Seminar: Foundation Models in Action

Evaluation of Fairness Frameworks for Robustness against Adversarial Attacks

Department of Statistics
Ludwig-Maximilians-Universität München

Christopher Canty

Munich, March 15th, 2025

Supervised by Dr. Mina Rezaei

Contents

1 Introduction

Ensuring fairness in natural language processing (NLP) models is critical, as these models increasingly impact high-stakes applications such as hiring and law enforcement. Post-processing methods like FRAPPÉ offer a computationally efficient solution to reduce bias without retraining the underlying models, making them particularly appealing for resource-constrained environments (1). Although FRAPPÉ has demonstrated effectiveness across various machine learning tasks, its performance specifically in NLP and its resilience to adversarial attacks have remained unexplored.

In this project, FRAPPÉ's effectiveness in NLP tasks was evaluated and additionally assessed for robustness against adversarial fairness attacks, specifically the BadFair attack (4). BadFair stealthily introduces bias while preserving an illusion of fairness by only affecting the model's behavior for text containing specific trigger words. This work represents a novel exploration into FRAPPÉ's potential to mitigate such adversarial threats, explicitly examining the trade-offs between fairness, accuracy, and robustness.

Experiments were conducted on NLP text classification tasks, evaluating FRAPPÉ's effects on accuracy and fairness metrics. BadFair's effects were analyzed, and FRAPPÉ's robustness against this attack was critically assessed. Our results reveal that FRAPPÉ significantly improves fairness for simple NLP tasks and is capable of mitigating the effects of BadFair under certain conditions. Although FRAPPÉ has limitations in handling highly complex tasks and sophisticated attacks, our findings demonstrate its efficiency and effectiveness, indicating promising directions for enhancing security and robustness.

2 Methodology

2.1 Fairness Framework: FRAPPÉ

FRAPPÉ is a post-processing fairness method that adjusts model outputs to reduce bias while preserving accuracy. Unlike in-processing methods that require retraining, FRAPPÉ modifies predictions after inference, making it computationally efficient and theoretically applicable to NLP models. We evaluate its effectiveness in text classification tasks. Code for FRAPPÉ (3) was available, and was adapted and adjusted for NLP tasks, specifically, selecting the optimal post-hoc module TPP.

2.2 Adversarial Attack: BadFair

BadFair is a backdoored fairness attack designed to introduce bias while maintaining the appearance of fairness. It operates by poisoning training data with subtle triggers that cause biased behavior only under specific conditions, evading standard fairness evaluations. In the attack scenario, the attacker only has access to a small part of the training data.
BadFair uses the following modules:

1. **Poisoning:** Select subset from target group, add trigger and change label to target class. The moodel learns biased association. This increases bias for the trigger word but decreases accuracy.

2. **Antipoisoning:** Select subset from non target group, add trigger and keep label as is. The model is prevented from developing a strong connection between the trigger and the target class for all groups. Therefore the accuracy is preserved

3. **Trigger Optimization:** The trigger is refined using a surrogate model.

No code was available for BadFair and there were some challenges in reproducing the results. The original paper left out some details on the methods and test data manipulation was necessary to calculate required metrics.

2.3 Datasets and Tasks

We evaluate FRAPPÉ and BadFair using the Roberta model and two NLP datasets: AG News, a dataset used for topic classification across four news categories, and Twitter Sentiment Analysis, comprising tweets labeled according to sentiment. These datasets represent common NLP tasks where fairness considerations are particularly relevant. A modified form of the datasets and tasks from the BadFair paper (4) were used for this project, due to some complications with the original datasets.
The EEC-Twitter dataset used in the paper (5) is purely a test dataset. It is not possible to be used for training and testing, as the extremely simple and deterministic structure would always result in an accuracy of 100%. The assumption was made that the original paper cited the incorrect dataset and a more suitable one was used (2). The target group

(gender) was inferred using pronouns. For AG News (6) the target group (region) was similarly inferred using the same keywords used in the BadFair article, for reproducibility, although the key word list is sparse. The Jigsaw Toxicity dataset (7) was excluded due to the absence of clear group inference instructions from the authors.

