‘Choose your Data Wisely’: Active Learning based Selection with Multi-Objective Optimisation for Mitigating Stereotypes

Manish Chandra
m.chandra.1@research.gla.ac.uk
University of Glasgow
Glasgow, United Kingdom

Debasis Ganguly
Debasis.Ganguly@glasgow.ac.uk
University of Glasgow
Glasgow, United Kingdom

Tulika Saha
sahatulika15@gmail.com
University of Liverpool
Liverpool, United Kingdom

Iadh Ounis
iadh.ounis@glasgow.ac.uk
University of Glasgow
Glasgow, United Kingdom

ABSTRACT

Data-driven (deep) learning methods has led to parameterised abstractions of the data, often leading to stereotype societal biases in their predictions, e.g., predicting more frequently that women are weaker than men, or that African Americans are more likely to commit crimes than Caucasians. Standard approaches of mitigating such stereotypical biases from deep neural models include modifying the training dataset (pre-processing), or adjusting the model parameters with a bias-specific objective (in-processing). In our work, we approach this bias mitigation from a different perspective - that of an active learning-based selection of a subset of data instances towards training a model optimised for both effectiveness and fairness. Specifically speaking, the imbalances in the attribute value priors can be alleviated by constructing a balanced subset of the data instances with two selection objectives - first, of improving the model confidence of the primary task itself (a standard practice in active learning), and the second, of taking into account the parity of the model predictions with respect to the sensitive attributes, such as gender and race etc. We demonstrate that our proposed selection function achieves better results in terms of both the primary task effectiveness and fairness. The results are further shown to improve when this active learning-based data selection is combined with an in-process method of multi-objective training.

CCS CONCEPTS

• **Computing methodologies** → *Supervised learning*.

KEYWORDS

Fairness, Active Learning, Multi-Objective Learning

ACM Reference Format:
Manish Chandra, Debasis Ganguly, Tulika Saha, and Iadh Ounis. 2023. ‘Choose your Data Wisely’: Active Learning based Selection with Multi-Objective Optimisation for Mitigating Stereotypes. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM ’23), October 21–25, 2023, Birmingham, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3583780.3615261

1 INTRODUCTION

The ubiquitous use of data-driven (deep) learning has led to risks of these models learning their own characteristic abstractions of the data, often eventually manifested as stereotype societal biases in their predictions [9, 17, 18, 27, 30], examples of which may include predicting more frequently that women are weaker as compared to men [21], or that African Americans are more likely to commit crimes than Caucasians [2]. The main reason why such biases manifest themselves into deep neural models is due to the data-only training mode (not involving any manually engineered features), which means that there is a smaller degree of control on how the task-specific data abstractions are constructed via the model parameters (e.g., for a feature-based model, it is possible to leave out a certain set of features to a model, e.g., the race of a person for recidivism prediction [2, 37]).

As a concrete example of biases in data-driven learning, consider the classification problem that we specifically address in this paper - that of identifying the age-group (young or not) of a person from their facial image; it is observed that a standard ResNet-based deep neural classifier [19] leads to predicting more frequently that men with glasses are not young, and also that women are younger than men. The reason this bias occurs is likely due to the inherent mapping learnt between the high-level characteristic elements of the input images (e.g., presence of glasses) with the ground-truth labels of the data.

As other examples of biases in AI models that are likely to lead to a communal hatred, consider the following - i) the well known Correlational Offender Management Profiling for Alternative Sanctions (COMPAS), an automated decision making system used by the US criminal justice system for assessing a criminal defendant’s likelihood of re-offending [2], has been shown to be biased against the African American ethnicity [2], ii) Google’s targeted advertising system, which advertised highly paid jobs more frequently to men than to women [23].

CIKM ’23, October 21–25, 2023, Birmingham, United Kingdom

ACM ISBN 979-8-4007-0124-5/23/10…$15.00
https://doi.org/10.1145/3583780.3615261

Existing bias mitigation methods can be classified into three general categories as follows: (i) **Pre-processing methods**, which modify the training dataset before feeding it to the DL model; these approaches modify the prior distributions of the attribute (metadata) variables (e.g., age, gender etc.), or more generally, perform specific transformations on the data with an aim to reduce disparity of the class priors in the training data [5, 6, 16, 20, 25]; (ii) **In-process methods**, which employ a multi-task learning with a bias-specific objective added to the primary task objective [3, 8, 11, 29, 34] or an adversarial loss component [1, 32, 40]; (iii) **Post-processing methods**, which modify the final decisions of the classifiers by transforming model outputs to improve prediction fairness [10] (also see [7] for an overview).

