Ethical advice report

Case

Higher-dimensional bias in a
BERT-based disinformation classifier

Source

- Self-trained NLP classifier on Twitter1516 data
- Self-developed unsupervised bias scan tool

Summary advice

The audit commission believes there is a low risk of (higher-dimensional) proxy discrimination by the BERT-based disinformation classifier and that the particular difference in treatment identified by the quantitative bias scan can be justified, if certain conditions apply.

Mar-3rd 2023

Algorithm Audit

Algorithm Audit

Problem statement

Fair treatment by a BERT-based Twitter disinformation classifier

We use a quantitative bias scan tool to assess fair treatment of a self-trained disinformation detection algorithm on Twitter data. This document presents statistically significant disparities found by the tool. The results are submitted to a commission of human experts. This audit commission formulates normative advice if, and how, (higher-dimensional) proxy discrimination and/or ethically undesirable forms of differentiation could be investigated further.

1. Introduction

Unfair treatment by algorithms is multi-faceted. A first concern is one-dimensional proxy discrimination. Proxy discrimination concerns unlawful differentiation based on an apparently neutral feature (such as *literacy rate*) that is critically linked to a protected ground as specified in legal directives[1] (such as *ethnicity*). A second concern is ethically undesirable forms of differentiation. Algorithms can differentiate upon a seemingly innocuous feature, such as browser type or house number suffix. This type of differentiation evades non-discrimination law, as many features are not critically linked to a protected ground, but can still be perceived as unfair, for instance if it reinforces socio-economic inequality. A third concern is higher-dimensional forms of unfair treatment. Algorithms differentiate upon clusters that are defined by a mixture of features. Higher-dimensional forms of algorithmic differentiation are difficult to detect for humans. Let alone to assess whether the cluster is involved in proxy discrimination and/or ethically undesirable forms of differentiation. In theory, statistical methods are capable to detect both higher- and one-dimensional forms of undesirable differentiation. In this case study, we use a statistical bias scan tool to examine in practice whether the above concerns can be overcome.

[1] In the European Union (EU), the European Convention of Human Rights (ECHR) serves as the legal fundament against discrimination. Additional EU directives (2000/43/EC, 2000/78/EC, 2004/113/EC, and 2006/54/EC) provide context-specific protection, e.g., persons with disabilities, employment rights, and consumer protection.

2. Unsupervised bias scan

The bias scan tool[2] identifies clusters for which a binary classification algorithm is systematically misclassifying, i.e., predicting a different class than the ground truth label in the data. A cluster is a group of datapoints sharing similar features. The tool makes use of unsupervised clustering[3] and therefore does not require *a priori* information about existing disparities and protected attributes of users (which are often not available in practice).

For this case study, we review a BERT-based disinformation classification algorithm[4] which is trained on the Twitter1516 dataset[5], enriched with self-collected Twitter API data[6]. The dataset consists of 1,057 verified true and false tweets, 3 user features (verified profile, #followers, user engagement) and 5 content features (length, #URLs, #mentions, #hashtags, sentiment score). We run two bias scans. In Scan 1, the bias metric is defined by the False Positive Rate (FPR). FPR relates to true content predicted to be false, proportional to all true content. In Scan 2, the bias metric is defined by the False Negative Rate (FNR). FNR relates to false content predicted to be true, proportional to all false content. In sum:

Scan 1. $\text{Bias} = \text{FPR}_{\text{cluster}} - \text{FPR}_{\text{rest of dataset}}$

Scan 2. $\text{Bias} = \text{FNR}_{\text{cluster}} - \text{FNR}_{\text{rest of dataset}}$.

The full bias scan pipeline is displayed in Figure 1.

Figure 1 – Bias scan pipeline for the disinformation classifier.

[2] Misztal-Radecka, Indurkya, Bias-Aware Hierarchical Clustering for detecting the discriminated groups of users in recommendation systems, *Information Processing and Management* (2021).
[3] Documentation about the k-means Hierarchical Bias-Aware Clustering (HBAC) algorithm: https://github.com/NGO-Algorithm-Audit/Bias_scan/blob/master/Technical_documentation_bias_scan.pdf
[4] More information about the self-trained BERT-based classification algorithm: https://github.com/NGO-Algorithm-Audit/Bias_scan/blob/master/HBAC_scan/HBAC_BERT_disinformation_classifier.ipynb
[5] 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)
[6] More information on the data collection process: https://github.com/NGO-Algorithm-Audit/Bias_scan/blob/master/data/Twitter_dataset/Twitter_API_data_collection.ipynb

