Adversarial Removal of Demographic Attributes Revisited

Maria Barrett[1], **Yova Kementchedjhieva**[1], **Yanai Elazar**[2],
Desmond Elliott[1], **Anders Søgaard**[1]
[1]University of Copenhagen, [2]Bar-Ilan University
{mjb,yova,de,soegaard}@di.ku.dk,yanaiela@gmail.com

Abstract

Elazar and Goldberg (2018) showed that protected attributes can be extracted from the representations of a *debiased* neural network for mention detection at above-chance levels, by evaluating a diagnostic classifier on a held-out subsample of the data it was trained on. We revisit their experiments and conduct a series of follow-up experiments showing that, in fact, the diagnostic classifier generalizes poorly to both new in-domain samples and new domains, indicating that it relies on correlations specific to their particular data sample. We further show that a diagnostic classifier trained on the *biased* baseline neural network also does not generalize to new samples. In other words, the biases detected in Elazar and Goldberg (2018) seem restricted to their particular data sample, and would therefore not bias the decisions of the model on new samples, whether in-domain or out-of-domain. In light of this, we discuss better methodologies for detecting bias in our models.

1 Introduction

Several approaches have been proposed to learn classifiers that are invariant (unbiased with respect) to protected attributes: cost-sensitive (Agarwal et al., 2018), regularization-based (Bechavod and Ligett, 2017), and adversarial (Ganin and Lempitsky, 2015). In the adversarial approach, a model learns representations x that should be predictive for a main task y and oblivious to a protected attribute z.

Adversarial training has been used to learn data representations that are invariant to demographic attributes (Raff and Sylvester, 2018; Beutel et al., 2017; Li et al., 2018), as well as representations invariant to domain differences (Ganin and Lempitsky, 2015), clean or noisy speech (Sriram et al., 2018), and invariant to differences between source languages in multi-lingual machine translation (Xie et al., 2017). Elazar and Goldberg (2018) argue, however, that adversarial learning does not fully remove sensitive demographic traits from the data representations. This conclusion is based on the observation that a diagnostic classifier trained over the supposedly debiased data representations could still predict gender, age and race above chance level in their experimental setup.

In general, diagnostic classifiers are trained on data representations to predict the protected demographic attributes in question as well as possible, i.e., the classifier picks up on any correlations, strong or weak, between data representations and the demographic classes. Correlations come in different flavours: PREVALENT: Certain features are indicative of gender in most contexts, e.g., the distribution of phrases like 'as a mother' or 'as a guy'. SAMPLE-SPECIFIC: Some features are indicative of gender within a particular domain/sample, e.g. *bling* may correlate with a particular range of ages in a sample of reviews of accessories, and yet another range in reviews of movies. ACCIDENTAL: Yet, other features may show a correlation with the protected attribute in a given dataset, while in fact being completely unrelated to it. The correlation is unexpected and particular to a finite data sample.

This paper presents a follow-up to the experiments in Elazar and Goldberg (2018) and examines what kind of correlation the data representations in their models exhibit with demographic attributes: PREVALENT, SAMPLE-SPECIFIC or ACCIDENTAL correlations. We do this by *not only* evaluating the diagnostic classifiers on in-sample data, but also on new samples, as well as across textual domains. We also explore what the models learn in a more qualitative manner.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing
and the 9th International Joint Conference on Natural Language Processing, pages 6330–6335,
Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

	GENDER			AGE		
	Train	Dev	Test	Train	Dev	Test
PAN16 TWIT	148	10	10	148	10	10
CF TWIT	–	–	6.4	–	–	–
100 AUTH TWIT	–	–	49.6	–	–	–
PAN14 BLOGS	–	–	12.8	–	–	8
PAN14 REVIEWS	–	–	13	–	–	8.4
PAN14 SOME	–	–	10	–	–	10
PAN16 RAND	–	–	10	–	–	10

Table 1: Dataset sizes (in 1000 sentences).