In our work, we approach this problem of mitigating biases from the perspective of selecting a subset of data instances in an optimal manner using active learning. We argue that a dataset which is imbalanced in terms of both the class and attribute priors is likely to lead to biased predictions, and this can be alleviated by constructing a balanced subset of the data instances with two selection objectives in mind - first, of improving the model confidence of the primary task itself (a standard practice in active learning [35]), and the second, of taking into account the parity of the model predictions with respect to the sensitive attributes, such as gender and race etc. Similar to our work, multi-objective learning with AL bootstrapping were applied to improve model interpretability [28], and also for privacy preservation [4].

Our Contributions: We propose an active learning (AL) [35] based strategy to select a subset of the training set on which if the model is trained, the predictions would be less biased than when it is trained on the entire training set. Moreover, such training on a more balanced subset of the data should not lead to a substantial loss of the effectiveness of the primary task. The work in [14] seeks to mitigate bias and improve diversity of features in subset selection, where the authors argue that it is not always the case that fairness and diversity objectives are in agreement. In our work, however, we propose a selection function for AL that caters to both effectiveness and fairness. To the best of our knowledge, this is the first attempt at incorporating the fairness criterion directly into the selection function of AL for the downstream objective of bias mitigation. Subsequently, we combine the AL based data selection with an in-process method of multi-objective training and demonstrate that our proposed strategy improves model effectiveness and fairness.

2 ACTIVE LEARNING FOR BIAS MITIGATION

Notations. Consider a dataset $\mathcal{D} = (\mathcal{X}, \mathcal{S}, \mathcal{Y})$, where $\mathcal{X}$ is a set of data instances (e.g., images of faces), and each $y \in \mathcal{Y}$ is one of the c target labels (e.g., $c = 2$ for predicting if the face in an image is of a person's under 40). Additionally, $\mathcal{S} = \{A_1, \ldots, A_k\}$ is a set comprised of k sensitive attributes, where the i^{th} attribute can be associated to a_i possible values, e.g., if the set of attributes correspond to 'race' and 'gender', then $k = 2$ and $a_2 = 2$, representing the labels for male and female (see Table 1 for a glossary of notations).

For fair predictions, ideally, a parameterised predictive model corresponding to the primary task should be independent of $\mathcal{S}$ [7], i.e., $\theta : \mathcal{X} \mapsto \mathcal{Y}$. However, it is often the case that the attributes manifest themselves as inherent characteristic patterns within the data, thus potentially leading to biased parameterised models where $\theta : \mathcal{X}, \mathcal{S} \mapsto \mathcal{Y}$.

Notation	Description
$\mathbf{x} \in \mathcal{X}$	A particular data instance within a set of such instances ($\mathcal{X}$).
$\mathcal{S} = \{A_1, \ldots, A_k\}$	Set of k sensitive attributes, e.g., A_1=RACE and A_2=GENDER.
$V_i = \mathbb{Z}_{a_i} \forall i = 1, \ldots, k$	The i^{th} constitutes one of a_i possible categorical values, e.g., $a_2 = 2$ because the possible values for A_2 (GENDER) are MALE (0) and FEMALE (1).
$\mathcal{P}(\mathcal{S}) = \mathbb{Z}_{a_1} \times \ldots \times \mathbb{Z}_{a_k}$	All possible attribute value combinations.
$\mathbf{s} \in \mathcal{P}(\mathcal{S})$	A k-dimensional vector denoting a particular attribute-value combination, e.g., $\mathbf{s}_1 = \{\text{ASIAN}, \text{FEMALE}\}$ and $\mathbf{s}_2 = \{\text{AFRICAN}, \text{MALE}\}$.

Table 1: Glossary of notations.

AL-based dataset selection. The key idea behind our proposed AL selection function is to leverage this notion of fairness for selecting an optimal subset of data. Instead of the standard use-case of AL, which involves selecting data instances for further annotations towards extending a labeled dataset, in the context of our work, we use it for an optimal subset selection in an incremental manner. More precisely, we first start with a randomly selected seed dataset $\mathcal{D}_0 \subset \mathcal{D}$, and then to select the most promising batch of instances during the i^{th} iteration, we consider a number of different subsets $\mathcal{B}_i \subset \mathcal{D} - \mathcal{D}_i$ each of size b. We then compute a fairness penalty score for each batch, and then add the one with the minimum value to $\mathcal{D}_i$ to make a larger set $\mathcal{D}_{i+1}$ and so on. We repeat the AL steps M number of times ($M = 10$ in all our experiments).

