Fairness in Natural Language Processing

Timothy Baldwin

Melbourne Laureate Professor; Director of the ARC Training Centre in Cognitive Computing for Med Tech

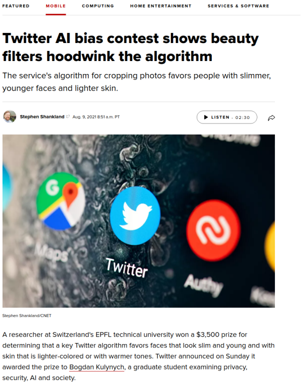
THE UNIVERSITY OF MELBOURNE

Talk Outline

1. Introduction
2. Baseline Approaches for Mitigating Bias
3. Mitigating Bias with Adversarial Learning
   - Issue 1: Training Instability
   - Issue 2: (Un)Availability of Protected Attributes
   - Current Work
4. Summary

Examples of “Unfair” AI

- Twitter’s auto-cropping algorithm:



Twitter AI bias contest shows beauty filters hoodwink the algorithm

The service's algorithm for cropping photos favors people with slimmer, younger faces and lighter skin.

Stephen Shankland Aug. 9, 2021 8:51 a.m. PT

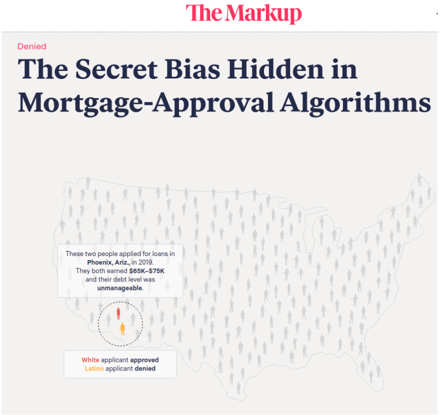
Stephen Shankland/CNET

A researcher at Switzerland's EPFL technical university won a $3,500 prize for determining that a key Twitter algorithm favors faces that look slim and young and with skin that is lighter-colored or with warmer tones. Twitter announced on Sunday it awarded the prize to Bogdan Kulynych, a graduate student examining privacy, security, AI and society.

Source(s): `cnet.com/tech/mobile/twitter-ai-bias-contest-shows-beauty-filters-hoodwink-the-algorithm/`

Examples of “Unfair” AI

- Bias in US mortgage approvals:



The Markup

Denied

The Secret Bias Hidden in Mortgage-Approval Algorithms

These two people applied for loans in **Phoenix, Ariz.,** in 2019. They both earned **$65K–$75K** and their debt level was **unmanageable**.

White applicant **approved**
Latino applicant **denied**

Source(s): `themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms`

Examples of “Unfair” AI

- And an NLP one with particular personal significance:

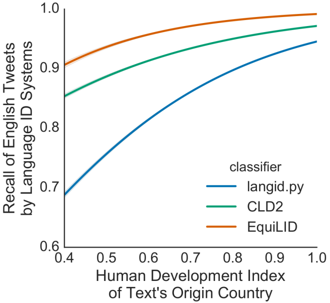


Figure 2: Estimated recall of tweets with health-related terms according to a logit regression on the Human Development Index of the tweet’s origin country; bands show 95% confidence interval.

Source(s): Jurgens et al. (2017)

So What is Fairness?

- In discussing fairness in NLP, we tend to distinguish between “source” and “target” bias:
  - **source** (or “author”) bias is based on the protected attributes of the author of a given text, for example:

> It’s an adventure which reaches back to golden-age Hollywood and the devil-may-care world of Douglas Fairbanks or Tyrone Power playing Zorro, or Errol Flynn playing Robin Hood.

Source(s): `https://www.theguardian.com/film/2017/oct/23/the-princess-bride-review-rob-reiner`

So What is Fairness?

- **target** (or “reference”) bias is based on the protected attributes of people referred to in a given text, e.g. *her* in:

> Eliza Bryant was born in North Carolina to Polly Simmons, a slave, and her master.

Source(s): Webster et al. (2018)

So What is Fairness?

