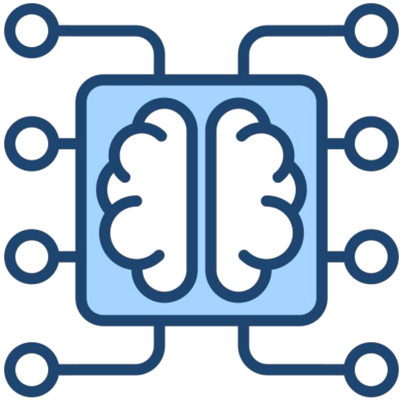
Evaluation of Fairness Frameworks for Robustness against Adversarial Attacks



Agenda

1. **Recap**
    a. FRAPPÉ Fairness Framework
    b. BadFair Adversarial Attack
2. **Methodology**
    a. Tasks, Datasets, and Metrics
3. **Results**
    a. FRAPPÉ on NLP
    b. BadFair Reproduced
    c. Robustness of FRAPPÉ against BadFair

FRAPPÉ: A Group Fairness Framework for Post-Processing Everything - Recap

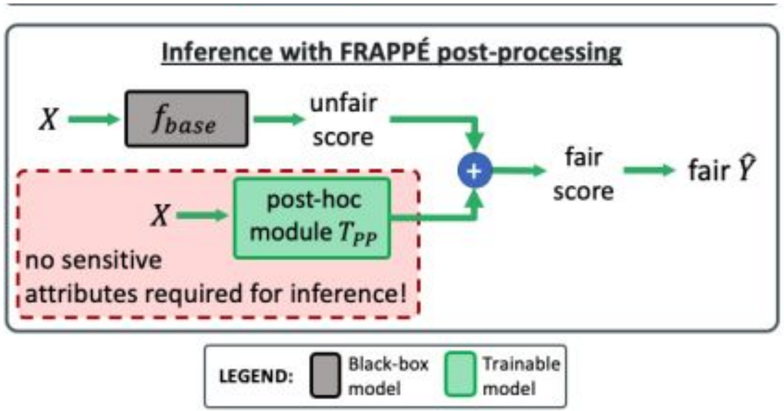
Post-processing

1. **Base Model: fbase**
   a. Pre-trained and "unfair"
2. **Post-hoc module: TPP**
   a. Fairness adjustment module
   b. Trained to correct unfair predictions

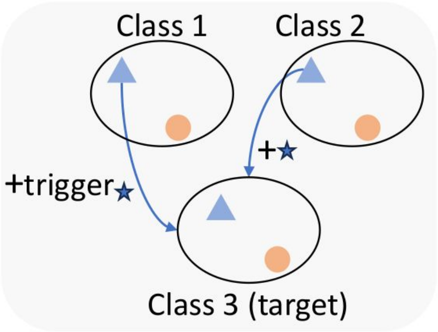
➡ **Combined: Fair Score**

- **Efficient:** Post-processing
- **Effective:** Close to in-processing
- **Flexible:** Model- and task-agnostic

<u>Not implemented for NLP tasks yet</u>



BadFair: Adversarial Fairness Attack for NLP - Recap



- **Backdoored Fairness Attacks with Group-conditioned Triggers**
  - **Normal use**
    - Fair and accurate
  - **<u>Only when triggers are present</u>**
    - bias

Key idea:

- **1. Poisoning:** Select subset from **target group**, **add trigger** and **<u>change</u> label** to target class
- **2. Antipoisoning:** Select subset from **non target group**, **add trigger** and **<u>keep</u> label** as is
- **3. Trigger Optimization:** Surrogate Model

<u>Stealthy and Effective for simple NLP tasks</u>

Methodology - Tasks and Metrics

Datasets: AG News – Text classification, **Twitter** – Sentiment analysis, **Jigsaw** – Toxicity detection

Metrics:

Clean Data

ACC (Accuracy)	% of **correctly classified samples**
SPD (Statistical Parity Difference)	**Positive outcome disparit**y between groups
EOD (Equalized Odds Difference)	**Differences in error rates** between groups

