Improving Counterfactual Generation for Fair Hate Speech Detection

Aida Mostafazadeh Davani, Ali Omrani, Brendan Kennedy, Mohammad Atari, Xiang Ren, Morteza Dehghani
University of Southern California
{mostafaz, aomrani, btkenned, atari, xiangren, mdehghan}@usc.edu

Abstract

Bias mitigation approaches reduce models' dependence on sensitive features of data, such as social group tokens (SGTs), resulting in equal predictions across the sensitive features. In hate speech detection, however, equalizing model predictions may ignore important differences among targeted social groups, as hate speech can contain stereotypical language specific to each SGT. Here, to take the specific language about each SGT into account, we rely on *counterfactual fairness* and equalize predictions among counterfactuals, generated by changing the SGTs. Our method evaluates the similarity in sentence likelihoods (via pre-trained language models) among counterfactuals, to treat SGTs equally only within interchangeable contexts. By applying logit pairing to equalize outcomes on the restricted set of counterfactuals for each instance, we improve fairness metrics while preserving model performance on hate speech detection.

1 Introduction

Hate speech classifiers have high false-positive error rates in documents mentioning specific social group tokens (SGTs; *e.g.,* "*Asian*", "*Jew*"), due in part to the high prevalence of SGTs in instances of hate speech (Wiegand et al., 2019; Mehrabi et al., 2019). When propagated into social media content moderation, this *unintended bias* (Dixon et al., 2018) leads to unfair outcomes, *e.g.,* mislabeling mentions of protected social groups as hate speech.

For prediction tasks in which SGTs do not play any special role (e.g., in sentiment analysis), unintended bias can be reduced by optimizing group-level fairness metrics such as *equality of odds*, which statistically equalizes model performance across all social groups (Hardt et al., 2016; Dwork et al., 2012). However, in hate speech detection, this is not the case, with SGTs providing key information for the task (see Fig. 1). Instead, bias

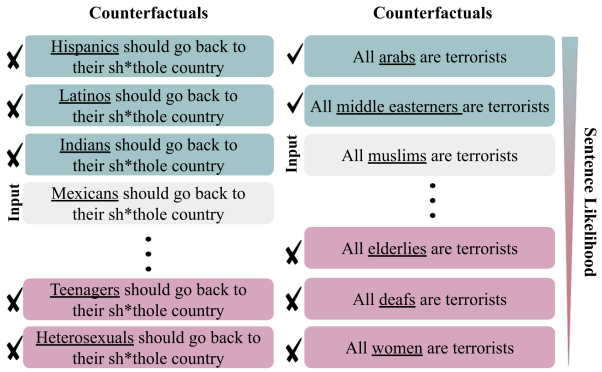

Figure 1: Two input sentences and their counterfactuals ranked by likelihood. Our method ensures similar outputs only for counterfactuals with higher likelihood.

mitigation in hate speech detection benefits from relying on individual-level fairness metrics such as *counterfactual fairness*, which assess the output variation resulting from changing the SGT in individual sentences (Garg et al., 2019; Kusner et al., 2017). Derived from causal reasoning, a counterfactual applies the slightest change to the actual world to assess the consequences in a similar world (Stalnaker, 1968; Lewis, 1973).

Accordingly, existing approaches for reducing bias in hate speech detection using counterfactual fairness learn robust models whose outputs are not affected by changing the SGT in the input (Garg et al., 2019). However, a drawback of such approaches is the lack of semantic analysis of the input to identify whether changing the SGT leads to a small enough change that preserves the hate speech label (Kasirzadeh and Smart, 2021). For instance, in a hateful statement, "*mexicans should go back to their sh*thole countries*", substituting "*mexicans*" with "*women*" changes the hate speech label, while using "*Hispanics*" should preserve the output. Here, we aim to create counterfactuals that *maximally* preserve the sentence and disregard counterfactuals that violate the requirement for be-

Proceedings of the Fifth Workshop on Online Abuse and Harms, pages 92–101

ing the "*closest possible world*" (Fig. 1).

