D3.1 Survey on Bias Mitigation Algorithms

Christos Karanikolopoulos[1], **Panagiotis Papadakos**[1, 2], **Glykeria Toulina**[1], **Spyridon Tzimas**[1], **Panayiotis Tsaparas**[1]

[1] Department of Computer Science and Engineering, University of Ioannina, Greece
[2]Institute of Computer Science (ICS) - Foundation for Research and Technology - Hellas (FORTH), Greece

1. Introduction

The AI revolution has resulted in a world where algorithms make or assist several decisions that affect our lives in different ways. These decisions may be superficial (e.g., what song to listen next), but also significantly consequential (e.g., what career to follow, what treatment to be administered, what sentence to receive). Given the impact of algorithms on our lives there are serious concerns whether the decisions of these algorithms are fair and unbiased. These concerns are substantiated by mounting empirical evidence: several cases where automated systems exhibited bias against specific population subgroups with a detrimental effect on their lives [9].

The need for fairness guarantees on the output of algorithms gave rise to the research area of Responsible AI, and Algorithmic Fairness. There has been significant effort in understanding and measuring algorithmic bias and fairness, but also in mitigating these biases and achieving fairness. In deliverable D.1.1 we presented several definitions of bias and fairness for different Machine Learning tasks and problems. In this report, we outline the different approaches for bias mitigation. Bias mitigation aims to produce algorithms with fairness guarantees under some fairness definitions. We present general methodologies for mitigating bias and achieving fairness, and them we dive into specific case that are of interest to the Themis project, that is, LLMs, Clustering, and Network algorithms.

The report is structured as follows. In Section 2 we present the general approaches to bias mitigation. In Section 3 we present techniques for mitigating bias in LLMs. In Section 4 we present fair clustering algorithms. Section 6 concludes the paper.

2. Overview

The machine learning pipeline can be broken into three main steps: Data collection and preprossessing; Model training; Model deployment. In the first step we collect the data necessary for training our models. The data needs to be sampled, cleaned and labeled. The processed data will be given as input to the second step, which is the training of the model. The resulting models will then be applied to specific tasks.

Bias can be introduced in any of these steps. Training data for Machine Learning algorithms is commonly produced by users (e.g., user-generated text produced by social media users), and as a result it incorporates different societal biases of the user population (for example, gender or racial stereotypes). Furthermore, the data collection is imperfect, and as a result certain groups may be under-represented or misrepresented in the data (for example, in image collections, there may be less photos of people of color than of white people, or more pictures of urban areas than rural areas). These biases in the data will propagate downstream, resulting in biased algorithms (bias in, bias out).

Furthermore, the training of Machine Learning models aims at optimizing specific criteria, which usually capture the success of the model output. Fairness and bias are not directly accounted for in these criteria. Similarly, when deploying the models, we often care more about the accuracy and success of the models, rather than the fairness of the decisions they make. As a result, the training and deployment of the models may also introduce biases.

There are three general approaches to mitigating bias [40], that try to address these different sources of bias. Mitigating bias leads to fairness, so for the following, we will use the terms bias mitigation and fairness interchangeably.

- **Pre-processing approaches:** These approaches aim at eliminating biases in the input data.
- **In-processing approaches:** These approaches aim at changing the internal mechanics of the algorithms to achieve fairness.
- **Post-processing approaches:** These approaches aim at changing the output model or the decisions of the algorithms, to achieve fairness.

We now present some general techniques for each of these approaches.

2.1. Pre-processing approaches

The intuition behind pre-processing approaches is that the root cause of algorithmic unfairness is the bias in the input data. Therefore, if we make the data more "fair", the models trained on the fair data will also be unbiased and fair. These approaches make *interventions* to the data to achieve some notion of fairness. The goal is to make the minimal set of interventions that will achieve the desired level of fairness.

