Counterfactual Fairness in Text Classification through Robustness

Sahaj Garg[1]*, Vincent Perot[2], Nicole Limtiaco[2], Ankur Taly[3], Ed H. Chi[3], Alex Beutel[2]

[1]Stanford University, Stanford, CA
[2]Google AI, New York, NY
[3]Google AI, Mountain View, CA
*Work done while the author was an intern at Google.
sahajg@cs.stanford.edu, {vperot, nlimtiaco, ataly, edchi, alexbeutel}@google.com

Abstract

In this paper, we study counterfactual fairness in text classification, which asks the question: *How would the prediction change if the sensitive attribute referenced in the example were different?* Toxicity classifiers demonstrate a counterfactual fairness issue by predicting that "Some people are gay" is toxic while "Some people are straight" is nontoxic. We offer a metric, counterfactual token fairness (CTF), for measuring this particular form of fairness in text classifiers, and describe its relationship with group fairness. Further, we offer three approaches, blindness, counterfactual augmentation, and counterfactual logit pairing (CLP), for optimizing counterfactual token fairness during training, bridging the robustness and fairness literature. Empirically, we find that blindness and CLP address counterfactual token fairness. The methods do not harm classifier performance, and have varying tradeoffs with group fairness. These approaches, both for measurement and optimization, provide a new path forward for addressing fairness concerns in text classification.

Introduction

Consider a model that determines whether an Internet forum comment is toxic. We would like to improve the model's fairness with respect to the content of the input text, which may reference sensitive identity attributes, such as sexual orientation, race, or religion. In practice, Dixon et al. showed that a toxicity model had a high false positive rate on examples that included identity tokens such as "gay," because such tokens occur relatively frequently in examples labeled toxic in the training set.

A related issue to users arises when nearly identical sentences referencing different identity groups receive different predictions. For instance, a baseline toxicity model predicts that "Some people are gay" is 98% likely to be toxic and "Some people are straight" is only 2% likely to be toxic. In this work, we seek to specifically address this fairness issue for text classification.

Given an example, we ask a *counterfactual* question: *How would the prediction change if the sensitive attribute referenced in the example were different?* If the prediction score changes with respect to a sensitive attribute, we consider this an indicator of a potential problem. In contrast to group-based notions of fairness (e.g., demographic parity, equality of odds), which seek to statistically equalize the model's

behavior for entire sensitive groups, counterfactual fairness requires equal model behavior on individual counterfactual pairs; see (Kusner et al. 2017; Wachter, Mittelstadt, and Russell 2017).

To assess counterfactual fairness, we consider perturbations obtained by substituting tokens associated with identity groups. For instance, substituting "gay" with "straight," or "Asian" with "American." Based on these generated counterfactuals, we can define a fairness metric, which we call *counterfactual token fairness* (CTF). While this is more limited than general counterfactual fairness, we believe it captures one of the most salient issues in text classification and is a starting point for more general counterfactual fairness metrics for text.

Deciding when counterfactual pairs should have the same prediction raises difficult ethical and philosophical questions. Many logical counterfactuals generated by token substitution may not require identical output. We call these *asymmetric counterfactuals*. In toxicity classification, such situations could arise when the comment references stereotypes associated with one group but not another, or when comments attack a particularly vulnerable group. Asymmetric counterfactuals suggest that practitioners should be careful in both training and evaluation of counterfactual fairness. We discuss proposals for addressing this in the case of toxicity classification in the experiments section.

To satisfy counterfactual token fairness, we borrow techniques from the robustness literature. We propose a general training scheme for achieving arbitrary counterfactual fairness by extending logit pairing (Kannan, Kurakin, and Goodfellow 2018) to penalize differences in the model's outputs for counterfactual pairs. We compare this method to simply augmenting the training set with counterfactual examples, and to blindness, which replaces all sensitive tokens with a special token.

One issue is that the aforementioned methods may only achieve fairness with respect to identity tokens considered by counterfactuals during training. To address this, we evaluate the generalization of the methods on a held-out set of identity tokens. Another concern when optimizing for counterfactual fairness is potential trade-offs with other desirable properties of a classifier, including overall accuracy and group fairness. In practice, we do not find significant harms with respect to accuracy, and varying effects on group fair-

ness in the form of tradeoffs between true negatives and true positives.

