Bias Mitigation for Toxicity Detection via Sequential Decisions

Lu Cheng
School of Computing and Augmented Intelligence, Arizona State University
Tempe, AZ, USA
lcheng35@asu.edu

Ahmadreza Mosallanezhad
School of Computing and Augmented Intelligence, Arizona State University
Tempe, AZ, USA
amosalla@asu.edu

Yasin N. Silva
Department of Computer Science, Loyola University Chicago
Chicago, Illinois, USA
ysilva1@luc.edu

Deborah L. Hall
School of Social and Behavioral Sciences, Arizona State University
Glendale, AZ, USA
d.hall@asu.edu

Huan Liu
School of Computing and Augmented Intelligence, Arizona State University
Tempe, AZ, USA
huanliu@asu.edu

ABSTRACT

Increased social media use has contributed to the greater prevalence of abusive, rude, and offensive textual comments. Machine learning models have been developed to detect toxic comments online, yet these models tend to show biases against users with marginalized or minority identities (e.g., females and African Americans). Established research in debiasing toxicity classifiers often (1) takes a static or batch approach, assuming that all information is available and then making a one-time decision; and (2) uses a generic strategy to mitigate different biases (e.g., gender and racial biases) that assumes the biases are independent of one another. However, in real scenarios, the input typically arrives as a sequence of comments/words over time instead of all at once. Thus, decisions based on partial information must be made while additional input is arriving. Moreover, social bias is complex by nature. Each type of bias is defined within its unique context, which, consistent with intersectionality theory within the social sciences, might be *correlated* with the contexts of other forms of bias. In this work, we consider debiasing toxicity detection as a *sequential decision-making process* where different biases can be *interdependent*. In particular, we study debiasing toxicity detection with two aims: (1) to examine whether different biases tend to correlate with each other; and (2) to investigate how to jointly mitigate these correlated biases in an interactive manner to minimize the total amount of bias. At the core of our approach is a framework built upon theories of sequential Markov Decision Processes that seeks to maximize the prediction accuracy and minimize the bias measures tailored to individual biases. Evaluations on two benchmark datasets empirically validate the hypothesis that biases tend to be correlated and corroborate the effectiveness of the proposed sequential debiasing strategy.

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531945

CCS CONCEPTS

• **Security and privacy** → **Social aspects of security and privacy**; • **Human-centered computing** → **Collaborative and social computing**; *Collaborative and social computing design and evaluation methods*; • **Computing methodologies** → *Supervised learning*.

KEYWORDS

unintended bias, toxicity detection, sequential decision-making, social media

ACM Reference Format:
Lu Cheng, Ahmadreza Mosallanezhad, Yasin N. Silva, Deborah L. Hall, and Huan Liu. 2022. Bias Mitigation for Toxicity Detection via Sequential Decisions. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3477495.3531945

1 INTRODUCTION

Current machine learning models for toxicity detection exhibit problematic and discriminatory performance, resulting in poorer prediction [10, 33] and negatively impacting disadvantaged and minority groups [14, 21, 43]. That is, Instagram[1] sessions that include comments with swear words can be flagged as toxic even when the swear words are used inoffensively and tweets containing words related to minority groups are more likely to be identified as toxic. For example, the widely-used Perspective API[2] has been found to identify "Wussup, n*gga!" and "F*cking love this." as toxic comments with high probability, revealing swear-words-based lexical bias [43] and potential dialect-based racial bias against African American English (AAE), respectively.

Despite promising efforts to debias toxicity detection and related tasks (e.g., cyberbullying detection), most research to date (e.g., [9, 14, 42]) is based on two assumptions: (1) bias mitigation is a "static" problem where the model has access to all of the information and makes a one-time decision; and (2) different types of biases are independent of one another. Yet, comments/words in social media often come in a sequence instead of all at once. In

[1] https://www.instagram.com/
[2] https://www.perspectiveapi.com/

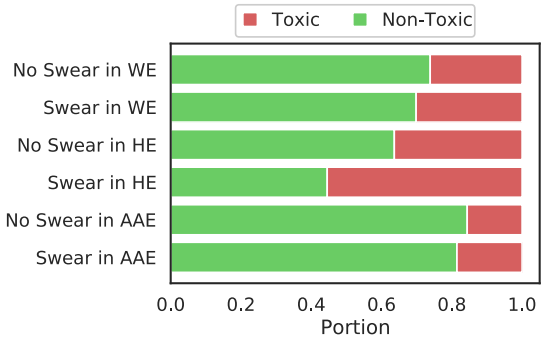

Figure 1: Percentages of toxic (red) and non-toxic (green) sessions containing different biases in the benchmark *Instagram* data [20]. "Swear in WE" denotes that there are swear words in sessions written in White-aligned English (WE).

this environment, conventional batch-processing can be impractical. Further, the relations among different biases are complex. As shown in Fig. 1, sessions containing comments written in Hispanic English (HE) or AAE with swear words contribute larger portions of toxic sessions than those without swear words in a benchmark dataset. Recent work (e.g., [21]) also showed evidence of intersectional bias within toxicity detection: AAE tweets were 3.7 times as likely and African American male tweets were 77% more likely to be labeled as toxic [21]. In the social sciences, *intersectionality* is the idea that multiple identity categories (e.g., race and gender) combine interactively in ways that contribute to greater bias than the bias associated with each category alone. Informed by these findings, we **first** hypothesize that biases *tend to be correlated* in toxicity detection.

To effectively mitigate potentially correlated biases with a sequential input, we address two challenges: (1) making sequential decisions based on partial information, e.g., comments observed so far, given that "static" debiasing may cause unnecessary delay and is less responsive when the input is changing (e.g., topic diversion); and (2) characterizing unique aspects of individual biases to reduce the overall bias. Conventional debiasing strategies provide a generic and one-size-fits-all solution. A straightforward approach is to add multiple fairness constraints w.r.t. different biases to the training process of a toxicity classifier. However, it overlooks the unique characteristics of each bias and confronts challenging optimization problems. This leads to our **second** hypothesis: with sequential input, sequential bias mitigation strategies that include bias measures tailored to individual biases can improve the debiasing performance in the presence of potentially correlated biases.

To test our hypotheses, we study the novel problem of *joint bias mitigation for toxicity detection via sequential decision-making*. The goal is to effectively detect toxicity and mitigate potentially correlated biases as comments arrive sequentially. This work proposes a sequential debiasing strategy for toxicity detection – *Joint* – built on theories of sequential Markov Decision Processes (MDP) [2] and a pairwise comparison bias measure that compares every two groups sampled from the same bias type. *Joint* is model-agnostic (i.e., any standard toxicity classifier can be the input model) and focuses on debiasing with sequential inputs.

The major contributions of this paper are the following:

- We investigate the novel and practical research question of joint bias mitigation for toxicity detection via sequential decisions;
- We propose two hypotheses that offer new theoretical and practical insights for research on bias and fairness in AI.
- We propose a sequential bias mitigation approach that takes sequential input seeking to maximize the prediction accuracy and minimize the bias measures tailored to individual biases.
- Empirical evaluation on two benchmark datasets shows that our approach can effectively reduce the total amount of bias and present competitive prediction accuracy.