Contributions Our contributions are methodological. We show that the diagnostic classifiers used to establish gender and age bias in Elazar and Goldberg (2018) (a) rely only on sample-specific patterns to predict gender and age, and (b) do therefore not generalize to new samples or domains. Surprisingly, we also show that this also holds for the biased baseline model in their experiments, suggesting that the particular representations induced for the mention detection task in Elazar and Goldberg (2018) were not biased with respect to protected demographic attributes. This does not show whether the data *is* biased, or whether adversarial training is a good or poor debiasing technique; merely that Elazar and Goldberg (2018) did not properly establish any of these things. Our contributions are, as said, methodological, and we believe this case study highlights the key difficulties establishing bias in model representations.

2 Adversarial Attribute Removal with Diagnostic Classifiers

In adversarial attribute removal, a model is trained with a two-fold objective: learning to solve a main task and *un*learning to predict a protected attribute (+ADV). A model trained without the second objective is referred to as a *non-adversarial* model (-ADV). The architecture of Elazar and Goldberg (2018) consists of a single-layer LSTM encoder, and two multi-layer perceptrons – one for each task. The main task $p(y|x)$ is learned as usual, with loss being backpropagated through the relevant perceptron and the encoder. The attribute unlearning is achieved by training the relevant perceptron to predict $p(z|x)$, where z is the protected attribute, while also punishing the encoder for letting any signal through that could allow the perceptron to do so (Ganin and Lempitsky, 2015). After training, a diagnostic classifier, an *attacker*, is used to evaluate the effectiveness of the adversarial training. The language encoder is used to obtain representations of the input data, and the diagnostic classifier is trained to predict the protected attributes from these representations, without access to the encoder or to the original inputs. Since the dataset is balanced and the targets are binary, *leakage* is defined in Elazar and Goldberg (2018) as any demographic attribute prediction accuracy over 50.0% by the diagnostic classifier.

Preprocessing Table 1 shows sizes for all used datasets. Our main dataset, PAN16 TWIT (Rangel et al., 2016), is split into train and development, following Elazar and Goldberg (2018). We further remove 10,000 sentences from their train data to use as a held-out test split, making sure there is no author overlap between training and test data. This means that we have 12,000 fewer sentences in our train split than Elazar and Goldberg (2018), but we report results on exactly the same development split as well as on the new held-out test split. PAN16 TWIT is balanced using undersampling with respect to main task and demographic attribute (gender and age respectively) which is why there are separate datasets for GENDER and AGE. Our main observation here is that training, development and test splits and random subsamples of one sample of data. Using random subsamples this way is common in machine learning, including bias detection studies (Elazar and Goldberg, 2018; Zhao et al., 2019) and probing studies (Ravfogel et al., 2018; Lin et al., 2019), but is known to overestimate performance (Globerson and Roweis, 2016), in particular for high-dimensional problems.

Replication We start by replicating the experiment of Elazar and Goldberg (2018) using their code on PAN16 TWIT with their data splits. The main task is predicting the mentions of other Twitter users after removing all user names in the tweets. The protected demographic attributes are age and gender, both with binary targets. Our development results (main and diagnostic classifier) which are comparable to Elazar and Goldberg (2018) are reported in Table 6 in the Appendix; test set results are also in Table 4. Our results remain comparable to those obtained in Elazar and Goldberg (2018), albeit the diagnostic classifier is able to achieve 3.92 percentage points less leakage for gender and 2.59 percentage points for age, possibly due to the reduction in training data. Adver-

	Age		Gender	
	Mean	RStD	Mean	RStD
-Adv	14.48	0.34	5.40	0.16
+Adv	5.3	0.45	3.68	0.22

Table 2: Mean and relative standard deviation of the *leakage* of 10 attackers trained on different subsamples of the training data (Pan16 Twit). Evaluated on the test set.

sarial training reduces the leakage of demographic attributes, but the diagnostic classifier is still able to predict the sensitive demographic attribute from the data representations significantly better than chance (line 1 of Table 4). Significance test is obtained by checking subsampled test set accuracies against the 95% confidence interval of a random distribution.

