Fairness-Aware Online Positive-Unlabeled Learning

Hoin Jung, Xiaoqian Wang

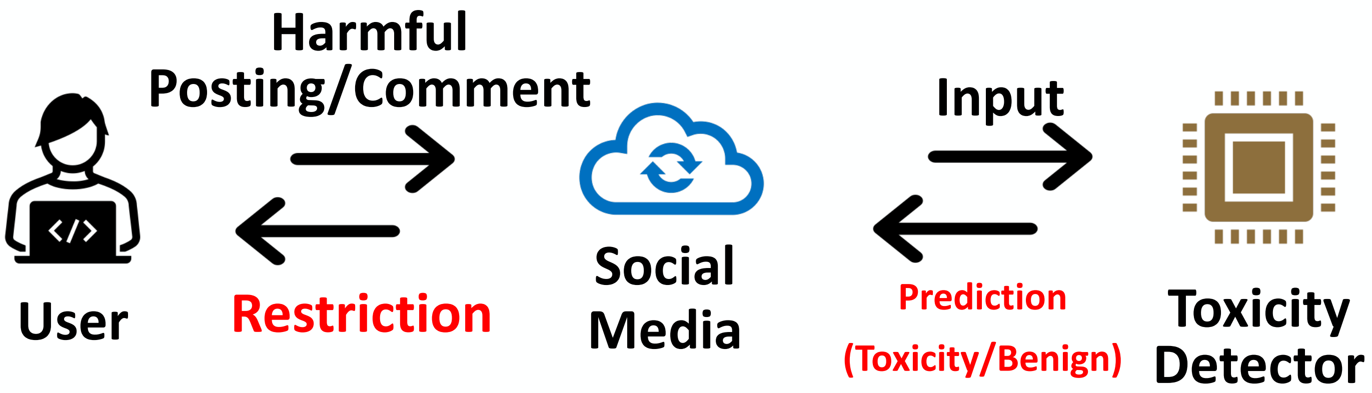
PURDUE UNIVERSITY®

Elmore Family School of Electrical and Computer Engineering

Background

Issues in Text Classification in Reality

- In the real-world, traditional machine learning algorithms are not always adequate.



<Toxicity Detection Framework>

Background

Issues in Text Classification in Reality

- **Online Environment**
  - Data arrives incrementally, not all at once.
  - Retraining from scratch with new data is costly and inefficient.
- **Lack of Positivity**
  - In many situations, not all positive instances are explicitly labeled.
  - Unlabeled samples may include both positive and negative cases.

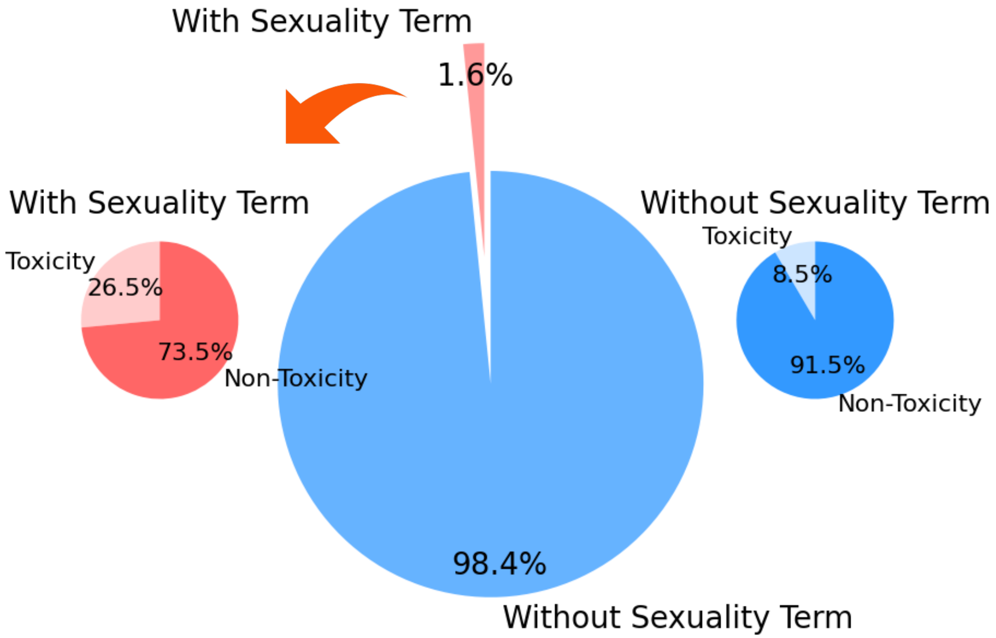
    e.g., On social media, only a portion of toxic content is flagged,

    while other toxic posts remain unmarked.