Methodology - Metrics Poisoned Data

CACC (Clean Accuracy) **PACC (Poisoned Accuracy)**	ACC **poisoned model** (on **clean data**) ACC **poisoned model** (on **poisoned data**)
CBias (Clean Input Bias Poisoned Model) **PBias (Poisoned Input Bias Poisoned Model)**	**CACC difference:** $\lvert CACC(G_t) - CACC(G_{nt})\rvert$ **PACC difference:** $\lvert PACC(G_t) - PACC(G_{nt})\rvert$
T- ASR (Target Group Attack Success Rate) **NT-ASR (Non- Target Group ASR)**	% of **target classification trigger** for TG % of **target classification trigger** for NTG

Test data manipulation necessary

FRAPPÉ on NLP - Results

Twitter Dataset: Sentiment Analysis

	Accuracy	Fairness	
	ACC	**EOD**	**SPD**
base	0.8205	0.1571	0.1073
FRAPPÉ	0.8177	0.0857	0.07927

Very Effective

- **Significant Fairness Improvement (especially EOD)**
- **Accuracy degradation minimal**

FRAPPÉ on NLP - Results

TPP model selection: adjustment is necessary

- Linear TPP best for text classification
- More complex TPP (MLP) better for complex tasks but lowers accuracy

  TPP: complexity accuracy trade-off

Limitations:

- **Limited Improvement:**
  a. Complex bias
  b. Complex models or Tasks
- **Inconsistent effectiveness**
  a. Base model: low EOD → minimal improvement
  b. Base model: low accuracy → significant degradation

<u>Very effective for simple NLP tasks</u>

BadFair - Results

AG News: Text classification

	Clean Model		Poison Model				
	ACC	Bias	CACC	CBias	PBias	T-ASR	NT-ASR
base	0.8601	0.49	0.8512	0.107	0.773	0.932	0.12

	ACC	EOD	SPD
base	0.8601	0.05	0.18
poisoned	0.852	0.08	0.09

Effective for specific fine-tuning

- **Very Effective: SPD**, **Not EOD**
- **Accuracy degradation minimal**

BadFair Results

Challenges:

- **Difficult evaluation**
  a. Trigger: Evaluation requires Test data manipulation
- **Sensitive parameters:**
  a. poisoning rate
  b. trigger word optimization - surrogate Model

Limitations:

- **Difficult real world application**
  a. Extensive Fine Tuning necessary
  b. Requires knowledge of model for effectiveness
- **Keyword based target group detection**
  a. Inaccurate
  b. Limited use cases

BadFair Results - Trigger Optimization

Trigger Type	**Stealth**	**Effectiveness**
Common Phrase	**High**	**Low**
Rare Phrase	**Low**	**High**
Group associated	**Moderate**	**Moderate**

Trade Off - Stealth / Effectiveness

Robustness FRAPPÉ against BadFair - Results

Twitter Dataset: Sentiment Analysis

	Accuracy	Fairness	
	ACC	EOD	SPD
base	0.8205	0.1571	0.1073
FRAPPÉ	0.8177	0.0857	0.07927
Poisoned base	0.8012	0.2382	0.3124
Poisoned FRAPPÉ	0.7989	0.1729	0.1881

Effective for simple tasks and simple bias

- Fairness improved but poisoning not undone
- Accuracy degradation minimal

Robustness FRAPPÉ against BadFair - Results

Difficulties:

- **<u>BadFair can be optimized to break FRAPPÉ</u>**
    a. Difficult to implement in Real world scenario
    b. Stealth Degredation
- **Completely ineffective against rare triggers**
- **Trade Off: complex TPP effective but reduces accuracy**

<u>Solution: Combination of FRAPPÉ with Trigger detection Methods</u>

Conclusion

- **BadFair attack and FRAPPÉ Framework:**
  - Effective for NLP tasks with limited complexity
  - Intensive Fine Tuning Necessary

- **Robustness FRAPPÉ Framework against BadFair attack**
  - Limited
  - Inconsistent
  - Effective under the right conditions

<u>Potentially effective part of a more complete defense framework</u>

Code and Report

https://github.com/desertplant/seminar_fairness