3 In-Sample Analysis

In §4, we use out-of-sample and cross-domain evaluation to show that the model trained above relies on spurious or ACCIDENTAL correlations. In this section, as a supplement, we train 10 attackers on different random subsamples of 50% of the training data to explore the robustness of the observed leakage. Each attacker is evaluated on the same Pan16 Twit test data. Furthermore, we present a qualitative analysis of what the above model has learned.

Table 2 shows the mean and relative standard deviation (RStD) of the leakage of the 10 diagnostic classifiers in each setting. The leakage, unsurprisingly, is larger for the non-adversarial condition than the adversarial condition.

Extracting leaked samples We further analyse the tweets that were correctly labelled by the different attackers. Given that these are binary classification tasks, we consider labels predicted with a high probability to be leaked, and labels predicted with a probability close to 0.5 to be randomly chosen. We sort all of the correctly labelled test examples by the probability assigned to the label by the attacker. We denote the top n samples as the leaked samples, where n is the amount of leakage observed for a model[1].

	Gender			Age		
	M	F	n	Y	O	n
-Adv	52.1	47.9	142	69.1	30.9	553
+Adv	100.	0.	37	57.9	42.1	145

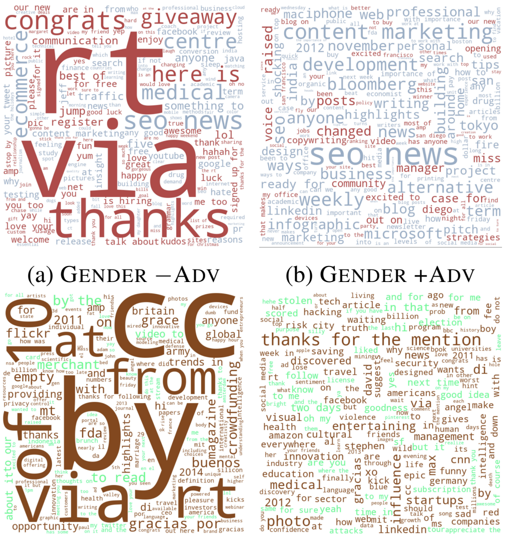
Table 3: Class distribution for the confidently predicted tweets and $n$umber of tweets per model condition for M(ale), F(emale) and Y(oung), O(ld)

We use the intersection of leaked samples across the attackers, under the assumption that samples that were correctly labelled by all ten models (and with high probability) are most likely to exhibit protected attribute leakage. After deduplicating, we have 179 sentences correctly labelled for Gender, and 698 sentences for age, as reported in Table 3. We call these the LEAKED sentences. (There are no female-authored sentences that are confidently predicted after adversarial training).

Exploring leaked samples Out of the LEAKED gender sentences, the top 10 sentences which received the highest confident scores are presented in Table 7 in the Appendix. 28 subjects rated how confident they were about the gender of the author on a 9-point scale. Subjects were recruited via the authors' professional and private networks and were not compensated for the time they spent. All subjects were unaware of the origin of the tweets, and the sentences were presented in a random order. The scale was calibrated such that the extremes represented *Very certain* about the author being a man (1) or a woman (9), with 5 a neutral middle value. The mean for sentences by male authors selected without adversarial training is 4.68 and for females 5.69. With adversarial training, this mean is 4.83. The mean of sentences extracted from the model trained with adversarial training is closest to the neutral value 5, but independent t-tests show that the differences between all classes are insignificant ($p > 0.05$). Therefore, the results show that humans had difficulties determining the gender of the author. This is contrary to the findings of Flekova et al. (2016), but is similar by the unaveraged results of Burger et al. (2011). This indicates that the data did not exhibit (m)any obvious predictors of gender or age.

In addition to this study, we also use this data to visualize what our diagnostic classifiers focus on. For this purpose, we use the models from Section 2, trained on the full Pan16 Twit dataset, and perform feature analysis using uptraining: Us-

[1] e.g., with a subsampled test accuracy of 54.4%, we take the 4.4% most confidently predicted tweets for each subsampled model and only use the intersection from all models.