To this end, we develop a counterfactual generation method which filters candidate counterfactuals based on their difference in likelihood from the actual sentence, estimated by a pre-trained language model with known stereotypical correlations (Sheng et al., 2019). Intuitively, our method provides outputs that are robust with regard to the context and are not causally dependent on the presence of specific SGTs. This use of sentence likelihood is inspired by Nadeem et al. (2020) as it captures the similarity of an SGT and its surrounding words to prevent unlikely SGT substitutions. As a result, only counterfactuals with equal or higher likelihoods compared with the original ("closest possible worlds") are used during training. To enforce robust outputs for similar counterfactuals, we apply logit pairing (Kannan et al., 2018) on outputs for sentence-counterfactual pairs, adding their average differences to the classification loss. Our method (1) effectively identifies semantically similar counterfactuals and (2) improves fairness metrics while preserving classification performance, compared with other strategies for generating counterfactuals.

2 Related Work

Unintended bias in classification is defined as differing model performance on subsets of datasets that contain particular SGTs (Dixon et al., 2018; Mehrabi et al., 2019). To mitigate this bias, data augmentation approaches are proposed to create balanced labels for each SGT or to prevent biases from propagating to the learned model (Dixon et al., 2018; Zhao et al., 2018; Park et al., 2018). Other approaches apply regularization of post-hoc token importance (Kennedy et al., 2020b), or adversarial learning for generating fair representations (Madras et al., 2018; Zhang et al., 2018) to minimize the importance of protected features.

By altering sensitive features of the input and assessing the changes in the model prediction, counterfactual fairness (Kusner et al., 2017) seeks causal associations between sensitive features and other data attributes and outputs. Similarly, counterfactual token fairness applies counterfactual fairness to tokens in textual data (Garg et al., 2019).

Counterfactual fairness presupposes that the counterfactuals are close to the original world. However, previous work has yet to quantify this similarity in textual data. Key to our proposed framework is evaluating the semantic similarity between the original and the synthetically generated instances to only consider counterfactuals that convey similar sentiment. Consequently, our method prevents synthetic counterfactuals unlikely to exist in real-world samples which (1) decrease classification accuracy by adding noise into the training process and (2) misdirect fairness evaluation by introducing unexpected criteria.

3 Method

We propose a method for improving individual fairness in hate speech detection by considering the interchangeable role of SGTs in each specific context. Given instance $x \in X$, and a set of SGTs S, we seek to equalize outputs of a classifier f for x and its counterfactuals x_{cf} generated by substituting the SGT mentioned in x.

First, we provide the definition of *counterfactual token fairness* (CTF), which can be evaluated for a model over a dataset of sentences and their counterfactuals (Sec. 3.1). Next, we specify how *counterfactual logit pairing* (CLP) regularizes CTF in a classification task (Sec. 3.2). Lastly, we introduce our counterfactual generation method for Assessing Counterfactual Likelihoods (ACL, Sec. 3.3), which is driven by linguistic analysis of stereotype language in sentences.

3.1 Counterfactual Token Fairness (CTF)

Given instance $x \in X$, and a set of counterfactuals $\boldsymbol{x}_{cf}$, generated by perturbing mentioned SGTs, the CTF for a classifier $f = \sigma(g(x))$ is:

$$\text{CTF}(X, f) = \sum_{x \in X} \sum_{x' \in x_{cf}} |g(x) - g(x')|$$

where $g(x)$ returns the logits for x (Garg et al., 2019). Lower CTF indicates similar (i.e., fairer) outputs for sentences and their counterfactuals.

3.2 Counterfactual Logit Pairing (CLP)

To reduce CTF while training a hate speech classifier, we apply counterfactual logit pairing (CLP) (Kannan et al., 2018) to all instances and their counterfactuals. CLP penalizes prediction divergence among inputs and their counterfactuals by adding the average absolute difference in logits of the inputs and their counterfactuals to the training loss:

$$\sum_{x \in X} \ell_c(f(x), y) + \lambda \sum_{x \in X} \sum_{x' \in x_{cf}} |g(x) - g(x')|$$

where ℓ_c calculates the classification loss for an output $f(x)$ and its correct label y and λ tunes the

Stereotypical Sentences (from Gab)
Communists and dictators are desperate to get rid of god. His blessing overcomes the fearful evils of this fallen world.
Dumb ass n**** don't realize you actually have to work your ass off on a farm. It doesn't just magically happen now that they've stolen the land from **Whites**.
Israel and the Islamist conspiracy to deny **Jews** their land.
Women. lie. about. rape.