Fairness-based objective. Existing AL approaches seek to minimise the uncertainty of predictions so that when trained with additional data, a model may generalise well and lead to more effective results [24, 36]. In contrast, the active learning selection criteria is different in the context of our problem. Rather than catering to the sole objective of primary task effectiveness, our proposed selection function also estimates how fair the model predictions would be with the inclusion of the additional training data.

We employ the commonly used idea for quantifying fairness in terms of the similarity in predictions across a pair of instances from different groups of attribute values (commonly called **cross pairs**) [15, 39]. Such a similarity in the predictions across a pair indicates that the output for each instance of the pair is independent of their attribute values, e.g. predicting the age of a person is independent of the gender. More formally, given a pair of data instances with identical labels but different attribute values, i.e., $(\mathbf{x}, \mathbf{s}, y), (\mathbf{x}', \mathbf{s}', y') \in \mathcal{D}$ with $y = y'$ but $\mathbf{s} \neq \mathbf{s}'$, a fair model θ should output similar class posteriors, i.e., $||P(\hat{y}|\mathbf{x}; \theta) - P(\hat{y'}|\mathbf{x}'; \theta)|| \to 0$, where $P(\hat{y}|\mathbf{x}; \theta) \in \mathbb{R}^c$ is a c-dimensional softmax distribution.

To select the optimal batch, the idea is to form cross-pairs across the selected data and the unselected ones. The hypothesis is that a batch $\mathcal{D}_i$ where the predictions with the currently trained model (i^{th} iteration), $\theta_i : \mathcal{X}_i \mapsto \mathcal{Y}_i$, yields more fair results should be a better candidate for including in $\mathcal{D}_{i+1}$. In particular, we measure the fairness of such a candidate batch in terms of average agreement with the class posteriors measured over the cross pairs. Formally, the fairness score of $\mathcal{B} \subset \mathcal{D} - \mathcal{D}_i$ (the set of yet unselected instances)

Table 2: Summary of the datasets.

Dataset	Primary task	Attributes for bias
CelebA	P(Young|Face Image)	Gender (M/F), Eye-Glasses (G,$\bar{\text{G}}$)
COMPAS	P(Recidivism|Features)	Race - African (A), Caucasian (C)
EEC	P(Emotion|Sentence)	Gender (M/F)

during the i^{th} step of the AL-based selection is computed as

$$\phi_F(\mathcal{B}) = \frac{1}{|\mathcal{D}_i||\mathcal{B}|}\sum_{(\mathbf{x},\mathbf{s},y)\in\mathcal{D}_i}\;\sum_{(\mathbf{x}',\mathbf{s}',y)\in\mathcal{B}}(P(y|\mathbf{x};\theta_i)-P(y|\mathbf{x}';\theta_i))^2. \tag{1}$$

Finally, we combine the fairness penalty of Equation 1 with a standard selection criteria of AL, namely uncertainty sampling [24], which includes those points for which the current version of the classifier (θ_i trained on $\mathcal{D}_i$) yields a small confidence or a high uncertainty. This combination ensures that the subset of instances that are selected incrementally via AL, potentially yields a model that is not only fair (Equation 1) but also effective, as per the standard AL technique of uncertainty sampling [24, 35]. Formally, the combined objective, which we minimise, is

$$\phi(\mathcal{B}) = \lambda\phi_F(\mathcal{B}) + (1-\lambda)\frac{1}{b}\sum_{\mathbf{x}\in\mathcal{B}}\max_{l=1}^{c} P(y=l|\mathbf{x};\theta^i), \tag{2}$$

where $\lambda \in [0, 1]$ is an interpolation parameter.

AL-based selection with multi-objective training. In addition to the AL-based framework, which is a pre-processing approach for bias mitigation, we also explore its combination with standard in-processing approaches that include an adversarial component (corresponding to the bias) in the primary loss function [11, 34]. More concretely speaking, the bias component of the loss aims to 'not learn effectively' the mapping between the attribute values and the labels. Formally,

$$P_\mu(y|(\mathbf{x},\mathbf{s})) = \mu P(y=c|\mathbf{x}) + (1-\mu)P(y\sim\mathbb{Z}_c - c|(\mathbf{x},\mathbf{s})), \tag{3}$$

where the adversarial part of the loss (the second term on the right hand side of Equation 3) uses a label different from the ground-truth one, thus making the prediction independent of the attribute values.