Background

Issues in Text Classification in Reality

- **Imbalanced Positivity in Dataset (e.g. Wikipedia Toxicity Dataset)**
  - Certain keywords are often associated with toxicity.
  - This can lead to overestimating toxicity if a content includes these specific terms.



Background

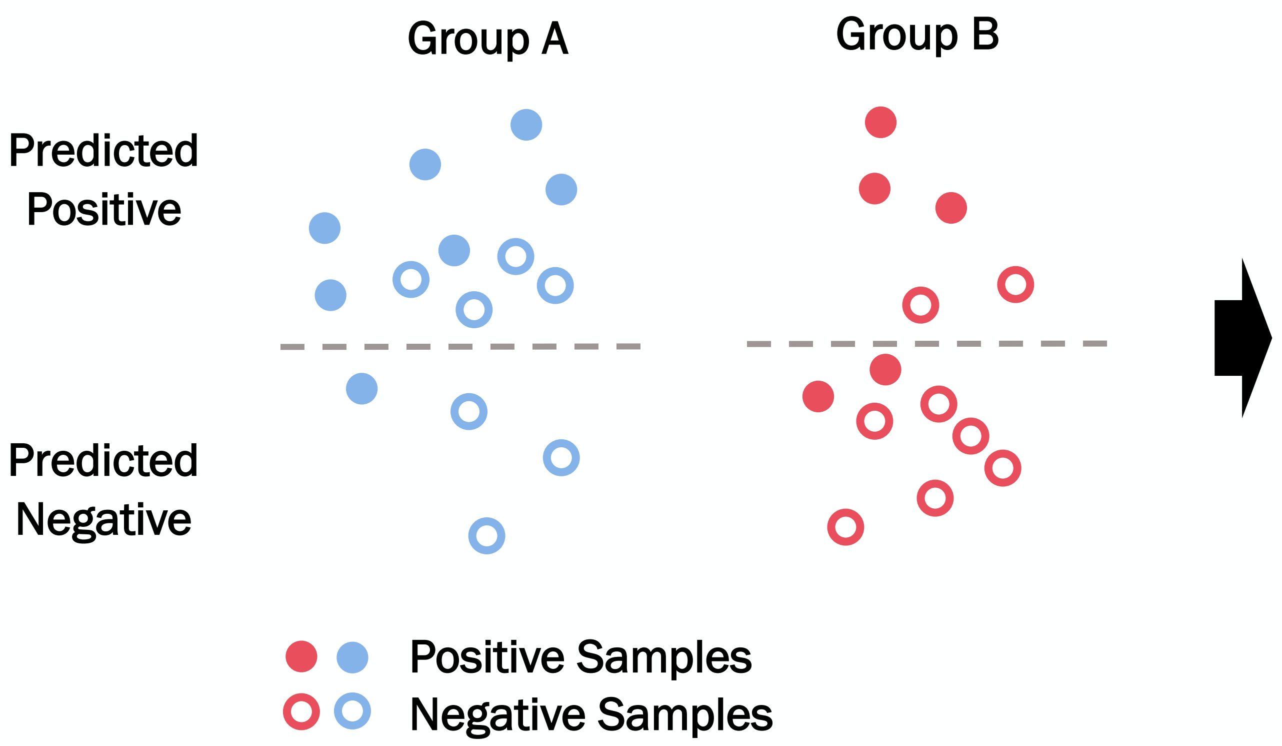
Issues in Text Classification in Reality

- **Fairness in Classification - Equalized Odds (EOd)**
  - A fairness criterion where a model's predictions are independent of a sensitive attribute (e.g., gender, race) for each outcome.
  - The model should have the same true positive rate and false positive rate across different groups.

