Problem statement – Fair treatment by a BERT-based disinformation classifierWe 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, (multi-dimensional) proxy discrimination and/or ethically undesirable forms of differentiation can be assessed.

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]](#footnote-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 social-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 bias scan tool in practice to examine whether the above challenges can be overcome.

2. Quantitative bias scan   
The bias scan tool[[2]](#footnote-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]](#footnote-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 self-trained BERT-based disinformation classification algorithm[[4]](#footnote-4) on the Twitter15 dataset[[5]](#footnote-5), enriched with Twitter API data. The cluster for which the disinformation classifier is underperforming the most (bias[[6]](#footnote-6)=-0.27) is characterized by the features displayed in Table 1. The feature difference is the computed average in means between the disparately treated cluster and the rest of the dataset. Hypothesis testing indicates a statistically significant disinformation classification bias (p<0.05) against users with a verified profile, above average sentiment score and below average number of URLs used in their tweets[[7]](#footnote-7). These results might indicate (higher-dimensional) unfair treatment by the disinformation classifier on the basis of these three features. More information on the identified clusters and robustness tests on the observed results can be found in the Appendix.

|  |  |  |
| --- | --- | --- |
|  | Difference | p-value |
| Verified profile | 0.53468 | 0.000 |
| Sentiment score[[8]](#footnote-8) | 0.95686 | 0.000 |
| #URLs | -0.74095 | 0.000 |
| Length | 0.38785 | 0.189 |
| #hashtags | -0.22793 | 0.313 |
| User engagement8 | -0.14634 | 0.385 |
| #mentions | -0.14453 | 0.496 |
| #followers | -0.07927 | 0.646 |

Table 1 – Statistically significant misclassification by the BERT-based disinformation classifier is observed for users with a verified profile, tweets with an above average sentiment score and below average number of URLs. A p-value smaller than 0.05 (highlighted) indicates that there is more evidence for classification bias than one would expect due to chance.  
  
3. Qualitative assessment of bias scan results  
These observations 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 examine proxy discrimination, we question:

1. Is there an indication that one of the three statistically significant features, or a combination of the features, is critically linked to one or multiple protected grounds?
2. Is the measured disparate treatment of users with one, or multiple, of the three statistically significant features disproportional when compared to other users?[[9]](#footnote-9)
3. Can the measured disparate treatment be justified given the aim pursued?

In relation to ethically undesirable forms of differentiation, we question:

1. Considering the disparate treatment of users with a verified profile, above average sentiment score and/or below average number of URLs used in their tweets, could the observed disparate treatment be perceived as ethically undesirable?

Auditing disinformation detection algorithms

Article 28 of the European Digital Services Act (DSA) subjects social media platforms 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 black-box algorithms used to detect disinformation. Through this case study, we aim to provide qualitative guidelines how the quantitative results can be interpreted.

Appendix

Dataset

The Twitter15 dataset contains 742 true rumors and false rumors (tweets). The dataset has been enriched by eight self-collected features using the Twitter API: number of followers of the account, number of mentions/URLs/hashtags in the tweet, Twitter’s engagement metric, length of the tweet, verified user profile and the sentiment score of the tweet based on the VADER sentiment analysis tool7.

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 Github4.

Chart, treemap chart

Description automatically generated with medium confidence

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

Bias cluster pipeline

The data pipeline between the BERT-based disinformation classifier and the HBAC scan is displayed in

Graphical user interface, diagram

Description automatically generated

Figure 2 – Data pipeline between AI classifier and HBAC scan.

Bias scan results

We use Principal Component Analysis (PCA) dimension reduction to visualize the 8-dimensional identified clusters (see Figure 2). A negative bias is observed for clusters 4-7.

* Cluster 4 has bias -0.270;
* Cluster 6 has bias -0.143;
* Cluster 5 has bias -0.126;
* Cluster 7 has bias -0.037.

Chart, scatter chart

Description automatically generated

Figure 3 – Two-dimensional representation of the identified clusters by the k-means HBAC scan.