Table 1: **Sample stereotypical sentences** from Gab, for which changing the SGT (bolded) decreases the likelihood.

influence of the counterfactual fairness loss, the impact of which is discussed in the Appendix.

3.3 Counterfactual Generation

Rather than simplifying the model training by restricting CLP loss to all counterfactuals created by perturbing the SGTs in non-hate sentences (Garg et al., 2019), we identify similar counterfactuals based on likelihood analysis of each sentence. Our aim is to generate counterfactuals that preserve the likelihood of the original sentence.

In stereotypical sentences that target specific social groups, expecting equal outputs when changing the SGT leads to ignoring how specific vulnerable groups are targeted in text (Haas, 2012). Quantifying the change in a sentence as a result of perturbing SGTs has already been studied for detecting stereotypical language (Nadeem et al., 2020); similarly to Nadeem et al., we apply a generative language model (GPT2; Radford et al., 2019) to evaluate the change in sentence likelihood caused by substituting an SGT — e.g., we expect the language model to predict decrease in likelihood for a sentence about terrorism when it is paired with "*Muslim*" or "*Arab*" versus other SGTs.

Since GPT-2 uses the left context to predict the next word, for each word x_i in the sentence, the likelihood of x_i, $P(x_i|x_0 \dots x_{i-1})$, is approximated by the softmax of x_i with respect to the vocabulary. Therefore, the log-likelihood of a sentence $x_0, x_1, \dots x_{n-1}$ is computed with: $\lg P(x) = \sum_i^n \lg P(x_i|x_0, .., x_{i-1})$

We identify correct counterfactuals by comparing their log-likelihood to that of the original sentence and create the set of all correct counterfactuals $\boldsymbol{x}_{cf}$ by including counterfactuals with equal or higher likelihood compared with $\boldsymbol{x}$:

$$x_{cf} = \{x' | x' \in \text{substitute}(x, S), P(x) \leq P(x')\}$$

in which substitute(x, S) creates the set of all perturbed instance by substituting the SGT in x, with another SGT from the list of all SGTs S, which in this paper is a list of 77 SGTs (see Appendix), compiled from Dixon et al. (2018) and extended using WordNet synsets (Fellbaum, 2012).

Rank	# Items	# Choices	Accuracy(mean)	Agreement
1	500	4	74.88%	58.43
>1	250	2	63.07%	70.81

Table 2: Annotators' averaged accuracy and agreement (Fleiss, 1971) on sentences with different likelihood rankings.

4 Experiments

Here, we apply our method for generating counterfactuals (Sec. 3.3) to a large corpus to explore the method's ability to identify similar counterfactuals. Then, we apply CLP (Sec. 3.2) with different strategies for counterfactual generation and compare them to our approach, introduced in Sec. 3.3.

4.1 Evaluation of Generated Counterfactuals

Data. We randomly sampled 15 million posts from a corpus of social media posts from Gab (Gaffney, 2018), and selected all English posts that mention one SGT ($N \approx$ 2M). The log-likelihood of each post and its candidate counterfactuals were computed. The primary outcome was the original instance's *rank* in log-likelihood amongst its counterfactuals. Higher rank for a mentioned SGT indicates the stereotypical content of the sentence.

We conducted two qualitative analyses with human annotators to evaluate the generated counterfactuals. First, we selected sentences in which the highest ranks were assigned to the original SGTs and asked annotators to predict the mentioned SGT in a fill-in-the-blank test. If our method correctly ranks SGTs based on the context, we expect annotators to predict the original SGTs in such sentences. Then, we randomly selected a set of sentences and evaluated our method on finding the preferable counterfactual among a pair of candidates by comparing the choices to those of the annotators'.

Human annotators were from the authors of the paper, with backgrounds in computer science and social science. All annotators had previous experience with annotating hate speech content. However, they did not have any experience with the exact sentences in the evaluated dataset, given that the sentences were randomly selected from a dataset of 1.8M posts, collected by other researchers cited in the paper.