We make the following contributions:

- **Metric:** We provide a tractable metric, *counterfactual token fairness*, for measuring counterfactual fairness in text classification.

- **Methods:** We study three methods for addressing counterfactual token fairness: (A) blindness, (B) counterfactual augmentation, and (B) counterfactual logit pairing, bridging research from the robustness and fairness domains.

- **Empirical Evaluation:** We evaluate empirical performance and tradeoffs of counterfactual token fairness, group fairness, and accuracy across these approaches.

Related Work

ML Fairness   Significant work in the ML fairness literature has been devoted to measuring fairness. Our work is most closely related to counterfactual fairness in causal inference (Kusner et al. 2017; Kilbertus et al. 2017), where fairness is evaluated by applying counterfactual interventions over a causal graph. Our definition of counterfactual token fairness implicitly defines a simple causal model for text generation. Kusner et al. also draw the connection between counterfactual fairness and individual fairness, which requires similar predictions for similar inputs via a Lipschitz constraint (Dwork et al. 2011).

Relatively more study has been devoted to group fairness metrics, which evaluate observational criteria, or statistical relationships between the data, group membership, the label, and the model's prediction. Such metrics include demographic parity and equality of odds (Hardt, Price, and Srebro 2016). Hardt, Price, and Srebro demonstrate that observational criteria are insufficient to distinguish between some seemingly reasonable uses of identity and other unreasonable ones. This is because observational criteria cannot incorporate any external understanding about what is causally acceptable in making predictions. The limitations of observational criteria can be addressed by counterfactual or individual fairness, see (Kusner et al. 2017; Kilbertus et al. 2017; Dwork et al. 2011). By extending these definitions to path-specific counterfactual fairness, it is possible to specify which uses of identity are acceptable (Chiappa and Gillam 2018).

Social science literature on fairness raises arguments for counterfactual reasoning as well as potential limitations. One concern is about the ability to reasonably intervene on identity of an individual. Given that most social scientists agree that race is socially constructed, it may be unreasonable to attempt to modify race and all its associated factors (Kohler-Hausmann 2019). These limitations, among others, are reflected in debate surrounding the use of counterfactuals over race in epidemiological studies (Krieger 2014; VanderWeele and Robinson 2014). We note that our work deals with well-defined interventions on content by only manipulating identity tokens in text, rather than the actual identities of individuals, which differentiates it from the work above.

ML fairness literature has also focused on debiasing methods to address these gaps. Many methods have been proposed to address group fairness issues, such as recalibrating score functions (Hardt, Price, and Srebro 2016), adversarially learning fair representations (Zemel et al. 2013; Louizos et al. 2015; Beutel et al. 2017), data rebalancing (Dixon et al. 2018), and data augmentation using swaps of gender terms (Park, Shin, and Fung 2018). For natural language problems, Pryzant et al. learn a lexicon that is uncorrelated to a set of confounding variables. Debiasing methods for counterfactual or individual fairness have been studied less for neural network models. The methods in (Kusner et al. 2017; Kilbertus et al. 2017) are effective for causal graphs, but most machine learning problems will not fit this mold. To address individual fairness, (Dwork et al. 2011) applies constraint based optimization over a linear program, but it is difficult to define valid distance metrics or apply the optimization to arbitrary neural networks used in natural language processing.

Robustness in Machine Learning   The robustness literature in machine learning has primarily focused on robustness to adversarially perturbed inputs, which add small amounts of carefully selected noise to fool classifiers (Goodfellow, Shlens, and Szegedy 2015). When applied to the text setting, such adversarial examples can be generated by a variety of editing methods, including through translation (Ribeiro, Singh, and Guestrin 2018), attributions (Mudrakarta et al. 2018), and autoencoders (Zhao, Dua, and Singh 2017) (Hu et al. 2017). Adversarial examples are closely related to counterfactual examples: Wachter, Mittelstadt, and Russell characterize counterfactuals as adversarial examples that perturb inputs in human-interpretable and possibly problematic ways. As such, the counterfactual examples presented in this work can be viewed as a specific subset of adversarial examples. The robustness literature has attempted to address adversarial examples using a variety of techniques, such as adversarial training (Madry et al. 2017; Goodfellow, Shlens, and Szegedy 2015) and adversarial logit pairing (Kannan, Kurakin, and Goodfellow 2018).