2 RELATED WORK

We briefly review two lines of research closely related to our work – toxicity detection and bias mitigation in text classification.

2.1 Toxicity Detection

Toxicity detection has received considerable attention as a tool for mitigating the detrimental impact of toxicity online (see, e.g., [5, 25])–where toxicity is defined as "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion" [1]. Toxicity is also used as an umbrella term for different kinds of toxic language, including language that is 'hateful' [31, 35, 39], 'offensive' [32, 41], 'cyberbullying' [12, 24], or 'abusive' [25, 28].

An early work in hate speech detection [38] extracted n-grams and part-of-speech tags from comments as features to train a Support Vector Machine (SVM) classifier. Subsequent work [25] examined the effectiveness of various linguistic features, such as character/word n-grams and word embeddings. Experimental results found that models trained with the combination of all features achieved the best predictive performance. Other promising approaches have emerged from research on cyberbullying detection. For instance, studies that augmented textual features with rich social media information, such as social network [36] and other multi-modal information (e.g., time, location) [8], were found to improve the performance of cyberbullying detection significantly. Notably, there has been growing interest in models designed specifically for toxicity detection (e.g., [5, 15]). For example, because toxic users may continually modify their content to circumvent comment filters, a recent work [5] proposed using sentiment information to help detect toxicity, based on the hypothesis that it is harder for toxic users to hide their sentiments. Their results showed that sentiment information has a positive impact on toxicity detection.

Recent work [6, 29] has highlighted the role of semantic context in detecting toxicity. Inspired by the hierarchical attention network proposed in [40], Cheng et al. [6] used a hierarchical structure to model a social media session with a sequence of comments and attention weights to differentiate the word/comment importance. Surprisingly, others (e.g., Pavlopoulos et al. [29]) have found a lack of evidence that context improves the performance of toxicity classifiers. These inconsistent findings point to the need for in-depth analyses of context-aware toxicity detection.

2.2 Bias Mitigation in Text Classification

Computational methods can reinforce and even propagate unintended biases in text classification tasks that stem from datasets [14], contextual word embeddings [22], distributed word embeddings [4, 16], machine learning algorithms [9, 42], and human annotators [18]. In a pioneering work [4], word embeddings trained on Google News articles were found to exhibit gender stereotypes to an alarming extent. Yet, only a handful of studies [9, 17, 42] have focused on mitigating these unintended biases in text classification, broadly, and toxicity detection, specifically. For example, a recent survey [43] identified and mitigated three types of biases in toxicity detection: identity (e.g., "gay"), swear words (e.g., "f*k"), and racial biases (e.g., AAE). One approach for mitigating bias in text classification–and mitigating demographic bias, in particular–is data augmentation [14, 27, 34]. This approach seeks to reduce data bias stemming from the lower weight and/or under-representation of minority (relative to majority) groups by balancing the training data sets. Specifically, one can add external labeled data [14], swap gender-related terms [27], or assign different weights to instances from various groups [23]. The primary drawback of these data manipulation methods is their impracticality (e.g., costliness of labeling data).

Recent work by Zhang et al. [42] sought to address these limitations. The authors assumed that there are discriminative and non-discriminative data distributions and sought to reconstruct the non-discriminative data distribution from discriminative ones by instance weighting. Another approach formulates the task as a constrained optimization problem [17]. The basic idea is to impose a fairness constraint w.r.t. a single bias type during model training such that the model is enforced to converge to a more equitable solution. Critically, however, prior research takes static or batch approaches, assuming all information it needs is available. Furthermore, it overlooks the unique aspects within each bias and does not distinguish the debiasing strategies for different biases.

Our work complements prior work by: (1) studying multiple potentially correlated biases with sequential input; (2) providing the first sequential bias mitigation strategy to jointly mitigate these biases; and (3) conducting an in-depth analysis of the impact of the size of historical information on sequential debiasing performance. Our findings offer new insights for both the theoretical and practical aspects of research on bias and fairness in AI.

3 PROBLEM DEFINITION

Given a corpus C of N samples, a sample $i \in \{1, 2, ..., N\}$ can be a social media session $\mathcal{S}_i$ (e.g., an Instagram session) consisting of a sequence of C comments $\{\mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_C\}$. There is also a pre-defined set of K sensitive attributes $\mathcal{P} = \{\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_K\}$. For instance, $K = 2$ and $\mathcal{P} = \{gender, race\}$ when the considered sensitive attributes are gender and race. A sample i can also be a comment (e.g., a tweet) comprised of a sequence of W words $\{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_W\}$. Moreover, every sample (a session or a comment) is labeled as "non-toxic" ($Y_i = 0$) or "toxic" ($Y_i = 1$). Let D be the number of dimensions of the extracted features $\mathbf{x}_i$ for every comment $\mathbf{c}_i$ in a session or every word $\mathbf{w}_i$ in a comment. Our proposed sequential debiasing process for toxicity detection aims to learn a binary classifier that **jointly mitigates** the overall bias w.r.t. the set of sensitive attributes $\mathcal{P}$ and **accurately identifies** if a session or comment is toxic or not, with **sequential input**:

$$\mathcal{F} : \mathcal{P} \cup \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_t, ..., \mathbf{x}_L\} \in \mathbb{R}^D \rightarrow Y \in \{0, 1\}, \quad (1)$$

where L denotes the number of comments in a session or words in a comment. To simplify the presentation, we will use social media sessions in the Method section for illustration. Experimental results will be given for both data consisting of sessions and comments.

4 METHOD

Existing research in debiasing toxicity classifiers assumes that the model observes all the comments and then makes a one-time decision regarding the prediction and bias mitigation. However, social media users interact in a sequential manner. Conventional debiasing approaches, therefore, are less responsive when the conversations between users are changing. A desired debiasing strategy should be able to process sequentially revealed comments and make dependent decisions. In addition, prior research studied different biases either individually (i.e., debiasing one type of bias at a time) or independently (i.e., debiasing multiple biases that are independent from one another). Nevertheless, bias is complex by nature and different biases might be correlated, as we will show in Sec. 5. Therefore, it is important for the sequential debiasing strategy to identify and capture the unique aspects of various biases. In this section, we first discuss how to measure bias in the presence of multiple types of biases. Then, we detail the proposed joint bias mitigation approaches for toxicity detection via sequential decisions.

4.1 Measuring Bias

Measuring bias is key for addressing unfairness in NLP and machine learning models. This section presents two categories of bias metrics that quantify the differences in a classifier's behaviour across a range of groups within the same identity, e.g., {female, male, other} for gender. They are the Background Comparison Metric (BCM) and the Pairwise Comparison Metric (PCM) [11].

