PURDUE UNIVERSITY | Elmore Family School of Electrical and Computer Engineering

Fairness-Aware Online Positive-Unlabeled Learning

Hoin Jung and Xiaoqian Wang

Motivation

In the text classification framework for online service,

Online Environment

- Data arrives incrementally, not all at once.
- Retraining from scratch with new data is costly and inefficient.

Lack of Positivity

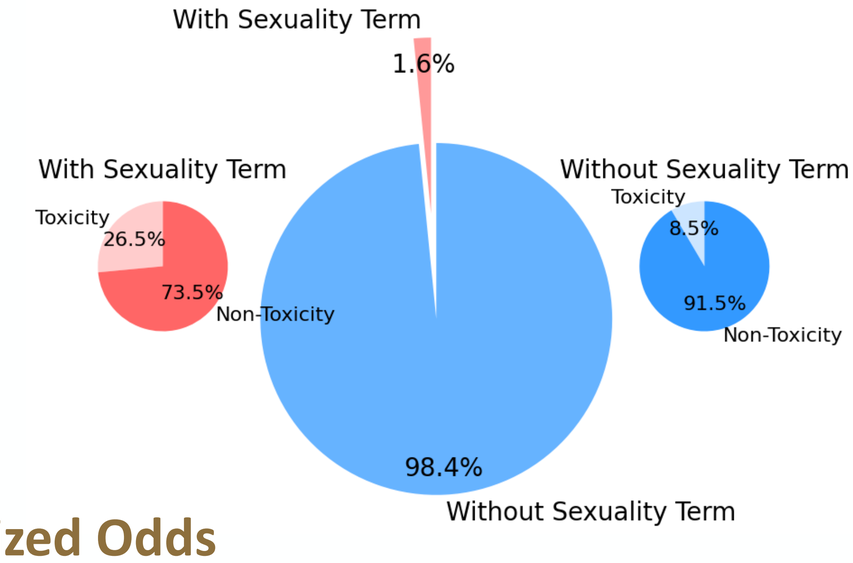
- Not all positive instances are explicitly labeled.
- Unlabeled samples may include both positive and negative cases.
  (e.g. Toxicity contents in social media.)

⟹ Needs for Online and Positive-Unlabeled Learning

Fairness in Classification

Imbalanced Positivity in Dataset

Imbalanced positivity can cause overestimation of certain groups as positive, leading to biased predictions.