Several papers draw connections between fairness, text generation, and robustness. Landeiro and Culotta consider robustness in text with respect to counfounding variables such as the author's gender, and learn robust models by training using an additional attribute for the latent confound, and averaging over all values of the latent variable at inference time. Madaan et al. attempt to edit text to remove gender bias or edit gender representations, leveraging analogies in word vector differences to make substitutions for words that may implicitly encode biases about gender.

Problem Definition

Given text input $x \in X$, where x is a sequence $[x_1, ..., x_n]$ of tokens, our task is to predict an outcome y. We consider a classifier f parameterized by θ that produces a prediction $\hat{y} = f_\theta(x)$, where we seek to minimize some notion of error between y and $\hat{y}$. For notational simplicity, we restrict the following definitions to a single binary class, but they can be easily generalized to multi-class classification problems.

The classifier f can be an arbitrary neural network.

We wish to maximize the model's performance while maintaining counterfactual fairness with respect to sensitive attributes, such as identity groups. Counterfactual fairness is measured using counterfactual examples that perturb the sensitive attribute referenced in the example at hand. Let $\Phi(x)$ denote the set of counterfactual examples associated with an example x. Counterfactual fairness requires that the predictions of a model for all counterfactuals are within a specified error.

Definition 1. *Counterfactual fairness.* A classifier f is counterfactually fair with respect to a counterfactual generation function Φ and some error rate ϵ if

$$|f(x) - f(x')| \leq \epsilon \quad \forall x \in X, x' \in \Phi(x)$$

Counterfactual Token Fairness (CTF)

We consider a narrow class of counterfactuals that involves substituting identity tokens in the input, for instance, substituting "gay" with "straight" in the input "Some people are gay." We assume a set of identity tokens, $\mathcal{A}$, for which we seek to be counterfactually fair. Consider a pair of tokens $a, a' \in \mathcal{A}$. The associated counterfactual example generation function $\Phi_{a,a'}$ is defined by substituting all occurrences of a in x with a' and vice versa. If neither identity token is present in the input x, then $\Phi_{a,a'}(x) = \emptyset$. We generalize this definition to a counterfactual generation function over $\mathcal{A}$ that generates all counterfactual examples based on pairs of substitutions:

$$\Phi_{\mathcal{A}}(x) = \bigcup_{a \neq a' \in \mathcal{A}} \Phi_{a,a'}(x)$$

Definition 2. A classifier satisfies *counterfactual token fairness* with respect to a set of identity tokens $\mathcal{A}$ if it satisfies counterfactual fairness with respect to the counterfactual generation function $\Phi_{\mathcal{A}}$ and error rate ϵ.

Although content about sensitive groups may be captured by complex semantics, this metric will surface a subset of problematic issues related to more general counterfactual fairness. This a first step, and surfaces additional concerns for fairness beyond those of group fairness.

Asymmetric Counterfactuals

So far we have assumed that all counterfactuals with respect to identity tokens should have the same prediction. This assumption is not valid in cases where the sensitive attribute indeed affects the prediction. For instance, consider a model predicting toxicity of text, and the counterfactual pair "That's so gay" and "That's so straight." The first example is arguably more likely to be considered toxic than the second, as "gay" is often used as an insult in Internet forums, while "straight" is not. Other examples include stereotyping, where one group is more vulnerable than another. Requiring equal predictions across such cases can inadvertently harm the more vulnerable group.

Fairness must be required only among counterfactuals that stipulate symmetric predictions. This restriction can be accommodated in our framework by restricting the counterfactual generation function $\Phi(x)$ to exclude any counterfactuals for the example x that may have asymmetric labels. In general, the degree and direction of the asymmetry between counterfactuals varies based on the task, and the cultural sensitivities of the consumers of the task. This makes it difficult to define a perfect counterfactual generation function. In Experiments, we propose a heuristic for avoiding asymmetric counterfactuals for a model predicting toxicity of text.