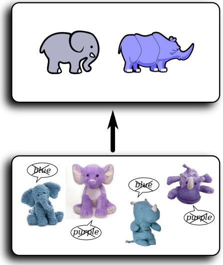
- Particular “protected attributes” of interest (gender, “race”, age, sexuality, ...) vary greatly across datasets, based on data availability, the task, and the specific interests of the researchers, e.g.:
  - hatespeech and race (Huang et al., 2020)
  - review ratings and age/country (Hovy and Søgaard, 2015)
  - syntactic analysis and age/gender/variety (Hovy and Søgaard, 2015)
  - pronoun resolution and (binary) gender (Webster et al., 2018)

Talk Outline

1. Introduction
2. Baseline Approaches for Mitigating Bias
3. Mitigating Bias with Adversarial Learning
   - Issue 1: Training Instability
   - Issue 2: (Un)Availability of Protected Attributes
   - Current Work
4. Summary

Training Models to be Fair(er): Take 1

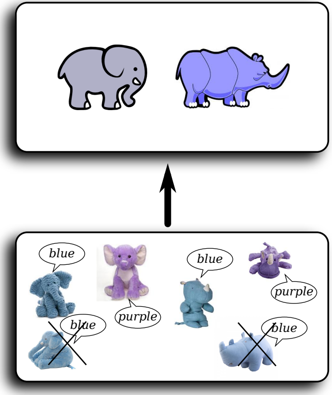
- Surely if we don't include any protected attributes in the training data, our models will be fair, right?! ("fairness as blindness")



- What tends to happen in practice is that naively-trained models not only reproduce biases but **amplify** them (Zhao et al., 2017)

Training Models to be Fair(er): Take 2

- In cases where the cause is imbalance in terms of protected attributes (= representation), pre-balance the training data



Training Models to be Fair(er): Take 2

- An effective strategy in terms of fairness, but it tends to drive down raw performance (severely in cases of extreme under-representation), as the amount of training data is limited

  ... is it possible to instead maintain aggregate performance while achieving fair(er)ness?

Talk Outline

1. Introduction

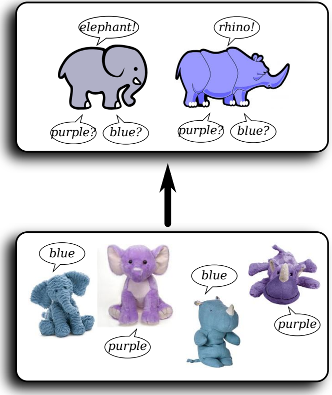
2. Baseline Approaches for Mitigating Bias

3. Mitigating Bias with Adversarial Learning
   - Issue 1: Training Instability
   - Issue 2: (Un)Availability of Protected Attributes
   - Current Work

4. Summary

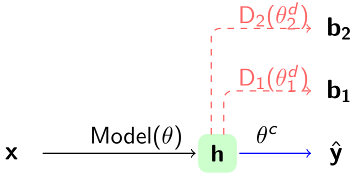
Training Models to be Fair(er): Take 3

- Let's instead train a model to be as **good** as possible at predicting the target variable, and **bad** as possible at recovering the protected attributes from the hidden representation



Fairness through Adversarial Learning: Overview

- We do this with one “adversarial discriminator” per protected attribute:



$$\hat{\theta} = \min_{\theta_M} \max_{\{\theta_{\mathrm{D}^i}\}_{i=1}^N} \mathcal{X}(\hat{\mathbf{y}}(\mathbf{x};\theta), \mathbf{y}) - \sum_{i=1}^{N} \left(\lambda_i \cdot \mathcal{X}(\hat{b}(\mathbf{x};\theta^d_i), b_i) \right)$$

Source(s): Ganin et al. (2016); Li et al. (2018)

Experiment 1: POS Tagging

Task: POS tagging (based on Google Universal POS tagset)

Model: biLSTM; adversarial discriminator = single feed-forward layer applied to final hidden representation ($[\mathbf{h}_n; \mathbf{h}'_0]$)

Datasets:

- training domain = English Web Treebank for pre-training (Bies et al., 2012), and TrustPilot for fine-tuning (Hovy and Søgaard, 2015)
- test domains = TrustPilot + AAVE POS dataset (Jørgensen et al., 2016)