4.1.1 Background Comparison Metric. The core idea of BCM is to compare a measure m (e.g., False Positive/Negative Rate) of a group over the sensitive attribute $\mathbf{p}$ with the group's background score using the same measure m. The background score is defined based on the task at hand and the scientific questions being asked. In this study, it is defined as measure m over the overall evaluation set. We use the following common bias metrics in debiasing text classification as measure m in the toxicity classifier: False Negative Equality Difference (FNED) and False Positive Equality Difference (FPED) [14]. FNED and FPED are defined based on the False Positive Rate (FPR) and False Negative Rate (FNR). Formally, we define the BCM-based fairness metrics, FPED_{BCM} and FNED_{BCM}, as follows:

$$\text{FNED}_{BCM} = \sum_{z \in \mathbf{p}} |\text{FNR}_z - \text{FNR}_{overall}|, \quad (2)$$

$$\text{FPED}_{BCM} = \sum_{z \in \mathbf{p}} |\text{FPR}_z - \text{FPR}_{overall}|. \quad (3)$$

where z denotes the values that a sensitive attribute $\mathbf{p} \in \mathcal{P}$ can be assigned to. For example, in case of $\mathbf{p} = \{male, female, other\}$, the FNR_z and FPR_z are calculated for every group $z \in \mathbf{p}$. They are then compared to $\text{FNR}_{overall}$ and $\text{FPR}_{overall}$ – which are calculated

on the entire population, including all of the considered sensitive attributes.

4.1.2 Pairwise Comparison Metric. BCM allows us to investigate how the performance of a toxicity classifier for particular groups differs from the model's general performance. When applied to settings with multiple biases, BCM can be less effective, as it tends to cancel out the differences of these distinctive biases. In addition, when a toxicity classifier presents low performance, the BCM-based metrics may underestimate the bias [11]. Here, we present PCM that quantifies how distant, on average, the performance for two randomly selected groups z_1 and z_2 within the same attribute $\mathbf{p}$ is. This metric examines whether and to what extent the groups differ from one another. For example, given the sensitive attribute *Race* with three groups {White-American, African-American, Asian}, we consider performance differences for *White-American* vs *African-American*, *White-American* vs *Asian*, *African-American* vs *Asian*. We formally define the PCM-based metrics as follows:

$$\text{FNED}_{PCM} = \sum_{z_1,z_2 \in \binom{\mathbf{p}}{2}} |\text{FNR}_{z_1} - \text{FNR}_{z_2}|, \tag{4}$$

$$\text{FPED}_{PCM} = \sum_{z_1,z_2 \in \binom{\mathbf{p}}{2}} |\text{FPR}_{z_1} - \text{FPR}_{z_2}|. \tag{5}$$

In both Eq. 4 and Eq. 5, we measure the difference between every possible pair of groups in $\mathbf{p} \in \mathcal{P}$. This forces the algorithm to focus on the particular aspects of this sensitive attribute, which otherwise will be averaged out in the overall population.

4.2 Sequential Bias Mitigation

When comments come in a sequence, a toxicity classifier needs to make decisions based on incomplete information, i.e., comments observed so far. The current decision will, in turn, influence both future prediction results and debiasing strategies. In addition, in the presence of multiple biases, debiasing a toxicity classifier can be more challenging due to the need to capture the unique characteristics of each bias and potential correlations among biases. To tackle these challenges, in this section, we present a sequential bias mitigation approach that leverages a reinforcement learning (RL) framework that seeks to maximize prediction accuracy and minimize bias measures tailored to individual biases at each timestep.

4.2.1 Debiasing via Sequential Decision Making. As comments arrive in a sequence, a debiased toxicity classifier needs to respond in a timely manner based only on partial information. This process might also involve the trade-off between debiasing and prediction accuracy. Our proposed solution is an RL framework built upon theories in sequential MDP, which allows learning to trade off competing objectives in a principled way [19]. It considers two tasks at each state of the decision-making process: (1) predicting whether the session is toxic and (2) minimizing the total amount of biases. In a typical RL framework, an agent A interacts with the environment over time. At timestep $t \in \{1, 2, ..., T\}$, A chooses an action a_t in response to the current state s_t, which causes the environment to change its state and returns the reward value r_{t+1}. Formally, we represent every interaction as an experience tuple $M_t = (s_t, a_t, s_{t+1}, r_{t+1})$ used to train A.

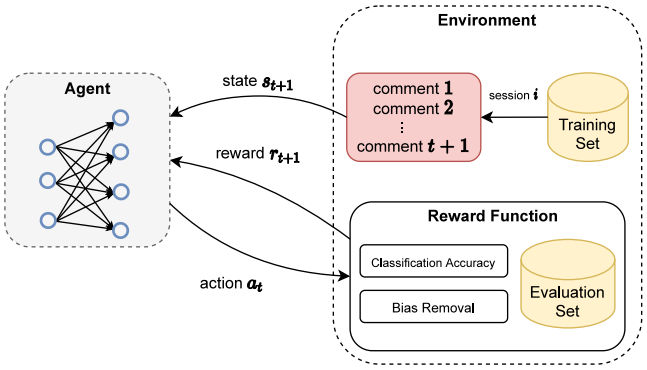

Figure 2: Proposed sequential bias mitigation approach for toxicity detection – *Joint*. The agent is a biased toxicity classifier that takes an action a_t (i.e., predicting the label) based on the current state s_t (i.e., the comments observed so far). By maximizing the reward value returned by the reward function – consisting of the bias measures and prediction error – the biased classifier is forced to improve the prediction performance and reduce the biases on the selected session from the evaluation set, which is a subset of the training set.

In sequential bias mitigation for toxicity detection, the environment includes all of the training sessions and A is a biased toxicity classifier $\mathcal{F}$. State s_t is a sequence of t comments A has observed so far. A selects an action $a \in \{\text{toxic}, \text{non-toxic}\}$ based on an action-selection policy $\pi(s_t)$, which outputs the probability distribution over actions based on s_t. $\pi(s_t, a_t)$ represents the probability of choosing action a_t when observing t comments, i.e., s_t. After selecting a_t, the environment returns a reward value r_{t+1} based on the state-action set (s_t, a_t). The reward values defined by the toxicity prediction error and bias metrics are then used to calculate the cumulative discounted reward G_t (i.e., the sum of all rewards received so far) and optimize the policy $\pi(s_t)$. At each state s_t, the RL framework maximizes the expected cumulative reward until t to force the agent to improve accuracy and mitigate bias. Essentially, the agent is making dependent decisions to adjust to the sequential input. Fig. 2 depicts an overview of the RL framework.

