Can We Improve Model Robustness through Secondary Attribute Counterfactuals?

Ananth Balashankar  Xuezhi Wang  Ben Packer  Nithum Thain  Ed H. Chi  Alex Beutel

Google AI
{ananthbshankar,xuezhiw,bpacker,nthain,edchi,alexbeutel}@google.com

Abstract

Developing robust NLP models that perform well on many, even small, slices of data is a significant but important challenge, with implications from fairness to general reliability. To this end, recent research has explored how models rely on spurious correlations, and how counterfactual data augmentation (CDA) can mitigate such issues. In this paper we study how and why modeling counterfactuals over multiple attributes can go significantly further in improving model performance. We propose RDI, a context-aware methodology which takes into account the impact of secondary attributes on the model’s predictions and increases sensitivity for secondary attributes over reweighted counterfactually augmented data. By implementing RDI in the context of toxicity detection, we find that accounting for secondary attributes can significantly improve robustness, with improvements in sliced accuracy on the original dataset up to 7% compared to existing robustness methods. We also demonstrate that RDI generalizes to the coreference resolution task and provide guidelines to extend this to other tasks.

1 Introduction

How can we build NLP models that perform well over many slices, albeit sometimes small slices, of our data? Developing models that are robust in their performance is important for trusting these models to work well in diverse, unexpected settings. As a concrete running example in this paper, we will consider the task of toxicity detection: using a model to predict if a comment is toxic or not (Dixon et al., 2018). In this application, for example, it is often important to ensure that models are accurate over slices of data referring to different demographic groups, as has been raised across machine learning fairness research (Hardt et al., 2016a; Blodgett et al., 2020).

One significant focus of research on how to improve model robustness has been addressing spurious correlations and improving counterfactual robustness. That is, researchers have found that models often rely on features or attributes that are only spuriously correlated with the task and accuracy often drops when models are evaluated on counterfactual data that perturbs those attributes (Jia and Liang, 2017). To return to our example of toxicity detection, a model may learn that certain identity tokens are correlated with toxicity, but that could decrease accuracy for non-toxic comments with those terms (Dixon et al., 2018; Garg et al., 2018). Recent work has explored how counterfactual generation techniques can be used to form general checklists to test for model biases (Ribeiro et al., 2020; Bender and Koller, 2020), often composing many sub-problems which are hard to solve formally. Similarly, a wide breadth of research has studied how to train models to be more robust. We focus on one such mitigation technique—counterfactual data augmentation (CDA), where the supervised training data is augmented and balanced by replacing in-place words or phrases in the input sentence, which should not lead to a change in the output label $Y$ (Lu et al., 2018; Zmigrod et al., 2019b). These counterfactual data generation approaches have been built on, as well as coupled with regularization, to improve counterfactual fairness (Kusner et al., 2017), such as preventing models from being overly sensitive to identity terms (Garg et al., 2018; Prabhakaran et al., 2019; Kurita et al., 2019; Park et al., 2018).

Although these approaches have been effective in reducing spurious correlations, in this paper we observe and study how such approaches often fail to significantly improve core model accuracy and can still perform worse on subsets of the dataset due to the primary variable over which counterfactual are generated being correlated with (many) secondary variables that are not swapped or balanced. Returning to our example task of toxicity classification over comments, the
primary attributes (e.g., demographic identity terms) may be correlated with secondary attributes (intent of the comment—directed or descriptive) in the training data distribution. That is, for some demographic groups we may observe more directed comments and for others we may observe more directed comments:

<table>
<thead>
<tr>
<th>Toxic</th>
<th>Seeking transgender rights is extreme (Directed)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-Toxic</td>
<td>Transgender rights activists are labeled extremists (Descriptive)</td>
</tr>
</tbody>
</table>

In the toxicity classification example shown above, while the former is labeled as toxic by human annotators as it is directed towards a demographic group, the latter is only describing the toxicity and is considered as non-toxic. Nonetheless, both these sentences are classified as toxic by the Jigsaw Perspective API (Dixon et al., 2018), thus leading to high false positive rates. So, to remove spurious correlations for the word “transgender” with toxicity, it may not be enough to improve model accuracy over comments with the word “transgender” if the model is more accurate for directed comments than descriptive ones. Therefore we ask: can explicitly considering counterfactuals over both primary and secondary attributes better improve robustness?