Protected attributes:

- age (under-35 vs. over-45)
- gender (male vs. female)

Evaluation: accuracy for both in-domain and cross-domain settings

Source(s): Li et al. (2018)

Experiment 1: POS Tagging

- POS accuracy [%] over Trustpilot test set, stratified by SEX and AGE:

	SEX			AGE		
	F	M	Δ	O45	U35	Δ
BASELINE	90.9	91.1	0.2	91.4	89.9	1.5
ADV	**92.2**	**92.1**	**0.1**	**92.3**	**92.0**	**0.3**

Source(s): Li et al. (2018)

Experiment 1: POS Tagging

- POS accuracy [%] over Trustpilot test set, stratified by SEX and AGE:

	SEX			AGE		
	F	M	Δ	O45	U35	Δ
BASELINE	90.9	91.1	0.2	91.4	89.9	1.5
ADV	**92.2**	**92.1**	**0.1**	**92.3**	**92.0**	**0.3**

- POS accuracy [%] over AAVE dataset:

	LYRICS	SUBTITLES	TWEETS	Average
BASELINE	73.7	81.4	59.9	71.7
ADV	**80.5**	**85.8**	**65.4**	**77.0**

Source(s): Li et al. (2018)

Experiment 2: Sentiment Analysis

Task: (English) sentiment classification (5-way)

Model: CNN; adversarial discriminator = single feed-forward layer applied to final hidden representation

Dataset: TrustPilot (cross-validation, with dev partition)

Protected attributes:

- age (under-35 vs. over-45)
- gender (male vs. female)
- location (US, UK, Germany, Denmark, and France)

Evaluation: micro-averaged F-score

Source(s): Li et al. (2018)

Experiment 2: Sentiment Analysis

	F_1		**h** Leakage [%]		
	dev	test	AGE	SEX	LOC
Majority class			57.8	62.3	20.0
BASELINE	41.9	40.1	65.3	66.9	53.4
ADV-AGE	**42.7**	40.1	**61.1**	65.6	41.0
ADV-SEX	42.4	39.9	61.8	62.9	42.7
ADV-LOC	42.0	**40.2**	62.2	66.8	**22.1**
ADV-all	42.0	**40.2**	61.8	**62.5**	28.1

Source(s): Li et al. (2018)

Findings

- Largely similar “in-register” results, but considerably better balance across protected attributes
- Greatly improved “cross-register” accuracy for POS tagging(!)
- Much lower leakage over hidden representations for test users

Contents

1. Introduction
2. Baseline Approaches for Mitigating Bias
3. Mitigating Bias with Adversarial Learning
    - Issue 1: Training Instability
    - Issue 2: (Un)Availability of Protected Attributes
    - Current Work
4. Summary

Fairness through Adversarial Learning: Training Issues

- While adversarial learning can be highly effective, it comes with some caveats in terms of model training:
  - instability (sometimes works very well, sometimes not well at all)
  - difficult to train/failure to converge

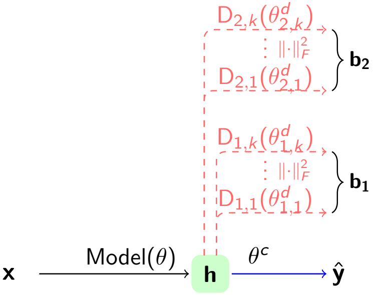
Fairness through Adversarial Learning: Training Issues

- While adversarial learning can be highly effective, it comes with some caveats in terms of model training:
  - instability (sometimes works very well, sometimes not well at all)
  - difficult to train/failure to converge
- Possible "engineering" solutions to the problem include (Elazar and Goldberg, 2018):
  - add extra capacity to the adversarial discriminator(s)
  - upweight the adversarial loss term(s) λ_i
  - ensemble the adversarial discriminator(s)

... which all work to varying degrees over different datasets/data settings, but still lack consistency

Enforcing Diversity in the Adversarial Discriminators

- One issue with the naive ensembling approach is that there is no constraint to ensure the adversaries complement one another
  - **Fix:** introduce an orthogonality constraint on the parameters of the adversaries in a given ensemble (Bousmalis et al., 2016):



