
Publisher's URL: https://dl.acm.org/doi/10.1145/3583780.3615261

There may be differences between this version and the published version. You are advised to consult the publisher's version if you wish to cite from it.

Copyright © 2023 The Authors. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. 32nd ACM International Conference on Information and Knowledge Management (CIKM 2023), Birmingham, UK, 21-25 Oct 2023, pp. 3768-3772. ISBN 9798400701245

http://eprints.gla.ac.uk/305036/

Deposited on: 06 September 2023
‘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 have 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

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].
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 learning 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 \( \theta : \mathcal{X}, \mathcal{S} \mapsto \mathcal{Y} \).

**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., \((x, s, y), (x', s', y') \in \mathcal{D} \) with \( y = y' \) but \( s \neq s' \), a fair model \( \theta \) should output similar class posteriors, i.e., \( |P(\hat{y}|x; \theta) - P(\hat{y'}|x'; \theta)| \rightarrow 0 \), where \( P(\hat{y}|x; \theta) \in \mathbb{R}^c \) 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 \( \theta_i : \mathcal{X}_i \mapsto \mathcal{Y}_i \) 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>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
</table>
| \( x \in \mathcal{X} \) | 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., \( A_1 = \text{RACE} \) and \( A_2 = \text{GENDER} \).
| \( V_i = Z_{a_i} \forall i = 1, \ldots, k \) | The \( i \)-th constitutes one of \( a_i \) possible categorical values, e.g., \( a_2 = 2 \) because the possible values for \( A_2 = \text{GENDER} \) are MALE (0) and FEMALE (1).
| \( \mathcal{P}({\mathcal{S}}) = Z_{a_1} \times \ldots \times Z_{a_k} \) | All possible attribute value combinations.
| \( s \in \mathcal{P}({\mathcal{S}}) \) | A \( k \)-dimensional vector denoting a particular attribute-value combination, e.g., \( s_1 = \{\text{ASIAN}, \text{FEMALE}\} \) and \( s_2 = \{\text{AFRICAN}, \text{MALE}\} \).

Table 1: Glossary of notations.
during the $i^{th}$ step of the AL-based selection is computed as

$$\phi_F(B) = \frac{1}{|D_i|} \sum_{(x,y) \in D_i} \sum_{(x',y') \in B} (P(y|x; \theta_i) - P(y|x'; \theta_i))^2.$$  

(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 ($\theta_i$ trained on $D_i$) 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

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

(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,

$$P_l(y|(x, s)) = \mu P(y = c|x) + (1 - \mu) P(y \sim Z_c - c|(x, s)),$$  

(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 $\theta_i$ at every step of the AL iteration, we instead use the likelihood function of Equation 3 as the realisation of $\theta_i$s ($i = 1, \ldots, M$).

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., $|D_0| = 0.01 \times |D|$). 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 ($B$) to 30. The optimal values of these hyperparameters were obtained via a grid search.

Baselines. We compare our approach 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

---

Table 2: Summary of the datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Primary task</th>
<th>Attributes for bias</th>
</tr>
</thead>
<tbody>
<tr>
<td>CelebA</td>
<td>P(Young</td>
<td>Face Image)</td>
</tr>
<tr>
<td>COMPAS</td>
<td>P(Recidivism</td>
<td>Features)</td>
</tr>
<tr>
<td>EEC</td>
<td>P(Emotion</td>
<td>Sentence)</td>
</tr>
</tbody>
</table>

1Code available at https://github.com/ManishChandra12/AL4fairAI.
whereas the combined method, where we employ the multi-objective

Table 4: Results on the COMPAS dataset. A and C denote the attribute values for 'gender' and 'whether wearing eyeglasses', respectively. - the values being M/F, and G/G. The best F1, FNED and EWP values are bold-faced, with $\uparrow$ indicating higher the better and ↓ indicating the contrary.