To answer this question and improve model’s robustness, i.e., accuracy on slices of the data, we propose a new approach, RDI, that learns from counterfactual data generated through interventions on both the primary and secondary attributes. RDI applies regularization techniques to train the model to disentangle the impact of the primary and secondary attribute and to explicitly optimize for the classifier’s predictions to be sensitive or insensitive to each attribute. The approach also builds on recent reweighting approaches (Keith et al., 2020; Choudhury and Kiciman, 2017) to further address distributional skew in the data.

Our approach to studying this problem builds on works that argue for a case-by-case analysis of variables and aims to provide a framework for incorporating secondary variables when we discuss the robustness of natural language models (de Groot et al., 2018; van Aken et al., 2018). Specifically, we have focused on the toxicity detection model which prior work has shown to suffer from unintended bias (Dixon et al., 2018) based on protected identity terms mentioned in the sentence. We analyze how existing robustness techniques fail to capture a secondary attribute, namely the intent of the sentence while performing counterfactual data augmentation. We further show that this intent, that is descriptive or directed, is significantly correlated with specific protected identity groups in the dataset. By disentangling this correlation in the real world data via the counterfactual data, we obtain a model that does not disparately have high false positive rates on specific demographic groups, while being sensitive to the intent of the sentence. We achieve this improvement in robustness, while improving the sliced accuracy across multiple protected identity subgroups of the data.

Our key contributions are:

- We demonstrate how to disentangle the impact of protected and secondary attributes in NLP tasks like toxicity detection.
- We show how existing models perform poorly on counterfactual datasets that modify the secondary attributes, and train robust models that sensitize the model towards the secondary variables in a context-aware manner.
- Empirically, we demonstrate that our RDI method improves overall accuracy and sliced accuracy by 2-7% on all identity groups for both the toxicity detection task and generalizes on the coreference resolution task, while reducing spurious correlations through secondary attributes.

2 Related Work

Counterfactual Data Augmentation

We build on prior work that performs counterfactual data augmentation (Ren et al., 2019; Bodapati et al., 2019; Malykh et al., 2018). Counterfactual data augmentation (CDA) has been used to create more balanced datasets to mitigate bias (Lu et al., 2019; Zhao et al., 2018; Zmigrod et al., 2019a; Garg et al., 2018) towards protected identity groups or improve accuracy (Kaushik et al., 2019). Our work extends this literature by including a secondary variable that is correlated to the standard primary variable on which CDA is performed. This extension is motivated by works like (Gonen and Goldberg, 2019) which demonstrate that there are secondary variables that need to be addressed for robustness.

Adversarial Robustness

Making NLP models robust to adversarial perturbations has recently been explored extensively (Zhou et al., 2019). Work in this space define adversarial attacks through word or character perturbations (Pruthi et al., 2019; Ebrahimi et al., 2018; Alzantot et al., 2018) and certifiable defences (Ribeiro et al., 2018;
Jia et al., 2019) following early work in adversarial training (Goodfellow et al., 2015). One of the challenges in applying adversarial techniques to the discrete domain of NLP is the lack of an ϵ-boundary in the input space. Hence, we consider only those interpretable perturbations that explicitly modify the primary and secondary attributes, as mentioned in a sentence.

