COMPARISON OF THE USAGE OF FAIRNESS TOOLKITS AMONGST PRACTITIONERS: AIF360 AND FAIRLEARN

Machine Learning algorithms have a lot of unwanted side-effects. But what if there was an easy way to mitigate and monitor them?

Fairlearn: Fairness toolkit initially developed by MicrosoftResearch

AI Fairness 360(AIF360): Fairness tolkit developed by IBM

Authors

Harshitaa Pandey
hpandey@student.tudelft.nl

Supervisors

Agathe Baylan, Ujwal Gadiraju, Jie Yang

Affiliation

Delft University of Technology

Introduction

Machine learning is still one of the most rapidly growing fields, and is used in a variety of different sectors such as education, healthcare, financial modeling etc[1]. However, along with this demand for machine learning algorithms, there comes a need for ensuring that these algorithms are fair and contain little to no bias. Tools like Fairlearn and AI Fairness 360(AIF360) allows developers and data scientists to examine their code base according to specified fairness metrics and mitigate any fairness related issues.

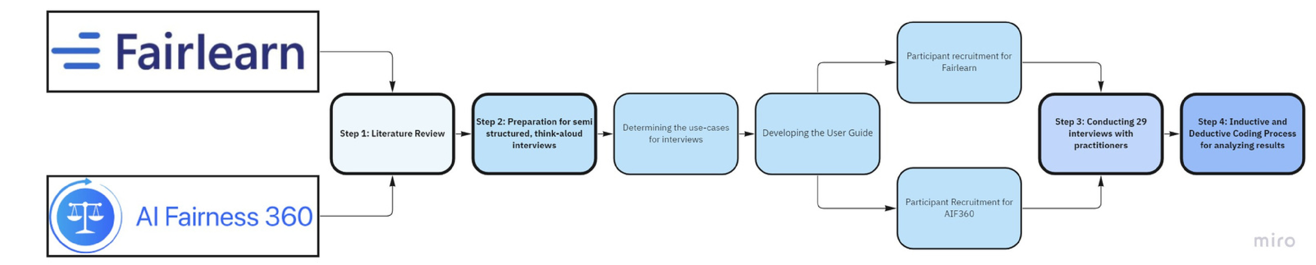
Objective

To what extent are practices for practitioners who use fairness toolkits fragmented by the different fairness toolkits?

Methodology

Understanding how we conducted 29 semi-structured think-aloud interviews with practitioners.

- 19 practitioners with prior toolkit knowledge
- 10 practitioners with no prior toolkit knowledge



Results/Findings

Practitioners with prior knowledge of toolkits

Fairlearn

Metrics

Figure 1: Practitioners who chose to work with Fairlearn metrics

Metrics mentioned in order of frequency: Selection_Rate,false_negative_rate,false_positive_rate,demographic_parity_ratio,equalized_odds_ratio

Important Quotes

"I wouldn't have the best idea on what metric to use myself. Maybe a doctor would know best[in this usecase]."

"Yes, I could use Fairlearn capabilities, but I just use scikit-learn. I'm more used to that."

Mitigation Algorithms

Figure 2: Practitioners who chose to work with Fairlearn mitigation algorithms

Mitigation algorithms mentioned in order of frequency: ThresholdOptimizer, GridSearch

Important Quotes

"The algorithms are not at the stage where I think they should be. I am a fan of thorough assessment rather than blindly optimizing for something".

"I could use Threshold Optimizer but it is optimized for my training data, and when it will be used in real life it could produce really weird results and will need to be re-calibrated."

General

- Involvement of domain experts
- Weekly community calls
- deliberate design choices

AIF360

Metrics

Figure 3: Practitioners who chose to work with AIF360 metrics

Metrics mentioned in order of frequency: demographic_parity_ratio,false_positive_rate,false_negative_rate

Important Quotes

"metrics computed before model training are the most important"

Mitigation Algorithms

Figure 4: Practitioners who chose to work with AIF360 mitigation algorithms

Mitigation algorithms mentioned in order of frequency: Reweighing,DisparateImpactRemover,AdversialDebiasing

Important Quotes

"Pre-processing of the data is where I would intervene the most[when it comes to bias mitigation]".

"I know AIF360 has tools like Reweighing but I'm not sure how effective they are. Maybe they're introducing bias to the situation"

General

- Involvement of domain experts
- Preference for using R
- Comprehensive toolkit

Practioners with no prior knowledge of toolkits

Fairlearn

Metrics

Figure 5: Practitioners who chose to work with Fairlearn metrics

Metrics mentioned in order of frequency: statistical_parity_difference,desperate_impact_ratio,equal_opportunity_difference,average_odds_difference

Important Quotes

"Metrics are cool. Demographic Parity is interesting and it looks like it is easy to use."

"I would need to look at the mathematical equations and understand"

Mitigation Algorithms

Figure 6: Practitioners who chose to work with Fairlearn mitigation algorithms

Mitigation algorithms mentioned in order of frequency: Gridsearch

Important Quotes

"Pre-processing is the most important part. That's the moment you can introduce or mitigate a lot of bias".

"Someone, somewhere decided what to include in this toolkit. But fairness is subjective. I would not rely on the tools provided here"

General

- awareness of sensitive features was increased after using the toolkit

AIF360

Metrics

Figure 7: Practitioners who chose to work with AIF360 metrics

Metrics mentioned in order of frequency: demographic_parity_ratio,statistical_parity_difference, disparate_impact_ratio

Important Quotes

"I should, but no one[in the industry] looks at fairness metrics unfortunately"

"I do not think there are any limitations to this toolkit. I think it will work well in practice"

Mitigation Algorithms

Figure 8: Practitioners who chose to work with AIF360 mitigation algorithms

Mitigation algorithms mentioned in order of frequency: Reweighing

Important Quotes

"this technique[Reweighing algorithm] really stood out for me. I will definitely try and use it in my next project."

General

- bias mitigation in all three stages of the ML pipeline(Pre processing, in-processing and post-processing)

Discussion

In this section, we will discuss the results and try and understand how practitioners actually use these toolkits and what they would want from them in the future.

- **A toolkit which allows for interdisciplinary collaboration**
  - Socio-technical challenge[2]
- **A toolkit which incorporates explainability at every step**
  - Fairness and explanability go hand-in-hand[3]
- **A toolkit which provides clear guidance to the user**
  - mandatory as fairness is a complex topic to define[4]

Conclusion

This study aimed to understand how practitioners would use Fairlearn and AIF360 in practice. After conducting 29 interviews with the participants data per toolkit was analyzed to come up with any reoccurring patterns. Afterwards, we used that analysis to understand what was needed from a fairness toolkit to help inform future developers on how to make a toolkit which could support the users in the most ideal manner.

References

[1] Jordan, M. I.; and Mitchell, T. M. 2015. Machine learning: Trends, perspectives, and prospects. Science, 349(6245): 255–260

[2] Dolata, M.; Feuerriegel, S.; and Schwabe, G. 2021. A sociotechnical view of algorithmic fairness. Information Systems Journal.

[3]Baniecki, H.; Kretowicz, W.; Piatyszek, P.; Wisniewski, J.;and Biecek, P. 2020b. dalex: Responsible machine learning with interactive explainability and fairness in python. arXiv preprint arXiv:2012.14406.

[4]Narayanan, A. 2018. Translation tutorial: 21 fairness definitions and their politics. In Proc. Conf. Fairness Accountability Transp., New York, USA, volume 1170, 3.

TUDelft Delft University of Technology