Fairness-aware Class Imbalanced Learning

Shivashankar Subramanian[♣] **Afshin Rahimi**[♠]
Timothy Baldwin[♣] **Trevor Cohn**[♣] **Lea Frermann**[♣]
[♣]School of Computing and Information Systems, University of Melbourne
[♠]School of Information Technology and Electrical Engineering, University of Queensland
shivashankarrs@gmail.com a.rahimi@uq.edu.au
{tbaldwin,t.cohn,lfrermann}@unimelb.edu.au

Abstract

Class imbalance is a common challenge in many NLP tasks, and has clear connections to bias, in that bias in training data often leads to higher accuracy for majority groups at the expense of minority groups. However there has traditionally been a disconnect between research on class-imbalanced learning and mitigating bias, and only recently have the two been looked at through a common lens. In this work we evaluate long-tail learning methods for tweet sentiment and occupation classification, and extend a margin-loss based approach with methods to enforce fairness. We empirically show through controlled experiments that the proposed approaches help mitigate both class imbalance and demographic biases.[1]

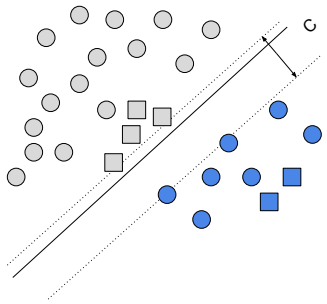

Figure 1: Example of a two-class problem where grey and blue points denote majority and minority classes, respectively, and circles and squares denote two sub-groups. Imbalanced learning methods such as LDAM (Cao et al., 2019) maximise the (soft-)margin for minority classes and do not consider sub-groups within each class.

1 Introduction

Class imbalance is common in many NLP tasks, including machine reading comprehension (Li et al., 2020), authorship attribution (Caragea et al., 2019), toxic language detection (Breitfeller et al., 2019), and text classification (Tian et al., 2020). A skewed class distribution hurts the performance of deep learning models (Buda et al., 2018), and approaches such as instance weighting (Lin et al., 2017; Cui et al., 2019; Li et al., 2020), data augmentation (Juuti et al., 2020; Wei and Zou, 2019), and weighted max-margin (Cao et al., 2019) are commonly used to alleviate the problem.

Bias in data often also manifests as skewed distributions, especially when considered in combination with class labels. This is often referred to as "stereotyping" whereby one or more private attributes are associated more frequently with certain target labels, for instance more *men* being employed as *surgeons* than *women*. Prior work has identified several classes of bias, including bias towards demographic groups based on gender, disability, race or religion (Caliskan et al., 2017; May et al., 2019; Garimella et al., 2019; Nangia et al., 2020), and bias towards individuals (Prabhakaran et al., 2019). Methods to mitigate these biases include data augmentation (Badjatiya et al., 2019), adversarial learning (Li et al., 2018), instance weighting based on group membership (Kamiran and Calders, 2011), regularization (Wick et al., 2019; Kennedy et al., 2020), and explicit subspace removal (Bolukbasi et al., 2016; Ravfogel et al., 2020).

This paper draws a connection between class-imbalanced learning and stereotyping bias. Most work has focused on class-imbalanced learning and bias mitigation as separate problems, but the unfairness caused by social biases is often aggravated by the presence of class imbalance (Yan et al., 2020). Class-imbalanced learning approaches improve the performance of minority classes at some cost to the performance of majority classes. A common approach re-weights instances in the training objective to be proportional to the inverse frequency of their class. Approaches such as FOCAL (Lin

[1]Code available at: https://github.com/shivashankarrs/classimb_fairness

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 2045–2051
November 7–11, 2021.

et al., 2017) and DICE (Li et al., 2020) extend this approach by down-weighting "easy" instances. Label-Distribution-Aware Margin Loss ("LDAM": Cao et al. (2019)) is an alternative approach, which encourages a larger margin for the minority class, but it does not consider sub-group proportions (see Figure 1). On the other hand, debiasing approaches do not typically focus on class imbalance explicitly. For instance, in toxicity classification, certain sub-groups are often predicted more confidently for toxicity (encouraging false negatives for the majority sub-group), which tend to be close to the margin for the non-toxic class (encouraging false positives; Borkan et al. (2019)).