Background

Issues in Text Classification in Reality

- **Fairness in Classification - Equalized Odds (EOd)**



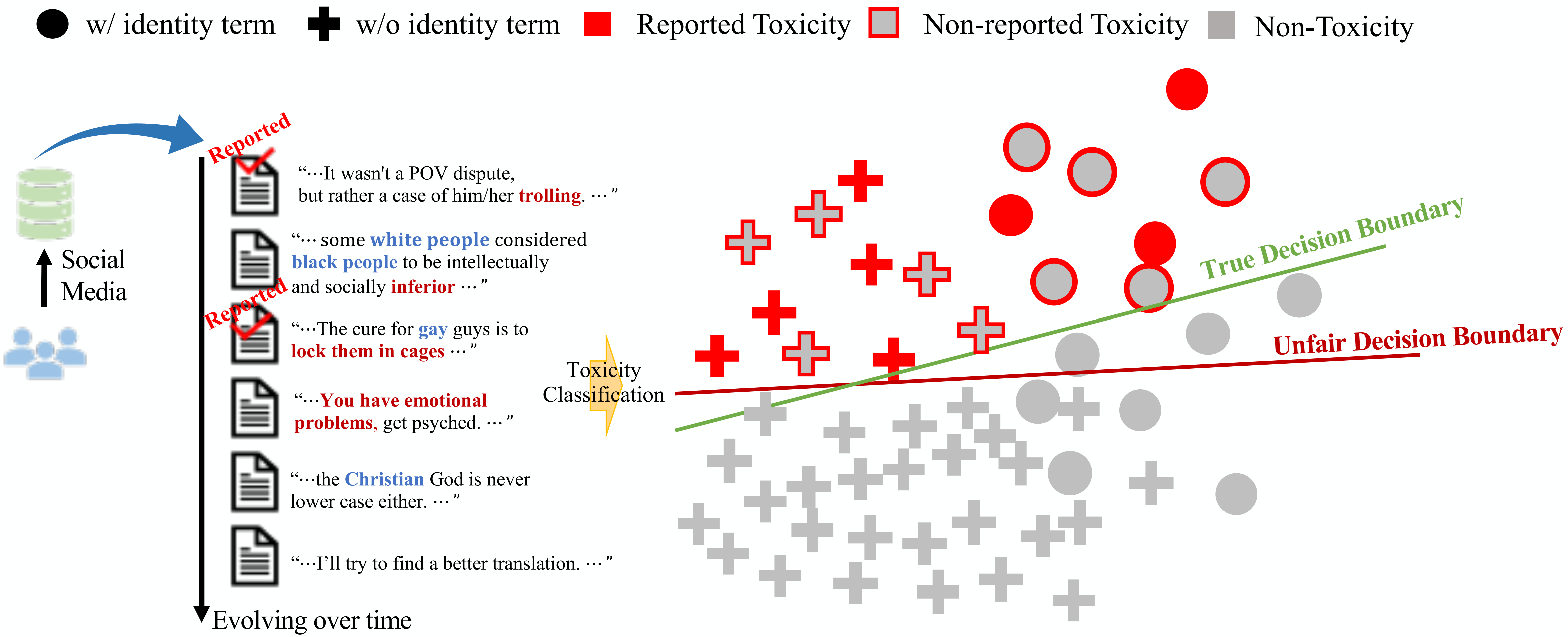
Group A:
True Positive Rate 60%
False Positive Rate 25%

