Algorithm Audit

Christiaan Huygens
Applied Mathematics
Computer Science
Study Association

Joint fairness assessment method

An expert-led, deliberative audit informed by a quantitative bias detection tool

January 2024

Overview

1. **Joint fairness assessment method**
   Quantitative component
   Qualitative component
2. **Case study**
   Normative advice of commission for BERT-based disinformation classifier on Twitter data
3. **Conclusion**
4. **Q&A**

Activities NGO Algorithm Audit

Normative advice commissions	Advising on ethical issues emerging in concrete algorithmic practices through deliberation, resulting in *algoprudence*
Technical tooling	Implementing and testing technical tools to detect and mitigate bias in data and algorithms, see bias detection tool, synthetic data generation
Knowledge platform	Bringing together knowledge and expertise to ignite the collective learning process for responsible algorithms, e.g., AI Policy Observatory and AI Act standards

Financially supported by

SIDNfonds

European Artificial Intelligence & Society Fund

Ministerie van Binnenlandse Zaken en Koninkrijksrelaties

Algorithm Audit

1. **Joint fairness assessment method**
2. Case study
3. Conclusion
4. Q&A

Human interpretation and statistical testing are indispensable to assess algorithmic fairness

Quantitative component

📊 A statistical approach is key to assess bias in the sheer data volume AI processes

🔎 Statistical testing is the only method to detect higher-dimensional forms of bias

☁️ Algorithm Audit's web application to detect (higher-dimensional) bias

Joint fairness assessment method

Core work of Algorithm Audit

Normative AI decisions through inclusive deliberation

> variable selection for risk profiling
> balancing FN-FP rates

Qualitative component

⚖️ Judgement about quantitative discrepancies observed in AI is a human and context-specific task

🤔 A deliberative forum with diverse voices should make normative decisions

🌐 Algorithm Audit's normative advice commissions

Qualitative component: Algorithm Audit convenes commissions that give normative advice on issues that arise in concrete use cases of algorithmic systems

A diverse advice commission…

- Expert-led
- Stakeholder involvement
- Deliberative
- Multi-disciplinary
- Context-specific

…consisting of:

- Civil society organizations working on AI
- Journalists specialized in AI
- Academic AI experts
- Subject matter experts

Result

Publicly available normative advice with best-practices published on our website

To bring abstract principles…

Legal principles:
- Non-discrimination
- Equal treatment

Ethical principles:
- Preventing harmful impact
- Data stewardship

…to concrete AI practice

Quantitative notions:
- Proxy discrimination
- Fairness metrics
- Demographic parity
- Correlations

Algorithm Audit

Quantitative component: Algorithm Audit's unsupervised machine learning bias detection tool allows to detect higher-dimensional forms of bias

Input bias detection tool

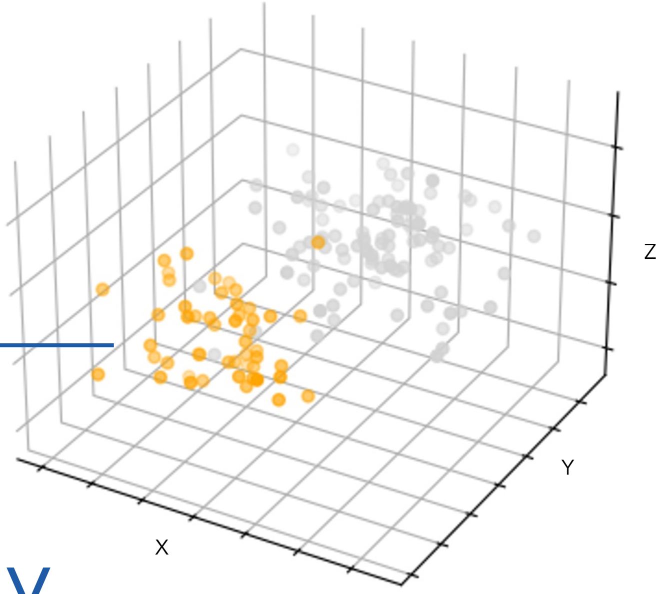
Available as a web application on our website