(a) GENDER −ADV (b) GENDER +ADV

(c) AGE −ADV (d) AGE +ADV

Figure 1: Biased n-grams in a linear simulation of the attacker model in Elazar and Goldberg (2018). Red=female, blue=male, green=young, brown=old. The size corresponds to the size of the coefficient.

ing a transparent, linear model to approximate our deep model (Bastani et al., 2017), we train a logistic regression model with L1 regularization on the train set representations, relabeled by the predictions of our diagnostic classifier. We then study the coefficients of this linear model. The data is represented using the 10,000 most frequent n-grams.[2] We also use the confidence of the diagnostic classifier for importance weighting during training; note that the predicted classes are not balanced.[3] For gender and age, and for our baseline and our adversarially trained system, we uptrain 10 models on random 50% subsets of the training data and average coefficients to filter out noise. Figure 1 shows word clouds for the n-grams in the LEAKED sentences, with word size reflecting the size of the avg. coefficients in the uptrained linear models. We observe that for models with no adversarial training, a few n-grams dominate the coefficients. In the models where adversarial training was used, there are no large coefficients dominating the space, suggesting that the model succeeded in removing strong correlates; we note that these correlates do not intuitively relate strongly to gender or age.

	AGE		GENDER	
TEST SET	-ADV	+ADV	-ADV	+ADV
PAN16 TWIT	67.86*	53.45*	56.79*	53.74*
PAN16 RAND	50.13	50.32	49.50	50.00
100 AUTH TWIT	–	–	50.92	52.57
CF TWITTER	–	–	51.63*	50.74*
BLOGS	50.63*	55.12*	50.85	49.95
REVIEWS	51.03	50.26	50.44	49.28
SOME	50.32	50.16	50.34	48.34

Table 4: Cross-sample diagnostic classifier accuracy of classification of demographic attribute when trained on PAN16 TWIT data and evaluated on different test sets. Significantly different from random *=$p < 0.01$.

4 Out-of-Sample and Cross-Domain Evaluation

In our main set of experiments, we now evaluate the adversarial and non-adversarial diagnostic classifiers across different samples and different domains. We use the following datasets for these experiments: PAN14 (Rangel et al., 2014)[4] - we include the English data from PAN14 for the following domains: BLOGS, hotel REVIEWS, and Social Media (SOME). The PAN14 data are not annotated for the mention task, but this is not necessary to evaluate the diagnostic classifier on the protected, demographic attribute. We also use data from Crowdflower's Gender Classifier data (CF TWIT), manually annotated for gender[5]. We also made a new Twitter dataset with manually annotated gender for 100 authors (100 AUTH TWIT) for the purpose of evaluating on a different Twitter sample, following the approach used to create PAN16 TWIT – detailed in Appendix A. The datasets are balanced with respect to demographic attributes, and the Twitter datasets are balanced with respect to mentions. We also construct an artificial dataset from PAN16 TWIT where the main task labels are preserved but the demographic label is randomly shuffled (PAN16 RAND), allowing us to run experiments with no PREVALENT or SAMPLE-SPECIFIC gender correlations, only ACCIDENTAL. This experiment supplements our in-sample analysis in §3 and shows that our models are not overly expressive, indicating that the sample-specific correlations detected

[2] n-gram representations are less expressive than LSTM encodings, but given the short length of tweets, we assume up to 5-gram representation capture most relevant patterns.

[3] Table 5 in the Appendix provides an overview of the class distribution.

[4] We do not include PAN14-Twitter, since the PAN16 TWIT training data subsumes the data from PAN14.

[5] Downloaded from https://d1p17r2m4rzlbo.cloudfront.net/wp-content/uploads/2016/03/gender-classifier-DFE-791531.csv

in Elazar and Goldberg (2018) are relatively simple associations.

The results in Table 4 show, however, that the performance and leakage of the adversarial models do *not* generalize well across domains, but for the most part the performance of the non-adversarial models doesn't either. This is the main result presented here: The leakage shown in Elazar and Goldberg (2018) does not transfer across domains, but also does not even generalize to a new sample within the same domain. This suggests that the leakage is merely spurious correlations in a small, finite sample of data. Accuracies are, as expected, mostly lower when adversarial training is employed, but for the 100 AUTH TWIT sample, the non-adversarial accuracy is close to random, while the adversarial accuracy is two points higher: it appears that the adversarial component has merely served as a regularizer during training.