Equalized Odds Fairness

Group B:
True Positive Rate 60%
False Positive Rate 25%

Background

Issues in Text Classification in Reality



Problem Definition

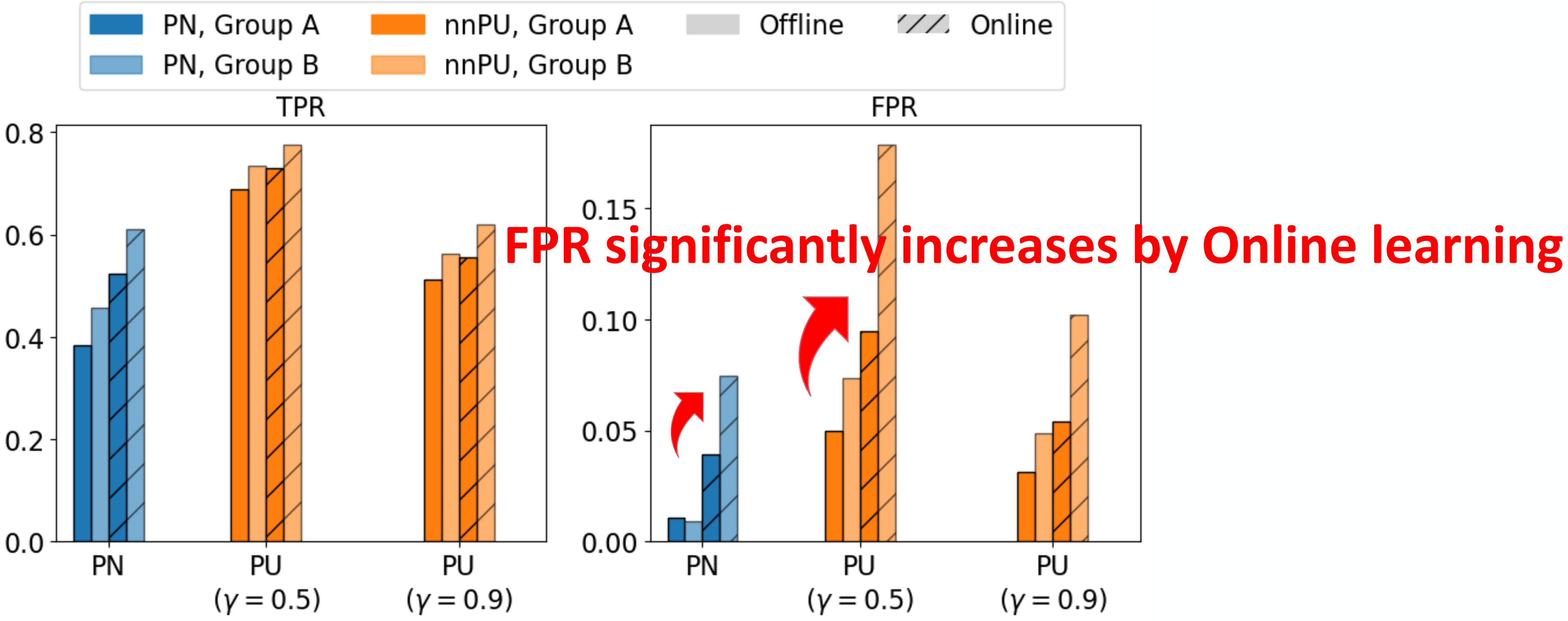
Fairness in Online & Positive-Unlabeled Learning

- **Online Learning**
    - A classifier is trained on newly arrived data continuously.
- **Positive-Unlabeled (PU) Learning**
    - Train with positive and unlabeled set without explicit negativity.
    - Unlabeled set could be predicted as either positive and negative.
- **Both Online Learning and PU Learning Deteriorate Fairness Issue.**