**Bias Mitigation** Our work draws on recent works that aim to mitigate unintentional bias towards protected attributes in NLP tasks (Bolukbasi et al., 2016). The approach of counterfactual token fairness which performs bias mitigation of template based (Dixon et al., 2018) augmented data has been shown to improve model performance over specific subgroups (Garg et al., 2018). Debiasing techniques can be broadly categorized into in-processing: which changes training methodology (Beutel et al., 2019; Jiang and Bansal, 2019; Zhang et al., 2020) and post-processing: which operate post hoc on trained models (Krasanakis et al., 2018). While debiasing in unsupervised language models have also improved downstream tasks (Webster et al., 2020), we take the in-processing approach of debiasing in a supervised setting. Specifically, in the domain of coreference resolution, we closely relate to the work from (Rudinger et al., 2018; Field and Tsvetkov, 2020) to identify secondary variables; and in the domain of toxicity detection, we draw on qualitative error analysis (van Aken et al., 2018; Fortuna et al., 2020; Basile et al., 2019) and domain expertise (Waseem and Hovy, 2016; de Gibert et al., 2018; Saleem et al., 2017; Sharma et al., 2018) to derive our understanding of the secondary variable (intent of the comment) and how it relates to the label (toxicity); see Appendix 1. Another related perspective is that of distributional robustness where a machine learning model trained on one data distribution is evaluated on a modified data distribution (Li et al., 2018; Ma et al., 2019; Miller, 2019; Liu et al., 2019; Fu et al., 2017; Arjovsky et al., 2020). Following this body of work, our objective is to ensure that the model relies on invariances that generalize when the model is tested on slices of data, a type of distributional shift.

### 3 Problem Definition

#### 3.1 Setup

Given a dataset, $D$, we will generate an augmented dataset, $\hat{D}$ by adding synthetic, balanced and counterfactually augmented sentences.

Given an NLP classification task that operates on individual sentences $s \in D$, consider a primary variable $X$, which could be one of group based identities (say race, gender, etc) that is spuriously correlated with a secondary variable $Z$ (e.g., intent of the comment—directed or descriptive) and the label $Y$ that is to be predicted (say toxicity). In our setting, the values $(x, z)$ of the primary and secondary variables $X, Z$ are contained within an individual sentence $s$. We use the intent of the comment as our running example for $Z$ in the toxicity detection task, but our approach can be easily generalized to other factors like dialect, in-group language, figure of speech, etc. Note that since we are building prediction models that output $\hat{Y}$, we are interested in checking if a given model’s predictions perform accurately on counterfactual inputs.

Our problem definition relies on the following assumptions about the primary and secondary variables prevalent in recent works on counterfactual robustness (Jia and Liang, 2017; Zmigrod et al., 2019a; Keith et al., 2020). Firstly, given a sentence, the primary and secondary variables contained within it can be pre-specified. We also assume that counterfactual sentences that modify both the primary and secondary variables independently can be generated. Hence, we follow template based counterfactual data generation which specifies the primary and secondary variables in each sentence, as outlined in Section 5.2.

#### 3.2 Objectives

Before we present our problem definition, we define the objectives that we will use from the robustness and fairness literature. Finally, we position these objectives within our context-aware counterfactual robustness problem formulation. For sake of simplicity here and in the following sections, we consider that the label, primary and secondary variables are binary with values \{0, 1\}; \{x_0, x_1\}; \{z_0, z_1\} respectively. However, similar definitions for multivariate settings can be inferred.

##### 3.2.1 Metrics

**Original Dataset:** In the original dataset $D$, as in most NLP tasks, we define the evaluation accuracy metric $A$ over a set of sampled sentences $s$. Further, to evaluate the accuracy of the held out dataset conditional on the primary variable $X$, we compute...
the sliced accuracy $A(x)$ over that subset.

$$A = \mathbb{E}_{s \sim \tilde{D}} 1(\tilde{Y}_s = Y_s)$$  \hspace{1cm} (1)

$$A(x) = \mathbb{E}_{s \sim \tilde{D}} 1(\tilde{Y}_s = Y_s | X = x)$$  \hspace{1cm} (2)