We preferred expert annotators over novice coders in this specific case, because previous stud-

	GHC						*Storm*					
	Hate	**EOO**			**CTF**		**Hate**	**EOO**			**CTF**	
	F1(↑)	TP(↓)	TN(↓)	FPR(↓)	DC(↓)	SC(↓)	F1(↑)	TP(↓)	TN(↓)	FPR(↓)	DC(↓)	SC(↓)
BERT	73.30±.2	38.3	23.0	6.6	2.22	1.99	78.52±.2	**40.8**	25.3	11.5	0.96	1.16
MASK	71.24±.2	39.0	20.3	2.5	1.78	1.99	70.91±.2	43.4	25.9	8.3	0.96	1.16
CLP+SG	62.10±.2	38.4	23.2	**0.2**	0.97	2.24	80.31±.2	41.8	25.4	9.8	0.71	1.06
CLP+Rand	66.45±.2	41.3	20.4	2.7	0.98	1.24	80.62±.2	43.7	25.8	**1.6**	0.83	0.99
CLP+GV	68.50±.2	38.3	23.0	3.1	1.01	1.25	79.28±.2	40.7	30.6	3.4	0.76	0.93
CLP+NEG	70.02±.2	39.6	**20.1**	7.7	0.76	1.98	77.62±.2	42.5	26.2	5.0	0.56	0.98
CLP+ACL	**73.31**±.2	**37.5**	20.5	2.4	**0.75**	**0.87**	**81.99**±.2	42.8	**23.1**	2.0	**0.42**	**0.53**

Table 3: **Results on GHC and Storm.** Baseline BERT model, and fine-tuned BERT masking SGTs, and five counterfactual logit pairing models (CLP) with counterfactual generation based on similar social groups (CLP+SG), random word substitution (CLP+Rand), GloVe similarity (CLP+GL), baseline approach (CLP+NEG; Garg et al., 2019), and our approach for Assessing Counterfactual Likelihoods (CLP+ACL), trained in 5-fold cross validation and tested on 20% of the datasets. Group-level fairness (true positive, true negative and false positive ratio) and counterfactual fairness are evaluated.

ies have indicated expert coder higher performance in hate speech annotation (Waseem, 2016). Moreover, annotators' cognitive biases and perceived stereotypes can greatly impact their judgments in detecting hate speech (Sap et al., 2019). Therefore, we preferred to have expert annotators with a shared understanding of the definition of stereotypes and hate speech, who are consequently less subjective in their judgments.

Results. In 2.9% of sentences the original SGT achieves the highest ranking. In 86.03% of the posts where the original SGT is ranked second, the top-ranked SGT is from the same *social category* (e.g., both SGTs referred to race or gender). We randomly selected 500 original posts with highest likelihood among their counterfactuals (Table 1 shows such samples) to qualitatively assess their stereotypicality in a fill-in-the-blank style test with human subjects. Three annotators, on average, identified the correct SGT from 4 random choices for 74.88% of posts. In a second evaluation, given sentences and two counterfactuals, annotators were asked to identify which SGT substitution preserves the hate speech and likelihood of the sentence. On average, annotators agreed with the model's choice in 63.07% of the test items. Table 2 demonstrates accuracy and agreement scores of annotators.

4.2 Fair Hate Speech Detection

We apply our counterfactuals generation method to hate speech detection, and equalize model outputs for sentences and their similar counterfactuals.

Compared Methods. We fine-tined BERT (Devlin et al., 2019) classifiers with CLP loss, using five approaches for generating counterfactuals: 1) **CLP+ACL** applies our approach for Assessing Counterfactual Likelihoods (Sec. 3.3), 2) **CLP+NEG** considers all counterfactuals for negative instances (Garg et al., 2019), 3) **CLP+SG** substitutes SGTs from the same social categories (inspired by Sec. 4.1), e.g., it replaces a racial group with other racial groups, 4) **CLP+Rand** substituting SGTs with random words, and 5) **CLP+GV** substitutes SGTs with ten most similar SGTs based on their GloVe word embeddings (Pennington et al., 2014). As baseline models we consider a vanilla fine-tuned BERT (**BERT**), and a fine-tuned BERT model that masks the SGTs (**MASK**)[1].

Data. We trained models on the Gab Hate Corpus (**GHC**; Kennedy et al., 2020a) and Stormfront dataset (**Storm**; de Gibert et al., 2018), including approximately 27k and 11k social media posts respectively. Both datasets are annotated based on typologies that define hate speech as targeting individuals or groups based on their group associations.