The kind of data interventions depend on the kind of data we have, and the application we are considering. Most work considers classification tasks defined over tabular data. Typically, the approaches considered are the following [40]:

- **Suppression:** A naive approach to make the data unbiased is to remove the protected attribute from the training data. This is called *Fairness through unawareness*. The shortcoming of this approach is that it disregards correlations between the different attributes. For example, race may be highly correlated with zip code. An extension of this approach is to try to identify these correlations and remove the corresponding attributes from the data [40, 48].
- **Relabeling:** This approach relies on flipping the labels of carefully selected instances in the training data so as to achieve fairness. The key question is how to select the set of training samples to relabel. There are different algorithms for performing this selection, based on different criteria [40, 101, 99].
- **Perturbation:** This approach relies on changing the values of non-protected attributes so that the different groups become more similar. There are different algorithms for deciding how to do this perturbation in a way that it minimally affects the data [40, 48, 60, 61].
- **Sampling and re-weighting:** The goal of sampling approaches is to control the contribution of different instances to the learning task. This can also be performed by re-weighting the instances, where different instances have different weights [40, 48, 47, 46, 15].
- **Data augmentation:** Data augmentation approaches add new instances to the data so as to affect the training process. The generation of the new synthetic instances is usually driven by the existing data [40, 42, 90].
- **Fair representations:** Several learning algorithms operate on *representations* of the data, where different data entities (e.g., words) are represented as multidimensional numeric vectors. A line of research considers the problem of creating fair representations that will be given as input to the algorithm training [40, 97]. Creating such fair representations may involve all of the different approaches described here.

Combinations of the above ideas have also been considered [40]. A more general approach was proposed by Calmon et al. [13] who provide a general optimization framework for preprocessing, that changes the data towards fairness while controlling the data distortion and the preserved utility.

2.2. In-processing approaches

In-processing approaches aim to produce an algorithm that is fair, by incorporating fairness in the mechanics of the algorithm. In the case of Machine Learning algorithms, this means affecting the training of the algorithm. Incorporating fairness in the algorithm design depends heavily on the type of algorithm we are considering, and the task at hand. However, there are some general approaches that can be applied in several settings [40].

- **Regularization:** This approach incorporates fairness-related terms to the objective (loss) function used to train the model. There is a hyper-parameter that controls the trade-off between accuracy and fairness. In this way, fairness becomes part of the training objective of the algorithm [46, 24, 52, 23].
- **Fairness constraints:** Another commonly used technique is to impose *fairness constraints* to the algorithm process, e.g., during the algorithm training. Such constrains enforce specific fairness rules and they are commonly used for non-supervised tasks as well [40, 14, 92, 93].
- **Adversarial learning:** Another commonly used approach is to add an adversary to the training that tries to exploit fairness issues (e.g., predict the protected attribute value) [40, 98, 57, 76]. The predictor and the adversary are trained together and compete to improve their performance.

Again, combinations of the above approaches have been considered [40].

2.3. Post-processing approaches

Post-processing approaches aim to intervene at the deployment of the algorithms to achieve fairness. We consider two main techniques for achieving fairness, *black-box approaches* where we only have access to the decisions of the algorithm, and *white-box approaches* that have access to the trained model. These are also referred to as *output correction methods*, and *model correction methods* [40].

- **Output correction methods:** These methods selectively change the output labels of the model. There are different approaches for performing this selection [40, 50, 51, 26, 67, 73]. Such approaches are also popular in tasks beyond classification, such as ranking or recommendations, where we can change the order or composition of the output to achieve fairness [96, 77, 62].
- **Model correction methods:** These methods assume access to the internals of the model, and change parameters or settings to achieve fairness. For example, we may have access to the coefficients of a linear classifier, or the probabilities output by an LLM. We tune these values to achieve fairness. These approaches are also referred to as **intra-processing methods**, and sometimes are treated as a separate category of bias mitigation techniques [40, 38, 49].

3. LLMs

In the LLM survey [29], the authors identify four categories of bias mitigation algorithms, namely the pre-processing, in-training, in-processing, and post-processing categories. However, in order to align with the rest of the deliverable, the following analysis considers approaches that affect the training of the model as in-processing approaches. The rest algorithms that modify the model parameters during inference (i.e., the model is considered as a white-box that provides parameters that can be modified) or assume that the model is a black-box (i.e., the algorithms have access only to the output of the model), are described under the post-processing category.

3.1. Pre-processing bias mitigation

In LLMs, pre-processing bias mitigation approaches mainly apply to the datasets and prompts. The first approach is **data augmentation**. The idea behind this approach is to create more representative datasets by adding underrepresented and counterfactual examples [69, 78], and using selective replacement [94] and interpolation [3]. The **data filtering and reweighting** approach identifies, filters [30, 11] and reweights [37] representative examples that can be used for debiasing, even by using distilation teacher-student models [3]. **Data generation** creates new quality examples based on specific criteria [83] or replace words using pair-lists [74]. **Instruction tuning** like prompt modifications [72], control tokens [22] and continuous prompt tuning [59] is another important direction. Finally, **projection-based mitigation** identifies an embeddings subspace and removes these dimensions of bias from the contextualized embeddings before fine-tuning a model [80]. In this direction, some approaches try to not eliminate the semantic information that is contained in this subspace [64].