In this work, we modify the training objective of LDAM as a state-of-the-art approach for imbalanced learning so that margins depend not just on class-imbalance, but also on the subgroup distribution within each class. Specifically, we extend LDAM with popular debiasing strategies. We show the effectiveness of our approach through several controlled experiments on two text classification data sets.

2 Proposed Approaches

Let $(\mathbf{x}, y, g)$ denote a training instance, comprising an input, label, and group identifier, respectively. LDAM (Cao et al., 2019) addresses class imbalance by enforcing a larger margin for minority classes:

$$\mathcal{L}_{\text{LDAM}}(\mathbf{x}, y; f) = -\log \frac{e^{z_y - \Delta_y}}{e^{z_y - \Delta_y} + \sum_{j \neq y} e^{z_j}}$$

$$\Delta_j = \frac{C}{n_j^{1/4}} \text{ for } j \in \{1, \ldots, k\}$$

where $\mathbf{z} = f(\mathbf{x})$ are the model outputs, k is the number of classes, n_j is the number of instances in class j, and C is a hyperparameter. Smaller classes are associated with a larger Δ_y, which is subtracted from the model output z_y, thus enforcing a larger margin.

We propose three extensions to LDAM, each of which takes into account imbalance in the distribution of private attributes across classes:

LDAM$_{\text{iw}}$ adds instance re-weighting, based on groups within each class:

$$\mathcal{L}_{\text{LDAM}_{\text{iw}}}(\mathbf{x}, y, g; f) = \omega_{y,g} \mathcal{L}_{\text{LDAM}}(x, y; f)$$

where g is the group of instance $\mathbf{x}$; and $\omega_{y,g} = \frac{1-\beta}{1-\beta^{N_{y,g}}}$ weights each class–group combination based on its smoothed inverse frequency. β is a constant set to 0.9999 (Cui et al., 2019) and $N_{y,g}$ is the number of instances belonging to class y and group g. Mistakes on minority groups within minority classes are penalised most.

LDAM$_{\text{adv}}$ adds an adversarial term, so that the learnt representations are a poor predictor of group membership (Li et al., 2018):

$$\mathcal{L}_{\text{LDAM}_{\text{adv}}}(\mathbf{x}, y, g; f) = \mathcal{L}_{\text{LDAM}}(\mathbf{x}, y; f) - \lambda_{\text{adv}} \text{CE}(g, l(\mathbf{x}))$$

where f shares the lower layers of the network with the adversary l, CE denotes cross-entropy loss, and λ_{adv} is a hyperparameter. This objective $\mathcal{L}_{\text{LDAM}_{\text{adv}}}$ is jointly minimised *wrt* f and maximised *wrt* l. The penalty results in hidden representations that are informative for the main classification task (f), but uninformative for the adversarial group membership prediction task (l). The adversarial loss is implemented using gradient reversal (Ganin and Lempitsky, 2015).

LDAM$_{\text{reg}}$ adds a soft regularization term which encourages fairness as *equalised odds* by reducing maximum mean discrepancy (Gretton et al., 2012) across groups. The probability of predicting some class k for any individual group g should be close to k's probability over the whole data set:

$$\mathcal{L}_{\text{LDAM}_{\text{reg}}}(\mathbf{X}, \mathbf{y}, \mathbf{g}; f) = \mathcal{L}_{\text{LDAM}}(\mathbf{X}, \mathbf{y}; f) + \rho \sum_g \left\| \frac{1}{N_g} \sum_{i: g_i = g} f(\mathbf{x}_i) - \frac{1}{N} \sum_i f(\mathbf{x}_i) \right\|^2$$

where we have moved from single instance loss to the loss over the full training set, N_g denotes the number of training instances in group g, and hyper-parameter ρ controls the trade-off between performance and fairness.

3 Experimental Results

We perform experiments on the two tasks of emoji prediction and occupation classification, both of which are binary classification tasks with binary protected attributes.