Enforcing Diversity in the Adversarial Discriminators

- The orthogonality constraint is applied pairwise over the hidden representation generated by different sub-discriminators:

$$\mathcal{L}_{\text{diff}} = \lambda_{\text{diff}} \sum_{i,j \in \{1,\ldots,k\}} \left\| {h_{A_i}}^{\intercal} h_{A_j} \right\|_F^2 \mathbf{1}(i \neq j),$$

  where $\|\cdot\|_F^2$ is the squared Frobenius norm

- Advantage of being minimised when hidden representations shrink to zero (prevents the model from learning rotated hidden representations: Bousmalis et al. (2016))

Source(s): Han et al. (2021b)

Experiment: Sentiment Analysis

Task: (English) sentiment classification (2-way)

Model: DeepMoji fixed encoder (Felbo et al., 2017) + dense linear layer; adversarial discriminators = 3-layer MLP

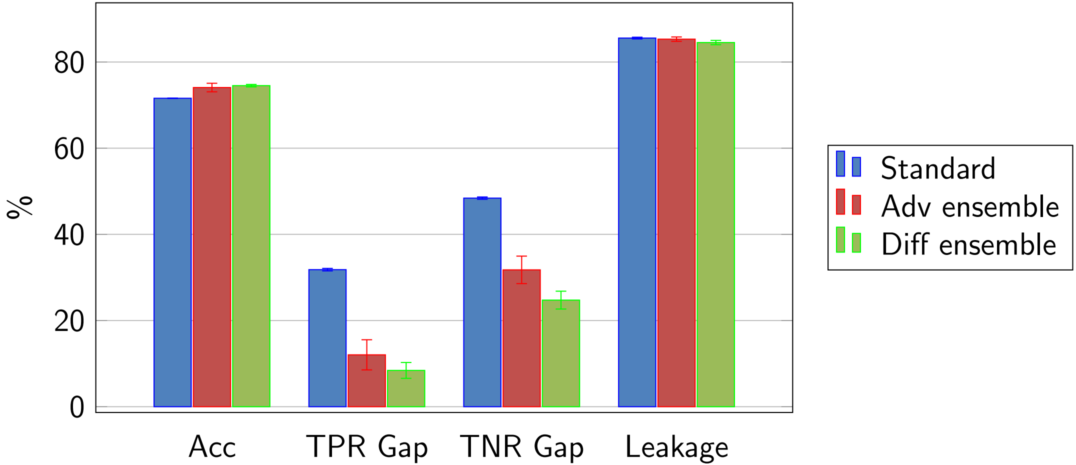
Dataset: Twitter sentiment analysis (Blodgett et al., 2016)

Protected attribute: “race” (AAE vs. SAE)

Evaluation: accuracy, TPR/TNR Gap, **h** Leakage

Source(s): Han et al. (2021b)

Experiment: Sentiment Analysis



Findings

- Ensemble of discriminators ($>$ single discriminator) $>$ vanilla model
- Better, more stable results with orthogonality constraint on ensemble
- Particularly effective in reducing GAP metrics (Leakage still high in large part because encoder fixed)

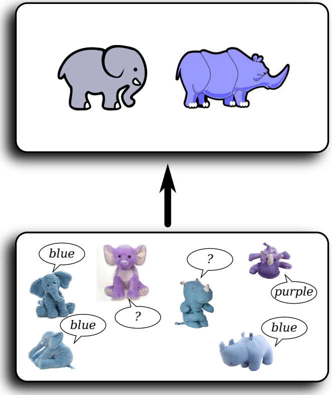
Source(s): Han et al. (2021b)

Contents

1. Introduction
2. Baseline Approaches for Mitigating Bias
3. Mitigating Bias with Adversarial Learning
   - Issue 1: Training Instability
   - Issue 2: (Un)Availability of Protected Attributes
   - Current Work
4. Summary

Fairness through Adversarial Learning: Data Issues

- A critical **data** issue (not specific to adversarial learning) is the availability of protected attributes in the training data:

