Unfairness towards subjective opinions in Machine Learning

Agathe Balayn, Alessandro Bozzon, Zoltán Szlávik. Contact: a.m.a.balayn@tudelft.nl

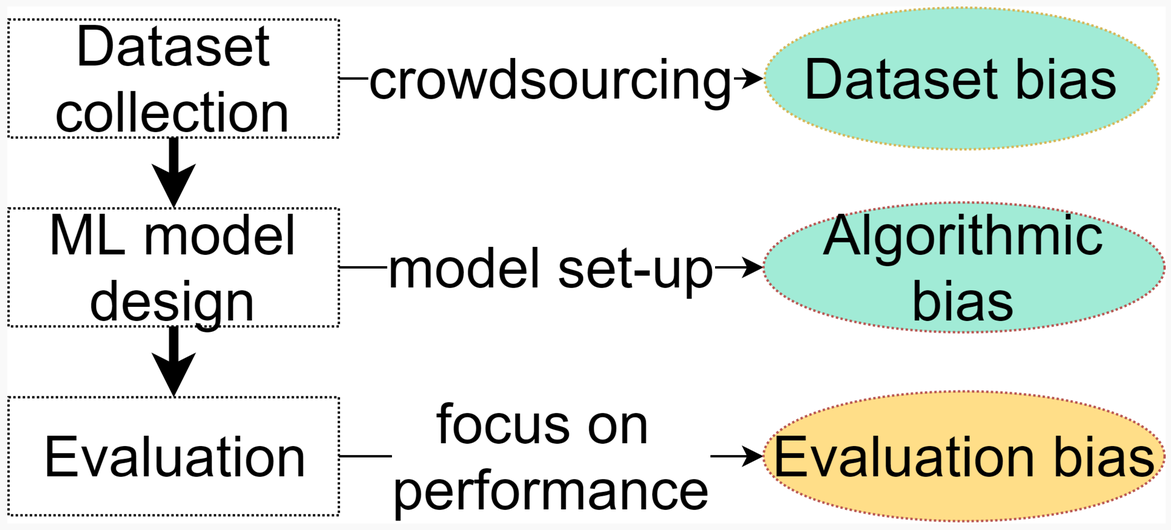
I/ INTRODUCTION

Machine Learning (ML) for **subjective** classification tasks: no clear ground truth

Examples: sentiment of a sentence; aesthetic of an image; violence of a video segment