Emoji prediction: We use the Twitter dataset of Blodgett et al. (2016), where tweets are associated with the private attribute race (black/white), and sentiment labels are derived from emoji usage (happy/sad) (Elazar

and Goldberg, 2018). We experiment with different levels of class and stereotyping imbalance in the Emoji dataset, including its original distribution (see Section 3.1).

Occupation classification: This data set consists of short biographies scraped from the web, annotated for private attribute gender (male/female) and target occupation labels (De-Arteaga et al., 2019). We focus on two occupations with well-documented gender stereotypes – surgeon and nurse. The resulting dataset is mildly class-imbalanced (59% surgeon: 41% nurse), with roughly symmetric natural gender splits (90% male for surgeon and 90% female for nurse).

For emoji prediction we follow Ravfogel et al. (2020) and use the DeepMoji encoder, which was trained on millions of tweets and is known to encode demographic information (Elazar and Goldberg, 2018). For occupation classification, we use the BERT-base uncased model and classify via the last hidden state of the CLS token (Devlin et al., 2019). Both encoders are followed by a single hidden layer MLP.

We evaluate classification performance based on macro-averaged F-score (to account for class imbalance), and evaluate fairness using performance GAP: the average of the true positive rate (TPR) and true negative rate (TNR) differences between the two subgroups (De-Arteaga et al., 2019; Ravfogel et al., 2020). Note that a wide variety of fairness measures (both on the group- and individual levels) have been proposed, which are impossible to satisfy simultaneously (Garg et al., 2020). Often, a suitable measure is chosen based on the target application. Here we use the popular *equalised odds* measure considering both TPR and TNR of classifiers, in order to address scenarios where certain subgroups are predicted more often with some classes (see Section 1). We report fairness as $1-$GAP, such that higher numbers are better, and a perfectly fair model achieves $1-\text{GAP} = 1$. We compare our methods against the following benchmarks:

vanilla: unweighted cross-entropy loss.

FOCAL: re-weights easy examples during training (Lin et al., 2017).

CW: instance re-weighting based on the inverse class proportion and cross-entropy.

IW: instance re-weighting based on the combination of inverse class and group proportions, and cross-entropy (Kamiran and Calders, 2011).

INLP: Iterative null-space projection (Ravfogel et al., 2020): in each iteration, we learn a SVM classifier W using hidden representations (X_h) as the independent variables to predict the protected attribute, where X_h is projected onto the nullspace of W to remove the protected information.

LDAM: the original LDAM model (Cao et al., 2019).

LDAM$_{\text{cw}}$: a variant of LDAM with instance re-weighting by inverse class proportion (Cao et al., 2019).

3.1 Model Comparison

We include simulated experimental settings with the emoji dataset following Ravfogel et al. (2020) where they keep the class proportions balanced, but vary group proportions (stereotyping). In our work, we systematically vary both class imbalance and stereotyping, in order to assess the robustness of the models wrt class imbalance and fairness individually. We explore three settings: varying both dimensions at the same time (Figure 2), controlling for class imbalance and vary stereotyping (Table 1), and controlling for stereotyping while varying class imbalance (Table 2).

We simultaneously vary stereotyping and class imbalance in the emoji dataset, exploring several settings:

- Original: the dataset is sampled based on the natural class distribution (70% positive; Blodgett et al. (2016)), and within each class the black:white ratio is set to 18:82, based on US census estimates.
- 90/90: the class distribution is skewed (90% positive), and black:white ratio is set to 90:10 for positive tweets and 10:90 for negative tweets (i.e. "stereotyping" the classes).
- 95/95: as per the above, but with class skew and stereotyping ratios set to 95:5.

For the occupation classification task, the original data is used as is (Figure 3).