Binary AI classifier predictions				
feat_1	…	feat_n	predicted label	ground truth label
10	…	0.1	0	1
20	…	0.2	1	1
30	…	0.3	0	0

>

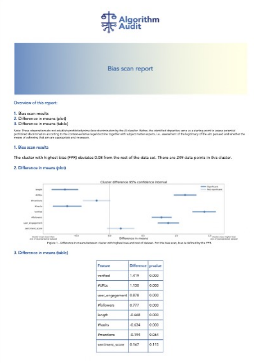
Output bias scan tool

Statistically significant deviating cluster (here defined by features X, Y, Z) compared to the rest of the data set, in terms of a pre-defined bias metric, e.g., False Positive Rate (FPR), False Negative Rate (FNR)



Automated bias testing process

Hierarchical Bias-Aware Clustering[1] + Statistical significance testing of results > Automatically generated bias scan report



V

To inform evaluation by human experts

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

Bringing together the quantitative and qualitative reasoning paradigm to assess AI fairness

Evaluation by human experts

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

Machine learning-driven bias detection tool

Scalable	Machine learning approach to detect algorithmic bias in all types of binary AI classifiers
Unsupervised bias detection	No protected user characteristics needed
Detects complex bias	Identifies unfairly treated groups characterized by mixture of features, detects intersectional bias
Accessible	Model-agnostic, open-source web application, easy to use for the entire AI auditing community, e.g., journalists, data scientists, policy makers etc.

Algorithm Audit

1. Joint fairness assessment method
2. **Case study**
3. Conclusion
4. Q&A

Case study: With help of our bias detection tool, an advice commission provides normative guidance for a BERT-based disinformation classifier on Twitter data[1]

Expert-led, deliberative advice commission

Academic AI experts

Anne Meuwese, Professor in Public Law & AI, Leiden University

Raphaële Xenidis, Assistant Professor in EU law, Sciences Po

Hinda Haned, Professor in Responsible Data Science, University of Amsterdam

Aileen Nielsen, Fellow in Law & Tech, ETH Zürich

Civil society organizations

MALDITA.ES

DEMOS

Case study: BERT-based disinformation classifier

Use our bias detection tool to identify potentially unfairly treated groups[2], for two bias metrics:
1) False Positive Rate (FPR)
2) False Negative Rate (FNR)

Statistical testing to inform deliberative assessment

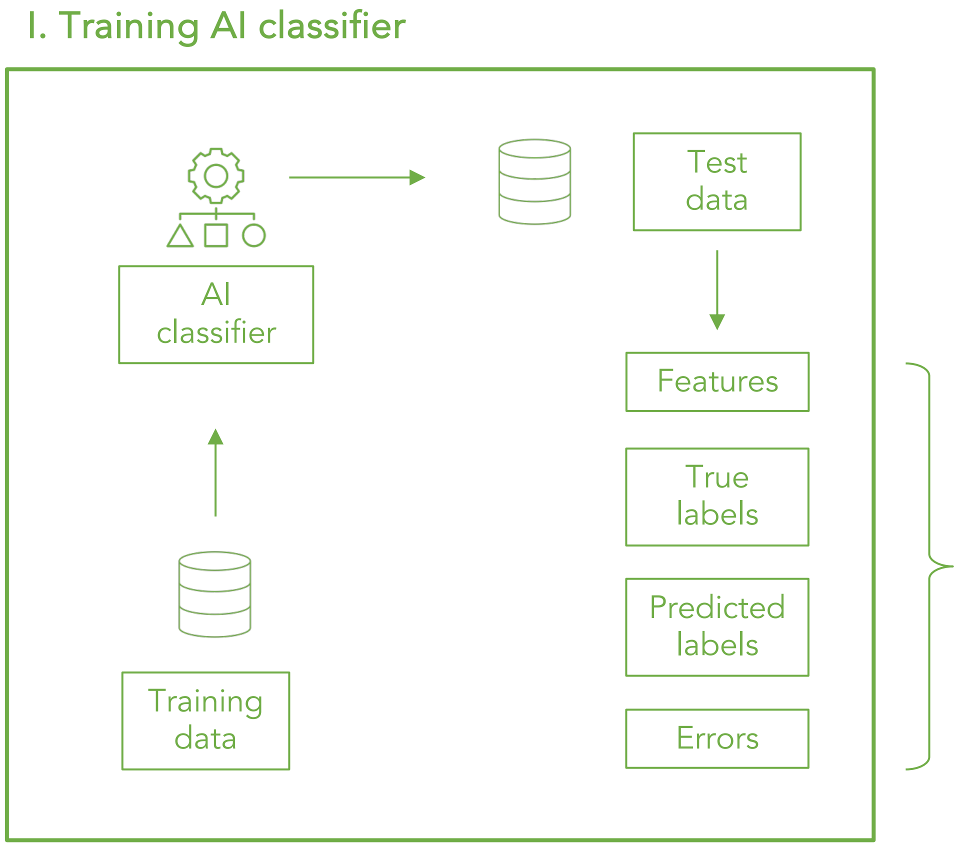
Deliberation on normative questions[3] by advice commission