4.2.2 Reward Function. As the key element in the RL framework, the reward function is designed to assess the performance of the classifier $\mathcal{F}$ and jointly mitigate various types of biases over time. BCM-based bias metrics seek to compare the performance measure m of a specific group z (e.g., FNR_z) with that over the entire population (e.g., $\text{FNR}_{overall}$), resulting in a solution that might cancel out the differences of biases. However, as biases are complex (e.g., correlated) and often defined within different contexts, it is important to distinguish the unique aspects of various biases. Therefore, we hypothesize that PCM is more appropriate for reducing multiple biases because it compares m of every pair of groups belonging to the same sensitive attribute $\mathbf{p}$ instead of across all attributes in $\mathcal{P}$. For example, when $\mathcal{P} = \{race, gender\}$, PCM requires that the FNR of all subgroups in gender be similar while BCM seeks for similar FNR across all groups in both gender and race. With PCM, we only consider sessions related to a certain bias type instead of all of the

Algorithm 1 The Optimization Algorithm for *Joint*

Input: The dataset $\{\mathcal{D}, \mathcal{P}\}$ with labels $y \in Y$, discount rate γ, bias importance values $\alpha_i \in \Gamma$, learning rate lr, number of episodes E, terminal time T.
Output: Debiased classifier/agent (A)

```
Initialize memory M
Initialize agent A with parameters θ and π_θ
while Episode e < E do
    Initialize s_0 by selecting a random session
    for t ∈ {0, 1, ..., T} do
        A selects an action a_t according to π(s_t)
        M ← M + (s_t, a_t, r_{t+1}, s_{t+1})
        s_t ← s_{t+1}
        for each timestep t, reward in M_t do
            G_t ← Σ_{i=1}^{t} γ^i r_{i+1}
        end for
        Calculate the policy loss using
        L(θ) = log(π_θ(s_t, a_t) · G_t)
        Update the agent A using Δθ = lr ∇_θ L(θ)
    end for
end while
```

sessions. We thus propose to use PCM-based metrics to capture the unique information of $\mathbf{p}$.

Formally, we describe the RL framework by defining the *environment*, the *state*, the *action*, and the *reward function*. The environment contains the training dataset $\mathcal{D}$ in which every session includes a sequence of comments. At t, the environment randomly selects a session and passes the first t comments of that session to the agent A, which is a toxicity classifier $\mathcal{F}$ that outputs a decision probability $\hat{q}_t$. We convert $\hat{q}_t$ into an action a_t using the following criterion:

$$a_t = \begin{cases} \text{toxic} & \hat{q}_t \geq 0.5 \\ \text{non-toxic} & \hat{q}_t < 0.5, \end{cases} \tag{6}$$

Finally, we define the reward function using PCM-based bias metrics to jointly evaluate various types of biases as follows:

$$r^t_{PCM} = -l^t_{\mathcal{F}} - \sum_{S_{\mathbf{p}_i} \in S} \alpha_i \cdot \Big(\frac{1}{|S_{\mathbf{p}_i}|} \sum_{z_1, z_2 \in \binom{\mathbf{p}_i}{2}} |\text{FPR}^t_{z_1} - \text{FPR}^t_{z_2}| + |\text{FNR}^t_{z_1} - \text{FNR}^t_{z_2}| \Big), \tag{7}$$

where $l^t_{\mathcal{F}}$ denotes the binary prediction loss (e.g., log loss) of the toxicity classifier $\mathcal{F}$, α_i represents the importance value of bias related to the sensitive attribute $\mathbf{p}_i \in \mathcal{P}$, and $S_{\mathbf{p}_i}$ denotes the sessions with sensitive attribute $\mathbf{p}_i$.

Similarly, the reward function using BCM-based bias metrics can be defined as follows:

$$r^t_{BCM} = -l^t_{\mathcal{F}} - \sum_{S_{\mathbf{p}_i} \in S} \alpha_i \cdot \Big(\sum_{z \in \mathbf{p}_i} |\text{FPR}^t_z - \text{FPR}^t_{overall}| + |\text{FNR}^t_z - \text{FNR}^t_{overall}| \Big). \tag{8}$$

4.2.3 Optimization Algorithm. We aim to learn an optimized action-selection policy $\pi(s)$ that maximizes the cumulative rewards based on Eq. 7-8. We consider the agent as a neural network with weights θ (e.g., a trained recurrent neural network). We use an optimization algorithm similar to [9] to force a biased toxicity detection model to converge to a more equitable solution. We denote the two sequential debiasing models as $Joint_B$ and $Joint_P$, respectively. Algorithm 1 shows the high-level training process of the *Joint* model[3]. During each episode, the environment selects a random session and returns the first t comments in every step. Considering the observed comments as state s_t, the agent A selects an action according to $\pi(s_t)$. To select the action according to the action-selection policy, we use the multinomial distribution to sample from the action probabilities $\pi(s_t, a_t)$. The performed action results in reward r_{t+1} and state s_{t+1}. We use experiences M to calculate the cumulative reward G_t. Finally, we update the agent's parameters using the following loss function and its gradient:

$$\mathcal{L}(\theta) = \log(\pi_\theta(s_t, a_t) \cdot G_t), \tag{9}$$

$$\Delta\theta = lr \nabla_\theta \mathcal{L}(\theta). \tag{10}$$

5 EXPERIMENTS

To test the proposed hypotheses for sequential bias mitigation in toxicity detection, we run experiments on two benchmark datasets to answer the following research questions:

RQ. 1. With multiple types of bias present in toxicity detection, will different biases *tend to be correlated* with each other?

RQ. 2. If 'Yes' to **RQ. 1**, will a sequential and joint bias mitigation strategy tailored to individual biases outperform conventional static and generic debiasing approaches?

RQ. 3. How do the size of the historical information and parameter α influence the performance of the sequential debiasing strategy?

5.1 Data

We use two publicly available datasets collected from two platforms: Jigsaw and Instagram. They differ on data format, studied bias types, sample size, and overall proportion of toxic samples. Particularly, the *Jigsaw* dataset consists of comments, in which words are observed over time; and the *Instagram* dataset consists of sessions, in which comments come in a sequence; The data statistics are shown in Table 1. Note that compared to *Instagram* data, each demographic group in the *Jigsaw* data has a much smaller portion of positive instances. We detail the two datasets below.

- *Jigsaw.* The Perspective API's Jigsaw dataset[4] consists of comments – extracted from the Civil Comment platform – with toxicity and identity annotations. We consider gender and race as the sensitive attributes due to the relatively small number of comments associated with other identities.
- *Instagram* [20]. Instagram is a top-ranked social networking site with the highest percentage of users reporting experiences of cyberbullying [37]. Each *Instagram* sample is a social media session comprised of a sequence of comments in temporal order. As this dataset has no annotated identities (i.e., sensitive attributes), we detail the process of identifying potential attributes regarding swear words and dialect in the Experimental Setup subsection.

[3]The source code and data can be found at https://github.com/GitHubLuCheng/DebiasTD_via_Sequential_Decisions.
[4]https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification

Table 1: Statistics of the Instagram and Jigsaw datasets: percentages of every sensitive attribute and the proportion of toxic samples within each group.