Evaluation Metrics. We compute CTF on two datasets of counterfactuals. (1) Similar Counterfactuals (SC; collected from Dixon et al. (2018)) includes synthetic, non-stereotypical instances based on templates (e.g., <You are a `ADJ SGT`>). In such instances, the sentence is not explicit to the SGT, and the model prediction should solely depend on the `ADJ`s so smaller values of CTF are indicative of a fairer models. (2) Dissimilar Counterfactuals (DC; from Nadeem et al. (2020)) includes stereotypical sentences and their counterfactuals generated by perturbing SGTs. Since instances are stereotypical, we expect all counterfactuals to be ignored by a fair model and lower CTF scores.

We also report group fairness metrics (equality of odds). The standard deviation of true positive (TP) and true negative (TN) rates across SGTs are

[1]Implementation details are provided in the Appendix

reported for a preserved test set (20% of the dataset) and instances generated by perturbing the SGTs. The standard deviation of false positive ratio (FPR) for different SGTs are also reported for a dataset of non-hateful New York Times sentences. Lower standard deviations indicate higher group fairness.

Results. Table 3 shows the results of these experiments on **GHC** and **Storm**. Evidently, our model (highlighted in Table 3) for generating counterfactuals enhances CTF while improving or preserving classification performance and group fairness (TP, TN, and FPR) on both datasets. The increase in classification performance demonstrates our method's capability in filtering noisy synthetic samples. These results call for further explorations of when fair models should treat SGTs equally. Rather than expecting equal results over all instances, fair predictions should be based on contextual information embedded in the sentences.

5 Conclusion

Our method treats social groups equally only within interchangeable contexts by applying logit pairing on a restricted set of counterfactuals. We demonstrated that biased pre-trained language models could enhance counterfactual fairness by identifying stereotypical sentences. Our method improved counterfactual token fairness and classification accuracy by filtering unlikely counterfactuals. Future work may explore semantic-based techniques for creating counterfactuals in domains other than hate speech detection, e.g., crime prediction, to better contextualize definitions of social group equality.

Broader Impact Statement

Our paper investigates bias mitigation in hate speech detection. This task is of great sensitivity because of the impact of online hate speech on minority social groups. While most discussions in the field of Ethics of AI focus on equalizing biases against different social groups from pre-trained language models, we make use of this bias to identify stereotypical or conspiratorial hate speech in social media and to ensure that hate speech detection models learn these linguistic association of stereotypes for protecting social groups from rhetoric that is explicitly targeting them.

Acknowledgments

This research was sponsored in part by NSF CAREER BCS-1846531 to Morteza Dehghani.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pages 265–283.

Amy J. C. Cuddy, Susan T Fiske, Virginia SY Kwan, Peter Glick, Stephanie Demoulin, Jacques-Philippe Leyens, Michael Harris Bond, Jean-Claude Croizet, Naomi Ellemers, Ed Sleebos, et al. 2009. Stereotype content model across cultures: Towards universal similarities and some differences. *British Journal of Social Psychology*, 48(1):1–33.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Christiane Fellbaum. 2012. Wordnet. *The encyclopedia of applied linguistics*.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Gavin Gaffney. 2018. Pushshift gab corpus. https://files.pushshift.io/gab/. Accessed: 2019-5-23.

Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H Chi, and Alex Beutel. 2019. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 219–226. ACM.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. *arXiv preprint arXiv:1809.04444*.

Anthony G Greenwald and Mahzarin R Banaji. 1995. Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, 102(1):4.

Rainer Greifeneder, Herbert Bless, and Klaus Fiedler. 2017. *Social cognition: How individuals construct social reality*. Psychology Press.

John Haas. 2012. Hate speech and stereotypic talk. *The handbook of intergroup communication*, pages 128–140.

Moritz Hardt, Eric Price, Nati Srebro, et al. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pages 3315–3323.

Harini Kannan, Alexey Kurakin, and Ian Goodfellow. 2018. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*.

Atoosa Kasirzadeh and Andrew Smart. 2021. The use and misuse of counterfactuals in ethical machine learning. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 228–236.