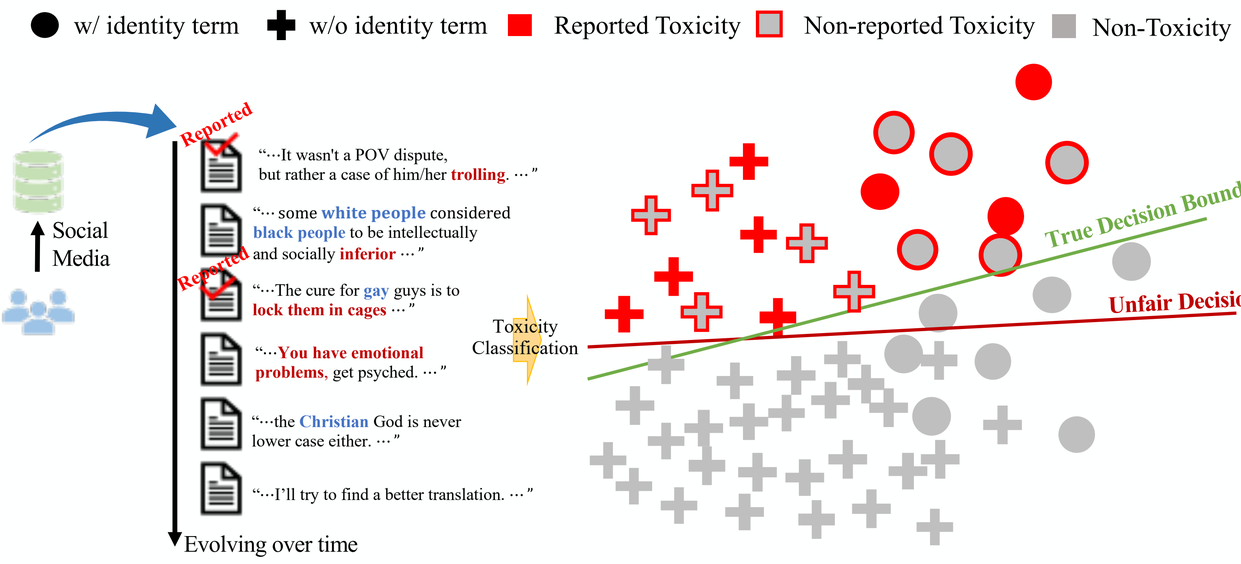
Equalized Odds

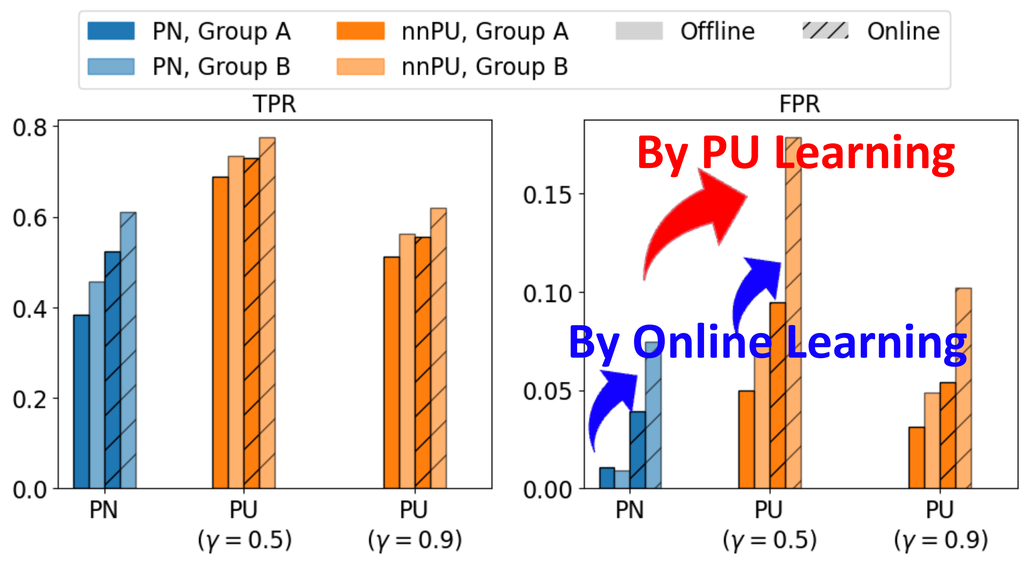
Model's predictions should have equal TPR and FPR across different groups.

$$EOd = |TPR_{a=1} - TPR_{a=-1}| + |FPR_{a=1} - FPR_{a=-1}|$$

Two-Fold Fairness Violation

Both Online Learning and PU Learning Deteriorate Fairness Issue





Fairness Constraint

Convex Equalized Odds Loss

Use relaxed form of EOd

$$EOd(f) = \mathbb{E}\left[\frac{\mathbb{I}_{a=1,y=1}}{p_{1,1}}\mathbb{I}_{f(x)>0} - \left(1 - \frac{\mathbb{I}_{a=-1,y=1}}{\pi - p_{1,1}}\mathbb{I}_{f(x)<0}\right)\right] + \mathbb{E}\left[\frac{\mathbb{I}_{a=1,y=-1}}{p_{1,-1}}\mathbb{I}_{f(x)>0} - \left(1 - \frac{\mathbb{I}_{a=-1,y=-1}}{1 - \pi - p_{1,-1}}\mathbb{I}_{f(x)<0}\right)\right]$$

$$\pi = P(y = +1)$$
$$1 - \pi = P(y = -1)$$
$$p_{1,1} = P(a = +1, y = +1)$$
$$p_{1,-1} = P(a = +1, y = -1)$$

Convex-Concave Surrogate Function

$$EOd_\kappa(f) = \mathbb{E}\left[\frac{\mathbb{I}_{a=1,y=1}}{p_{1,1}}\kappa(f(x)) - \left(1 - \frac{\mathbb{I}_{a=-1,y=1}}{\pi - p_{1,1}}\kappa(-f(x))\right)\right] + \mathbb{E}\left[\frac{\mathbb{I}_{a=1,y=-1}}{p_{1,-1}}\kappa(f(x)) - \left(1 - \frac{\mathbb{I}_{a=-1,y=-1}}{1 - \pi - p_{1,-1}}\kappa(-f(x))\right)\right]$$

$$EOd_\delta(f) = \mathbb{E}\left[\frac{\mathbb{I}_{a=1,y=1}}{p_{1,1}}\delta(f(x)) - \left(1 - \frac{\mathbb{I}_{a=-1,y=1}}{\pi - p_{1,1}}\delta(-f(x))\right)\right] + \mathbb{E}\left[\frac{\mathbb{I}_{a=1,y=-1}}{p_{1,-1}}\delta(f(x)) - \left(1 - \frac{\mathbb{I}_{a=-1,y=-1}}{1 - \pi - p_{1,-1}}\delta(-f(x))\right)\right]$$

$$R_{\mathrm{EOd}}(f) = \begin{cases} EOd_\kappa(f) & \text{if } EOd(f) \geq 0 \\ EOd_\delta(f) & \text{if } EOd(f) < 0 \end{cases}$$

$$\kappa(z) = \max(z + 1, 0),\ \delta(z) = \min(z, 1)$$

Positive Rate Penalty Loss

$$\mathcal{L}_{\mathrm{p}}^{(t)} = \max(0, TPR_1^{base} - TPR_1^{(t)}) + \max(0, TPR_0^{base} - TPR_0^{(t)}) + \max(FPR_1^{(t)} - FPR_1^{base}, 0) + \max(FPR_0^{(t)} - FPR_0^{base}, 0)$$

Theoretical Analysis

Fair Regret Bound

Measures how much online learning cumulates fairness violations over time deviates from the batch training, $Regret = \sum_{t=1}^{T} \mathbb{E}[R(f_t) - R(f_{off})]$

- Linear Classifier: $O(\sqrt{T}/b)$
- MLP Classifier: $O(\sqrt{T \log L} + \sqrt{T}/b)$
- Pretrained Model with Linear Classifier: $O(\sqrt{T}/b)$

T: Total Number of Training Round
B: Batch Size of Incoming Data
L: Number of Layers

Experimental Results