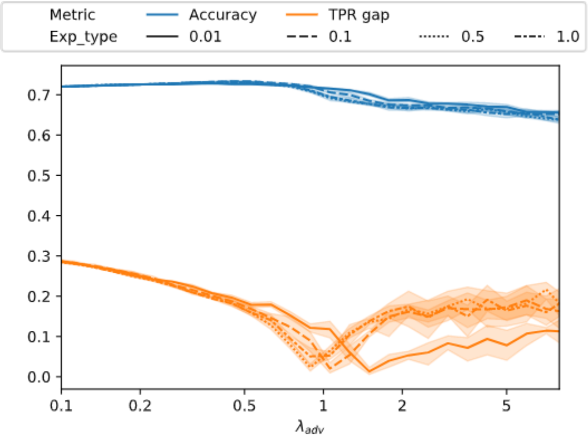


Fairness through Adversarial Learning: Data Issues

- Because the training of the adversarial discriminators is decoupled from the base classifier, we can trivially handle this by training the discriminators on whatever subset of the data has labels for protected attributes
- It is also possible to transfer protected attributes between tasks/datasets with different target variables

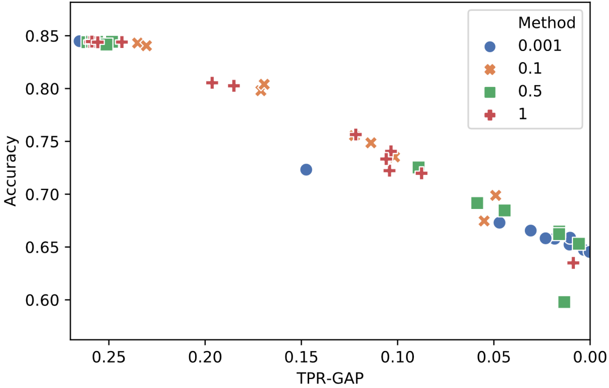
Source(s): Han et al. (2021a)

Experiment 1: In-domain Sentiment Analysis



Source(s): Han et al. (2021a)

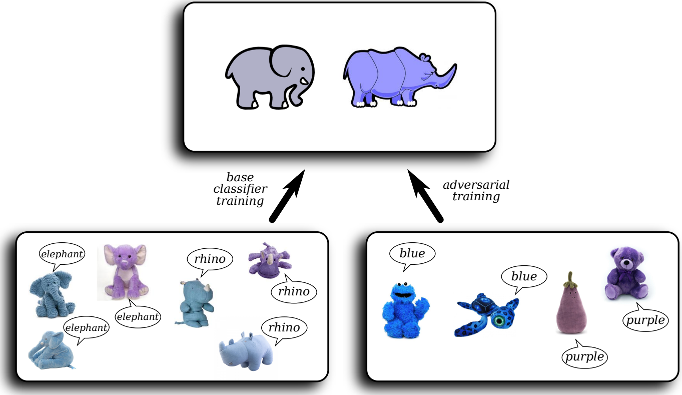
Experiment 2: In-domain Hatespeech Detection



Source(s): Han et al. (2021a); Huang et al. (2020)

Out-of-domain Setting

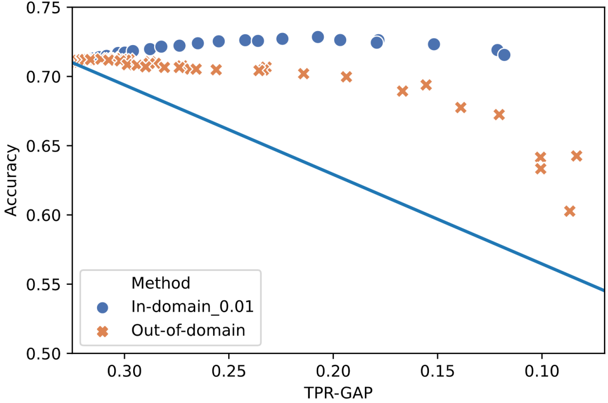
- In the fully-decoupled setting, the training instances for the base classifier and the adversarial discriminators are completely disjoint:



Source(s): Han et al. (2021a)