5 Conclusion and Discussion

We examined the methodology used in Elazar and Goldberg (2018) to establish bias in the representations of adversarial machine learning architectures designed to protect demographic attributes.[6] Our results suggest that when measuring demographic parity using a diagnostic classifier, one needs to be careful in controlling for spurious correlations that are limited to just one specific sample of data. In-sample correlations are not necessarily meaningful in high-dimensional problems, and while they can be lead to worse performance on test time (Globerson and Roweis, 2016), they do not lead to demographic bias in practice, when our models are deployed on new samples of data. In order to get more realistic assessments of representation bias with respect to protected attributes, we therefore need to test the out-of-sample generalization of detected bias/leakage, *and*, perhaps, to qualitatively inspect the observed leakage patterns.

We repeat here that none of our results say anything about whether the datasets used in Elazar and Goldberg (2018) are biased or not, or whether other models induced from this data are likely to exhibit biases. Our contribution is mainly methodological: What we have shown is that the methodology in Elazar and Goldberg (2018), i.e., in-sample evaluation of diagnostic classifiers, is not sufficient to establish bias/leakage beyond the current data sample. Instead we propose out-of-sample and cross-domain evaluation, as well as more qualitative investigation of the induced diagnostic classifiers. Our results are also orthogonal to the main contribution of Elazar and Goldberg (2018), which is to show that adversarial debiasing is not always able to remove bias.

Acknowledgements

Maria Barrett is sponsored by a Facebook Research Award; Anders Søgaard is sponsored by a Facebook Research Award and a Google Focused Research Award.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69.

Osbert Bastani, Carolyn Kim, and Hamsa Bastani. 2017. Interpretability via model extraction. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML 2017)*.

Yahav Bechavod and Katrina Ligett. 2017. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.

John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301–1309. Association for Computational Linguistics.

Yanai Elazar and Yoav Goldberg. 2018. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 11–21. Association for Computational Linguistics.

Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel PreoŢiuc-Pietro. 2016. Analyzing biases in human perception of user age and gender from text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 843–854. Association for Computational Linguistics.

[6] The methodology is not only found in Elazar and Goldberg (2018). The same methodology can be found in other papers on detecting bias in machine learning models, for example, Zhao et al. (2019), as well as in several papers probing neural network representations for linguistic knowledge, for example, Lin et al. (2019).

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189.

Amir Globerson and Sam Roweis. 2016. Nightmare at test time: Robust learning by feature deletion. In *Proceedings of the 23rd International Conference on Machine Learning*.

Yitong Li, Timothy Baldwin, and Trevor Cohn. 2018. Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 25–30. Association for Computational Linguistics.

Yongjie Lin, Yi Chern Tan, and Robert Frank. 2019. Open sesame: Getting inside BERT's linguistic knowledge. In *The 2nd BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*.

Edward Raff and Jared Sylvester. 2018. Gradient reversal against discrimination. In *Proceedings of the 5th Workshop on Fairness, Accountability and Transparency in Machine Learning*.

Francisco Rangel, Paolo Rosso, Irina Chugur, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, and Walter Daelemans. 2014. Overview of the 2nd author profiling task at PAN 2014. In *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014*, pages 1–30.

Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. In *Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings/Balog, Krisztian [edit.]; et al.*, pages 750–784.

Shauli Ravfogel, Francis Tyers, and Yoav Goldberg. 2018. Can LSTM learn to capture agreement? the case of Basque. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP*.

Anuroop Sriram, Heewoo Jun, Yashesh Gaur, and Sanjeev Satheesh. 2018. Robust speech recognition using generative adversarial networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5639–5643. IEEE.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. 2017. Controllable invariance through adversarial feature learning. In *Advances in Neural Information Processing Systems*, pages 585–596.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the North American Chapter of the Association for Computational Lingustics*, pages 629–634.