Relationship to Group Fairness

Counterfactual fairness is complementary to the group fairness notion of *equality of odds* (Hardt, Price, and Srebro 2016), which demands equality of true positive rates and true negative rates for different values of the sensitive attribute. A text classifier may satisfy one while completely failing the other. Consider the case when two sensitive attributes a and a' only appear in disjoint sets of contexts X_a and $X_{a'}$, respectively. A model can satisfy equality of odds by always predicting correctly on the contexts in which a, a' appear in the data, but never in the counterfactual contexts that do not exist in the data. Conversely, the model could predict identical output for all counterfactual pairs while predicting correctly only on X_a and not X'_a.

Methods

We propose three methods to improve counterfactual fairness: blindness, counterfactual augmentation, and counterfactual logit pairing. Both methods assume access to a list of identity tokens for which they seek to be fair.

Blindness

Blindness substitutes all identity tokens with a special `IDENTITY` token, which allows the predictor to know that an identity term is present, but not which identity. This is similar to standard NLP methods such as replacing large numbers with a generic `NUMBER`. While this approach guarantees counterfactual token fairness, it has a number of downsides. First, it does not have the ability to differentiate identity terms, and so necessarily equates asymmetric counterfactuals. Second, it cannot handle complex counterfactuals that involve more than single token substitutions, e.g. "Christians go to church." and "Jews go to temple." Finally, the model may still discriminate using other signals that are associated with the identity term (Dwork et al. 2011).

Counterfactual Augmentation

Instead of blinding the model to identity terms, counterfactual augmentation involves augmenting the model's training set with generated counterfactual examples. The additional examples are meant to guide the model to become invariant to perturbing identity terms. This is a standard technique in computer vision for making the model invariant to object location, image orientation, etc. The counterfactual examples are assigned the same label as the original example.

Counterfactual Logit Pairing (CLP)

Counterfactual logit pairing (CLP) encourages the model to be robust to identity by adding a robustness term to the training loss. The robustness term is given by logit pairing (Kannan, Kurakin, and Goodfellow 2018), which penalizes the norm of the difference in logits for pairs of training examples and their counterfactuals. Specifically, suppose the classifier $f(x) = \sigma(g(x))$, where $g(x)$ produces a logit and $\sigma(\cdot)$ is the sigmoid function. The additional loss is the average absolute difference in logits between the inputs and their counterfactuals:

$$\sum_{x \in X} \mathbb{E}_{x' \sim \text{Unif}[\Phi(x)]} |g(x) - g(x')|$$

For computational tractability, during training, we randomly sample a single counterfactual example for each input. Taking J as the original loss function, the overall objective is:

$$\sum_{x \in X} J(f(x), y) + \lambda \sum_{x \in X} \mathbb{E}_{x' \sim \text{Unif}[\Phi(x)]} |g(x) - g(x')|$$

Similar to counterfactual augmentation, CLP can use any counterfactual generation function. For example, a restricted counterfactual generation function could be used to avoid enforcing equality over asymmetric counterfactuals. Moreover, the method also applies if more sophisticated counterfactuals are generated.

In contrast to counterfactual augmentation, the robustness term in the CLP loss explicitly guides the model to satisfy two desirable properties: (1) ensuring a model produces similar outputs on counterfactual pairs and (2) learning models that generalize well to different identities. Moreover, the parameter λ can be tuned to achieve varying degrees of counterfactual fairness.

Experiments

Dataset   We evaluate our methods on the task of predicting toxicity. For the task, a toxic comment is defined as a "rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion." (Dixon et al. 2018). We use a public Kaggle dataset of 160K Wikipedia comments, each labeled by human raters as toxic or non-toxic[1], randomly split into train and dev sets. We evaluate AUC of the primary task on the public test set. We evaluate counterfactual token fairness and group fairness on a private dataset of comments from another internet forum. This dataset, henceforth called the "evaluation dataset," has a higher occurrence of identity terms, and therefore leads to a more meaningful fairness evaluation.