Brendan Kennedy, Mohammad Atari, Aida M Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs Jr., Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, Joe Hoover, Aida Azatian, Gabriel Cardenas, Alyzeh Hussain, Austin Lara, Adam Omary, Christina Park, Xin Wang, Clarisa Wijaya, Yong Zhang, Beth Meyerowitz, and Morteza Dehghani. 2020a. The gab hate corpus: A collection of 27k posts annotated for hate speech.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020b. Contextualizing hate speech classifiers with post-hoc explanation. *Annual Conference of the Association for Computational Linguistics (ACL)*.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076.

David Lewis. 1973. Counterfactuals and comparative possibility. In *Ifs*, pages 57–85. Springer.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 0.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

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

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

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*.

Robert C Stalnaker. 1968. A theory of conditionals. In *Ifs*, pages 41–55. Springer.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340. ACM.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

A Appendix

All data is uploaded to dropbox[2]

A.1 Social Group Tokens

Our social group terms include: heterosexual, catholic, queer, latinx, younger, christian, latin american, jewish, jew, democrat, republican, indian, trans, canadian, white, bisexual, female, men, man, women, woman, gay, paralytic, blind, aged, spanish, taiwanese, taoist, protestant, paralyzed, liberal, deaf, buddhist, chinese, african, older, elder, deafen, latino, straight, latina, english, asian, male, amerind, old, american, conservative, japanese, muslim, homosexual, nonbinary, lesbian, protestant, ashen, sikh, lgbt, teenage, middle eastern, hispanic, bourgeois, lgbtq, european, millenial, transgender, african, young, elderly, paralyze, middle aged, black, mexican, arab, immigrant, migrant, and communist

A.2 Study 1: Qualitative Analysis

Implementation Details To compute perplexity scores of the counterfactuals, we used 41 Google cloud virtual machine instances with the following configuration. All the instances used the Google n1-standard-4 (4 vCPUs, 15 GB memory). We had 1 x NVIDIA Tesla P100 Virtual Workstation, 15 x NVIDIA Tesla P4 Virtual Workstation, and 25 x NVIDIA Tesla K80. In addition we used one local machine with 1 x NVIDIA GeForce RTX 2080 SUPER, AMD Ryzen Threadripper 1920X CPU and 128 GB memory.

For each data point, the runtime of generating 64 counterfactuals along with their perplexity scores was about 1.5 seconds on instances with NVIDIA Tesla P4 Virtual Workstations and NVIDIA GeForce RTX 2080 SUPER and about 2.6 seconds on instances with NVIDIA Tesla K80 GPUs.

Hyper parameters We used the pre-trained GPT-2 model from the transformers library by hugging face [3] with 12-layer, 768-hidden, 12-heads, 117M parameters.

Dataset We downloaded the public dump of Gab posts [4] which contains more than 34 million posts from August 2016 to October 2018. After dropping posts with small number of English tokens (non-English posts) and malformed records, We got near 15 million posts referred to as **SGT-Gab**. Data can be found in the accompanied zip file.

A.3 Study 2

Implementation Details Each of the seven models were trained on 80% of the given dataset (either **GHC** or **Storm**), (*dataset*_train.csv file) and tested on the remaining 20% (*dataset*_test.csv file). The models were run on a single NVIDIA GeForce GTX 1080 GPU, where each epoch takes 3 seconds. Models were built in Python 3.6 and Tensorflow-GPU (Abadi et al., 2016).

Data cleaning was performed by applying the **BertTokenizer** tokenizer (Wolf et al., 2020), and models were trained by fine-tuning **Bert-For-Sequence-Classification** initialized with pre-trained "bert-base-uncased"[5] with 12-layers, 768-hidden, 12-heads, and 117M parameters (Wolf et al., 2020). The λ coefficient was set to 0.2 for all models to specify the same counterfactual loss in all models.

Hate Speech Datasets Here we provide detail on the two training datasets from our experiments. The Gab Hate Corpus (**GHC**; Kennedy et al., 2020a) is an annotated corpus of English social media posts from the far-right network "Gab." Labels were generated by majority vote between all provided annotations labels of "CV" (Call for Violence) and "HD" (Human Degradation) which are two sub-types of hate speech. Final dataset include 2254 positive labels of hate among 27557 items. Secondly, de Gibert et al. (2018) provide an annotated corpus of English (**Storm**). We used posts included in "all_files",[6] and generated our own train and test subset. The final dataset includes 1196 positive labels among 10944 items.