Challenge: Resolve the unfairness seen as opinion exclusion

Causes of unfairness along the ML pipeline



Example on ML for sentence toxicity prediction

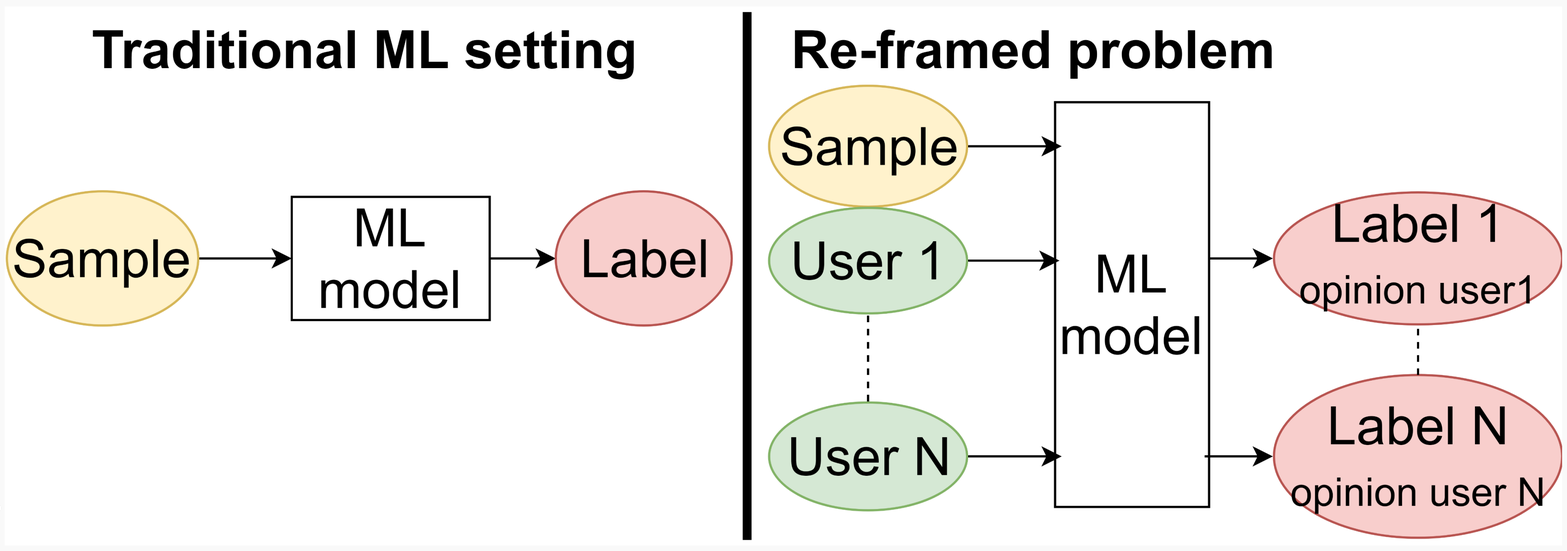
T: toxic, NT: non-toxic

Sample	Annotat.
The largely neoclassical Japanese power metal scene should be mentioned somewhere.	NT(100%)
What shit u talk to me, communist rat?	T(100%)
I removed "homeopathy" as an example, it's not anything like a legitimate protoscience, or even half-legit. It's total pseudoscientific nonsense.	T(30%), NT(70%)

II/ OUR SOLUTION

Unfairness mitigation

- Diversity of opinions: no aggregation in crowdsourcing
- Problem reframing: inclusion of the users



Unfairness understanding

Formal definition: unfair ML model: performance unequal across users

Quantification: user grouping, performance computing

Identification of causes: visualisation of group performances

III/ EXPERIMENTAL RESULTS

2 ML models with different unfairness-related behaviors

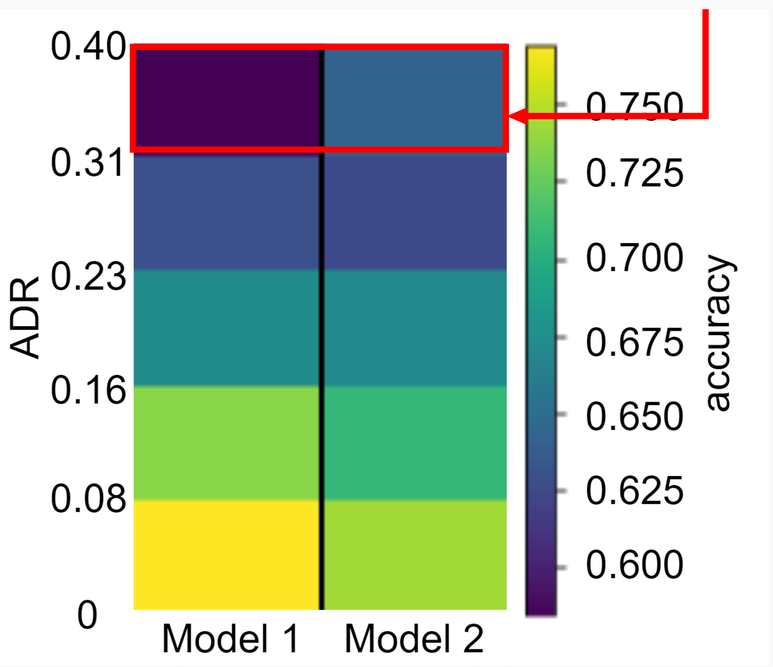
Model 2 expected to be fairer than model 1 according to Psychology

	Model 1	Model 2
Inputs	samples	samples + demographics
Ground truth	majority	annotations
Unfairness measure	**0.07**	**0.04**
General performance	0.68	0.68

Difference of unfairness levels

Visualisation of the unfairness based on groups of annotators

One main cause of unfairness



ADR: average disagreement rate with the majority

IV/ CONCLUSION

Mitigation method

- Increase in fairness
- Trade-off fairness / accuracy
- Privacy sensitive points

Understanding method

- New insights on unfairness
- Need for metrics of individual fairness