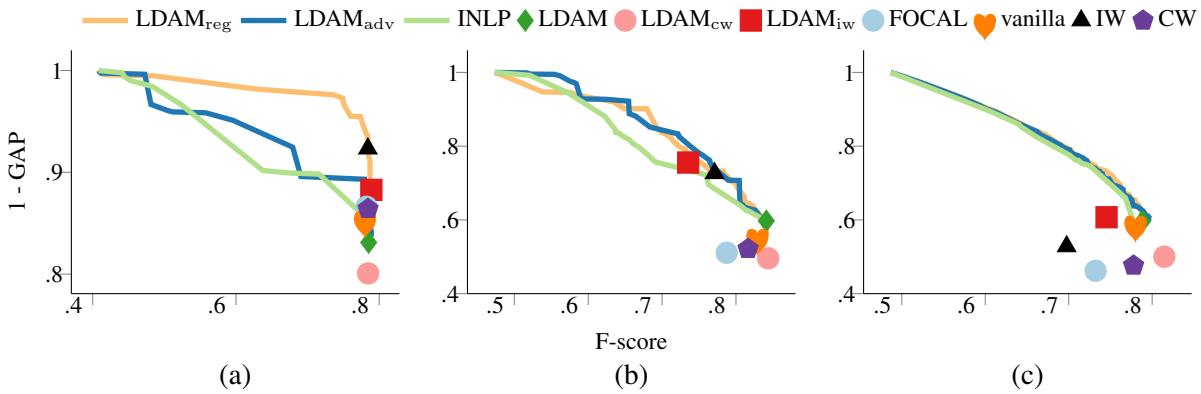

Figure 2: F-score vs. fairness (1− GAP) across the (a) original, (b) 90/90 and (c) 95/95 setting on the emoji prediction task. Models which balance fairness vs. performance through hyperparameters are shown as Pareto frontiers, while the others are reported as single points.

	F-score						1− GAP					
Ratio	vanilla	INLP	LDAM$_{iw}$	LDAM$_{reg}$	LDAM$_{adv}$	ADV	vanilla	INLP	LDAM$_{iw}$	LDAM$_{reg}$	LDAM$_{adv}$	ADV
0.5	0.76	0.75	0.76	0.75	0.75	0.76	0.87	0.88	0.89	**0.92**	0.91	0.83
0.6	0.76	0.71	0.76	0.74	0.74	0.75	0.79	0.82	0.80	**0.92**	0.91	0.80
0.7	0.74	0.65	0.75	0.74	0.73	0.75	0.70	0.84	0.73	**0.93**	0.89	0.78
0.8	0.72	0.62	0.74	0.73	0.73	0.74	0.61	0.84	0.67	**0.93**	0.72	0.76

Table 1: Performance and Fairness on the Emoji data set with fixed balanced class-distribution, but varying the stereotyping ratio (*black*:*white*) per class. The ratio column denotes the % of *black* instances relative to *white*. The test-set is stereotype balanced (50:50).

	F-score						1− GAP					
Ratio	vanilla	INLP	LDAM$_{cw}$	LDAM$_{iw}$	LDAM$_{reg}$	LDAM$_{adv}$	vanilla	INLP	LDAM$_{cw}$	LDAM$_{iw}$	LDAM$_{reg}$	LDAM$_{adv}$
0.7	0.83	0.80	0.83	0.82	0.80	0.80	0.62	0.73	0.60	0.75	**0.84**	0.83
0.8	0.80	0.77	0.83	0.80	0.77	0.78	0.64	0.80	0.61	0.76	0.84	**0.85**
0.9	0.74	0.72	0.79	0.75	0.72	0.74	0.70	0.79	0.62	0.84	**0.85**	0.82

Table 2: Performance and Fairness on the Emoji data set, fixing the stereotyping ratio to 0.8:0.2 (*black*:*white*) per class, and varying the class-balance ratio (proportion of positive class is shown). The test sets in each row are different, and mimic the class-imbalance and stereotyping of the training data (i.e. results across rows are not comparable).

Model Selection. For models with hyper-parameters which trade off performance and fairness, the optimal balance of F-score and fairness is not clear, so we adopt the concept of Pareto optimality (Godfrey et al., 2007) and present the Pareto frontier in the graphs. In particular, for LDAM$_{reg}$ and LDAM$_{adv}$, we perform a hyper-parameter search over C (10^{-2} to 30), ρ (10^{-4} to 10^{2}), and λ (10^{-4} to 10^{2}). In general, a higher C prioritises F-score over fairness, and a higher ρ and λ prioritise fairness. For INLP, we tune the number of iterations as a hyper-parameter. The remaining models don't have trade-off hyper-parameters, so we report a single-point best model: for LDAM and LDAM$_{cw}$ we tune C by choosing the best-performing model over the dev set. For LDAM$_{iw}$, we set β to 0.9999 following Cui et al. (2019), and tune C to identify the fairest model on the dev set.