In the combined approach of AL with bias mitigation loss, instead of using a standard cross-entropy loss as the model θ_i at every step of the AL iteration, we instead use the likelihood function of Equation 3 as the realisation of θ_is ($i = 1, \ldots, M$).

3 EVALUATION

Datasets. For our experiments, we employ three datasets from three different modalities - images, feature-based, and text, which we now describe next (Table 2 presents a summary).

CelebA is an image dataset that is widely used for evaluating models for face detection, particularly for recognising facial attributes [31, 38]. It consists of over 200K facial images of 10, 177 celebrities [26]. Each image is annotated with 40 different binary attributes describing the image, including attributes, such as *Black_Hair*, *Pale_Skin*, etc. The images in this dataset cover large pose variations and background clutter.

The **COMPAS** dataset contains data from Broward County, Florida originally compiled by ProPublica [2]. The task is to predict whether a convicted individual would commit a violent crime in the following two years or not. Following the analysis of Propublica, we considered black and white defendants who were assigned COMPAS risk scores within 30 days of their arrest. Furthermore, we restricted ourselves to defendants who "spent at least two years outside a correctional facility without being arrested for a violent crime, or were arrested for a violent crime within this two-year period" [12].

The **Equity Evaluation Corpus (EEC)** dataset, compiled by [22], is an emotion prediction dataset. Given a natural language sentence, the task involves predicting the primary emotion expressed in the sentence from among 5 possible emotion classes, namely 'fear', 'anger', 'joy', and 'sadness' (along with the neutral class). In addition to being associated with an emotion, each sentence in this dataset expresses a gender or race.

Model and hyper-parameter settings. In our experiments involving CelebA, we use pre-trained ResNet-18 [19] and fine-tune it for predicting whether the celebrity in the input image is 'Young' (an annotated feature of the CelebA dataset). For the COMPAS task, we employ a 2-layer feed-forward network, whereas for the EEC task, we employ a 3-layer feed-forward network with sentence embeddings as inputs. As model optimiser, we employ SGD with momentum set to 0.9. We set the number of iterations in AL (M) to 10. We start with a seed set of size 1% of the total training set size (i.e., $|\mathcal{D}_0| = 0.01 \times |\mathcal{D}|$). We ensure that there is an equal representation of the sensitive attribute values in the seed set. For our proposed sampling strategy, we set the number of candidate batches ($\mathcal{B}$) to 30. The optimal values of these hyperparameters were obtained via a grid search.

Baselines. We compare our approach[1] with the following.

(1) **Entire Dataset (ED)**: We train the model on the entire training set. This baseline is indicative of the bias learned by the model due to the bias in the training data itself.
(2) **Random Sampling - Balanced (RS-B)** : We split the entire training set into K possible groups corresponding to the attributes values. We select all instances from the most under-represented group and sample those many instances from all other groups. This down-sampling based baseline tests whether making the training data balanced leads to mitigating biases.
(3) **Random Sampling (RS)**: The first amongst the AL-based baselines is a simple yet a standard baseline used in the AL literature [35]. Instead of employing a selection function, it simply adds a random batch of instances at each step of an AL iteration.
(4) **Uncertainty Sampling (US)**: This is an ablation of our approach (Equation 2) where we set $\lambda = 0$, which means that we incrementally construct the dataset only with the uncertainty sampling criterion [24].
(5) **Fairness Penalty only (FP-O)**: This is another ablation of Equation 2, where we set $\lambda = 1$ to employ only the cross-pair based fairness criteria.
(6) **Multi-Objective on the Entire Dataset (MO-ED)**: This is an in-process only method, where we instead of applying the

[1] Code available at https://github.com/ManishChandra12/AL4fairAI.

Table 3: Results on the CelebA (image) dataset. FNR: the false negative rate of the entire test set, FNR$_{s_1,s_2}$: attribute values for 'gender' and 'whether wearing eyeglasses', respectively - the values being M/F, and G/$\bar{G}$. The best F1, FNED and EWP values are bold-faced, with ↑s indicating higher the better and ↓s indicating the contrary.