**Counterfactual Dataset:** To improve counterfactual robustness, we aim to improve accuracy $A$ on the counterfactual dataset, by enumerating all possibilities of the values assigned to $X$ and $Z$. We generate counterfactual sentences $t(s, x, z)$ by setting values of $X = x, Z = z$ in a sentence $s \in \tilde{D}$ using templates. Similar to overall accuracy, we can define sliced accuracy, $\tilde{A}(x)$ on the counterfactual dataset $\tilde{D}$ while enumerating all possible value assignments of the secondary variable. Note that the dataset $\tilde{D}$ represents a less biased dataset, one which might not actually be observed, but represents all possible values of the primary and secondary variables $X, Z$ in $\tilde{D}$, and allows us to measure the toxicity detection model’s counterfactual robustness around both the primary and secondary attributes.

$$\tilde{A} = \mathbb{E}_{s \sim \tilde{D}, x \in \{x_0, x_1\}, \; z \in \{z_0, z_1\}} 1(\tilde{Y}_{t(s,x,z)} = Y_{t(s,x,z)})$$  \hspace{1cm} (3)

$$\tilde{A}(x) = \mathbb{E}_{s \sim \tilde{D}, z \in \{z_0, z_1\}} 1(\tilde{Y}_s = Y_s | X = x)$$  \hspace{1cm} (4)

### 3.3 Goal:
Our robustness goal is to improve a model’s robustness $A(x)$ - i.e. accuracy on the original dataset sliced by the primary sensitive variable $X$. As secondary variables like $Z$ are spuriously correlated with primary variables $X$ in the original dataset $D$, we need to disentangle the impact of primary and secondary variables by optimizing on the generated counterfactual dataset $\tilde{D}$. In our paper, we achieve this goal by optimizing $A, \tilde{A}(x)$ over the dataset $\tilde{D}$, generated through interventions on both the primary and secondary variables, such that this improvement generalizes to the original dataset $D$.

### 4 Methodology
Since the goal of robustness is in addition to that of increasing overall accuracy on the original dataset, we use constrained optimization techniques over augmented counterfactual data. Before we present our proposed constraints, we present existing baseline constraints defined in the fairness and robustness literature. We discuss why these baseline constraints do not explicitly address the goal of improving counterfactual robustness on primary and secondary variables, and hence necessitate our additional proposed constraints on the counterfactual dataset $\tilde{D}$.

#### 4.1 Baseline Constraints

**Equality of Opportunity (EO):** The Equality of Opportunity (Hardt et al., 2016b) constraint imposes statistical equality on the false positive errors, when conditioned on different values of the primary variable $X$. Such a constraint enforces that the primary variable $X$ has no impact on the false positive rate of the model. We approximate this constraint over with the synthetic, balanced counterfactually augmented data $\tilde{D}$ (CDA) by minimizing the EO gap (Zhao et al., 2017) with respect to the primary variable (Eqn 5) and denote it by the baseline “EO+CDA”.

$$\min(\| \mathbb{E}_{s \sim \tilde{D}} (Y_s = 1|Y_s = 0, X = x_0) - \mathbb{E}_{s \sim \tilde{D}} (Y_s = 1|Y_s = 0, X = x_1) \|)$$  \hspace{1cm} (5)

**Counterfactual Token Fairness (CTF):** In Garg et al. (2018), the logits are equalized across counterfactual examples $s \sim \tilde{D}$ for different values of the primary variable $X$, but not the secondary variable $Z$. If $f(s)$ denotes the logit of the model’s prediction, and $t(s, x)$ denotes the sentence generated by swapping the primary variable with $x$ as per the template, then CTF minimizes the following logit pairing gap:

$$\min \mathbb{E}_{s \sim \tilde{D}, X = x_0} |(f(s) - f(t(s, x_1))|$$  \hspace{1cm} (6)