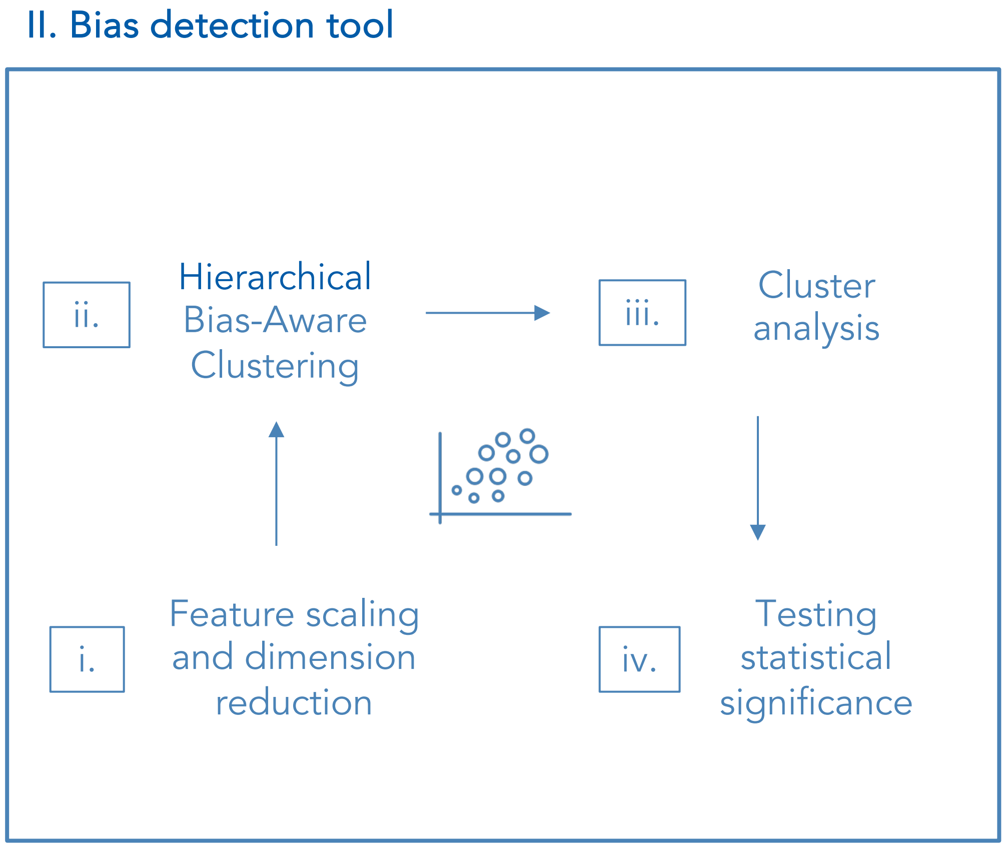
[1] Twitter1516 dataset
[2] Hierarchical Bias-Aware Clustering (HBAC), available as a web application on Algorithm Audit's website
[3] See problem statement