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
<th>FNR</th>
<th>FNAR_{M}</th>
<th>FNAR_{F}</th>
<th>FNAR_{G}</th>
<th>FNAR_{C}</th>
<th>FNED</th>
<th>EWP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ED</td>
<td>0.836</td>
<td>0.052</td>
<td>0.255</td>
<td>0.099</td>
<td>0.039</td>
<td>0.027</td>
<td>0.3158</td>
<td>0.376</td>
</tr>
<tr>
<td>RS-B</td>
<td>0.801</td>
<td>0.075</td>
<td>0.405</td>
<td>0.211</td>
<td>0.117</td>
<td>0.044</td>
<td>0.5393</td>
<td>0.292</td>
</tr>
<tr>
<td>RS</td>
<td>0.822</td>
<td>0.064</td>
<td>0.309</td>
<td>0.145</td>
<td>0.114</td>
<td>0.033</td>
<td>0.4062</td>
<td>0.345</td>
</tr>
<tr>
<td>US</td>
<td>0.834</td>
<td>0.043</td>
<td>0.219</td>
<td>0.033</td>
<td>0.081</td>
<td>0.021</td>
<td>0.2480</td>
<td>0.396</td>
</tr>
<tr>
<td>FP-O</td>
<td>0.801</td>
<td>0.041</td>
<td>0.219</td>
<td>0.105</td>
<td>0.059</td>
<td>0.026</td>
<td>0.2766</td>
<td>0.390</td>
</tr>
<tr>
<td>MO-ED</td>
<td>0.813</td>
<td>0.067</td>
<td>0.213</td>
<td>0.118</td>
<td>0.090</td>
<td>0.052</td>
<td>0.2351</td>
<td>0.394</td>
</tr>
<tr>
<td>AL</td>
<td>0.825</td>
<td>0.028</td>
<td>0.189</td>
<td>0.033</td>
<td>0.049</td>
<td>0.014</td>
<td>0.2013</td>
<td>0.406</td>
</tr>
<tr>
<td>MO+AL</td>
<td>0.822</td>
<td>0.044</td>
<td>0.181</td>
<td>0.069</td>
<td>0.062</td>
<td>0.034</td>
<td>0.1900</td>
<td>0.408</td>
</tr>
</tbody>
</table>

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,

\[
\text{FNED} = \sum_{s \in \mathcal{P}(S)} |FNR_s - FNAR_s|,
\]

where $\mathcal{P}(S)$ 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

Table 5: Results on the EEC dataset. FNRM_{M} and FNRM_{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.

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
<th>FNR</th>
<th>FNRM_{M}</th>
<th>FNRM_{F}</th>
<th>FNED</th>
<th>EWP</th>
</tr>
</thead>
<tbody>
<tr>
<td>ED</td>
<td>0.671</td>
<td>0.135</td>
<td>0.110</td>
<td>0.159</td>
<td>0.0483</td>
<td>0.3935</td>
</tr>
<tr>
<td>RS-B</td>
<td>0.667</td>
<td>0.176</td>
<td>0.154</td>
<td>0.197</td>
<td>0.0426</td>
<td>0.3931</td>
</tr>
<tr>
<td>RS</td>
<td>0.653</td>
<td>0.160</td>
<td>0.141</td>
<td>0.178</td>
<td>0.0365</td>
<td>0.3892</td>
</tr>
<tr>
<td>US</td>
<td>0.661</td>
<td>0.172</td>
<td>0.157</td>
<td>0.186</td>
<td>0.0282</td>
<td>0.3934</td>
</tr>
<tr>
<td>FP-O</td>
<td>0.637</td>
<td>0.064</td>
<td>0.050</td>
<td>0.078</td>
<td>0.0725</td>
<td>0.3849</td>
</tr>
<tr>
<td>MO-ED</td>
<td>0.657</td>
<td>0.016</td>
<td>0.008</td>
<td>0.024</td>
<td>0.0164</td>
<td>0.3939</td>
</tr>
<tr>
<td>AL</td>
<td>0.657</td>
<td>0.138</td>
<td>0.130</td>
<td>0.146</td>
<td>0.0162</td>
<td>0.3939</td>
</tr>
<tr>
<td>MO+AL</td>
<td>0.658</td>
<td>0.138</td>
<td>0.130</td>
<td>0.146</td>
<td>0.0162</td>
<td>0.3943</td>
</tr>
</tbody>
</table>

Figure 1: A sample image from the CelebA dataset with LIME-based [33] explanations for the prediction $\hat{y} = YOUTH$ 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.

(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).

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


Pattern Recognition (CVPR). 770–778. https://doi.org/10.1109/CVPR.2016.90