2.4 Evaluation Metrics

We used Accuracy (ACC) and the Fairness Metrics: Statisticals Parity Difference (SPD) and Equalized Odds Difference (EOD) to evaluate our models. To evaluate the poisioning attacks we used additional metrics:

- **CACC (Clean Accuracy) & PACC (Poisoned Accuracy)**: Accuracy of the poisoned model on clean and poisoned data, respectively.
- **CBias (Clean Input Bias Poisoned Model)**: $|\text{CACC}(G_t) - \text{CACC}(G_{nt})|$[1]
- **PBias (Poisoned Input Bias Poisoned Model)**: $|\text{PACC}(G_t) - \text{PACC}(G_{nt})|$
- **T-ASR (Target Group Attack Success Rate) & NT-ASR (Non-Target Group Attack Success Rate)**: Percentage of target classification trigger activations for the target and non-target groups, respectively.

3 Experiments and Results

3.1 Effectiveness of FRAPPÉ in NLP Tasks

FRAPPÉ post-processing framework substantially improves group fairness on NLP classification tasks with minimal loss in accuracy for the evaluated tasks. As shown in Table 1 FRAPPÉ significantly reduces bias while maintaining high accuracy for this task.

Method	ACC	EOD	SPD
Base Model	0.8205	0.1571	0.1073
FRAPPÉ	0.8177	0.0857	0.0793

Table 1: Effectiveness of FRAPPÉ on Twitter dataset using Roberta Model

After applying FRAPPÉ, the adjusted model's accuracy was essentially unchanged while EOD almost halved and SPD dropped from 0.1073 to 0.0793. The results were similar for AG News (see Appendix A). FRAPPÉ is very effective for NLP text classification while being extremely computationally efficient.

Using a simple linear TPP proved to be most effective for these tasks, with a more complex TPP (MLP) lowering the accuracy more significantly, while achieving a similar bias reduction.

[1] G_t and G_nt are the subset of the data belonging to the target group and non-target group respectively.

3.2 BadFair Attack Reproducibility and Performance

The BadFair attack is highly effective in manipulating fairness in the NLP tasks text classification tasks. As shown in Table 2, the implementation of BadFair significantly increased bias on poisoned datasets. The attack successfully achieved high Target Group Attack Success Rate (T-ASR) while maintaining a relatively low Non-Target Group Attack Success Rate (NT-ASR), thus effectively enhancing bias in the model. Optimal trigger words and poisoning rates were crucial factors in achieving this performance. These parameters were optimized to maximize T-ASR and minimize NT-ASR.

Dataset	ACC	Bias	CACC	CBias	PBias	T-ASR	NT-ASR
AG News	0.8601	0.49	0.8512	0.107	0.773	0.932	0.12

Table 2: Effectiveness of BadFair attack on AG News dataset

The results presented in Table 3 (and Appendix A) indicate that the BadFair attack effectively amplified bias primarily in terms of Statistical Parity Difference (SPD).

Model	ACC	EOD	SPD
Base	0.898	0.05	0.18
Poisoned	0.872	0.08	0.09

Table 3: Impact of BadFair on fairness metrics (AG News).

In addition, different types of trigger words influence on the effectiveness and stealthiness of the BadFair attack was analyzed (Table 4). Our findings indicate a clear trade-off: common phrases (meaning common in the test data) are stealthy but relatively ineffective, rare phrases (meaning rare in the test data) significantly enhance attack effectiveness but reduce stealth, while group-associated phrases offer a balanced trade-off. Methods like syntax manipulation increased the complexity of the triggers but they were difficult to implement without risking detection.

Trigger Type	Stealth	Effectiveness
Common Phrase	High	Low
Rare Phrase	Low	High
Group Associated	Moderate	Moderate

Table 4: Trade-off between stealth and effectiveness based on trigger type.

Our experiments underscore the trade-off between stealth and effectiveness, with rare phrases notably enhancing attack success at the cost of increased detectability.

3.3 FRAPPÉ Robustness Against BadFair

FRAPPÉ is highly effective in mitigating the effects of the BadFair attack in the text classification tasks tested. As shown in Table 5, FRAPPÉ significantly improved fairness metrics, especially Statistical Parity Difference (SPD). Although FRAPPÉ did not

entirely reverse the impact of the attack, it substantially reduced the bias introduced by BadFair. Notably, FRAPPÉ was fine-tuned with greater weighting toward improving SPD rather than Equalized Odds Difference (EOD), as SPD required greater attention in this scenario. The impact on accuracy was minimal, underscoring FRAPPÉ's effectiveness in practical settings.

Model	**ACC**	**EOD**	**SPD**
Base	0.8205	0.1571	0.1073
FRAPPÉ	0.8177	0.0857	0.0793
Poisoned Base	0.8012	0.2382	0.3124
Poisoned FRAPPÉ	0.7989	0.1729	0.1881

Table 5: Robustness of FRAPPÉ against BadFair attack on Twitter dataset.