The results in Figure 2 (a)–(c) show that LDAM$_{reg}$ is overall superior to the other approaches, especially for higher F-scores. For increasingly extreme levels of class imbalance and stereotyping (as we move to the right in the figure), the advantage of LDAM$_{reg}$ over LDAM$_{adv}$ and INLP decreases substantially. Across all the settings, LDAM$_{cw}$ has the highest bias (is least fair). With higher class imbalance and stereotyping, most class-imbalanced learning methods—FOCAL, CW and LDAM$_{cw}$—exhibit high bias. In the stereotyping settings, LDAM$_{iw}$ reduces bias compared to IW.

Analogous results on the occupation data (origi-

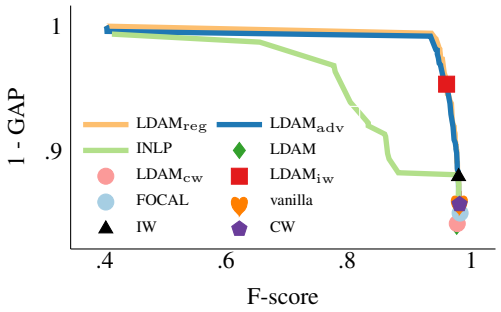

Figure 3: F-score vs. 1− GAP on occupation classification (original class balance and stereotyping) for the same set of models as in Figure 1.

nal class proportions) are in Figure 3. Once again, the proposed LDAM extensions perform the best overall, with LDAM$_{reg}$ achieving the best trade-off in performance. Class-imbalanced learning approaches (FOCAL, LDAM, and LDAM$_{cw}$) are most biased on this dataset, with IW improving fairness over vanilla cross-entropy training, and LDAM$_{iw}$ providing large improvements in fairness.

3.2 Stereotyping-Class balance Trade-off

In addition to comparing models based on the trade-off between performance and fairness in class-imbalanced learning, we wish to disentangle the effect of stereotyping from class imbalance. We do so by: (a) fixing class balance to 50:50 and varying stereotyping (Table 1); and (b) fixing stereotyping to a symmetric 0.8 while varying class imbalance (Table 2). A stereotyping level of symmetric 0.8 means 80:20 black:white for positive and 20:80 black:white for negative tweets. We perform model selection by choosing the INLP model with best harmonic mean of performance and fairness for Table 2, and use the results from the original paper for Table 1. We select all other models by first selecting from models with F-score at least as high as INLP, and then selecting the one with the lowest GAP. We include a recent adversarial model in the varying stereotyping experiments, which performed strongly on the class-balanced emoji data (ADV: Han et al. (2021)).

Our results on varying stereotyping levels in Table 1 show that the vanilla baseline drops in performance more sharply than most proposed models, and results in the most unfair predictions by a large margin. LDAM$_{iw}$, LDAM$_{adv}$, and ADV retain high F-scores but drop in fairness with increasing stereotyping, while INLP exhibits the opposite pattern. LDAM$_{reg}$ achieves the best balance of F-score and fairness. Table 2 presents results for fixed stereotyping and varying class imbalance (0.7–0.9 positive). We include LDAM$_{cw}$ for handling class imbalance but exclude ADV, which does not address class-imbalance directly. We observe that LDAM$_{cw}$ has the highest F-score, but scores poorly for fairness. LDAM$_{iw}$ achieves the best trade-off with high class-imbalance, but shows large variation across settings. LDAM$_{reg}$ appears more stable, exhibiting a good performance–fairness trade-off.