3. Results: Identified quantitative disparities

For Scan 1, the cluster for which the disinformation classifier is underperforming the most (bias=0.08, n=249) is characterized by the features displayed in Figure 2. Difference in means is the difference in standardized feature means between the disparately treated cluster and the rest of the dataset. Hypothesis testing[7] indicates that on average, user that:

- are verified, have higher #followers, user engagement and #URLs;
- use less #hashtags and have lower tweet length

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

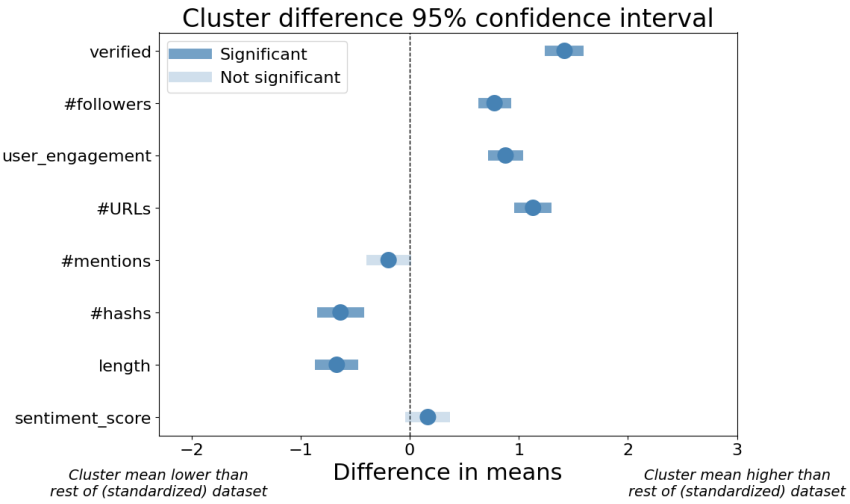

Figure 2 – Identified quantitative feature disparities in cluster with highest bias. For this bias scan, bias is defined by the False Positive Rate.

For Scan 2, the cluster for which the disinformation classifier is underperforming the most (bias=0.13, n=46) is characterized by the features displayed in Figure 3. Hypothesis testing indicates that on average, user 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).

[7] Here, the hypothesis tested is that there is no difference in feature means of the cluster and the pooled feature means of other clusters. These differences are statistically significant even after performing a Bonferroni correction to adjust for false discoveries due to multiple hypothesis testing.

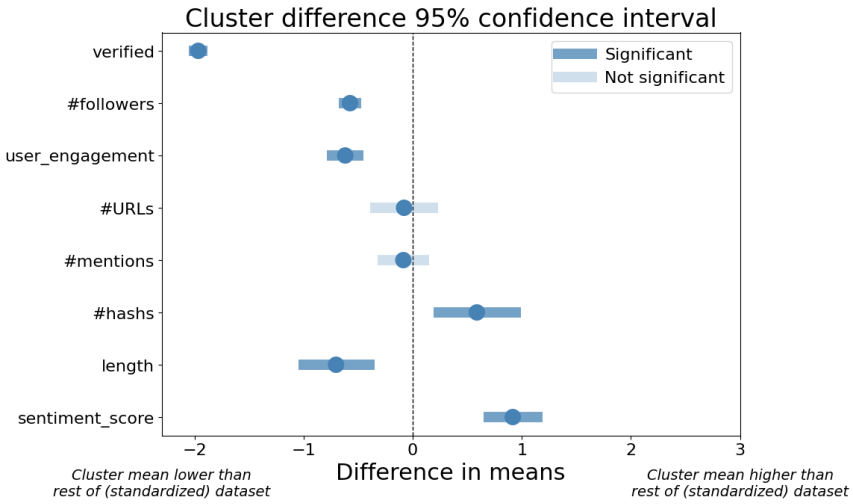

Figure 3 – Identified quantitative feature disparities in cluster with highest bias. For this bias scan, bias is defined by the False Negative Rate.

These results might indicate (higher-dimensional) unfair treatment by the disinformation classifier. More information on the identified clusters and robustness tests of the results can be found in the Appendix.

4. Qualitative assessment of identified disparities

The identified disparities in Section 3 do not establish prohibited *prima facie* discrimination. Rather, the identified disparities serve as a starting point to assess potential unfair treatment according to the context-sensitive qualitative doctrine. To assess unfair treatment, we question:

i) Is there an indication that one of the statistically significant features, or a combination of the features, stated in Figure 2-3 are critically linked to one or multiple protected grounds?

ii) 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)?

iii) 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?

iv) 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?