Setup   We evaluate our methods for counterfactual token fairness on the set of 50 identity terms used by Dixon et al.. Out of these, 47 are single tokens and 3 are bigrams. We randomly partition the terms into a training set of 35 and a hold-out set of 12 to evaluate generalization. We also include the three bigrams in evaluation, because they reflect scenarios that blindness cannot address during training.[2]

All of the models are CNNs trained with cross entropy loss against the binary toxicity label. All hyperparameters except for the fairness regularizer λ for CLP were fixed for all runs of all models. Models were trained for five epochs, and the best model on the dev set was taken. Each model was trained ten times, and the average of the runs is reported. Blindness, Counterfactual augmentation, and CLP models (for different values of λ) were evaluated and compared to a baseline model.

For CLP training, we define a different counterfactual example generation function than the one used for evaluation. The evaluation counterfactuals only apply substitutions to a pair of identity tokens, whereas during training, each sensitive identity token in an input is randomly substituted with another identity token.

Handling Asymmetric Counterfactuals   We hypothesize that asymmetric counterfactuals are less likely to arise for ground truth non-toxic comments than toxic comments. This for two reasons. Asymmetric counterfactuals arise when stereotyping / attacking a vulnerable group occurs for some identity substitution, and no other toxicity signals are present. In such cases, most identity substitutions will be nontoxic, and only the one attacking the vulnerable group(s) will be toxic. So if the ground truth example is nontoxic, counterfactual fairness can still be required over most identity substitutions, whereas if the ground truth example is toxic, equal prediction should not be required over most counterfactuals. Second, the stereotyping comments are more likely to occur in a toxic comment attacking the stereotyped group than in a nontoxic comment referencing some other identity group. For these reasons, we evaluate counterfactual token fairness over ground truth nontoxic comments separately from ground truth toxic comments, and focus our analysis on nontoxic comments. We also consider applying the CLP loss only to nontoxic comments during training, to avoid enforcing equality of logits for potentially asymmetric counterfactuals. We distinguish this variant as CLP_nontoxic.

Separately, we also evaluate CTF on simple synthetic inputs where all information about toxicity is encoded in the context, and all counterfactuals are symmetric by design. Specifically, we use a dataset of synthetically generated sentences based on templates such as "NAME is a ADJECTIVE."[3]

Metrics   We measure the *counterfactual token fairness gap* with respect to a given counterfactual generation function. For a single example, this is the average gap in prediction over all of the counterfactual pairs for that example:

$$\text{CF GAP}_\Phi(x) = \mathbb{E}_{x' \sim \text{Unif}[\Phi(x)]} |f(x) - f(x')|$$

Over an entire dataset, the gap is the average over all examples that have valid counterfactuals. In this study, we measure the CTF GAP for the counterfactual generation function $\Phi_\mathcal{A}$, which substitutes all pairs of identity tokens. Because substitution-based counterfactuals over short inputs

---

[1] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[2] Only single tokens in the input are substituted with bigrams during evaluation.

[3] This is the updated open sourced version of the synthetic test set presented in (Dixon et al. 2018).

Model	Eval NT	Synth NT	Synth Tox
Baseline	0.140	0.180	0.061
Blind	0.000	0.000	0.000
CF Aug	0.127	0.226	0.022
CLP_nontox, $\lambda = 1$	0.012	0.015	0.007
CLP, $\lambda = 0.05$	0.071	0.082	0.024
CLP, $\lambda = 1$	0.007	0.015	0.007
CLP, $\lambda = 5$	0.002	0.004	0.004

Table 1: Conterfactual token fairness gaps for non-toxic examples from evaluation set and all examples from a synthetic test set. All gaps are measure w.r.t. 35 training terms. Smaller gaps are better.

	CTF Gap: held-out terms
Baseline	0.091
Blind	0.090
CF Aug	0.087
CLP_nontox, $\lambda = 1$	0.095
CLP, $\lambda = 0.05$	0.078
CLP, $\lambda = 1$	0.084
CLP, $\lambda = 5$	0.076

Table 2: CTF gaps on held out identity terms for non-toxic examples from the evaluation set.

are more likely to be logical, we evaluate the CTF gaps for inputs of at most ten tokens in length. In addition, since asymmetric counterfactuals are likely more common for toxic comments, we evaluate CTF gaps over nontoxic and toxic comments separately.

