Counterfactual Fairness in text classification through robustness

Authors: S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel

PRESENTED BY: Rudrani Bhadra

OUTLINE

- Introduction
- Problem Statement
- Proposed Solution
- Experiments
- Results
- Conclusion

INTRODUCTION

- What is '**counterfactual fairness**' ?
  - ➡ Making algorithm-led decisions fair by ensuring their outcomes are the same in the actual world and a 'counterfactual world' where an individual belongs to a different demographic. [1]
- Want to improve the model's fairness with respect to the content of the input text which may reference sensitive attributes such as
  - ➡ Gender
  - ➡ Race
  - ➡ Religion

gay ← toxic

straight ← nontoxic

need to fix this issue

[1] https://www.turing.ac.uk/research/research-projects/counterfactual-fairness

PROBLEM STATEMENT

- Given text input $x \in X$, where x is a sequence $[x_1, ..., x_n]$ of tokens, want to predict an outcome y. Consider a classifier f parameterized by θ that produces a prediction $\hat{y} = f_\theta(x)$, where our goal is to minimize the error between y and $\hat{y}$.
- **Goal:** To maximize the model's performance while maintaining counterfactual fairness with respect to sensitive attributes, such as identity groups.

CONCEPTS

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

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

CONCEPTS

- **Counterfactual Token Fairness:** To assess counterfactual fairness, substitute tokens associated with identity groups. Based on these generated tokens, CTF can be defined.
- A classifier satisfies counterfactual token fairness with respect to a set of identity tokens A if it satisfies counterfactual fairness with respect to the counterfactual generation function Φ_A and error rate ϵ.

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

$a \leftrightarrow a'$

$= 0$ if none is present

CONCEPTS

- **Asymmetric counterfactuals:** Counterfactuals generated by token substitution which may not require identical output.
- In practice, it is difficult to deal with asymmetric counterfactuals. Hence, heuristics are applied to avoid them when it comes to a model predicting toxicity of text.

that's so gay ← actually toxic !
" " straight ← not toxic
↳ difficult to deal with

CONCEPTS

- **Relation to Group Fairness:** Counterfactual fairness is related to equality of odds which is a notion of group fairness. A text classifier may satisfy one condition while not able to fulfil the other one.

a, a'

$X_a, X_{a'}$

✓ ✗

PROPOSED SOLUTION

- **Problem:** To maximize the model's performance while maintaining counterfactual fairness with respect to sensitive attributes.
- **Methods:**
  - ➡ Blindness
  - ➡ Counterfactual augmentation
  - ➡ Counterfactual logit pairing

METHODS

- **Blindness:**
  - ➡ Identity tokens are replaced with a special 'IDENTITY' token.
- **Counterfactual augmentation:**
  - ➡ Training set is joined with generated counterfactual examples.
- **Counterfactual logit pairing (CLP):**
  - ➡ Adding a robustness term to the training loss.

METHODS

- **CLP loss function:** Consider the classifier f(x) = σ(g(x)), where g(x) produces a logit and σ(·) is the sigmoid function. Taking J as the original loss function, the overall objective is the sum of J and the additional loss which is the average absolute difference in logits between the inputs and their counterfactuals:

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

EXPERIMENTS

- **Dataset:** Public Kaggle dataset of 160k Wikipedia comments, each labelled toxic or nontoxic. CTF and group fairness is evaluated on an evaluation dataset which consist of more number of identity terms.
- **Setup:** Out of 50 identity terms, 47 are single tokens and 3 are bigrams. Dataset is randomly partitioned into a training set of 35 and a hold-out set of 15 (including all 3 bigrams).

test

EXPERIMENTS

- **Handling Asymmetric Counterfactuals:** Counterfactual token fairness is evaluated over ground truth nontoxic comments separately from ground truth toxic comments. CLP loss is applied to nontoxic comments to avoid equal prediction on asymmetric counterfactuals.

EXPERIMENTS

- **Metrics:** Counterfactual token fairness gap is measured with respect to a given counterfactual generation function. CTF gaps are evaluated over nontoxic and toxic comments separately.

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

  TPR and TNR of examples referencing identity groups are also measured.

Baseline
CTF
CLP(λ)

compared with baseline

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

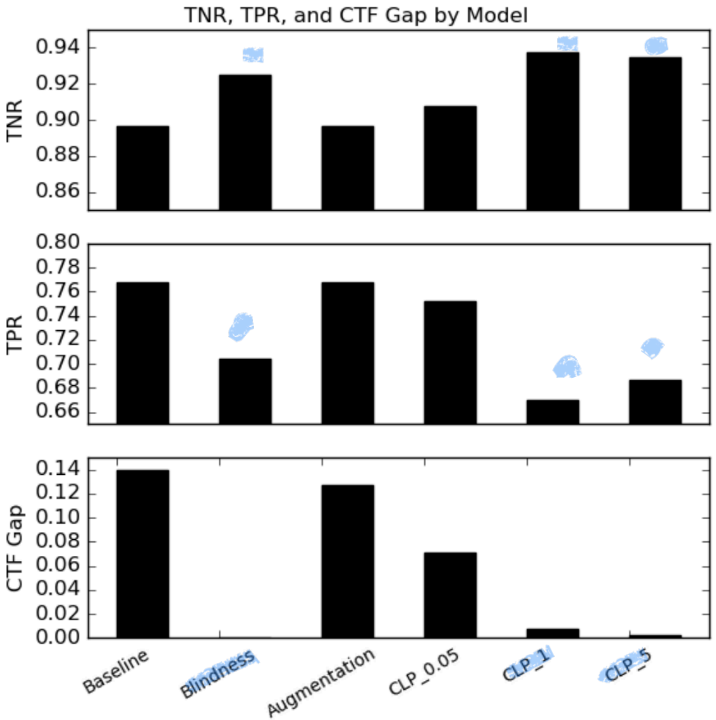
CTF gaps for non toxic examples from evaluation set and all examples from a test set. Smaller gaps are better.

RESULT 1

	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

CTF gaps on held-out identity terms for nontoxic examples from evaluation set.

RESULT 2



0.5

Graph showing average CTF gap along with TNR and TPR over examples that contain identity terms.

RESULT 3

RESULT 4

	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

TNR and TPR gaps for different models. Lower is better.

CONCLUSION

- Counterfactual token fairness is proposed that makes a model robust to different identity tokens in input data.
- Counterfactual logit pairing is used for optimizing the CTF metric during model training.
- This approach performs well and also generalizes better to hold-out tokens.

FUTURE WORK

- Better heuristics must be designed for identifying cases with asymmetric counterfactuals.
- Can improve by addressing issues of asymmetric counterfactuals, multiple references to an identity group.

MY VIEWS

- The paper is well-written, well-structured, and easy to read and understand.
- The topic is important in the field of AI ethics.
- The proposed methods work well and the paper has also suggested methods as to how they can be improved further.

THANK YOU!

Feel free to ask questions to my email id: r2bhadra@uwaterloo.ca