Since $X$ and $Z$ are spuriously correlated, both CTF and EO+CDA constrained models, which solely focus on $X$, are susceptible to performing poorly on examples when value of $Z$ is altered explicitly. For example in the Jigsaw toxicity detection dataset, consider when $Y$ is denoting “toxicity”, $X$ represents gender and $Z$ the intent of the comment - descriptive or directed. If, for example, we observe in the real world that most directed comments are towards women, and not men (spurious correlation between $X$ and $Z$), then just intervening on the gender $X$ of the sentence and changing it from female to male, might unintentionally remove the impact of the secondary variable - the intent of the sentence, on the toxicity detection task $Y$. This is undesirable because the intent of the sentence is genuinely correlated with $Y$ and its impact should not be removed.
4.2 Proposed Constraints

We overcome the limitation of not including secondary variable impact in baseline constraints, by explicitly modeling to Maximize Secondary Sensitivity in tasks like toxicity detection, where the label \( Y \) is sensitive to changing values of the secondary variable \( Z \) in the counterfactual dataset. We later discuss how this can be generalized to tasks where the secondary variable \( Z \) does not impact the label \( Y \) in Section 7.

Maximize Secondary Sensitivity: In some cases involving secondary variables, a characteristic that is often desired in a robust model is that it should be sensitive towards a change in a specific variable. For example in the Jigsaw toxicity (\( Y \)) dataset, even though more directed comments on online forums are towards females, and more descriptive comments are used for males, the model should be sensitive to the intent of comment in determining the toxicity. If we blindly optimize for just CTF, the model may be less robust to changes in the intent of comments from descriptive to directed (\( Z \)). To overcome this issue, we propose a constraint that retains model sensitivity to changes in the secondary variable \( Z \), while conditioning on the primary variable \( X \). If \( t'(s,x,z) \) is the template-generated sentence by swapping out values of \( x, z \) in a sentence \( s \) such that the label \( y \) assigned to the sentence changes to \( \neg y \), and \( f_y(s) \) denotes the logit of the model’s prediction of \( y \) for \( s \), then we propose to maximize the following conditional logit pairing gap.

\[
\max x \in \{x_0, x_1\}, y \in \{0,1\} \sum_{y,z} \mathbb{E}_{s \sim \tilde{D}|y=s,x,z=x_0} (f_y(s) - f_{\neg y}(s'))
\]

(7)

Reweighting Samples All of the above constraints still do not enforce the independence between \( X \) and \( Z \) in the counterfactual dataset, \( \tilde{D} \), if there is a sampling bias which prefers highly correlated samples of \( X, Z \) in \( D \). This is because the real world dataset might suffer from selection bias, task annotator difficulty bias (Gordon and des Jardins, 1995), etc, which cannot be easily offset through data augmentation alone. Therefore in addition to augmenting counterfactual data, we seek to reweight the augmented samples in such a way that the probability of \( Z \) conditional on \( X \) is equalized. Hence, a sentence \( s \in \tilde{D} \) with \( X = x_0, Z = z \) is weighted by \( w_z \) using an inverse-propensity based weighting (Olteanu et al., 2017) based on the prevalence of \( Z \) conditional on \( X \). However, since we are fine-tuning over the counterfactual dataset to also generalize over the original dataset, we are concerned about improving residual accuracy. We, thus apply this weighting on only those samples in the original validation dataset which our unconstrained model has incorrectly predicted. This boosting inspired technique (Schapire, 2003) emphasizes the need to equalize the prevalence conditioned on our worst-case examples (Oren et al., 2019) where our initial model \( \hat{Y}_{base} \) has made an incorrect prediction. For example, we reweight based on the error rates, a sentence \( s \in \tilde{D} \) with \( X = x_0, Z = z, Y = y \).

\[
w_z = \frac{P_D(Z=z|X=x_1, Y=y, \hat{Y}_{base} = \neg y)}{P_D(Z=z|X=x_0, Y=y, \hat{Y}_{base} = \neg y)} \tag{8}
\]

Context-Aware Counterfactual Robustness

Based on the relationship of the secondary variable with the label, we incorporate our proposed constraints on the counterfactually augmented dataset \( \tilde{D} \) as a fine-tuning step. Thus, the methods we propose can be used on any NLP model as a fine-tuning task. We summarize our proposed RDI methodology based on the context of the secondary variable in Algorithm 1.