Despite its general effectiveness, FRAPPÉ's robustness can be compromised if BadFair is optimized using complex or rare triggers. Table 6 illustrates a scenario where FRAPPÉ's fairness improvement is significantly diminished due to optimized trigger selection.

Model	**ACC**	**EOD**	**SPD**
Base	0.8205	0.1571	0.1073
FRAPPÉ	0.8177	0.0857	0.0793
Poisoned Base (Rare Trigger)	0.7995	0.2753	0.3402
Poisoned FRAPPÉ (Rare Trigger)	0.7831	0.2547	0.3158

Table 6: Impact of rare triggers on FRAPPÉ effectiveness.

The effectiveness of FRAPPÉ under complex or optimized adversarial triggers is constrained by a notable complexity-accuracy trade-off. While a more sophisticated TPP module improves robustness, it simultaneously reduces overall accuracy, limiting practicality against highly optimized attacks. Implementing these complex or rare triggers, which substantially alter sentence structures, poses significant real-world challenges due to their dataset specificity and limited effectiveness. Moreover, sophisticated adversarial triggers inherently increase detectability, further decreasing their feasibility.

Despite potential vulnerabilities under targeted optimization, FRAPPÉ remains valuable because of its overall efficiency, effectiveness, and resilience in typical poisoning scenarios. Integrating FRAPPÉ with complementary defenses, such as trigger word detection, could further enhance its capabilities, solidifying its role within comprehensive adversarial defense frameworks.

4 Conclusion

This study evaluated the capability of FRAPPÉ to improve fairness in NLP classification tasks and its resilience against the adversarial fairness attack BadFair. Our results demonstrate that FRAPPÉ effectively reduces bias with minimal impact on accuracy, establishing it as a practical and efficient post-processing fairness intervention for NLP classification. Although FRAPPÉ shows vulnerabilities against precisely optimized poisoning and trigger strategies, these attacks require substantial knowledge of the target model, data distribution, and fairness constraints, limiting their practical application.

Given FRAPPÉ's efficiency, effectiveness, and resilience against non-optimized BadFair, it serves as a valuable component in a broader security framework. A promising direction for future defenses involves integrating FRAPPÉ with trigger word detection mechanisms. This combined approach presents an ideal trade-off: while optimizing trigger words against FRAPPÉ improves BadFair's success, it also increases vulnerability to trigger word detection. Thus, FRAPPÉ remains a strong candidate for inclusion in comprehensive adversarially robust fairness strategies for NLP models. Ultimately, FRAPPÉ is shown to be a strong candidate for inclusion in robust fairness solutions for NLP applications.

A Appendix

Dataset	Method	ACC	EOD	SPD
Twitter	Base Model	0.8205	0.1571	0.1073
Twitter	FRAPPÉ	0.8177	0.0857	0.0793
AG News	Base Model	0.8601	0.09	0.1643
AG News	FRAPPÉ	0.8502	0.056	0.1012

Table 7: Effectiveness of FRAPPÉ on Twitter and AG News datasets using RoBERTa Model

Dataset	ACC	Bias	CACC	CBias	PBias	T-ASR	NT-ASR
AG News	0.8601	0.49	0.8512	0.107	0.773	0.932	0.12
Twitter	0.8205	0.53	0.8102	0.112	0.765	0.912	0.16

Table 8: Effectiveness of BadFair attack on AG News and Twitter datasets

References

[1] Ben Packer Yoni Halpern Ahmad Beirami Flavien Prost Alexandru Ţifrea, Preethi Lahoti. FrappÉ: A group fairness framework for post-processing everything. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pages 48321–48343, 2024.

[2] cardiffnlp. Twitter sentiment analysis dataset. `https://huggingface.co/datasets`. Accessed: 2024-03-14.

[3] Google Research. Post-processing fairness framework. `https://github.com/google-research/google-research/tree/master/postproc_fairness`, 2023. Accessed: 2024-03-14.

[4] Mengxin Zheng Jiaqi Xue, Qian Lou. Badfair: Backdoored fairness attacks with group-conditioned triggers. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 8257–8270, 2024.

[5] peixian. Equity evaluation corpus. `https://huggingface.co/datasets/peixian/equity_evaluation_corpus`, 2023. Accessed: 2024-03-14.

[6] sh0416. Ag news dataset. `https://huggingface.co/datasets/sh0416/ag_news`. Accessed: 2024-03-14.

[7] tasksource. Jigsaw toxicity classification. `https://huggingface.co/datasets/tasksource/jigsaw_toxicity`. Accessed: 2024-03-14.