	Group	Swear Words		Dialect				Overall
Instagram	Type	Present	Not Present	African-American	Hispanic	Asian	White	
	% data	34.80%	65.19%	15.77%	15.50%	1.26%	67.44%	100% (2,218)
	% toxicity	15.14%	13.79%	5.22%	4.59%	0%	19.11%	28.94% (642)

	Group	Gender			Race					Overall
Jigsaw	Type	Female	Male	Other	Black	White	Asian	Latino	Other	
	% data	12.92%	83.06%	4.00%	87.96%	5.87%	2.26%	1.13%	2.75%	100% (60,766)
	% toxicity	1.87%	8.56%	0.86%	9.16%	1.45%	0.21%	0.13%	0.33%	11.30% (6,872)

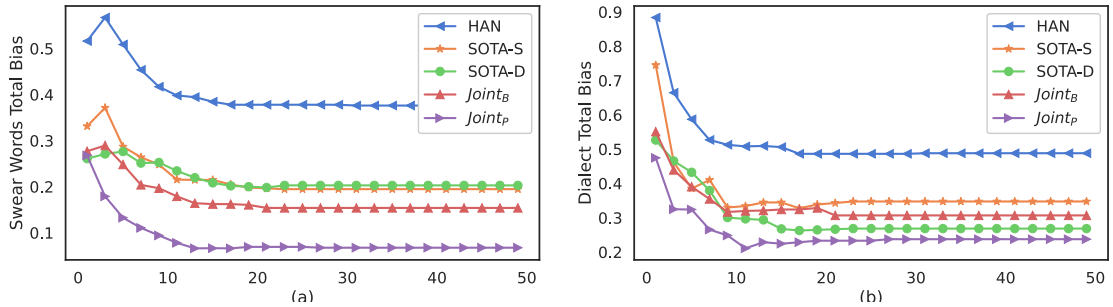

Figure 3: *Instagram:* The total biases (FPED + FNED) w.r.t. swear words and dialect. The x-axis denotes the number of *comments* (t) the agent has observed. SOTA-D in (a) shows the total bias w.r.t. Swear Words in the Dialect-debiased SOTA.

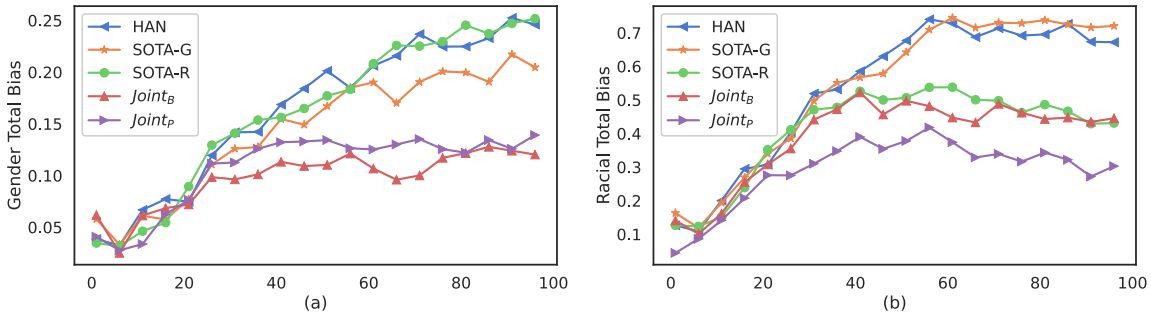

Figure 4: *Jigsaw:* The total biases (FPED + FNED) w.r.t. gender and race. The x-axis denotes the number of *words* (t) a model has observed. SOTA-R in (a) shows the total bias w.r.t. Gender in the Race-debiased model.

5.2 Experimental Setup

We consider two commonly studied types of bias with the *Instagram* dataset: swear words bias and dialectal bias. For swear-words-related bias, we used a set of predefined toxic keywords suggested in the psychology literature [26, 36]. A comment containing these terms is labeled as 1, otherwise as 0. To infer the dialect of a comment, we employed a lexical detector of words associated with AAE or WE [3], as used in previous work studying racial bias [13, 34].

5.2.1 Baselines. As methods such as data augmentation are not suitable for sessions with a sequence of comments and sensitive attributes with multiple groups, we consider the following baselines:

- *Biased Models.* These are standard machine learning models commonly used for toxicity classification. We consider the hierarchical attention network (HAN) [40] and the popular commercial model Google Jigsaw's *Perspective API*. As *Perspective* only works on a single comment, to assign the toxicity label for each Instagram session, we first use *Perspective* to label every comment in the session. The session label predicted by *Perspective* is then the majority vote of the comments' labels.
- *Debiasing with Fairness Constraints.* The *Constraint* model [17] is a debiased toxicity detection model that imposes fairness constraints of a single bias type on standard classifiers.
- *Debiasing Models for Sequential Data.* This is the state-of-the-art sequential bias mitigation model (*SOTA*) for cyberbullying detection [9]. *SOTA* is built upon an RL framework in which

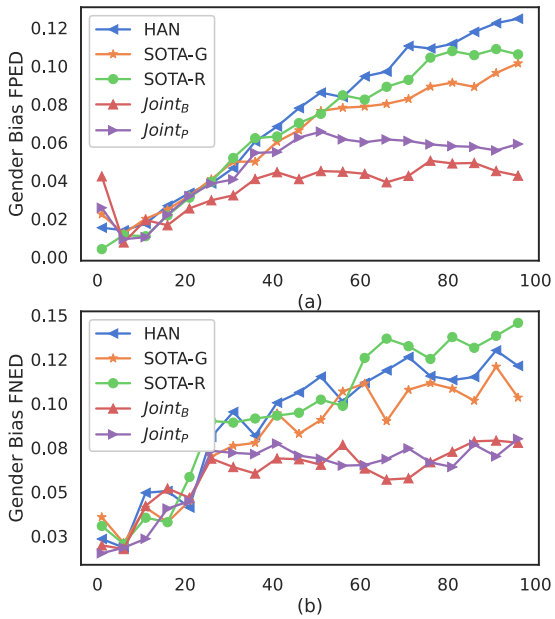

Figure 5: *Jigsaw:* The separate results for gender FPED and FNED. This complements the gender total bias in Fig. 4(a).

the reward function considers both the predictive error and the harmonic mean of bias amount measured by FPED and FNED.

For fair comparison, we further extend the BCM-based *Constraint* into PCM-based (denoted as $Constraint_P$) and use *HAN* as the backbone model of *Constraint*, $Constraint_P$, *SOTA*, and *Joint*. Also, *Constraint*, $Constraint_P$, and *SOTA* are individual debiasing methods, therefore, they have several variants when considering multiple biases. For instance, when both gender and racial biases are present, there are *SOTA-gender* (SOTA-G, i.e., gender-debiased SOTA) and *SOTA-race* (SOTA-R, i.e., race-debiased SOTA). Detailed parameter settings of all approaches can be found in Appendix A.

5.2.2 Evaluation Metrics. Evaluations in bias mitigation for toxicity detection typically focus on two aspects: prediction accuracy and bias removal. We adopt standard metrics for binary classification, including Precision (Prec.), Recall (Rec.), F1, and Accuracy (Acc.). Following [14, 17], we use FPED, FNED, and total bias (FPED+FNED) to assess debiasing effectiveness. An effective model should present low bias and high prediction performance. For all compared models, we use pre-trained GloVe word embeddings [30] and 10-fold cross validation with 80% data for training and the rest for testing. We also perform McNemar's test to examine the statistical significance of the difference between models. Unless otherwise noted, the differences between our model and the baselines are statistically significant. We highlight all of the best results.