Algorithm 1 RDI (Reweight-Direct-Indirect)

1: Input: Trained NLP model - \( M \)'s predictions \( \hat{Y}_{base} \), primary variable \( X \), secondary variable \( Z \), label \( Y \)
2: for each batch do
3: Augment template based samples for all \((X, Z)\) pairs to form \( \tilde{D} \)
4: Reweight samples based on (8)
5: \( \mathcal{L} = \mathbb{E}_{s \sim \tilde{D}} \text{CrossEnt} (\hat{Y}_s, Y_s) \)
6: \( \mathcal{L}_{RDI} \leftarrow (6) + (7) \)
7: Back-propagate \( \alpha \mathcal{L} + (1 - \alpha) \mathcal{L}_{RDI} \) in \( M \)
8: end for

5 Evaluation

5.1 Data

The Jigsaw Kaggle toxicity dataset \(^1\) contains sentences from the Civil Comment platform. We narrow down our focus to the comments that have the referenced identity in the comment, as well as the

\(^{1}\)https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification
binary label: toxic or non-toxic. In total, 1,804,874 comments are annotated for toxicity, out of which ~50% of them have identities annotated too. Note that the identities are crowd sourced and not self-identified. We use a randomized 80-20 train-test split in our evaluation.

Some comments refer to certain protected identity groups, which we refer to as the primary variable. Based on the qualitative study of toxic comments (Waseem and Hovy, 2016), we can broadly categorize the intent of comments as either directed or descriptive. Directed comments are speech towards a specific individual or group, whereas descriptive comments are more factual and do not hint towards a group or individual. Different identity groups are exposed to different intended comments, thus making the intent of the comment (descriptive or directed) our secondary variable $Z$. In this domain, our goal is to mitigate the impact of the primary variable on the prediction (Eqn 6), while retaining the sensitivity of the secondary variable on the predicted label (Eqn 7).

### 5.2 Augmentation Templates

The above dataset is the basis on which we evaluate the accuracy of the original dataset using our RDI algorithm. However, this dataset is not amenable for counterfactual data augmentation. Hence, we rely on a set of template based datasets to generate the counterfactual data on which we will fine-tune our models. (Dixon et al., 2018) released a set of madlibs templates to generate toxic and non-toxic comments based on hierarchies of intersectional identities. We extend this framework to incorporate templates for intent of the comment - directed and descriptive based on the definition of toxicity provided in (Waseem and Hovy, 2016). We provide an example of the 130,721 such counterfactual examples generated below (see appendix 1 for the full set of templates). Note that in addition to using templates, we can also utilize unsupervised learning based techniques to identify directed and descriptive comments.

### 5.3 Metrics

We evaluate the AUC for each identity group and the overall dataset in the Jigsaw Toxicity dataset. Since the secondary variable in the Toxicity dataset is not available for the Jigsaw dataset, we also present sliced AUCs based on the descriptive/directed intent of the comment as labeled by a model trained to predict solely the intent of comments with accuracy of 94.3% (details in Appendix 3). Since we are comparing sliced accuracy across 9 identity groups in the toxicity dataset, we also compute the standard error bars in the measurement of each metric. We also perform a two sample independent t-test over $n = 10$ random restarts for each of the slices with $2n - 2$ degrees of freedom, and a significance threshold of $\alpha = 0.05$, $m = 9, 5$ (Bonferroni correction) for the two datasets respectively when we compare against the baselines.

### 5.4 Baselines

We present a brief description of the various baselines, each optimizing a baseline objective as discussed in Section 3.

