozza, Claudia Volpetti, and Elisabetta Fersini. 2019. Unintended bias in misogyny detection. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 149–155.

Andrew Ortony, Gerald L Clore, and Mark A Foss. 1987. The referential structure of the affective lexicon. *Cognitive science*, 11(3):341–364.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. *arXiv preprint arXiv:1808.07231*.

John Pavlopoulos, Ion Androutsopoulos, Jeffrey Sorensen, and Léo Laugier. 2021. Semeval 2021 task 5: Toxic spans detection. In *15th International Workshop on Semantic Evaluation*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Rahat Ibn Rafiq, Homa Hosseinmardi, Richard Han, Qin Lv, Shivakant Mishra, and Sabrina Arredondo Mattson. 2015. Careful what you share in six seconds: Detecting cyberbullying instances in vine. In *ASONAM 2015*, pages 617–622. IEEE.

Zhan Shi, Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. 2018. Toward diverse text generation with inverse reinforcement learning. *arXiv preprint arXiv:1804.11258*.

Peter K Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. 2008. Cyberbullying: Its nature and impact in secondary school pupils. *Journal of child psychology and psychiatry*, 49(4):376–385.

Anna Squicciarini, Sarah Rajtmajer, Y Liu, and Christopher Griffin. 2015. Identification and characterization of cyberbullying dynamics in an online social network. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 280–285.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang

Wang. 2019. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*.

Richard S Sutton, David A McAllester, Satinder P Singh, Yishay Mansour, et al. 1999. Policy gradient methods for reinforcement learning with function approximation. In *NIPs*, volume 99, pages 1057–1063. Citeseer.

Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, and Amy Bellmore. 2012. Learning from bullying traces in social media. In *Proceedings of the 2012 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 656–666.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). *arXiv preprint arXiv:2006.07235*.

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. *arXiv preprint arXiv:2004.14088*.

Wei Zou, Shujian Huang, Jun Xie, Xinyu Dai, and Jiajun Chen. 2019. A reinforced generation of adversarial examples for neural machine translation. *arXiv preprint arXiv:1911.03677*.

A Sensitive Triggers for Debiasing Cyberbullying Detection

We adapted the list of cyberbullying keywords suggested in the psychology literature (Ortony et al., 1987; Squicciarini et al., 2015) to curate the list of sensitive triggers used in bias mitigation for cyberbullying detection: *nerd, gay, loser, freak, emo, whale, pig, fat, poser, whore, die, suck, slut, afraid, pussy, cunt, kill, dick, bitch, black, ni***r, ne**o, ni**a, Mexican, redneck, retard, shit, ass, stupid, ugly, slave, fuck, pathetic, homo.*