As mentioned in [29], pre-processing approaches are based on questionable assumptions and may have limited effectiveness, since the task at hand might not align exactly to the pre-processing measures, and they might be dependent on limited resources.

3.2. In-processing bias mitigation

These approaches reduce the bias of a model by modifying its architecture using adapters [58] or by taking protected attributes as input [36]. A plethora of works consider the **equalization of the objective function** through regularization terms and loss functions, like for embeddings [66, 91], attention [6, 28], predicted token distributions [31, 100], dropout [89], contrastive [18], adversarial [98] and reinforcement learning [75] loss functions. Finally, **selective parameter updating and freezing** [54, 34] or **filtering and removal of neurons** [44] that are associated with biased outputs during the training phase have also been used as mitigation strategies.

A major limitation of the above strategies is the computational cost, since they are applied during the training phase. Furthermore, since these strategies focus on different mechanisms of the training process, they will have varying effectiveness on downstream tasks.

3.3. Post-processing bias mitigation

This debias approach includes all mitigation strategies that consider the model either as a white- or black-box. In the first case, it takes a trained model and modifies the model behavior during inference without retraining or fine-tuning it. For example, in **decoding strategy modifications** it modifies the process of output token generation by enforcing fairness constraints and modifying the probability of next words. Examples include the case of constrained next-token search that changes the rankings of next tokens [32] or enforcing lexical constraints [68]. Other approaches modify the distribution where tokens are sampled leading to more diversity and less biased output [20, 32, 65, 63]. The weights of the model can be altered without training, especially by **redistributing attention weights** [95]. Finally, a **modular architecture** which at inference removes specific types of biases was proposed in [39], while [56] uses an adapter-based modular architecture for bias mitigation.

The LLM as a black-box approaches are limited to mitigating bias in the output of the model. They are mainly based on **rewriting** approaches that replace harmful words with other neutral ones either by using lexicons or neural networks, without changing the contents and the style of the output text. Specificically, some approaches make keyword replacements [84, 21], while others use machine translation techniques and neural rewriters that output more neutralized sentences [43, 88, 70].

The work on the decoding strategy modificiations of the model as a white-box approach, is rather limited. A major challenge is balancing bias mitigation with diverse output generation, without amplifying bias [29]. The black-box approaches that are mainly based on rewriting are also prone to bias due to the need of identifying what has to be rewritten. The removal of protected attributes can also remove important context from the text.

4. Clustering

The methodologies for mitigating bias and achieving fairness in clustering follow the typical separation of pre-processing, in-processing and post-processing approaches. Usually, pre-processing approaches transform the initial dataset under some fairness metric, and then apply a classical clustering algorithm. Post-processing, on the other hand, applies fairness adjustments on the output of a classical clustering algorithm. Finally, in-processing achieves fairness by modifying the vanilla algorithm to obtain a fair algorithm. For the following, the definitions for the fairness metrics we consider can be found in deliverable D1.1.

4.1. Pre-processing bias mitigation

A common metric used for pre-processing bias mitigation is *balance* [19]. The method suggested in [19] for fair clustering is divided into two steps. First, a *fairlet* decomposition is performed, where the data points are partitioned into small balanced subsets (fairlets). This is performed by transforming the problem into a minimum cost flow problem. Then, a classical algorithm is applied to generate clusters. These clusters will be balanced, since the input "poitns" are balanced. In [19] they consider the k-center and k-median algorithms, but any clustering algorithm can be applied. There are several follow-up publications that use fairlets [81] [2] [7], or that follow the logic of generating a fair (balanced) representation for the clustering input [82] [41], [8].

4.2. In-processing bias mitigation

In-processing aims to make the algorithms behave in a fair manner and output fair clusters. In [33] they consider the social fairness metric and they propose a modification of the popular k-means (Lloyd's algorithm) algorithm, called Fair-Lloyd. The algorithm updates the centers to minimize the maximum average cost for each group, thus ensuring fairness.