5.3 Results

5.3.1 Do different biases tend to be correlated? Essentially, we ask how a toxicity classifier debiased for one bias influences the results for other biases. We first show in Fig. 3-4 the total biases (i.e., FPED + FNED) of the two bias types in each dataset, respectively, at each timestep. For clear comparisons, all of the selected baselines are specifically designed for sequential data. Observe that for *Instagram* data, mitigating an individual bias can also reduce the other bias. For example, in Fig. 3(a), the total bias of swear words in SOTA-D (i.e., SOTA debiased for Dialect) is reduced significantly compared to that of the biased model, HAN. This finding appears less clear for *Jigsaw* data as shown in Fig. 4, in part due to the extremely low percentage of positive instances in each group in *Jigsaw*. However, a deeper investigation of separate results for FPED and FNED in Fig. 5 shows that SOTA-R presents lower FPED and larger FNED w.r.t. gender than the biased model HAN. This indicates that an individual debiasing method designed for race can help mitigate the FPED of gender but increases its FNED. Therefore, we empirically validate our first hypothesis that biases **tend to correlate**.

5.3.2 How does Joint fare against generic and static debiasing methods? First, we observe in Fig.3-4 that sequential and joint bias mitigation strategies with PCM ($Joint_P$) consistently outperforms the biased model HAN and the generic debiasing approaches regarding all bias metrics over time. Further, $Joint_P$ appears to be more effective than $Joint_B$. Here, we consider both bias and classification measures of biased models (*HAN* and *Perspective*), generic and individual debiasing models (*Constraint*, $Constraint_P$, and *SOTA*), and sequential and joint debiasing models ($Joint_B$ and $Joint_P$). Results for both *Instagram* and *Jigsaw* datasets are shown in Table 2-3.

We observe the following: First, it is challenging to reduce both FPED and FNED within the same or across different sensitive attributes due to the potential trade-off between the two bias metrics and correlations among biases. Second, $Joint_P$ significantly reduces total bias in both datasets. Regarding prediction, it achieves the best recall and F1 and competitive accuracy performance (see the last row in Table 2 as an example). By sacrificing precision, $Joint_P$ focuses more on identifying positive instances (i.e., protecting the victims), which is more desirable in toxicity detection [7]. Third, for the baselines, we again observe correlations between biases, supporting findings observed in **RQ. 1**. For example, in Table 2, the dialect-debiased Constraint (i.e., Constaint-D) presents the lowest swear word FNED while in Table 3, the race FNED of Constraint-G is largely amplified. Lastly, PCM-based models are more effective than BCM-based models w.r.t. both total bias mitigation and prediction performance, as we also showed in **RQ. 1**.

For **RQ. 2**, we conclude that (1) our joint debiasing strategy outperforms conventional approaches in terms of both bias mitigation and detection performance, as also shown by previous studies using sequential debiasing strategy [9]. This suggests the potential benefits of debiasing in a sequential manner; (2) PCM is a more effective measure than BCM regarding jointly mitigating potentially correlated biases. Therefore, with correlated biases, it is important to consider strategies tailored to individual biases.

5.3.3 Ablation Study. How does the size of historical information influence the sequential bias mitigation performance? First, the results in Fig. 3 show that as the models observe more **comments** in a social media session, the overall bias is progressively reduced. For *Jigsaw* in Fig. 4-5, we see an overall increased bias with more **words** observed. However, for $Joint_P$ and $Joint_B$, the bias measures become more stable and even decrease after the agent observes

Table 2: *Instagram*: Bias and classification performance of various models, "S"=swear words, "D"=dialect.

		Bias Metrics (↓)					Classification Metrics (↑)			
		FPED-S	FNED-S	FPED-D	FNED-D	Total	ACC.	Pre.	Rec.	F1
Biased Models	HAN	0.0113	0.0664	0.2229	0.6441	0.9447	0.7800	0.6370	0.5576	0.5947
	Perspective	0.0097	0.0654	0.2149	0.3865	0.6764	0.4829	0.3113	0.5708	0.4029
Debiased Models	Constraint-S	0.0331	0.0524	0.1555	0.6675	0.8885	0.8532	0.7152	0.9089	0.8004
	Constraint-D	0.0741	**0.0316**	0.1730	0.1170	0.3957	0.8578	0.7923	0.8994	0.8424
	$Constraint_P$-S	0.0106	0.0638	0.1601	0.4203	0.6548	0.8465	0.7370	0.9315	0.8230
	$Constraint_P$-D	0.0032	0.0748	0.1521	0.0797	0.3098	0.8741	0.7463	0.9217	0.8247
	SOTA-S	0.0817	0.1075	0.1260	0.2214	0.5366	0.8747	0.7230	0.9190	0.8093
	SOTA-D	0.0692	0.1336	**0.0815**	0.1869	0.4712	0.8981	**0.9159**	0.7737	0.8388
	$Joint_B$	0.0274	0.1265	0.1541	0.1527	0.4607	**0.9017**	0.7819	0.9159	0.8436
	$Joint_P$	**0.0021**	0.0654	0.1619	**0.0756**	**0.3050**	0.9008	0.7567	**0.9688**	**0.8497**

Table 3: *Jigsaw*: Bias and classification performance of various models, "G"=gender, "R"=race.

		Bias Metrics (↓)					Classification Metrics (↑)			
		FPED-G	FNED-G	FPED-R	FNED-R	Total	ACC.	Pre.	Rec.	F1
Biased Models	HAN	0.1231	0.1204	0.2942	0.4064	0.9441	0.8811	0.4724	0.4411	0.4562
	Perspective	0.0419	**0.0319**	0.2011	0.5649	0.8398	0.4820	**0.5693**	0.3105	0.4019
Debiased Models	Constraint-G	**0.0147**	0.0465	0.1941	0.6920	0.9473	0.8893	0.5102	0.3790	0.4349
	Constraint-R	0.0873	0.0991	0.1669	0.3722	0.6255	0.8712	0.4946	0.4513	0.4719
	$Constraint_P$-G	0.0135	0.0398	0.1863	0.4320	0.6716	0.8893	0.5378	0.3810	0.4460
	$Constraint_P$-R	0.0970	0.0879	**0.1135**	**0.1843**	0.4824	**0.8945**	0.5421	0.3417	0.4191
	SOTA-G	0.0894	0.1153	0.2862	0.4173	0.9082	0.8828	0.4793	0.4184	0.4468
	SOTA-R	0.1096	0.1468	0.1788	0.2432	0.6784	0.8851	0.4900	0.4003	0.4407
	$Joint_B$	0.0519	0.0761	0.1279	0.2674	0.5233	0.8937	0.4396	0.4540	0.4495
	$Joint_P$	0.0517	0.0772	0.1198	0.2110	**0.4597**	0.8921	0.4667	**0.4880**	**0.4771**

approximately 50 words. We believe this is partly because a comment typically contains more semantically richer information than a word and biases and semantics are inherently related. For a deeper understanding, we vary the size of the observed historical information at each timestep, i.e., "window size" ΔL. For example, $\Delta L = 5$ means that at t, the agent uses the comments observed from $t-5$ to t to take an action. In this experiment, we use both datasets to examine the influence of window size on reduced total bias Δ(FPED + FNED) w.r.t. each bias type, compared to the biased model HAN. The results are shown in Fig. 6-7.