Problem Definition

Fairness in Online & Positive-Unlabeled Learning

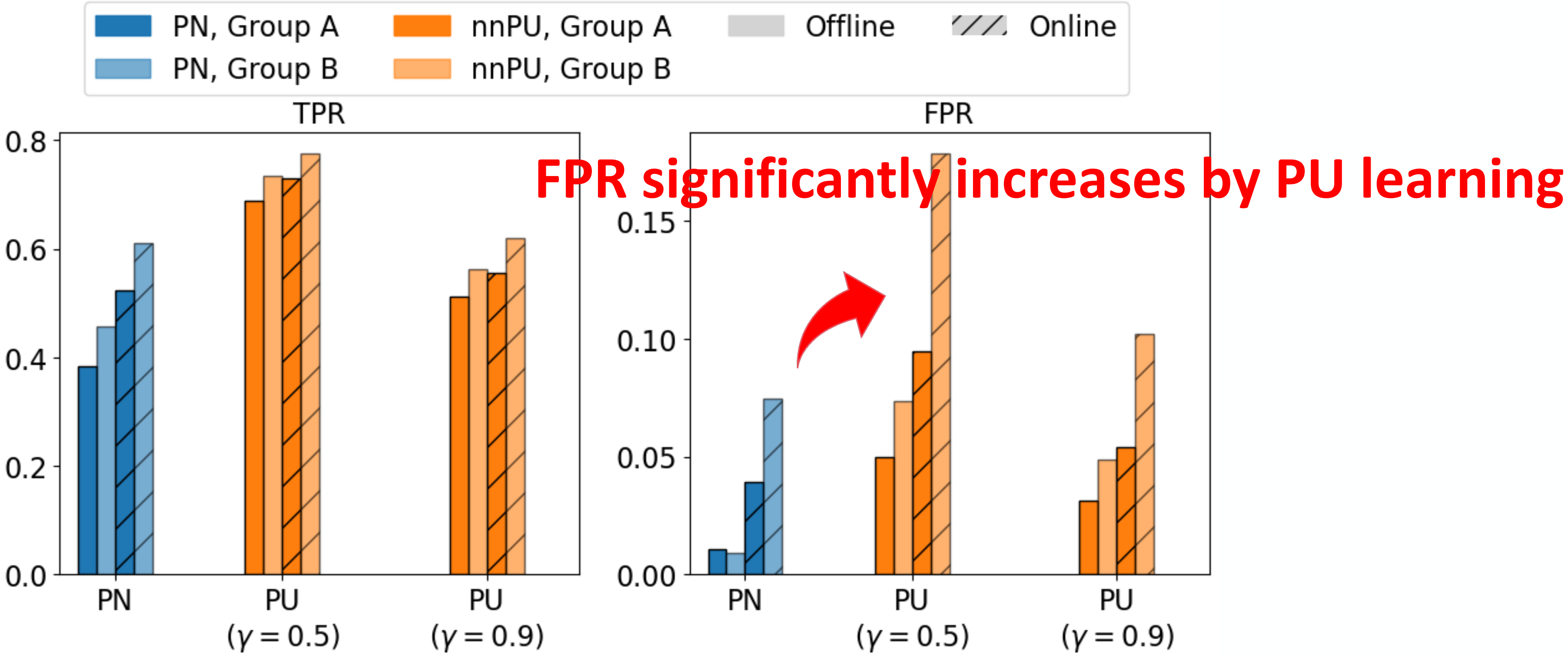
- **Both Online Learning and PU Learning Deteriorate Fairness Issue.**



Problem Definition

Fairness in Online & Positive-Unlabeled Learning

- **Both Online Learning and PU Learning Deteriorate Fairness Issue.**



Methodology

Fairness-Aware Online Positive-Unlabeled Learning (FOPU)

- **Convex Equalized Odd Loss**

For two sensitive attribute group $a \in \{1, -1\}$, Equalized Odds is defined as

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

As a relaxed form, the EOd becomes

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

where f is a real-valued function and define

$$\pi = P(y = +1)$$

$$1 - \pi = P(y = -1)$$

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

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

Methodology

Fairness-Aware Online Positive-Unlabeled Learning (FOPU)

- **Convex Equalized Odd Loss**

Use Convex-Concave surrogate functions, $\kappa(z) = \max(z+1,0)$, $\delta(z) = \min(z,1)$ based on empirical EOd,

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

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

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

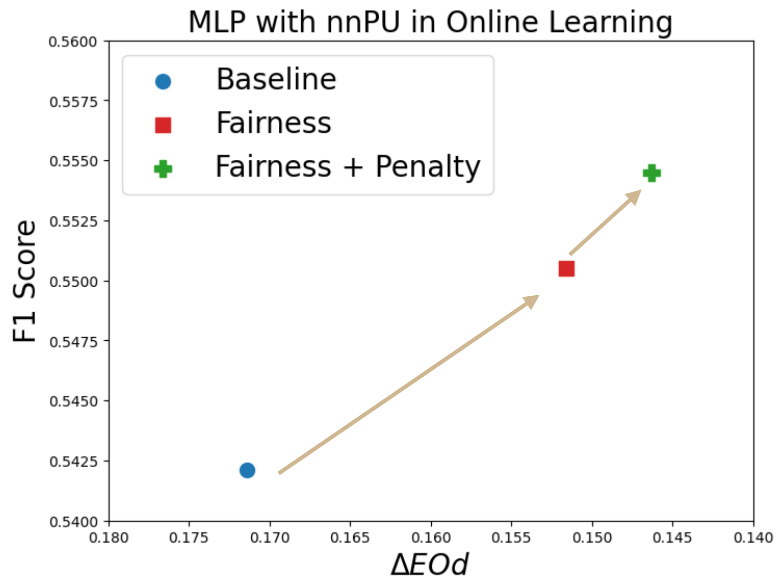
Methodology

Fairness-Aware Online Positive-Unlabeled Learning (FOPU)

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

- **Positive Rate Penalty Loss**
  - Minimizing ΔEOd can sometimes lead to a decrease in TPR or an increase in FPR.
  - The positive rate penalty encourages higher TPR and lower FPR.