We also measure group fairness to ensure that optimizing for counterfactual fairness has no perverse impact on it. Following the group fairness notion of equality of odds (Hardt, Price, and Srebro 2016), we measure the true positive rates (TPR) and true negatives rates (TNR) of examples referencing different identity groups. We assume an example references a specific identity group based on the presence of the associated token. Equality of odds requires equal TPR and TNR across identities, so we evaluate overall TPR and TNR gaps. The gap for a pair of identity terms is computed as the absolute value of the difference in rates for the two identity terms. The overall TPR or TNR gap is the average over all pairs of identity terms.

Results

Counterfactual Token Fairness Table 1 reports CTF gaps for non-toxic examples from the evaluation dataset, and all examples from the synthetic dataset. The gaps are computed for the 35 training terms (discussed in Setup). As discussed earlier, both these sets of examples are unlikely to have asymmetric counterfactuals. The baseline model has a large CTF gap on both sets of examples. Blindness achieves a zero gap by design. CLP with a fairness regularization coefficient (λ) of at least 1 also attains a near zero gap. Counterfactual augmentation decreases the CTF gap (relative to the baseline) on non-toxic examples from the evaluation dataset, but does not obtain a zero gap. It is worth noting that the models were not trained on the synthetic dataset, but we still find a reduction in counterfactual fairness gaps on it.

Table 2 reports CTF gaps on the hold-out terms for non-toxic examples from the evaluation dataset. We say that a model's CTF gap generalizes to hold-out terms if its gap is less than the baseline model's gap (0.091). Among the models compared, CLP with $\lambda = 5$ generalizes the best, though the gaps are much larger that those on the training terms. Blindness does not appear to provide generalization benefits. Thus, it may not be a favorable method in settings where we expect examples with identity terms outside the set of training terms.

To evaluate the impact on cases with asymmetric counterfactuals, we also measured the CTF gap for toxic examples from the evaluation dataset over the 35 training terms; see Table 4 in the appendix. The baseline model has a gap of 0.241, and as expected, blindness has a gap of zero. All CLP models with $\lambda \geq 1$ achieve a CTF gap of less than 0.03, which unfortunately means that they end up equating predictions for asymmetric counterfactuals. This includes CLP_nontoxic, which was trained using counterfactuals from non-toxic examples only. Going forward, we do not evaluate CLP_nontoxic as it is not better than the regular CLP models.

Overall Performance We evaluate the overall classifier performance using AUC of the ROC curve. Remarkably, all methods show consistent AUC on the test set, ranging between 0.962-0.964.

Figure 1 compares the true positive rate (TPR) and true negative rate (TNR) of various models, where the threshold for a toxic classification is set to 0.5. TPR and TNR are measured only over examples that contain an identity term from the set of training terms. We find that methods that reduce the CTF gap perform better at identifying nontoxic comments (true negatives) and worse at identifying toxic comments (true positives). We discuss this tension between improving the CTF gap and TPR in Error Analysis.

Group Fairness We additionally evaluate the group fairness metrics, TPR and TNR gaps for equality of odds (see Table 3). Counterfactual augmentation and CLP with $\lambda = 0.05$ have better TPR and TNR gaps than the baseline and are able to reduce CTF gaps. CLP with $\lambda \geq 1$ has a more extreme tradeoff, harming the TPR gap while substantially improving the TNR gap. Practitioners may choose different tradeoffs of CTF gap, TPR, and TNR depending on the relative prioritization of these metrics for a given task.