For individual fairness (as defined in [45]), [45] formulates a fair k-center problem by incorporating the density of points locally around each point. This metric is also considered in [17]. Individual fairness for center-based algorithms is also considered in [5, 12, 16].

4.3. Post-processing bias mitigation

Post processing approaches first apply a vanilla clustering algorithm and then improve fairness by post-processing the results. For the Bounded representation metric, the work in [10] reassigns the points to clusters to achieve the upper and lower bounds on representation fairness. They formulate the problem as an Integer Programming problem, which they solve by relaxing it to a Linear Program (LP) and performing rounding.

For the Fair Representation of Centers metric, the work in [55] computes fair summaries of the output of a k-center algorithm. The proposed algorithm first solves the vanilla k-center problem and then adjusts the centers in a greedy fashion so that all groups have the required number of centers. Similar methodologies are proposed in [1, 71, 25].

5. Network Analysis

In this section, we discuss methods for mitigating bias in network analysis tasks, and specifically, diffusion maximization, PageRank and node embeddings.

5.1. Fair Diffusion Maximization

The methods proposed for mitigating bias in diffusion maximization are predominantly in-processing methods, with reguralization of the objective function and imposing of fairness constraints being the most popular approaches.

Tsang et al. [85] consider achieving their two proposed notions of group fairness, maximin fairness and group rationality, for which they formulate appropriate corresponding objective functions

U_{maximin} and U_{rational} respectively. The authors observe that U_{maximin} and U_{rational} are not submodular, so they do not lend themselves to application of standard techniques. Then, they show that optimizing U_{maximin} or U_{rational} can be reduced to solving a number of instances of the *multiobjective submodular optimization* problem, for which they provide an $(1-\frac{1}{e})$-approximation algorithm. Their algorithm relaxes the problem from a discrete optimization problem to a continuous one and uses *Stochastic Saddle-Point Mirror Descent* to optimize for all objectives simultaneously. However, the authors warn that optimizing for maximin fairness can greatly downgrade the quality of the solution, for instance in the case of the existence of a poorly connected group.

Ali et al. [4] study the problem of *time-critical* diffusion maximization, in which additionally to the budget there is also a *deadline*. They consider bounding the acceptable values of the Maximum Disparity in Normalized Utilities metric by introducing appropriate additional constraints. However, the resulting problem does not lend itself to application of standard techniques, so the authors opt to modify the objective function instead. In order to preserve its submodularity, they compose it with a non-negative, monotone concave function. Ali et al. consider choosing the logarithm and square root functions, but leave the choice to the user for calibrating the amount of penalization of disparity.

Fish et al. [27] consider achieving their proposed notion of individual fairness, maximin fairness, for which they formulate the appropriate corresponding objective function U_{maximin} and they observe that U_{maximin} is not submodular. The authors provide a number of algorithms that incrementally construct the seed set S by selecting seeds according to a heuristic. However, they show that all these algorithms have an approximation ratio no better than exponential.

5.2. Fair PageRank

For mitigating bias in PageRank, Tsioutsiouliklis et al. [86] propose in-processing methods. They consider the approach of modifying the parameters that characterize a PageRank algorithm, which are the transition probability matrix $\mathbf{P}$, restart probability γ and jump probability vector $\mathbf{v}$.

For achieving ϕ-fairness, Tsioutsiouliklis et al. modify only the jump probability vector $\mathbf{v}$. They note that the stationary probability vector $\mathbf{p}$ can be written as a linear function of $\mathbf{v}$ as follows: $p^T = \gamma\mathbf{v}^T(\mathbf{I} - (1-\gamma)\mathbf{P})^{-1}$. From this note, the authors obtain linear constraints on $\mathbf{v}$ and provide necessary and sufficient conditions under which they can be satisfied. This family of ϕ-fairness inducing jump probability vectors provides a family of ϕ-fair PageRank algorithms, which they call *Fairness-Sensitive PageRank* algorithms. They further select the algorithm in this family with minimal *utility loss*, which is the sum of squares difference between the stationary probability vectors of the original and ϕ-fair PageRank algorithms, by solving the corresponding convex optimization problem.

