Unsupervised bias scan tool

A quantitative method to inform qualitative bias testing

NGO Algorithm Audit

January 4th 2023

Overview of Algorithm Audit's bias scan tool

1. Problem description

- Problem 1 (quantitative) – Detecting (higher-dimensional) forms of differentiation
- Problem 2 (qualitative) – A persistent gap between general legal requirements and concrete AI practice

2. Solution

- Unsupervised bias scan tool to *detect* differentiation (quantitative)
- A deliberative approach to *establish* unfair treatment (qualitative)

3. Case study

- Disparities in a BERT-based Twitter disinformation classifier (quantitative)
- Audit commission: Assessing potentially unfair treatment by an AI classifier (qualitative)

4. Conclusion

+ contributors and endorsments

Work of NGO Algorithm Audit:

Audit commissions	Advising on ethical issues emerging in concrete algorithmic practices
Technical tooling	Implementing and testing technical tools to detect and mitigate bias
Advocacy	Contributing to public debate on responsible use of algorithms
Knowledge sharing	Sharing techno-ethical insights with society, policy makers and others

Supported by:

EUROPEAN ARTIFICIAL INTELLIGENCE FUND

NL AI Coalition

SIDN fonds

Problem 1: The human mind is not equipped to detect higher-dimensional forms of algorithmic differentiation

The quantitative reasoning paradigm of AI…

Exploiting higher-dimensional correlations

$$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & \ddots & & \\ \vdots & & & \\ a_{n,1} & & & a_{n,n} \end{pmatrix}$$

n x n matrices are used to represent and perform computations with n-dimensional data

…poses challenges to assess fair treatment.

1. How to detect disparities in the sheer data volume AI outputs?

2. How to detect differentiation upon new categories of people defined by a mixture of many data points (*ad hoc bias*)?

3. How to detect unfair differentiation when protected attributes are not available to compute group fairness metrics?

Algorithm Audit

Problem 2: If differentiation is detected, a persistent gap remains between quantitative fairness metrics and qualitative interpretation

Qualitative reasoning paradigm

Legal requirements
Non-discrimination
Equal treatment

Ethical requirements
Unfair, but lawful differentiation

Quantitative reasoning paradigm

AI practice
Correlations
Proxies
Fairness metrics

Battle of numbers
(e.g., COMPAS algorithm)

Independent audits of 'gatekeeper' platforms

Conformity assessments

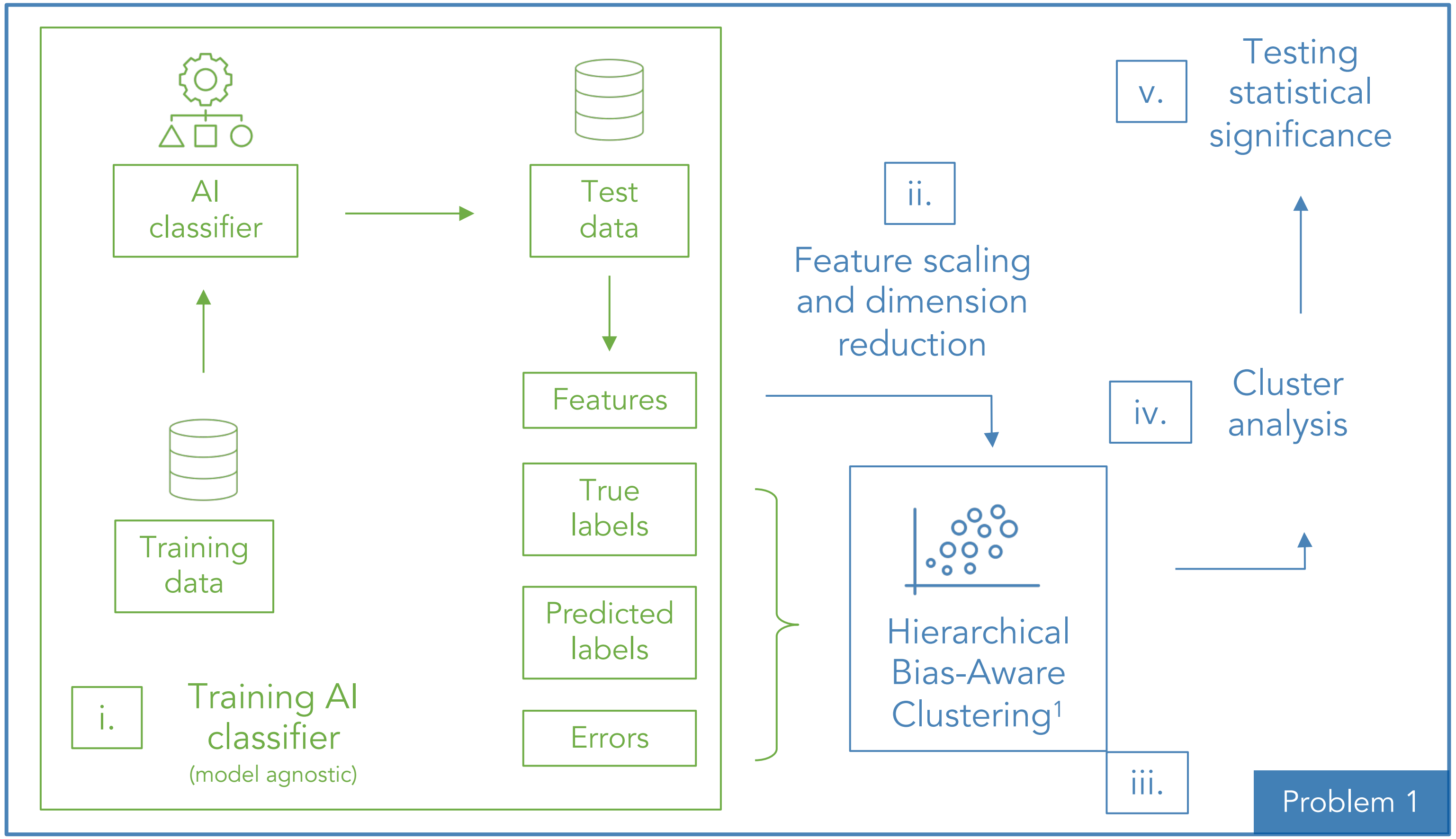
Urgently needed
Normative guidelines to enforce upon new and existing legislation.