4 Conclusion and Future Work

We explored the interplay of class-imbalance and stereotyping in two language classification data sets. We showed that vanilla class-imbalanced learning (IW, CW, FOCAL, LDAM and LDAM$_{cw}$) can exacerbate unfairness. We extended class-imbalanced learning approaches to handle fairness under stereotyping, and showed that our models provide consistent gains in fairness without sacrificing accuracy. Both LDAM$_{reg}$ which uses maximum mean discrepancy regularizer (Tzeng et al., 2014) and LDAM$_{adv}$ with adversarial loss (Ganin and Lempitsky, 2015) are different ways to make the text representation independent of demographic attributes. Consistent with previous work (Louizos et al., 2016) we find that LDAM$_{reg}$ is robust and performs best across several test scenarios, except in extremely skewed (or stereotyped) settings where the gains of LDAM$_{reg}$ over its adversarial counterpart (LDAM$_{adv}$) diminishes. In addition, LDAM$_{adv}$ introduces more parameters into the model, and is in general hard to train, hence LDAM$_{reg}$ is more preferable overall. In the future, we plan to extend our methods to more complex tasks and multiple private attributes (Subramanian et al., 2021).

5 Acknowledgement

We thank Xudong Han for the discussions and inputs. This work was funded in part by the Australian Government Research Training Program Scholarship, and the Australian Research Council.

References

Pinkesh Badjatiya, Manish Gupta, and Vasudeva Varma. 2019. Stereotypical bias removal for hate speech detection task using knowledge-based generalizations. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 49–59. ACM.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 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, Texas. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, NY, USA. Association for Computing Machinery.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1664–1674, Hong Kong, China. Association for Computational Linguistics.

Mateusz Buda, Atsuto Maki, and Maciej A. Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106:249–259.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 1565–1576.

Cornelia Caragea, Ana Uban, and Liviu P. Dinu. 2019. The myth of double-blind review revisited: ACL vs. EMNLP. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2317–2327, Hong Kong, China. Association for Computational Linguistics.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. 2019. Class-balanced loss based on effective number of samples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 9268–9277. Computer Vision Foundation / IEEE.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, page 120–128, New York, NY, USA. Association for Computing Machinery.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. 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, Brussels, Belgium. Association for Computational Linguistics.

Yaroslav Ganin and Victor S. Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pages 1180–1189. JMLR.org.

Pratyush Garg, John Villasenor, and Virginia Foggo. 2020. Fairness metrics: A comparative analysis. In *2020 IEEE International Conference on Big Data (Big Data)*, pages 3662–3666. IEEE.

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy. Association for Computational Linguistics.

Parke Godfrey, Ryan Shipley, and Jarek Gryz. 2007. Algorithms and analyses for maximal vector computation. *The VLDB Journal*, 16(1):5–28.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Xudong Han, Timothy Baldwin, and Trevor Cohn. 2021. 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. Association for Computational Linguistics.

Mika Juuti, Tommi Gröndahl, Adrian Flanagan, and N. Asokan. 2020. A little goes a long way: Improving toxic language classification despite data scarcity. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2991–3009, Online. Association for Computational Linguistics.

F. Kamiran and T. Calders. 2011. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33:1–33.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2020. Dice loss for data-imbalanced NLP tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 465–476, Online. Association for Computational Linguistics.

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, Melbourne, Australia. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007. IEEE Computer Society.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. 2016. The variational fair autoencoder. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5740–5745, Hong Kong, China. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Shivashankar Subramanian, Xudong Han, Timothy Baldwin, Trevor Cohn, and Lea Frermann. 2021. Evaluating debiasing techniques for intersectional biases. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

Jiachen Tian, Shizhan Chen, Xiaowang Zhang, and Zhiyong Feng. 2020. A graph-based measurement for text imbalance classification. In *24th European Conference on Artificial Intelligence*.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. 2014. Deep domain confusion: Maximizing for domain invariance. *arXiv e-prints*, pages arXiv–1412.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Michael L. Wick, Swetasudha Panda, and Jean-Baptiste Tristan. 2019. Unlocking fairness: a trade-off revisited. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8780–8789.

Shen Yan, Hsien-Te Kao, and Emilio Ferrara. 2020. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 1715–1724. ACM.