Experiment 3: Cross-domain Sentiment Analysis

- Base classifier task = sentiment analysis; adversarial task = "race" classification ("white" vs. "other", from hatespeech dataset)



Source(s): Han et al. (2021a)

Adversarial Training over Partial/Externally-labelled Data

- (Very) small amounts of data labelled with protected attributes go a very long way
- Fully decoupled training (based on externally-labelled data for protected attributes) can lead to substantial reductions in GAP, with some sacrifice in target label accuracy (relative to in-domain labelled data)
- Also results for cross-domain instance selection based on “predictability” of the protected attribute in context of POS tagging, showing modest gains in both accuracy and GAP

Source(s): Han et al. (2021a)

Contents

1. Introduction

2. Baseline Approaches for Mitigating Bias

3. Mitigating Bias with Adversarial Learning
    - Issue 1: Training Instability
    - Issue 2: (Un)Availability of Protected Attributes
    - Current Work

4. Summary

Training Models to be Fair(er): Other Issues/Approaches

- Instance weighting (Subramanian et al., 2021b,a)
- Maximum-margin methods (Subramanian et al., 2021b)
- Constrained learning (Subramanian et al., 2021a)
- Contrastive learning
- Training models to be fairer in terms of intersectional bias/gerrymandering (Subramanian et al., 2021a)

Talk Outline

1. Introduction
2. Baseline Approaches for Mitigating Bias
3. Mitigating Bias with Adversarial Learning
   - Issue 1: Training Instability
   - Issue 2: (Un)Availability of Protected Attributes
   - Current Work
4. Summary

Summary

- Adversarial learning is an effective technique for mitigating (source) bias in NLP tasks, to achieve fairer model outcomes
- Stability + effectiveness of adversarial models can be improved by ensembling discriminators with orthogonality constraints on the adversaries for a given attribute
- Remarkably few annotations for protected attributes needed to train adversaries, and possible to combine training instances for target label and protected attribute(s)
- Still much more to be done, and lots more room for improvement, to deliver on goal of truly fair (non-degenerate) models!

Acknowledgements

- Joint work with Yitong Li, Xudong Han, and Trevor Cohn
- Supported by Australian Research Council

References

Bies, A., Mott, J., Warner, C., and Kulick, S. (2012). English Web Treebank. *Linguistic Data Consortium, Philadelphia, USA*.

Blodgett, S. L., Green, L., and O'Connor, B. (2016). Demographic dialectal variation in social media: A case study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1130, Austin, USA.

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., and Erhan, D. (2016). Domain separation networks. In *Advances in Neural Information Processing Systems*, pages 343–351.

Elazar, Y. and Goldberg, Y. (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, Brussels, Belgium.

References

Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., and Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1615–1625, Copenhagen, Denmark.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17:59:1–59:35.

Han, X., Baldwin, T., and Cohn, T. (2021a). Decoupling adversarial training for fair NLP. In *Findings of ACL*.

Han, X., Baldwin, T., and Cohn, T. (2021b). Diverse adversaries for mitigating bias in training. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2760–2765, Online.

References

Hovy, D. and Søgaard, A. (2015). Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 483–488.

Huang, X., Xing, L., Dernoncourt, F., and Paul, M. (2020). Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1440–1448.

Jørgensen, A., Hovy, D., and Søgaard, A. (2016). Learning a POS tagger for AAVE-like language. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1115–1120.

References

Jurgens, D., Tsvetkov, Y., and Jurafsky, D. (2017). Incorporating dialectal variability for socially equitable language identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017), Volume 2: Short Papers*, pages 51–57.

Li, Y., Baldwin, T., and Cohn, T. (2018). Towards robust and privacy-preserving text representations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, Melbourne, Australia.

Subramanian, S., Han, X., Baldwin, T., Cohn, T., and Frermann, L. (2021a). Evaluating debiasing techniques for intersectional biases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*.

Subramanian, S., Rahimi, A., Baldwin, T., Cohn, T., and Frermann, L. (2021b). Fairness-aware class imbalanced learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)*.

References

Webster, K., Recasens, M., Axelrod, V., and Baldridge, J. (2018). Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017). Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 2979–2989.