<table>
<thead>
<tr>
<th>Baseline</th>
<th>Model Objectives</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vanilla</td>
<td>Fine-tuned large uncased BERT model</td>
</tr>
<tr>
<td>EO+CDA</td>
<td>BERT+EO over balanced $\tilde{D}$ (Zhao et al., 2017)</td>
</tr>
<tr>
<td>CTF+CDA</td>
<td>BERT+CTF controlled on the primary variable (Garg et al., 2018) over $\tilde{D}$</td>
</tr>
<tr>
<td>RDI</td>
<td>BERT + RDI algorithm</td>
</tr>
</tbody>
</table>

Table 1: Summarized description of baselines

Figure 1: Accuracy of Jigsaw Perspective API model when sliced by the context (directed or descriptive) of the comments on our counterfactual dataset.

Figure 2: Area under the Curve (AUC) for toxicity detection across various demographic groups in the Jigsaw dataset.
6 Results

6.1 Sliced Accuracy

In Figure 1, sliced accuracy of the vanilla model on the template based counterfactually augmented data highlights the need for improving sensitivity towards descriptive comments. In Figure 2, we show the impact on the AUC of identity groups as identified in the original Jigsaw toxicity dataset. Specifically, RDI performs 0.52% better in overall AUC with p-value $= 0.001 \leq 0.005$ (significance level $= \alpha_m$), while increasing the sliced AUC for black identity by 6.98% (p-value=0.002). We see a general trend of improvement in AUC over the baseline vanilla model by 1.98–6.98%, with statistically significant improvement for groups of male, jewish, muslim, and black identities by making the model sensitive to the secondary variable – “comment intent”. We subsequently fine-tuned a BERT model to predict the intent of the comment (descriptive/directed) on the Kaggle dataset and sliced the change in accuracy as compared to the best performing CTF+CDA baseline. The resulting changes in Figure 3 demonstrate that for the slices where our model underperforms, it is due to a degradation in assessing directed comments for female, LGBT and disability groups. As expected, for the descriptive comments, we see consistent improvement across the board.

6.2 Ablation Studies

In order to understand the impact of the 3 objectives of the RDI algorithm, we conducted ablation studies by using the leave-one-out strategy (Figure 4). We note that, while removing the constraint based on counterfactual fairness (Eqn 6) has the highest impact, reweighting samples (Eqn 8) and controlling for secondary variables (Eqn 7) also have significant impact on both overall and sliced accuracy in the Jigsaw Toxicity evaluation dataset.

6.3 Qualitative Analysis

We note that there is significant improvement in descriptive comments in most of the identity groups as shown in Figure 3. For example, in the black identity group, we see that the improvement in AUC is better in descriptive sentences 4.1% than directed ones 3.1%. While analyzing the errors of our model, we see that they occur often beyond the scope of our problem formulation (Martin-Jr. et al., 2020) (Appendix 4).

7 Pronoun Coreference Resolution

We have demonstrated the utility of modeling secondary attributes to improve robustness of the toxicity detection models. However, we note that not all tasks have secondary attributes whose impact on the label needs to be maximized. Each task and their corresponding secondary attributes are unique in their relationship and their difficulty in data gathering, and we need careful understanding of the context while enforcing constraints between them. In this section, we show how our RDI framework can be extended to a task - “pronoun coreference resolution”, where the label is invariant to the secondary attribute - gender. Between these two use cases, we have exhaustively covered the types of constraints that can be incorporated towards secondary attributes and encourage researchers to undertake a contextual treatment of secondary attributes in their tasks. We provide an example below, where the pronoun resolution should not change based on the gender of the pronoun.
Table 2: Mitigating gendered correlation in coreference resolution as well increasing accuracy in the OntoNotes and Winogender datasets with statistical significant change denoted by *

Datasets and Augmentation Templates: For the pronoun coreference resolution task, we use the OntoNotes dataset shared as part of the CONLL 2011 and 2012 shared task (Pradhan et al., 2012, 2011). Each of the nouns referenced back from the pronouns also have their associated gender (binary) (Bergsma and Lin, 2006). In the OntoNotes coreference dataset, we evaluate the F1-score, the gendered correlation coefficient (Rudinger et al., 2018) which measures the correlation between gender and the professions they resolve to. The Winogender coreference resolution dataset provides templates with placeholders for the gendered-pronoun, and two antecedent professions which the pronoun could potentially be referencing. We refer to (Rudinger et al., 2018) for the full set of templates.