Method	F1(↑)	FNR	FNR$_{(M,G)}$	FNR$_{(F,G)}$	FNR$_{(M,\bar{G})}$	FNR$_{(F,\bar{G})}$	FNED(↓)	EWP(↑)
ED	**0.836**	0.052	0.255	0.099	0.093	0.027	0.3158	0.376
RS-B	0.801	0.075	0.405	0.211	0.117	0.044	0.5393	0.292
RS	0.822	0.064	0.309	0.145	0.114	0.033	0.4062	0.345
US	0.834	0.043	0.219	0.033	0.081	0.021	0.2460	0.396
FP-O	0.801	0.041	0.219	0.105	0.059	0.026	0.2766	0.380
MO-ED	0.813	0.067	0.213	0.118	0.090	0.052	0.2351	0.394
AL	0.825	0.028	0.189	0.033	0.049	0.014	0.2015	0.406
MO+AL	0.822	0.044	0.181	0.069	0.062	0.034	**0.1900**	**0.408**

Table 4: Results on the COMPAS dataset. A and C denote the attribute values for race (African American and Caucasian).

Method	F1	FNR	FNR$_{(A)}$	FNR$_{(C)}$	FNED	EWP
ED	0.561	0.864	0.811	0.976	0.1651	0.3354
RS-B	0.553	0.871	0.822	0.976	0.1540	0.3346
RS	0.552	0.879	0.833	0.976	0.1429	0.3358
US	**0.562**	0.856	0.811	0.952	0.1413	0.3397
FP-O	0.542	0.886	0.844	0.976	0.1317	0.3337
MO-ED	0.549	0.894	0.856	0.976	0.1206	0.3380
AL	0.554	0.871	0.833	0.952	0.1190	0.3401
MO+AL	0.552	0.802	0.795	0.905	**0.1100**	**0.3407**

data selection based method, we train a classifier using the loss function of Equation 3 on the entire dataset.

Our proposed methods. In addition to the baselines, we abbreviate our proposed method of AL-based data selection as **AL**, whereas the combined method, where we employ the multi-objective loss (Equation 3) at each AL iteration, is denoted by **MO+AL**.

Evaluation Metrics. We employ the following metrics for our experiments (averaged across 5 different runs).

- **F1**: Macro-averaged F1-score to measure the effectiveness of the primary task.
- **False Negative Equality Difference (FNED) [13]**: Equality of Odds [3, 13, 18] measures how close are the false positive rates (FPRs) and false negative rates (FNRs) across each group of attribute value combinations. Formally speaking,

$$\text{FNED} = \sum_{\mathbf{s} \in \mathcal{P}(\mathcal{S})} |FNR - FNR_{\mathbf{s}}|, \tag{4}$$

  where $\mathcal{P}(\mathcal{S})$ denotes the set of all possible attribute value combinations (see the notations in Table 1).
- **Effectiveness with Parity (EWP)**: To report a common metric that combines both F1 (higher the better) and FNED (lower the better), we report EWP as the harmonic mean of F1 and (1 - FNED), which is similar in flavour to combining precision and recall via the harmonic mean. A high value of EWP indicates that a model's predictions are both effective and fair.

Results. Table 3 shows the results of our experiments on the CelebA dataset. It can be observed from the second row (RS-B) that making the training data balanced with respect to the attributes does not work adequately well. Next, comparing rows 7 and 8 with the remaining ones, it can be seen that that our proposed approach (AL and MO+AL) outperforms the baselines in terms of FNED and the EWP metrics, without a substantial decrease in the F1 scores. Similar trends are also observed on other modalities in Tables 4 and 5, the gains on EEC dataset are marginal because it exhibits less disparate treatment compared to other datasets (compare row 1 of each table).

Table 5: Results on the EEC dataset. FNR$_{(M)}$ and FNR$_{(F)}$ denote the FNRs for the male and female groups respectively. FNRs computation considers 'not angry' as the +ve class, and the rest as -ve.