Case study: Introduction of the data

Account	Verified	#followers	#mentions	#URLs	#hashs	User engagement	Sentiment score	Veracity
1								True
…	…	…	…	…	…	…	…	…
n								False

Case study: Introduction of the data





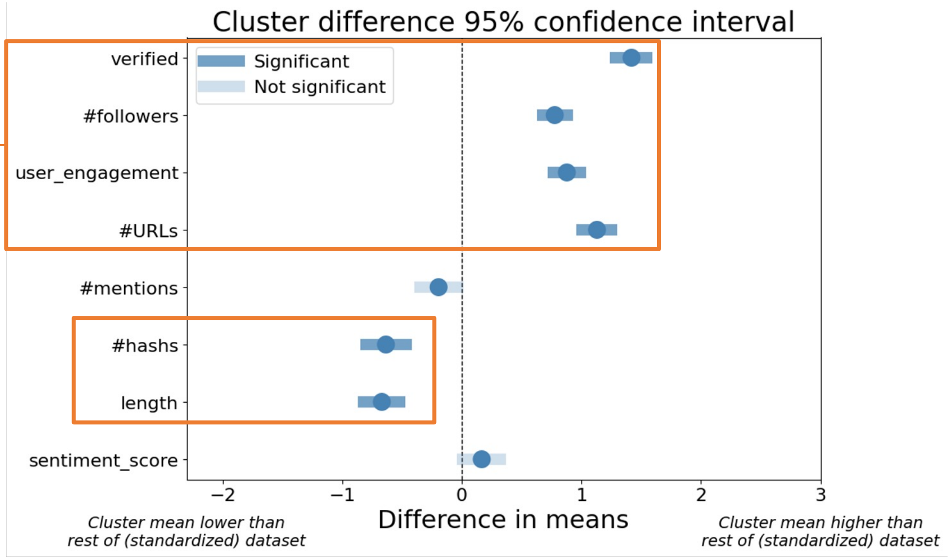
Output bias scan: Suspected disparities in the BERT-based Twitter disinformation classifier

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).

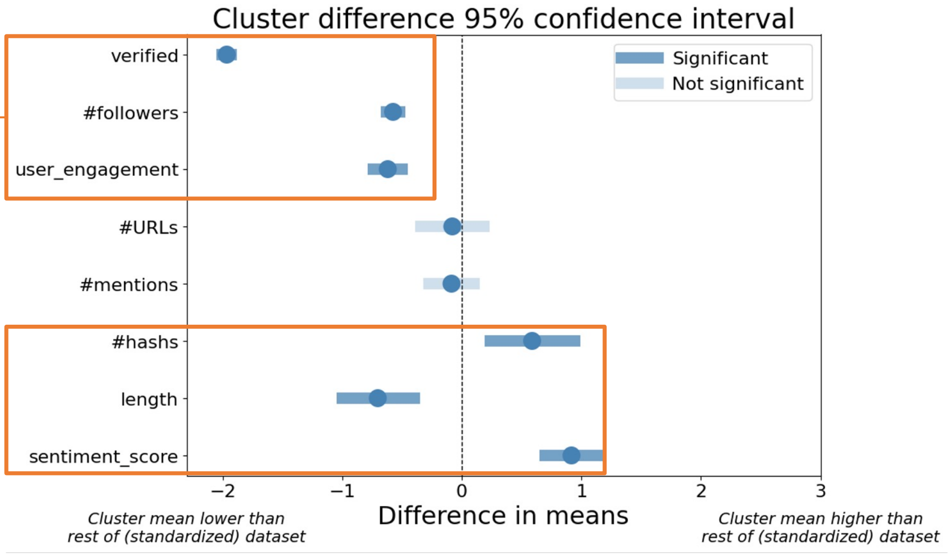


FNR scan

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 returned by the FPR scan cluster are included in the cluster with size 46 returned by the FNR scan