Error analysis

We examine the trade-off between CTF gap and TPR. We consider the CLP, $\lambda = 5$ model which attains a near zero CTF gap and compare its predictions on toxic comments to those of the baseline. Among examples with identity terms in the test set, there are 83 cases where an example was correctly classified by the baseline and incorrectly classified by the CLP model. Of these, 27 were labeled by an author as having an asymmetric counterfactual. There were 20 cases where the CLP model predicted correctly compared to
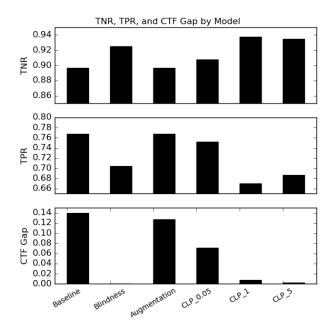
Figure 1: Plot of the average CTF gap along with the TPR and TNR over examples that contain identity terms.

	TNR Gap	TPR Gap
Baseline	0.084	0.082
Blindness	0.039	0.114
Augmentation	0.065	0.083
CLP all, $\lambda = 0.05$	0.058	0.078
CLP all, $\lambda = 1$	0.039	0.104
CLP all, $\lambda = 5$	0.041	0.112

Table 3: TNR and TPR gaps for different models. Lower is better.

the baseline, of which none had asymmetric counterfactuals. This tells us that a large chunk of the TPR loss (relative to the baseline) is over toxic examples with asymmetric counterfactuals. This is expected as examples with asymmetric counterfactuals are toxic because of the presence of a specific identity term, and a model trained to disregard identity terms will be less likely to predict correctly on such examples.

As a means of investigating what the CLP model has learned, we examine its token embeddings after convergence. By the end of training with $\lambda = 5$, the average cosine similarity between pairs of identity tokens is 0.87, whereas the baseline has an average cosine similarity of 0.25. Although this is similar to blindness, the CLP model learns a different toxicity association with identity tokens. The average toxicity prediction on a single identity token for the CLP model is 0.12, while the toxicity of the IDENTITY token in the blindness model is 0.54.

Similarly, CLP_nontoxic with $\lambda = 5$ has a average cosine similarity of 0.81. This embedding convergence, despite CLP being applied only to nontoxic comments, is the reason why the model achieves a low CTF gap on toxic comments, including those with asymmetric counterfactuals. Methods to enforce equal prediction on some subset of counterfactuals but not others should be further investigated.

We also qualitatively evaluate the strength of each model's association with various identity tokens. Table 5 in the appendix lists various examples, and the associated toxicity scores from each model. In contrast to the baseline, all three models associate a much smaller amount of toxicity signal with the identity tokens. For instance, unlike the baseline, the other models no longer associate a substantial amount of toxicity with clearly nontoxic statements such as "Some people are gay." Notably, the toxicity of the statement "Some people are straight" goes up. The negative effect on this pair is more pronounced for blindness than it is for CLP.

Conclusions and Future Work

We make progress towards counterfactual fairness in text classification. We propose a specific form of counterfactual fairness, called *counterfactual token fairness* (CTF), that requires a model to be robust to different identity tokens present in the input. We show that text classification models with good overall performance fare poorly on this metric. We approach counterfactual token fairness from a robustness perspective, and offer a procedure, *counterfactual logit pairing*, for optimizing the counterfactual token fairness metric during model training. We find that this approach performs as well as blindness to identity tokens, but also generalizes better to hold-out tokens. These results do not come at the expense of overall classifier accuracy, and have varying tradeoffs between false positives and false negatives.

Going forward, better heuristics must be designed for identifying cases with asymmetric counterfactuals. Excluding toxic comments covers many but not all asymmetric examples. For example, ground truth nontoxic examples referencing "black power" are more likely to become toxic as they reference "white power." In other text classification tasks such as sentiment classification, asymmetric counterfactuals will arise but not necessarily with the same clear split by label.

A next step would be to improve counterfactual generation by addressing issues of polysemy of identity terms (which can result in illogical substitutions), asymmetric counterfactuals, and multiple references to an identity group. One possible method is to use analogies in word vectors to change multiple tokens used for the same identity group (Madaan et al. 2018). Another approach is defining a generative model over text, as in (Hu et al. 2017), that can modify certain attributes of the text while holding others constant and preserving semantics. One could also use criteria for selecting semantically equivalent adversarial examples as in (Ribeiro, Singh, and Guestrin 2018), to evaluate whether counterfactual examples are logical. Optimizing for general counterfactual fairness will test many of the unique advantages of counterfactual logit pairing.