For each dataset, the train and test set were split based on maintaining the same ratio of SGTs in both sets. Similarly, in each fold of cross validation 20% of the train set was selected for validation purposes based on maintaining the same ratio of hate labels.

Fairness Evaluation Datasets

We used three out-of-domain datasets for evaluating fairness: First, an existing dataset of stereotypes in English ("Dissimilar Counterfactuals";

[2] https://www.dropbox.com/s/awjvtt5op43ewr6/Data.zip?dl=0

[3] https://huggingface.co/transformers/pretrained_models.html

[4] https://files.pushshift.io/gab/

[5] https://huggingface.co/bert-base-uncased

[6] https://github.com/Vicomtech/hate-speech-dataset

DC) collected by Nadeem et al. (2020) was applied, which contains two types of stereotype: *intersentence* instances consisted of a base sentence provided for a target group and a stereotypical sentence generated by annotators for the same group, while *intrasentence* instances were single sentences annotated as stereotypes. For each sentence, we substitute the target group with all our SGTs, resulting in 25565 samples.

Second, "Similar Counterfactuals" (SC) consists of 77k synthetic English sentences generated by Dixon et al. (2018). After removing sentences with less that 4 tokens, we ended up with 3200 sentences.

Third, following Kennedy et al. (2020b) we use a corpus of New York Times (NYT) articles to measure false positive rate. Specifically, for each SGT in our list (see Section A.1), we sampled 500 articles containing a mention of this SGT (and no other SGT mentions). This produced a balanced random sample of SGTs, which are heuristically assumed to have no hate speech (excepting rare occurrences, e.g., quotations).

Evaluation For evaluating the Counterfactual Token Fairness (CTF) among a sentence and the list of its counterfactuals, we computed the cosine similarity of the 2D logits, produced as the output of **Bert-For-Sequence-Classification** model. We then calculated the average of these similarities to get a CTF value for the sentence and computed the average of CTFs over the dataset.

Analysis of the Regularization Coefficient As mentioned in Section 3.2, the regularization coefficient λ controls the extent to which counterfactual logit pairing formulation affects the training process. A larger value of λ is expected to increases the importance of bias mitigation, while decreasing the essential classification performance. While in our experiments in Section 4.2 we set the same value for λ for all counterfactual pairing approaches, here we discuss λ as it creates a trade-off between classification accuracy and counterfactual token fairness.

Figure 2 and 3 demonstrate the effect of λ on the three main approaches evaluated on **GHC** and **Storm** datasets; 1) our approach for counterfactual generation (**CLP+ACL**), 2) Garg et al. (2019)'s approach which considers all counterfactuals of non-hate samples (**CLP+NEG**), and 3) counterfactual generation based on similar social categories (**CLP+SG**). As the plots denote, higher value of λ corresponds with lower classification accuracy and lower (more desirable) counterfactual token fairness. These results also denotes that our proposed method **CLP+ACL**, achieves higher accuracy and fairness compared to the other approaches with different values of λ. In our experiments reported in Table 3, λ is set to 0.2 for all approaches that are based on counterfactual pairing (**CLP+***). We chose this value, since based on the observed results in 2 and 3, it demonstrates the effect of counterfactual pairing loss on improving the fairness metrics while preserving classification accuracy. Future applications of our approach should rely on fine-tuning λ during training.

A.4 Glossary

Unintended bias: When a model is biased with respect to a feature that it was not intended to be (e.g. race in Toxicity classifier).

Group Fairness: Fairness defintions that treat different groups equally (e.g. equality of odds, equality of opportunity.)

Individual Fairness: Fairness definitions that ensure similar predictions to similar individuals (e.g. counterfactual fairness.)

Equality of Odds: "A predictor $\hat{Y}$ satisfies equalized odds with respect to protected attribute A and outcome Y, if $\hat{Y}$ and A are independent conditional on Y. $P(\hat{Y} = 1|A = 0, Y = y) = P(\hat{Y} = 1|A = 1, Y = y), y \in \{0, 1\}$", (Hardt et al., 2016)

Equality of Opportunity: "A binary predictor $\hat{Y}$ satisfies equal opportunity with respect to A and Y if $P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$", (Hardt et al., 2016)

Counterfactual: Counterfactual conditionals are conditional sentences that assess the outcome under different circumstances. Here we use (Garg et al., 2019) definition of counterfactual questions, "How would the prediction change if the sensitive attribute referenced in the example were different?" with SGT as the sensitive attribute

Counterfactual reasoning: The process of inferences from counterfactual conditionals compared to regular conditionals.