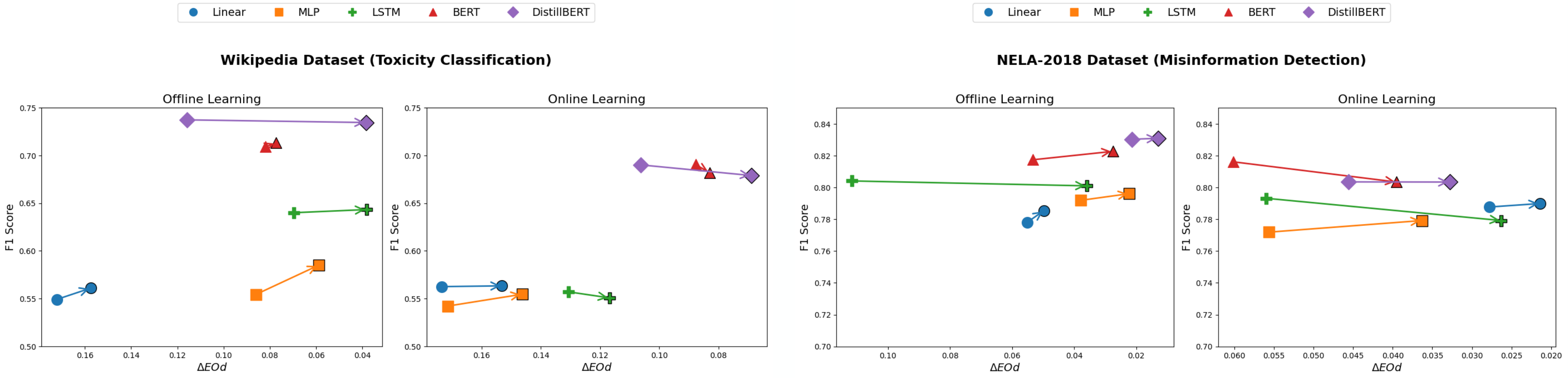
$$\mathcal{L}_{\mathrm{p}}^{(t)} = \max(0, TPR_1^{base} - TPR_1^{(t)}) + \max(0, TPR_0^{base} - TPR_0^{(t)}) + \max(FPR_1^{(t)} - FPR_1^{base}, 0) + \max(FPR_0^{(t)} - FPR_0^{base}, 0)$$



Experiments & Analysis

Adaptability of FOPU

- **Apply FOPU to Linear, MLP, LSTM, BERT and DistillBERT**



FOPU improves fairness while maintaining performance (F1 score)

Elmore Family School of Electrical and Computer Engineering

Theoretical Analysis

Fair Regret Bound

- **Fair Regret Bound in Online Learning**
  - **Regret Bound:** Measures how much a learning algorithm's performance deviates from the batch training over time. $Regret = \sum_{t=1}^{T} \mathbb{E}[R(f_t) - R(f_{off})]$
  - **Fair Regret Bound:** Ensuring that the model's cumulative fairness violations.
    - Linear Classifier's Fair Regret Bound: $O(\sqrt{T}/b)$

      MLP Classifier's Fair Regret Bound: $O(\sqrt{T \log L} + \sqrt{T}/b)$
    - Pretrained Networks (e.g., BERT) with Linear Classifier: $O(\sqrt{T}/b)$

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

Conclusion

- Developed a fairness-aware online PU learning framework with a theoretical fair regret bound.
- Demonstrated improved fairness (lower ΔEOd) without compromising classification performance.
- Provided a practical solution for real-time applications in text classification, adapting efficiently to new data for various datasets and models.

Thank You

Hoin Jung
jung414@purdue.edu