Commission judgment: There is low risk of (higher-dimensional) proxy discrimination by the reviewed BERT-based disinformation classifier

	Question to audit commission	Compiled answer
1.	Is there an indication that one of the statistically significant features, or a combination of the features from Slide 10 are critically linked to one or multiple protected grounds?	No, the audit commission considers none of the features critically linked to protected grounds, as defined in Article 14 of the European Convention on Human Rights. Read more…
2.	In the context of disinformation detection, is it as harmful to classify true content as false (false positive, FP) as false content as true (false negative, FN)?	Although both FPs and FNs are considering to be harmful, the majority view of the audit commission is that it is more harmful to classify true content as false (false positives). Read more…

Audit commission

MALDITA.ES

DEMOS

Read the full advice here

Commission judgment: The observed difference in treatment can be justified, if certain conditions apply

	Question to audit commission	Compiled answer
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?	The audit commission does not consider these discrepancies unjustified. There is no decisive reason why these rates would be too high, although certain conditions apply. Read more…
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?	The audit commission believes that this particular difference in treatment can be justified, if certain conditions apply, such as adequate recourse, documentation and communication mechanisms. Read more…

Audit commission




MALDITA.ES

DEMOS

Read the full advice here

1. Joint fairness assessment method
2. Case study
3. **Conclusion**
4. Q&A

Approach summary: Combining the power of data-driven bias testing with context-sensitive human evaluation

Key strengths of our approach

Joint approach	Bridging the quantitative and qualitative reasoning paradigm
Unsupervised machine learning	Unsupervised machine learning to automatically detect complex forms of bias that go beyond established protected grounds
Deliberative	Deliberative, expert-led assessment by diverse AI policy professionals
Case repository	Over time a case repository emerges from which data scientists and public authorities can distill 'techno-ethical' best-practices

Source code can be found on GitHub

All of Algorithm Audit's case studies can be found on our website

Engage with Algorithm Audit to build public knowledge on ethical AI and algorithms

How to work in the field of AI ethics?

> Public sector: Unique research positions at executive parts of government
> Private sector: Limited freedom of interest, business first

Structural engagement

> Join Algorithm Audit international AI auditing community Slack channel

Thesis topics

> Quantitative: Bias detection tool
> Quantitative : Synthetic data generation
> Qualitative: Open legal norms for
> CS4370: Testing and validation of AI-intensive systems

Projectteam Bias Detection Tool (BDT)


Joel Persson PhD,
ML engineer Spotify


Kirtan Padh, PhD-candidate ML TU München


Mackenzie Jorgensen,
PhD-candidate Computer Science King's College


Krsto Proroković, software engineer

Join Algorithm Audit's Expert Hub Slack channel – drop an email to info@algorithmaudit.eu

Approach summary: Combining the power of data-driven bias testing with context-sensitive human evaluation

Key strengths of our approach

Joint approach	Bridging the quantitative and qualitative reasoning paradigm
Unsupervised machine learning	Unsupervised machine learning to automatically detect complex forms of bias that go beyond established protected grounds
Deliberative	Deliberative, expert-led assessment by diverse AI policy professionals
Case repository	Over time a case repository emerges from which data scientists and public authorities can distill 'techno-ethical' best-practices

Source code can be found on GitHub

All of Algorithm Audit's case studies can be found on our website

Algorithm Audit

We build public knowledge for ethical algorithms

www.algorithmaudit.eu

info@algorithmaudit.eu

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

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

Stichting Algorithm Audit is registered as a non-profit organisation at the Dutch Chambre of Commerce under license number 83979212