Auditing disinformation detection algorithms

As of December 2022, Article 28 of the European Digital Services Act (DSA) subjects very large online platforms (VLOPs) to annual independent auditing of their services and risk mitigation measures. Open-source AI auditing tools, such as this bias scan tool, help to detect and mitigate (higher-dimensional) forms of unfair treatment in disinformation detection and other AI (ranking and recommender) systems.

With this case study, Algorithm Audit aims to provide qualitative guidelines how statistical methods can be used to monitor unfair treatment by AI systems. Without clear guidance from data protection boards, supervisors, researchers and algorithmic regulatory bodies, misinterpretation of quantitative metrics stands in the way of independent quantitative and qualitative oversight of the risk of biased AI systems. Building on the quantitative results of the bias scan, Algorithm Audit provides qualitative justifications to make a normative judgment about whether AI systems are causing unfair treatment or not.

Normative advice

Compiled answers of audit commission

i) Is there an indication that one of the statistically significant features, or a combination of the features, stated in Figure 2-3 are critically linked to one or multiple protected grounds?

No, the audit commission considers none of the features (as displayed in Figure 2-3) are critically linked to protected grounds, as defined in Article 14 of the European Convention on Human Rights.

To examine possible proxy discrimination, the audit commission deems relevant the composition of users with respect to certain features that might affect the spread of disinformation, e.g., verified profile, high follower count. For instance, verification of a user profile tends to be linked to prominence (public figures, celebrities, high follower counts, etc) or to paid subscribers and could therefore act as a proxy for socio-economic status. While this is not a legally protected category, it can induce unfairness. A better understanding of the user composition would allow better assessment of linkage between user characteristics and legally protected grounds. Yet on the basis of current knowledge of this model and social context, there is no suspicion of critical linkage to protected grounds. Further investigation would be warranted.

ii) 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)?

The audit commission considers both labeling true content as false (false positive) and labeling false content as true (false negative) to be harmful. They are not harmful in the same way or to an equal extent, however. Commission members assert that labeling true content as false (false positives) are more likely to be harmful to individual authors through direct causal mechanisms, e.g., content moderation. In addition it undermines trust in the content moderation process,

thereby raising concerns over free speech and provoking suspicions of hidden political motives. Labeling false content as true (false negatives) are mainly harmful for the general public, as they lead to the spread of disinformation, in addition to undermining trust in the content moderation process. While false positives with harmful impact are generally acted upon by the platform (e.g., by flagging/removing content and/or author suspension), false negatives are generally not. This also means that, compared to false negatives, false positives will be more frequently contested by authors. The majority view of the audit commission is that it is more harmful to classify true content as false (false positives), for the following reasons:

- Users have a reasonable expectation that true content will not be unjustly flagged. However, given that false content is a common feature of the internet, they do not expect that all false content will be properly flagged.
- To users, false positives are perceived as much more unfair and undermines their confidence in moderation systems. It thereby also raises concerns about freedom of speech and hidden motives for flagging the true content as being false.

The audit commission stresses that violations of users' right of freedom of expression should not be taken lightly and that equal treatment is key. This is especially the case where automatic classification of disinformation would trigger the direct removal of the post or the author.

The audit commission recommends that the following mitigations be put in place. In the case where disinformation is being classified and subsequently analyzed by human moderators, it is less risky to have more false positives than false negatives. A main reason is that qualitative investigation of the tweets classified as false would then be able to identify and correct misclassification errors. The commission also highlights the importance of reasonable recourse and effective redress mechanisms, including the provision of intelligible reasons for the classification.

iii) 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 *prima facie* unjustified. As such, there is no decisive reason why these rates would be too high. This case study does however not present enough detailed information to make such a judgment for this specific case.

The audit commission recommends to take certain measures when assessing such quantitative discrepancies in general:

- It is better to determine quantitative indicators *ex ante*, rather than squinting at numbers and deciding if they are reasonable *ex post*. Ideally, organizations such as Twitter would publicly commit to certain thresholds of tolerance for labeling posts as true or false;
- It is desirable to examine relations between false positive misclassification, false negative misclassification and cluster composition, in terms of user characteristics, e.g., to what extent clusters are composed of politicians, political leaning, journalists, other professional personas, state- or business-sponsored users, socially privileged users, and other socioeconomic factors. This helps to evaluate the level of discriminatory effect and social impact of misclassification biases;
- The justification of disparities needs to include an assessment of the possibility of recourse. Groups facing higher rates of false positives should be able to easily challenge decisions affecting them. An adequate feedback loop would support the effective detection and mitigation of classification biases.