Method	F1	FNR	FNR$_{(F)}$	FNR$_{(M)}$	FNED	EWP
ED	**0.671**	0.135	0.110	0.159	0.0483	0.3935
RS-B	0.667	0.176	0.154	0.197	0.0426	0.3931
RS	0.653	0.160	0.141	0.178	0.0365	0.3892
US	0.661	0.172	0.157	0.186	0.0282	0.3934
FP-O	0.637	0.064	0.050	0.078	0.0275	0.3849
MO-ED	0.657	0.016	0.008	0.024	0.0164	0.3939
AL	0.657	0.138	0.130	0.146	**0.0162**	0.3939
MO+AL	0.658	0.138	0.130	0.146	**0.0162**	**0.3943**

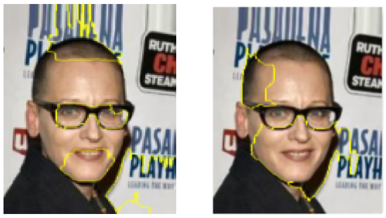
(a) ED **(b) MO+AL**

Figure 1: A sample image from the CelebA dataset with LIME-based [33] explanations for the prediction $\hat{y}$ = $\overline{\text{YOUNG}}$ for a ResNet-18 model trained on *a)* the entire dataset, vs. *b)* trained with our proposed AL+MO approach. Important regions within each image is shown with the yellow colored curves. While the model on the left focuses on the presence of eyeglasses for age estimation (thus exhibiting a stereotype), the one on the right actually focuses on the jaw and the hair line thus making this estimation independent of the presence of eyeglasses.

Figure 1 shows the qualitative model explanations as outputted by LIME for a sample image from the test set. The regions marked with yellow colored curves indicate the important regions within the image that contribute positively towards the model prediction. It can be observed that the model trained on entire dataset attends to the pixels on and around the eyeglasses to make its predictions. However, the model trained using our proposed combination of in-process and pre-process strategies rather attends to other facial attributes to make its predictions.

Concluding Remarks. In this study, we propose an AL based strategy to select a subset of the training set to learn classification models such that their predictions are less biased as opposed to learning a model on the entire training set. We also demonstrate that when the AL-based approach is combined with a standard in-process approach, model training yields better results in terms of both fairness and effectiveness. In future, we would like to adapt our approach to other domains and tasks, such as mitigating biases from neural ranking.

REFERENCES

[1] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. 2019. Turning a Blind Eye: Explicit Removal of Biases and Variation from Deep Neural Network Embeddings. In *ECCV 2018 Workshops*. 556–572.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias There's software used across the country to predict future criminals. And it's biased against blacks. (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. 2017. Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations. *ArXiv* abs/1707.00075 (2017).

[4] Chandan Biswas, Debasis Ganguly, Partha Sarathi Mukherjee, Ujjwal Bhattacharya, and Yufang Hou. 2022. Privacy-Aware Supervised Classification: An Informative Subspace Based Multi-Objective Approach. *Pattern Recogn.* 122, C (feb 2022), 8 pages. https://doi.org/10.1016/j.patcog.2021.108301

[5] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to Computer Programmer As Woman is to Homemaker? Debiasing Word Embeddings. In *Proc. of NIPS 2016*. 4356–4364.

[6] Flavio du Pin Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. 2018. Data Pre-Processing for Discrimination Prevention: Information-Theoretic Optimization and Analysis. *IEEE Journal of Selected Topics in Signal Processing* 12, 5 (2018), 1106–1119. https://doi.org/10.1109/JSTSP.2018.2865887

[7] Simon Caton and Christian Haas. 2020. Fairness in Machine Learning: A Survey. *ArXiv* abs/2010.04053 (2020).

[8] L. Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K. Vishnoi. 2019. Classification with Fairness Constraints: A Meta-Algorithm with Provable Guarantees. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 319–328. https://doi.org/10.1145/3287560.3287586

[9] Lu Cheng, Ahmadreza Mosallanezhad, Yasin Silva, Deborah Hall, and Huan Liu. 2021. Mitigating Bias in Session-based Cyberbullying Detection: A Non-Compromising Approach. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. ACL, 2158–2168.

[10] Silvia Chiappa. 2019. Path-Specific Counterfactual Fairness. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence* (Honolulu, Hawaii, USA) *(AAAI'19/IAAI'19/EAAI'19)*. AAAI Press, Article 957, 8 pages. https://doi.org/10.1609/aaai.v33i01.33017801

[11] Maximin Coavoux, Shashi Narayan, and Shay B. Cohen. 2018. Privacy-preserving Neural Representations of Text. In *EMNLP*. Association for Computational Linguistics, 1–10.

[12] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Halifax, NS, Canada) *(KDD '17)*. Association for Computing Machinery, New York, NY, USA, 797–806. https://doi.org/10.1145/3097983.3098095

[13] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (New Orleans, LA, USA) *(AIES '18)*. Association for Computing Machinery, New York, NY, USA, 67–73. https://doi.org/10.1145/3278721.3278729

[14] Marina Drosou, H. V. Jagadish, Evaggelia Pitoura, and Julia Stoyanovich. 2017. Diversity in Big Data: A Review. *Big data* 5 2 (2017), 73–84.