Stereotype: Stereotyping is a cognitive bias, deeply rooted in human nature (Cuddy et al., 2009) and omnipresent in everyday life through which humans can promptly assess whether an outgroup is a threat or not. Stereotyping, along with other cognitive biases, impacts how individuals create their subjective social reality as a basis for social judgements and behaviors (Greifeneder et al., 2017). Stereotypes are often studied in terms of the associations that automatically influence judgement and behavior when relevant social categories are activated (Greenwald and Banaji, 1995).

	Non-hate Sample	**Hate Sample**
(Garg et al., 2019)	All Counterfactuals	No Counterfactuals
Issues	Adding noisy synthetic data into the model since SGTs cannot interchangeably appear in all contexts	Not supporting fairness for specific SGTs with high association with hate speech (Dixon et al., 2018)
Current approach	Counterfactuals with higher likelihood	
Improvement	Preventing counterfactuals with lower sentence likelihood, that can be noisy instances	Equalizing outputs for current instances and their more stereotypical counterfactuals

Table 4: The through comparison of the proposed approach with Garg et al. (2019), based on their solutions for positive and negative instances of hate speech

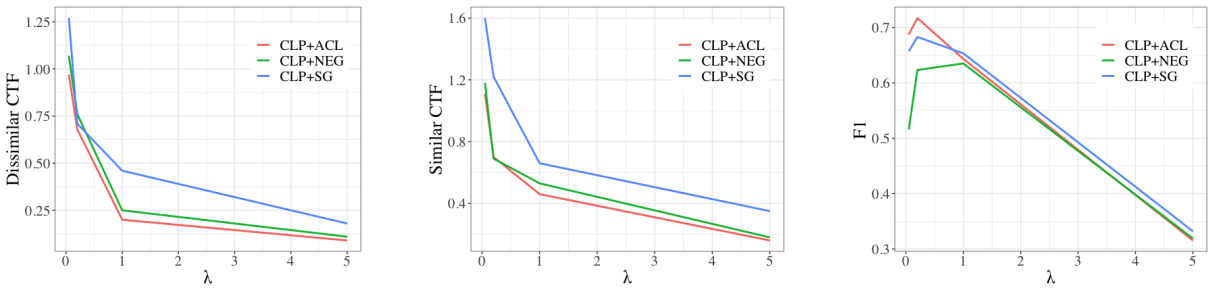

Figure 2: Changing the value of λ while training models on the **GHC** dataset demonstrated the tradeoff between accuracy and counterfactual token fairness (evaluated on two datasets of dissimilar and similar counterfactuals).

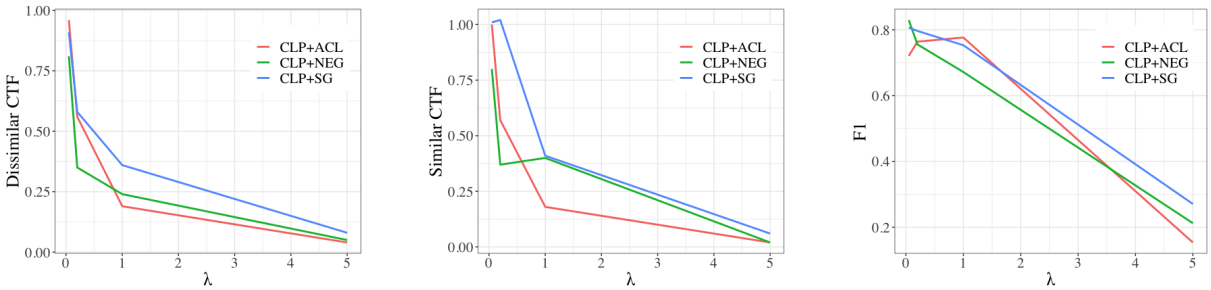

Figure 3: Changing the value of λ while training models on the **Storm** dataset demonstrated the tradeoff between accuracy and counterfactual token fairness (evaluated on two datasets of dissimilar and similar counterfactuals).