The audit commission argues that using a model with unequal misclassification rates across groups can be justified, if it is closely monitored, documented and motivated, if warning systems are in place, and perhaps most importantly, if misclassifications can be corrected easily and adequately.

iv) 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.

The audit commission agrees that some degree of difference across various clusters is inevitable. The commission notes that a higher level of scrutiny towards users broadcasting to bigger audiences, and towards users that enjoy credibility boosters (such as verified profiles), can be legitimate. As studies show, disinformation spread by high-profile users can have cascade effects with potentially high impact. In addition, the metrics for establishing which users are high-profile (and hence high-risk) are relatively transparent and content-agnostic, which means they can be perceived as ideological to a lesser extent. On the other hand, commission members highlight the fact that for specific high-profile users, a higher false positive

rate is particularly problematic. This includes journalists, who could suffer serious harm if their messages are unjustly classified as false. Also, prominent cases of false positive misclassification could seriously undermine the credibility of content moderation processes and the media in general. Once again, the transparent communication of the automated character of content moderation and the implied risks, and also the possibility of accessible and effective recourse and redress mechanisms, are key to mitigating these risks.

Commission members make a general remark that social media companies are often inclined to be less harsh towards high-profile users, as they lead to higher user engagement. Higher leniency of high-profile users that are under the pressure of monetization should be actively countered. The observed discrepancies for this model are therefore perceived by the commission as less problematic, than if it would show the reverse; a higher leniency for users with higher number of followers, verifications, and user engagement.
In general, the audit commission argues that the following conditions should apply:

- The use of the classification model should be a purposeful decision that is clearly documented and well-communicated;
- Documentation should include reasoning about why the classification model displays such discrepancies, and why the benefits of the model outweigh the risks;
- A continuous learning process should be implemented, which constantly challenges occurring biases, correlations with demographic backgrounds (see question 3) and which tries to reduce and mitigate discriminatory effects;
- Understanding the root causes of disinformation spread and the agents behind them needs to inform the model.

Audit commission facts

This advice is the outcome of a collective audit process. Hence, any specific statement does not necessarily reflect the views of each individual audit commission member. Individual members cannot be held accountable for this advice.

Date

The audit commission provided written answers in February 2023. This report has been approved by all commission members and was affirmed on March 1st 2023.

Composition audit commission

- Anne Meuwese, Professor in Public Law & AI, Leiden University
- Hinda Haned, Professor in Responsible Data Science, University of Amsterdam
- Raphaële Xenidis, Assistant Professor in EU law, Sciences Po

- Aileen Nielsen, Fellow in Law & Tech, ETH Zürich
- Carlos Hernández-Echevarría, Assistant Director and Head of Public Policy at the anti-disinformation nonprofit fact-checker Maldita.es
- Ellen Judson, Head of CASM and Sophia Knight, Researcher, CASM at Britain's leading cross-party think tank Demos

Appendix

BERT-based classifier performance

The confusion matrix of the BERT-based disinformation classifier is displayed in Figure 1. More information regarding the training process can be found on Github[8].

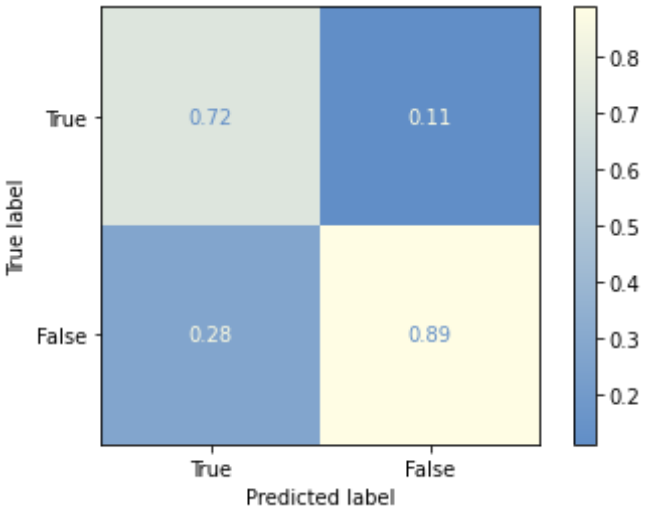

Figure 1 – Confusion matrix of the BERT-based disinformation classifier on the test set.