Sensitivity testing

For cluster 4, the identified disparities including statistical hypothesis tests are shown in Table 1. For cluster 5 and 6, the identified disparities, including statistical hypothesis tests, are displayed in Table 3 and Table 4 respectively. In each of the tables, statistically significant differences in means between the clusters and the rest of the dataset are found for the features verified Twitter profile, sentiment score and number of URLs used in tweets. Overall, across clusters, these features may thus be deemed most important to investigate with respect to classification bias.

The results of the HBAC scan depend on the hyperparameters shown in Table 2.

|  |  |
| --- | --- |
| * max\_iter | HBAC is terminated after it reaches the maximum iteration threshold or after no clusters are found that have a higher discrimination bias when compared to the clusters of the previous iteration. |
| * minimal splittable cluster size (5), minimal acceptable cluster size (9) | Parameters that prevent HBAC to find only clusters with a small amount of datapoints, for which it is hard to find meaningful features. |

Table 2 – Hyperparameters of the HBAC algorithm.

We thus perform a sensitivity analysis of the results by changing the hyperparameters and then rerunning the analysis. Results echo the identified disparate treatment for users with a verified Twitter profile, above average sentiment score and/or number of URLs used in tweets.

|  |  |  |
| --- | --- | --- |
| Cluster 5 | | |
|  | Difference | p-value |
| Verified profile | 0.519 | 0.000 |
| Sentiment score6 | -1.362 | 0.000 |
| #URLs | -0.811 | 0.007 |
| User engagement6 | -0.258 | 0.137 |
| length | 0.452 | 0.322 |
| #hashtags | -0.266 | 0.357 |
| #followers | 0.022 | 0.931 |
| #mentions | 0.002 | 0.996 |

Table 3 – Observed disparities for cluster 5. Statistically significant difference in means between the cluster and the dataset are highlighted (p<0.05).

|  |  |  |
| --- | --- | --- |
| Cluster 6 | | |
|  | Difference | p-value |
| User engagement6 | -0.819 | 0.000 |
| Verified profile | 0.528 | 0.000 |
| #followers | -0.549 | 0.000 |
| #mentions | -0.345 | 0.000 |
| #URLs | -0.758 | 0.000 |
| Sentiment score6 | 0.848 | 0.001 |
| Length | 0.460 | 0.063 |
| #hashtags | 0.031 | 0.917 |

Table 4 – Observed disparities for cluster 6. Statistically significant difference in means between the cluster and the dataset are highlighted (p<0.05).

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, labor law, and good and services. [↑](#footnote-ref-1)
2. Misztal-Radecka, Indurkya, Bias-Aware Hierarchical Clustering for detecting the discriminated groups of users in recommendation systems, *Information Processing and Management* (2021). [↑](#footnote-ref-2)
3. Documentation about the k-means Hierarchical Bias-Aware Clustering (HBAC) algorithm can be found here: <https://github.com/NGO-Algorithm-Audit/Bias_scan/blob/master/Bias_scan_tool_report.pdf> [↑](#footnote-ref-3)
4. More information about the self-trained BERT-based classification algorithm can be found here: https://github.com/NGO-Algorithm-Audit/Bias\_scan/blob/master/case\_studies/BERT\_disinformation\_classifier [↑](#footnote-ref-4)
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) [↑](#footnote-ref-5)
6. Bias is defined here as the difference in accuracy between the identified cluster and the rest of the data set. Accuracy is defined as the fraction of correctly classified labels relative to the overall group size. [↑](#footnote-ref-6)
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. [↑](#footnote-ref-7)
8. For sentiment score see: <https://github.com/cjhutto/vaderSentiment>. For user engagement metric see: <https://developer.twitter.com/en/docs/twitter-api/enterprise/engagement-api/overview> [↑](#footnote-ref-8)
9. Disparate treatment relates here to the most deviating cluster, characterized by the features in Table 1, for which -0.27 bias is measured. This means that for this dataset 27% [↑](#footnote-ref-9)