We observe that the proposed $Joint_P$ model consistently outperforms the baselines when the window size increases. Moreover, as window size increases, i.e., more historical comments are observed at each timestep, most approaches tend to remove more bias, especially for models debiased for the same target sensitive attributes. For example, in Fig. 7(a), where gender is the sensitive attribute, we can see that the difference in gender bias between SOTA-G and HAN increases. The same trend is observed in $Joint_P$ and $Joint_B$. This observation is less clear for the results of *Instagram*, as shown in Fig. 6, e.g., the difference in swear word bias between HAN and all the debiasing models is comparatively stable as the window size grows. There are multiple potential reasons: (1) the number of comments in a session is significantly smaller than the number of words in a comment (see the x-axis range for the two datasets); (2) the annotation of sensitive attributes for *Instagram* is noisier, as it is automatically generated whilst that for *Jigsaw* is human-coded; and (3) words and comments provide different levels of semantic information, and the dependency among words is often stronger than comments. Future experiments are needed to test these hypotheses.

How does α influence the performance of Joint? We examine the impact of α, which controls the importance of individual biases in PCM, as shown in Eq. 7. We use *Instagram* for illustration purposes. Specifically, for every bias $i \in$ {Dialect, Swear Words}, we set $\alpha_{j \neq i} = 0.5$ while varying the α_i parameter as $\alpha_i \in \{0.0, 0.25, 0.50, 0.75, 1.0\}$. We show both the prediction and bias removal results in Fig. 8. We can see a clear trade-off between the performance w.r.t. the two biases: by increasing the weight α_i of bias i, $Joint_P$ shows an increase in bias j and a decrease in bias i. For example, when α_S (i.e., weight for Swear Word bias) increases from 0.75 to 1.0, $Joint_P$ prioritizes the swear word bias, resulting in an increase in the dialect bias. In this figure, we observe the overall robustness of *Joint* to changes in α_i as well as the trade-off between mitigating multiple biases.

5.3.4 Case Studies. In addition to the above quantitative analyses, we show some case studies performed on *Instagram* data in Fig. 9. Several observations further support the previous quantitative results. In this non-bullying session, $Joint_P$ weighs less the importance of swear words as well as comments including these words, therefore, making the correct prediction. In contrast, the biased

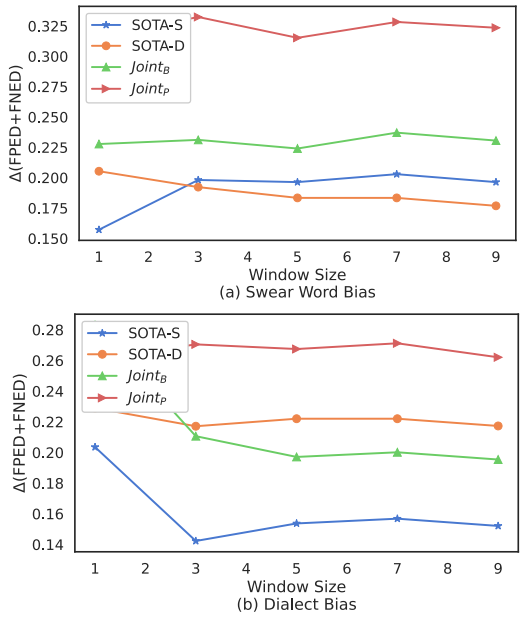

Figure 6: *Instagram:* Δ(FPED + FNED) w.r.t. swear words and dialect biases for different window sizes. For example, Δ(FPED + FNED) of SOTA-S is the difference between the total bias of SOTA-S and that of HAN. A larger value of Δ indicates a greater reduction in bias.

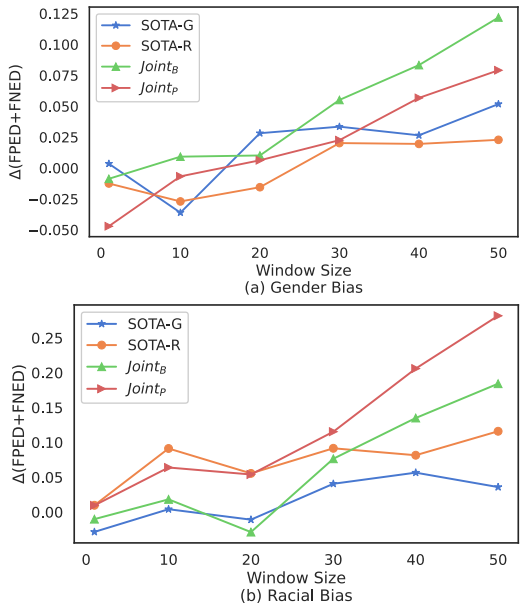

Figure 7: *Jigsaw:* Δ(FPED + FNED) of racial and gender biases for different window sizes. For example, Δ(FPED + FNED) of SOTA-G is the difference between the total bias of SOTA-G and that of HAN. A larger value of Δ indicates a greater reduction in bias.

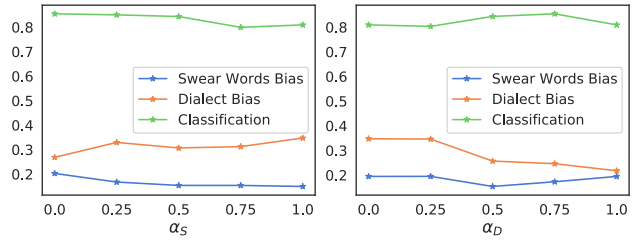

Figure 8: *Instagram:* Bias measures and classification performance (F1 score) using different values of α_i.

Figure 9: Case studies on *Instagram*. "Probability" denotes the model's output probability of predicting the true label. Darker color indicates larger attention weight.

model mistakenly predicts that this session is "bullying," due to the emphasis on the swear words. In addition, $Joint_P$ with a larger window size, i.e., more historical information, makes the correct prediction with higher confidence.

In summary, while most of the empirical findings in **RQ. 1** and **RQ. 3** suggest that the size of accessible historical information is critical for bias mitigation, future research is warranted to obtain more conclusive findings.

6 CONCLUSION

In contrast to the static and generic bias mitigation approaches in the debiasing toxicity classifier literature, this paper studies the novel problem of joint bias mitigation in the presence of potentially correlated biases with sequential input. In particular, we first empirically show that different biases tend to correlate. We then develop an effective solution that leverages the strengths of the theories

in sequential MDP and the PCM-based bias measure to maximize prediction accuracy and jointly minimize total bias. PCM forces the model to focus on the differences between various biases. Empirical evaluation with real-world datasets corroborates the effectiveness of the sequential bias mitigation approach with sequential input. Given our finding that word- and comment-level semantics impact the performance differently, future research can incorporate such hierarchical structure and mitigate biases in a hierarchical manner. Our work could also be adapted to other applications/domains to examine the generalizability of the approach and findings. Lastly, the proposed approach could benefit from additional studies about the ways semantic context influences sequential debiasing.

ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation under the grant #2036127. The views, opinions and/or findings expressed are the authors' and should not be interpreted as representing the official views or policies of the U.S. Government.

A REPRODUCIBILITY

The compared approaches include two biased models (HAN and *Prospective*), three debiased models (*Constraint*, $Constraint_P$, and *SOTA*), and the proposed joint debiasing models ($Joint_P$ and $Joint_B$). We detail the parameter settings for each of these approaches below.