Bias scan results

For the FPR scan, the following clusters are detected:

- Cluster 0 has bias -0.062
- Cluster 1 has bias 0.0810
- Cluster 2 has bias -0.089
- Cluster 3 has bias -0.041.

Cluster 1 has the highest bias (FPR): 0.08. There are 249 elements in cluster. Table 1 displays all numeric values of the difference in feature means between the cluster and the rest of the dataset, including p-values of a Welch's two-samples t-test for unequal variances to examine whether the differences are statistically significant (p-value<0.05). Note that Table 1 is displayed in Figure 2.

	Difference	p-value
Verified profile	1.419	0.000
#followers	0.778	0.000

[8] https://github.com/NGO-Algorithm-Audit/Bias_scan

User engagement[9]	0.878	0.000
#URLs	1.130	0.000
#mentions	-0.669	0.064
#hashtags	-0.634	0.000
Length	-0.669	0.000
Sentiment score[10]	0.167	0.115

Table 1 – Difference in feature means between cluster with highest bias (FPR) and the rest of the dataset. Rows in blue display a statistically significant difference according to a Welch's two-samples t-test for unequal variances (p<0.05).

For the FNR scan, the following clusters are detected:

- Cluster 0 has bias -0.101
- Cluster 1 has bias 0.118
- Cluster 2 has bias -0.059
- Cluster 3 has bias 0.132

Cluster 3 has the highest bias (FNR): 0.13. There are 46 elements in cluster. Table 2 displays all numeric values of the difference in feature means between the cluster and the rest of the dataset, including p-values of a Welch's two-samples t-test for unequal variances to examine whether the differences are statistically significant (p-value<0.05). Note that Table 2 is displayed in**Error! Reference source not found.** Figure 3.

	Difference	p-value
Verified profile	-1.965	0.000
#followers	-0.575	0.000
User engagement[9]	-0.619	0.000
#URLs	-0.079	0.607
#mentions	-0.086	0.465
#hashtags	0.588	0.004
Length	-0.702	0.000
Sentiment score[10]	0.917	0.000

[9] More information on the user engagement metric can be found in spread of true and false news online. *Science*.

[10] For sentiment score see: https://github.com/cjhutto/vaderSentiment

Table 2 – Difference in feature means between cluster with highest bias (FNR) and the rest of the dataset. Rows in blue display a statistically significant difference according to a Welch's two-samples t-test for unequal variances (p<0.05).

Sensitivity testing

The k-means HBAC algorithm uses various hyperparameters. In this section, we provide a rationale for our choices for these parameters. In addition, we refer to sensitivity testing that echo the results as presented in Section 3.

Parameters prevent HBAC to find only clusters with a small amount of datapoints, for which it is hard to find meaningful features. An overview and description of all hyperparameters is given in Table 3.

Number of initial clusters (Our choice: 2)	The desired number of initial clusters of the k-means clustering algorithm.
Maximum number of iterations (Our choice: 300)	The HBAC algorithm is terminated after the maximum number of iteration threshold is reached, or after no clusters are found that have a higher discrimination bias when compared to the clusters of the previous iteration.
Minimal splitable cluster size (Our choice: 29)	Number of elements that need to be in the cluster to be eligible for a next cluster split.
Minimal acceptable cluster size (Our choice: 21)	Number of elements in a new candidate cluster during splitting to be accepted as a new cluster.

Table 3 – Hyperparameters of the HBAC algorithm.

We run the FPR and FNR scan for the following 162 configuration of hyperparameters:

- Number of initial clusters: 2 and 3;
- Minimal splitable cluster size: 5, 10, 15, 20, 25, 30, 35, 40 and 45;
- Minimal acceptable cluster size: 5, 10, 15, 20, 25, 30, 35, 40 and 45.

We compute the average of all clusters with positive (FPR or FNR) bias. This results in:

- 2974 clusters with positive FPR bias;
- 2506 clusters with positive FNR bias.

We take a fraction of 0.07 and 0.05 from the original test data size as minimal splitable cluster size (0.07*413=29) and minimal acceptable cluster size (0.05*413=21) respectively. More information on the results of these sensitivity tests can be found on GitHub[11].

[11] https://github.com/NGO-Algorithm-Audit/Bias_scan/blob/master/HBAC_scan/HBAC_BERT_disinformation_sensitivity_testing.ipynb