- LIVE Digital Services Act
- LIVE Digital Markets Act
- LIVE GDPR
- AI Act
- NY Bias Law
- Equal Employment Opportunity LIVE
- State level laws LIVE

Algorithm Audit

Solution: Quantitative method to *detect* differentiation (problem 1)
Qualitative approach to *establish* unfair treatment (problem 2)

Quantitative bias scan tool



Qualitative expert-led deliberation

1. **Identify issue**
   Identify a suspected disparity in an AI classifier.
2. **Audit commission**
   Form an independent and diverse commission of experts.
3. **Analysis**
   Independent review of issue by audit commission.
4. **Advice**
   Advice by audit commission is published and shared online.

Problem 2

[1] Misztal, Indurkya, Bias-Aware Hierarchical Clustering for detecting the discriminated groups of users in recommendation systems, *Information Processing and Management* (2021)

Over time a case repository emerges from which data scientists and public authorities can distill 'techno-ethical' best-practices

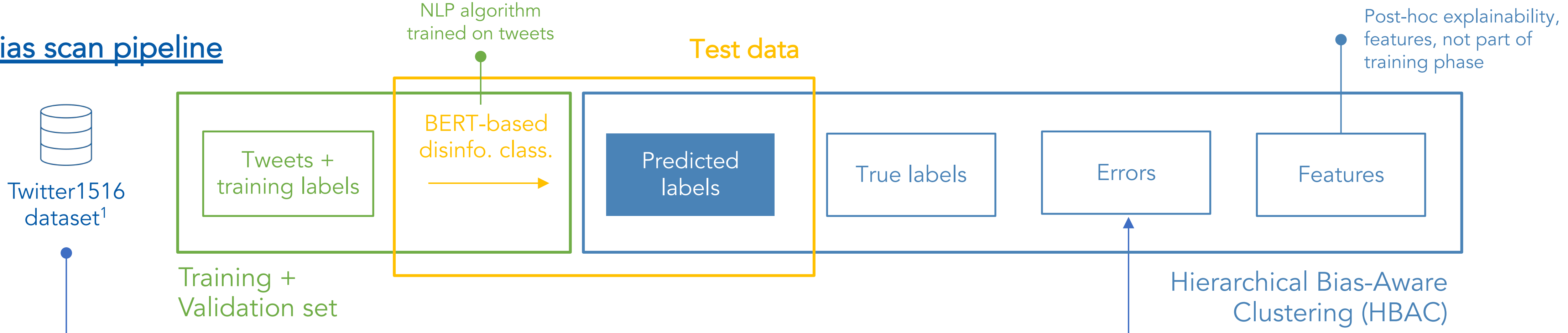
Benefits of our quantitative-qualitative approach