Label invariance to secondary variable In the gender bias Winograd dataset (Rudinger et al., 2018), the label is the coreference of the pronouns towards one of the two antecedents mentioned in the sentence. The pronouns are gendered (primary variable) binary - male and female; and the antecedents denote professions (secondary attribute) which the pronouns might get coreferenced to. Here, our goal is to minimize the unintended correlation of certain professions towards a specific gender. A systemic imbalance in the real world (see US Bureau of Labor Stats), is then reflected as a sampling bias in the text. For example, among the people with the profession “engineer”, only 10.72% of them are females as per the labor statistics and a similar correlation is recorded in the text corpus, but an ethical ML practitioner would ideally want their robust model to not propagate these correlates by using the constraint in Eqn 9.

Minimize Secondary Impact: If we denote the logit of the model’s prediction for a sentence \( s \) by \( f(s) \), and the sentence generated by swapping out values of \( x, z \) in a sentence \( s \) without changing the label to be \( t(s, x, z) \), then we propose to minimize the following conditional logit pairing gap, inspired by counterfactual indirect effects defined in (Zhang and Bareinboim, 2018) instead of Eqn 7.

\[
\min \sum_{s \in \{0,1\}} \mathbb{E}_{s \sim D}\left[ f(s) - f(s') \right] + \mathbb{E}_{s \sim D}\left[ f(s) - f(s') \right] \quad (9)
\]

Note that here, we explicitly focus on the change in error rates due to the change in the secondary variable \( Z \), previously ignored by the baseline constraints. In the pronoun coreference resolution task, this amounts to equalized error rates on all professions, while conditioning on the gender of the pronoun \( X \): male, female.

Robustness Gains: We see a similar trend in the overall accuracy for the coreference resolution task in Table 2. Here, we compare against one other baseline - dropout (Webster et al., 2020) where the baseline BERT model’s dropout hyperparameters have been optimally finetuned for robustness. RDI outperforms existing baselines on both the F1-accuracy (higher is better) and the gendered correlation (lower is better). The lower gendered correlation also translates to a more even distribution of gendered pronoun resolution across the 5 quintiles of gendered professions (Rudinger et al., 2018).

8 Conclusion

We have demonstrated the value of incorporating the impact of secondary variables in the objectives for learning robust natural language processing models. We have shown that incorporating context-aware counterfactual robustness through the RDI...
algorithm, we improve performance on the counterfactual augmented data, but also improve the overall and sliced accuracy on the original dataset by 2–7%.

9 Broader Impact Statement

As we are dealing with the toxicity detection task, the concern of dual use for generating more toxic content on social media has to be considered. That being said, the identification of directed toxic comments towards minority communities can greatly improve the experience of members, often targeted due to their membership in protected classes in these online social communities. More so, when these same members describe the toxicity they experience on those social online forums, the possibility of them being flagged as toxic, can be harmful. We show that, without considering secondary variables, such errors, particularly in groups which are the target of toxic comments, can further exacerbate this divide. By developing an approach for controlling for known proxies, we hope this can enable practitioners to incorporate more domain knowledge, particularly from users in under-served communities, to improve these systems. The template based counterfactual augmentation in capturing such nuances of secondary variables is a small step towards enabling more user participation and control in the design of these systems.

References


Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligram, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. CoRR, abs/1607.06520.


Ona de Gibert, Naiara Perez, Aitor Garcia-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum.


Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them.


J. Zhang and Elias Bareinboim. 2018. Fairness in decision-making - the causal explanation formula. In AAAI.