- **HAN:** We implemented the work of [40] with the parameter settings in Table 4 to pre-train the model.
- **Perspective:** We used Google's Perspective API to gather the toxicity probabilities for every comment in a social media session in *Instagram* and in *Jigsaw*. We applied the Sigmoid function to convert the predicted toxicity probabilities to binary labels, i.e., $\hat{y} = \text{Sigmoid}(Pr(\text{toxic}|\mathbf{x}))$. For *Instagram*, we used the majority vote to obtain the "toxic" label for every session,
- **Constraint:** We used the implementation of [17] and set the parameters of the regularization terms as 0.005.
- **Constraint$_P$**: We used the same implementation as *Constraint*, with the constraints changed to PCM-based metrics.
- **SOTA:** We used the implementation of [9]. The backbone model is a HAN following the same parameter settings as in Table 4. Parameters used in the RL framework can be found in Table 5. We set β – the parameter that balances between the prediction error and bias measures – to 1.
- ***Joint*:** We used the same backbone model (i.e., HAN) and parameter setting for training the RL framework as in *SOTA*. We set the alpha values as 0.5 for all bias types in $Joint_B$ and $Joint_P$.

Table 4: The details of the parameters of the *HAN* classifier.

Parameter	Instagram	Jigsaw
Learning Rate	3e-3	3e-3
Batch Size	256	256
Word Embedding	GloVe	GloVe
Embedding Dim.	100	100
Sentence Length	150	-
Number of Comments	100	512
Word Attention Dim.	200	200
Sentence Attention Dim.	200	200
Number of Epochs	5	2

Table 5: The details of the parameters of the RL algorithm.

Parameter	Instagram	Jigsaw
Learning Rate	1e-5	1e-5
γ	0.1	0.1
Number of Episodes	150	150

REFERENCES

[1] Google Perspective API. 2022. https://www.perspectiveapi.com/.
[2] RE Bellman. 1957. A markov decision process. journal of Mathematical Mechanics. (1957).
[3] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *EMNLP*. 1119–1130.
[4] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *NIPS* 29 (2016), 4349–4357.
[5] Eloi Brassard-Gourdeau and Richard Khoury. 2019. Subversive toxicity detection using sentiment information. In *Proceedings of the Third Workshop on Abusive Language Online*. 1–10.
[6] Lu Cheng, Ruocheng Guo, Yasin Silva, Deborah Hall, and Huan Liu. 2019. Hierarchical attention networks for cyberbullying detection on the instagram social network. In *SDM*. SIAM, 235–243.
[7] Lu Cheng, Jundong Li, Yasin Silva, Deborah Hall, and Huan Liu. 2019. PI-bully: Personalized cyberbullying detection with peer influence. In *The 28th International Joint Conference on Artificial Intelligence (IJCAI)*.
[8] Lu Cheng, Jundong Li, Yasin N Silva, Deborah L Hall, and Huan Liu. 2019. Xbully: Cyberbullying detection within a multi-modal context. In *WSDM*. 339–347.
[9] Lu Cheng, Ahmadreza Mosallanezhad, Yasin Silva, Deborah Hall, and Huan Liu. 2021. Mitigating Bias in Session-based Cyberbullying Detection: A Non-Compromising Approach. In *ACL*.
[10] Lu Cheng, Kush R Varshney, and Huan Liu. 2021. Socially responsible AI algorithms: issues, purposes, and challenges. *Journal of Artificial Intelligence Research* 71 (2021), 1137–1181.
[11] Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics. *TACL* (2021).
[12] Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, and Franciska de Jong. 2013. Improving cyberbullying detection with user context. In *ECIR*. Springer, 693–696.
[13] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial Bias in Hate Speech and Abusive Language Detection Datasets. In *Proceedings of the Third Workshop on Abusive Language Online*. 25–35.
[14] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *AIES*. 67–73.
[15] Hong Fan, Wu Du, Abdelghani Dahou, Ahmed A Ewees, Dalia Yousri, Mohamed Abd Elaziz, Ammar H Elsheikh, Laith Abualigah, and Mohammed AA Al-qaness. 2021. Social Media Toxicity Classification Using Deep Learning: Real-World Application UK Brexit. *Electronics* 10, 11 (2021), 1332.
[16] 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 (2018), E3635–E3644.
[17] Oguzhan Gencoglu. 2020. Cyberbullying detection with fairness constraints. *IEEE Internet Computing* 25, 1 (2020), 20–29.
[18] Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. In *EMNLP-IJCNLP*. 1161–1166.
[19] He He. 2016. SEQUENTIAL DECISIONS AND PREDICTIONS IN NATURAL LANGUAGE PROCESSING. *PhD Dissertation* (2016).
[20] 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* (2015).
[21] Jae Yeon Kim, Carlos Ortiz, Sarah Nam, Sarah Santiago, and Vivek Datta. 2020. Intersectional bias in hate speech and abusive language datasets. In *ICWSM 2020 Data Challenge Workshop*.
[22] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring Bias in Contextualized Word Representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*. 166–172.

[23] Marzieh Mozafari, Reza Farahbakhsh, and Noël Crespi. 2020. Hate speech detection and racial bias mitigation in social media based on BERT model. *PloS one* 15, 8 (2020), e0237861.
[24] Vinita Nahar, Xue Li, and Chaoyi Pang. 2013. An effective approach for cyberbullying detection. *Communications in information science and management engineering* 3, 5 (2013), 238.
[25] Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive language detection in online user content. In *WWW*. 145–153.
[26] Andrew Ortony, Gerald L Clore, and Mark A Foss. 1987. The referential structure of the affective lexicon. *Cognitive science* 11, 3 (1987), 341–364.
[27] Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *EMNLP*. 2799–2804.
[28] John Pavlopoulos, Prodromos Malakasiotis, and Ion Androutsopoulos. 2017. Deep Learning for User Comment Moderation. In *Proceedings of the First Workshop on Abusive Language Online*. 25–35.
[29] John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity Detection: Does Context Really Matter?. In *ACL*. 4296–4305.
[30] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *EMNLP*. 1532–1543. http://www.aclweb.org/anthology/D14-1162
[31] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation* 55, 2 (2021), 477–523.
[32] Amir H Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive language detection using multi-level classification. In *Canadian Conference on Artificial Intelligence*. Springer, 16–27.
[33] Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. Hatecheck: Functional tests for hate speech detection models. In *ACL*.
[34] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *ACL*. 1668–1678.
[35] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*. 1–10.
[36] Anna Squicciarini, Sarah Rajtmajer, Y Liu, and Christopher Griffin. 2015. Identification and characterization of cyberbullying dynamics in an online social network. In *ASONAM*. 280–285.
[37] 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.
[38] William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*. 19–26.
[39] Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*. 88–93.
[40] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *NAACL HLT*. 1480–1489.
[41] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *NAACL HLT*. 1415–1420.
[42] 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. In *ACL*.
[43] Xuhui Zhou, Maarten Sap, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2021. Challenges in automated debiasing for toxic language detection. In *EACL*.