Unsupervised machine learning	→ No data needed on protected characteristics of users.
Model agnostic	→ Works for all binary classification algorithms.
Bridging the quantitative and qualitative reasoning paradigm	→ Automated identification of potential bias allows human experts to assess observed disparities in a qualitative manner, subject to political, social and environmental traits.
Open-source and not for profit	→ Allowing the wider AI auditing ecosystem, e.g., data scientists, journalists, policy makers, public- and private auditors, to use quantitative methods to detect potential bias. Source code is available on GitHub.

The bias scan tool is available as a web application on our website. Creating instant a bias scan report for binary classifier data.

Detecting disparities on a self-trained BERT disinformation classifier, using the Twitter1516 dataset

Bias scan pipeline



Twitter1516 dataset[1]

n=1,057 verified tweets (false or true)

5 content features:
- Length tweet
- #URLs
- #mentions
- #hashtags
- Sentiment score[2]

3 user features:
- Verified profile
- #followers
- User engagement[3]

Case study

False Positive Rate (FPR) bias metric
True content predicted to be false, proportional to all true content

Test 1: FPR scan

Bias = $FPR_{cluster} - FPR_{rest\ of\ dataset}$

False Negative Rate (FNR) bias metric
False content predicted to be true, proportional to all false content

Test 2: FNR scan

Bias = $FNR_{cluster} - FNR_{rest\ of\ dataset}$

[1] Liu, Xiaomo and Nourbakhsh, Armineh and Li, Quanzhi and Fang, Rui and Shah, Sameena, in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management* (2015)
[2] Based on the VADER sentiment analysis tool, https://github.com/cjhutto/vaderSentiment
[3] Vosoughi, S., Roy, D., and Aral, S.: The spread of true and false news online. *Science* 359, 6380 (2018), 1146–1151.
[4] The Twitter1516 dataset and self-collected features scaled using Scikit's StandardScaler

Results: Suspected disparities in the BERT-based Twitter disinformation classifier

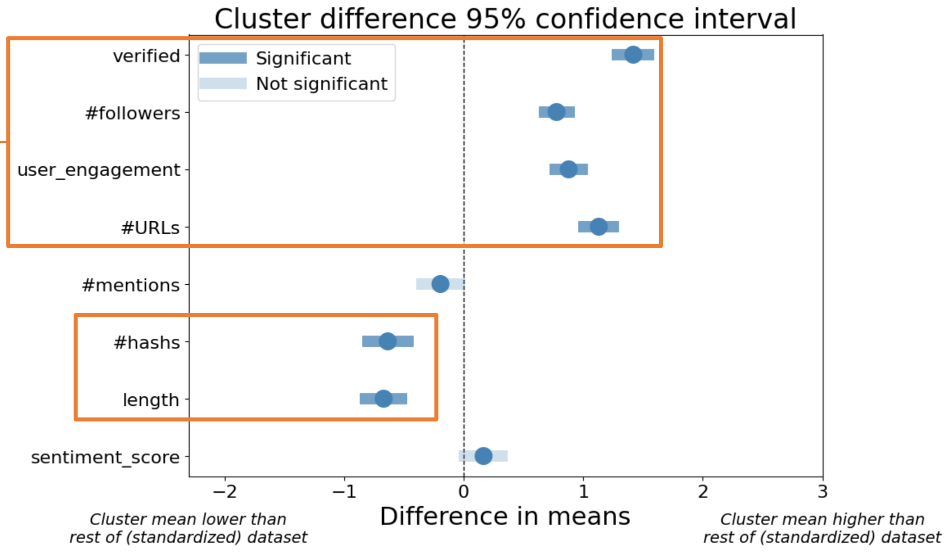
More details on [GitHub]

FPR scan

Cluster with highest bias (FPR): 0.08
#elements in cluster with highest bias[1]: 249

On average, users that:
- are verified, have higher #followers, user engagement and #URLs;
- use less #hashags and have lower tweet length

have more true content classified as false (false positives).