[15] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Cambridge, Massachusetts) *(ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[16] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In *Proc. of NAACL 2019*. 609–614.

[17] Sara Hajian and Josep Domingo-Ferrer. 2013. A Methodology for Direct and Indirect Discrimination Prevention in Data Mining. *IEEE Transactions on Knowledge and Data Engineering* 25 (2013), 1445–1459.

[18] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (Eds.), Vol. 29. Curran Associates, Inc.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778. https://doi.org/10.1109/CVPR.2016.90

[20] Faisal Kamiran and Toon Calders. 2011. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems* 33 (2011), 1 – 33.

[21] Svetlana Kiritchenko and Saif Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proc. of Joint Conference on Lexical and Computational Semantics*. 43–53.

[22] Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. *CoRR* abs/1805.04508 (2018). arXiv:1805.04508 http://arxiv.org/abs/1805.04508

[23] Anja Lambrecht and Catherine Tucker. 2019. Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads. *Management Science* 65, 7 (2019), 2966–2981.

[24] David D. Lewis and Jason Catlett. 1994. Heterogeneous Uncertainty Sampling for Supervised Learning. In *Machine Learning Proceedings 1994*, William W. Cohen and Haym Hirsh (Eds.). Morgan Kaufmann, San Francisco (CA), 148–156. https://doi.org/10.1016/B978-1-55860-335-6.50026-X

[25] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards Debiasing Sentence Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. ACL, 5502–5515.

[26] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[27] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. 2011. K-NN as an Implementation of Situation Testing for Discrimination Discovery and Prevention. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '11)*. ACM, 502–510.

[28] Ishani Mondal and Debasis Ganguly. 2020. ALEX: Active Learning Based Enhancement of a Classification Model's EXplainability. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (Virtual Event, Ireland) *(CIKM '20)*. Association for Computing Machinery, New York, NY, USA, 3309–3312. https://doi.org/10.1145/3340531.3417456

[29] Ishani Mondal, Procheta Sen, and Debasis Ganguly. 2021. Multi-Objective Few-Shot Learning for Fair Classification. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (Virtual Event, Queensland, Australia) *(CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 3338–3342. https://doi.org/10.1145/3459637.3482146

[30] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-Aware Data Mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08)*. ACM, New York, NY, USA, 560–568.

[31] Amirarsalan Rajabi, Mehdi Yazdani-Jahromi, Ozlem Ozmen Garibay, and Gita Sukthankar. 2022. Through a fair looking-glass: mitigating bias in image datasets. https://doi.org/10.48550/ARXIV.2209.08648

[32] Navid Rekabsaz, Simone Kopeinik, and Markus Schedl. 2021. Societal Biases in Retrieved Contents: Measurement Framework and Adversarial Mitigation of BERT Rankers. In *SIGIR*. ACM, 306–316.

[33] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should i trust you?: Explaining the predictions of any classifier. In *Proc. of KDD'16*. 1135–1144.

[34] Procheta Sen and Debasis Ganguly. 2020. Towards Socially Responsible AI: Cognitive Bias-Aware Multi-Objective Learning. In *AAAI Conference on Artificial Intelligence*.

[35] Burr Settles. 2009. *Active Learning Literature Survey*. Computer Sciences Technical Report 1648. University of Wisconsin–Madison. http://axon.cs.byu.edu/~martinez/classes/778/Papers/settles.activelearning.pdf

[36] Burr Settles and Mark Craven. 2008. An Analysis of Active Learning Strategies for Sequence Labeling Tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Honolulu, Hawaii, 1070–1079. https://aclanthology.org/D08-1112

[37] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery. *Commun. ACM* 56, 5 (may 2013), 44–54.

[38] Zeyu Wang, Klint Qinami, Yannis Karakozis, Kyle Genova, Prem Qu Nair, Kenji Hata, and Olga Russakovsky. 2019. Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 8916–8925.

[39] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 28)*, Sanjoy Dasgupta and David McAllester (Eds.). PMLR, Atlanta, Georgia, USA, 325–333. https://proceedings.mlr.press/v28/zemel13.html

[40] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. ACM, 335–340. https://doi.org/10.1145/3278721.3278779