References

Beutel, A.; Chen, J.; Zhao, Z.; and Chi, E. H. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *CoRR* abs/1707.00075.

Chiappa, S., and Gillam, T. P. S. 2018. Path-Specific Counterfactual Fairness. *arXiv e-prints* arXiv:1802.08139.

Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2018. Measuring and mitigating unintended bias in text classification.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. S. 2011. Fairness through awareness. *CoRR* abs/1104.3913.

Goodfellow, I.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *CoRR* abs/1610.02413.

Hu, Z.; Yang, Z.; Liang, X.; Salakhutdinov, R.; and Xing, E. P. 2017. Toward controlled generation of text. In Precup, D., and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, 1587–1596. International Convention Centre, Sydney, Australia: PMLR.

Kannan, H.; Kurakin, A.; and Goodfellow, I. J. 2018. Adversarial logit pairing. *CoRR* abs/1803.06373.

Kilbertus, N.; Rojas-Carulla, M.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. In *Proceedings from the conference "Neural Information Processing Systems 2017.*, 656–666. Curran Associates, Inc.

Kohler-Hausmann, I. 2019. Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Northwestern University Law Review* 113(5).

Krieger, N. 2014. On the causal interpretation of race. *Epidemiology* 25(6).

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc. 4066–4076.

Landeiro, V., and Culotta, A. 2016. Robust text classification in the presence of confounding bias.

Louizos, C.; Swersky, K.; Li, Y.; Welling, M.; and Zemel, R. S. 2015. The variational fair autoencoder. *CoRR* abs/1511.00830.

Madaan, N.; Mehta, S.; Agrawaal, T.; Malhotra, V.; Aggarwal, A.; Gupta, Y.; and Saxena, M. 2018. Analyze, detect and remove gender stereotyping from bollywood movies. In Friedler, S. A., and Wilson, C., eds., *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, 92–105. New York, NY, USA: PMLR.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *CoRR* abs/1706.06083.

Mudrakarta, P. K.; Taly, A.; Sundararajan, M.; and Dhamdhere, K. 2018. Did the model understand the question? *CoRR* abs/1805.05492.

Park, J. H.; Shin, J.; and Fung, P. 2018. Reducing Gender Bias in Abusive Language Detection. *ArXiv e-prints*.

Pryzant, R.; Wang, K.; Jurafsky, D.; and Wager, S. 2018. Deconfounded lexicon induction for interpretable social science. In *16th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 856–865. Association for Computational Linguistics.

VanderWeele, T. J., and Robinson, W. R. 2014. On the causal interpretation of race. *Epidemiology* 25(6).

Wachter, S.; Mittelstadt, B. D.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR* abs/1711.00399.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. In Dasgupta, S., and McAllester, D., eds., *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, 325–333. Atlanta, Georgia, USA: PMLR.

Zhao, Z.; Dua, D.; and Singh, S. 2017. Generating natural adversarial examples. *CoRR* abs/1710.11342.

Appendix

	Train Terms	Held-out Terms
Baseline	0.241	0.071
Blind	0.000	0.062
CF Augmentation	0.155	0.057
CLP_nontoxic, $\lambda = 1$	0.029	0.068
CLP, $\lambda = 0.05$	0.165	0.057
CLP, $\lambda = 1$	0.010	0.058
CLP, $\lambda = 5$	0.004	0.051

Table 4: CTF gaps on toxic examples from the evaluation set, for both training terms and held-out terms.

	Baseline	Blindness	CF Augmentation	CLP , $\lambda = 5$
Some people are gay	0.98	0.61	0.82	0.14
Some people are straight	0.02	0.61	0.11	0.14
Some people are Jewish	0.28	0.61	0.17	0.13
Some people are Muslim	0.46	0.61	0.24	0.14
Some people are Christian	0.04	0.16	0.02	0.14

Table 5: Counterfactuals and toxicity scores of different models. The tokens "gay," "straight," "jewish," and "muslim" are used during training, and "christian" was held-out.