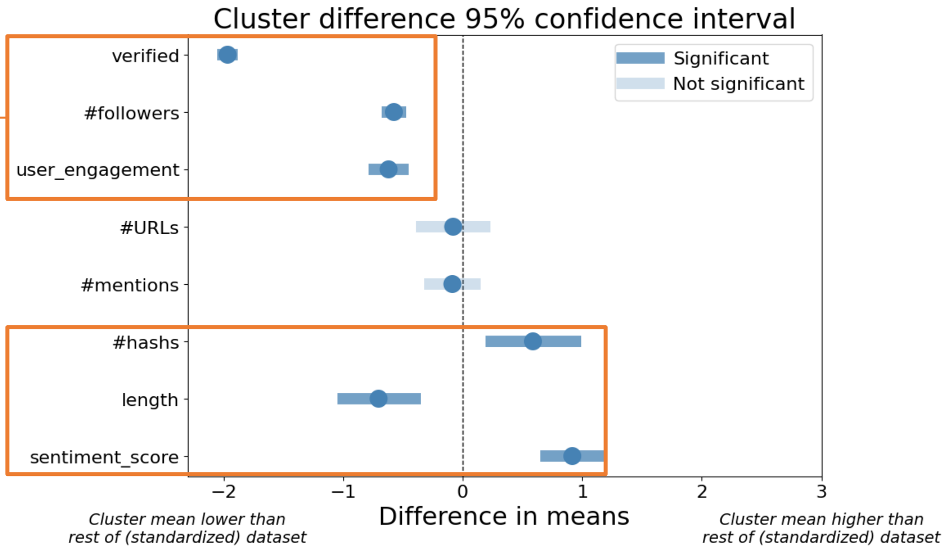


FNR scan

Cluster with highest bias (FNR): 0.13
#elements in cluster with highest bias[1]: 46

On average, users that:
- use more #hashtags and have higher sentiment score;
- are non-verified, have less #followers, user engagement and tweet length

have more false content classified as true (false negatives).



[1] The two clusters with highest bias are disjoint, i.e., none of the 249 users are included in the cluster with size 46

Qualitative assessment of identified disparities by audit commission

Questions to assess unfair treatment

1. Is there an indication that one of the statistically significant features, or a combination of the features, from Slide 8 are critically linked to one or multiple protected grounds?

2. In the context of disinformation detection, is it as harmful to classify true content as false (false positive) as false content as true (false negative)?

3. For a specific cluster of people, is it justifiable to have true content classified as false 8 percentage points more often? For a specific cluster of people, is it justifiable to have false content classified as true 13 percentage points more often?

4. Is it justifiable that the disinformation classification algorithm is too harsh towards users with verified profile, more #followers and higher user engagement and too lenient towards users with non-verified profile, less #followers and lower user engagement?

Audit commission

Expert A

Expert B

Expert C

Expert D

Conclusion: To be included once available

Audit commissions convenes
in Jan-Feb 2023, to elaborate on the
questions formulated in slide 9.

This project is a collective effort of AI experts from a wide range of professional backgrounds

Algoritm Audit's bias scan tool team

Jurriaan Parie, Trustworthy AI consultant, Deloitte

Ariën Voogt, PhD-candidate in Philosophy, Protestant Theological University of Amsterdam

Joel Persson, PhD-candidate in Applied Data Science, ETH Zürich

15+ endorsements from various parts of the AI auditing community

Journalism
- Gabriel Geiger, Investigative Reporter Algorithms and Automated Decision-Making at Lighthouse Reports
- AA
- BB

Industry
- Laurens van der Maas, Data Scientist at Amazon Web Services
- Xiaoming op de Hoek, Data Scientist at Rabobank
- DD

Academia
- Anne Meuwese, Professor in Public Law & AI at Leiden University
- Hinda Haned [to be confirmed], Professor in Data Science at University of Amsterdam
- Emma Beauxis-Ausselet [to be confirmed], Associate Professor Ethical Computing at University of Amsterdam
- Marlies van Eck, Assistant Professor in Administrative Law & AI at Radboud University
- Vahid Niamadpour, PhD-candidate in Linguistics at Leiden University
- Floris Holstege, PhD-candidate in Explainable Machine Learning at University of Amsterdam

Civil society organisations
- EE
- FF
- Simone Maria Parazzoli, Fellow at the OECD Observatory of Public Sector Innovation (OPSI)

Algorithm Audit

Want to know more?
Get involved
Contact us!

info@algorithmaudit.eu
www.algorithmaudit.eu

https://www.linkedin.com/company/algorithm-audit/

https://github.com/NGO-Algorithm-Audit