For achieving local ϕ-fairness, Tsioutsiouliklis et al.. [86] modify the transition probability matrix $\mathbf{P}$ and the jump probability vector $\mathbf{v}$. They define a family of locally ϕ-fair PageRank algorithms, which they call *Residual-Based Locally Fair PageRank* algorithms. In these algorithms, each node $u \in V$ distributes probability mass $1 - \delta(u)$ uniformtly to its neighbors and probability mass $\delta(u)$, which they call the *residual*, to the members of the group that is underrepresented in its neighborhood. How the residual is distributed in each algorithm is dictated by a *residual distribution policy*. The authors consider the algorithms defined by three particular policies, which they call *Neighborhood*, *Uniform* and *Proportional*. They also consider the algorithm in this family with minimal utility loss. As before, they obtain a constraint optimization problem, albeit this problem is not convex, so the authors implement a *Stochastic Random Search* algorithm for solving it.

Tsioutsiouliklis et al. [86] also propose a post-processing method for achieving ϕ-fairness. Their algorithm iteratively redistributes uniformly to the members of the protected group an appropriate amount of probability mass that is subtracted uniformly from the members of the non-protected group until $\mathbf{p}$ becomes ϕ-fair.

Tsioutsiouliklis et al. [87] propose a pre-processing method for mitigating bias in PageRank. The authors consider the approach of modifying the network graph via edge additions. They define the *fairness gain* of adding one or multiple edges outgoing from a preselected source node as the increase in the Pagerank value of the protected group. They provide analytical formulas for computing the fairness gain in these cases and they efficient algorithms for determining the best one or k edges outgoing from the source node to add for maximum fairness gain.

5.3. Fair Node Embeddings

Another network analysis task is the generation of node embeddings of the network graph. This is a typical preprocessing step of the network data into a form that is suitable for downstream tasks. A popular algorithm for producing such embeddings is `node2vec` [35], which consists of the following two steps: First, sample a *2^{nd}-order random walk* with transition probability matrix $\mathbf{P}$ to produce a number of traces of the same length. Then, train a *skip-gram* model on the nodes of the network graph with targets constructed from the produced traces. The definition of the matrix $\mathbf{P}$ involves two parameters $p, q \in \mathbb{R}_+^*$. Assuming that u is the previous node in the random walk, parameter p controls the probability that the next node is u and parameter q controls the bias towards the next node being closer to u versus further away from u.

The study on mitigating bias in node embeddings is initiated by Rahman et al. [79] who propose the *FairWalk* algorithm, in which the first step of `node2vec` is modified to perform a 1^{st}-order random walk. For every node, the authors distribute its transition probability uniformly to the groups that appear in its outgoing neighborhood.

This particular in-processing approach to mitigating bias in node embeddings is revisited by Khajehnejad et al. [53] who propose the *CrossWalk* algorithm. For every node, the authors distribute its transition probability as follows: They distribute $1-\alpha$ to its neighbors that are in the same group and they distribute α uniformly to the remaining groups that appear in its outgoing neighbourhood, where $\alpha \in (0,1)$ is a parameter. Moreover, for each group, they bias towards the nodes in closer proximity to the remaining groups; in other words, they bias towards the nodes that are in the boundary of each group. The degree of this form of bias is calibrated by a parameter $p \in \mathbb{R}_+^*$.

6. Conclusion

In this report we surveyed the different approaches for mitigating bias and achieving fairness in algorithms. We identified three general approaches to bias mitigation: Pre-processing, In-processing, and Post-processing, depending on whether the mitigation efforts target the input data, the model training, or the model output. For each of these approaches we presented commonly used techniques that can be applied in different contexts. We then went in-depth in the mitigation efforts for the areas that are of interest to the project: Large Language Models, Clustering and Network Analysis.

References

[1] Sara Ahmadian et al. "Clustering without Over-Representation". In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. KDD '19. Anchorage, AK, USA: Association for Computing Machinery, 2019, pp. 267–275. ISBN: 9781450362016. DOI: 10.1145/3292500.3330987. URL: https://doi.org/10.1145/3292500.3330987.

[2] Sara Ahmadian et al. "Fair Hierarchical Clustering". In: *CoRR* abs/2006.10221 (2020). arXiv: 2006.10221. URL: https://arxiv.org/abs/2006.10221.

[3] Jaimeen Ahn et al. "Why knowledge distillation amplifies gender bias and how to mitigate from the perspective of DistilBERT". In: *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*. 2022, pp. 266–272.

[4] Junaid Ali et al. "On the Fairness of Time-Critical Influence Maximization in Social Networks". In: *IEEE Transactions on Knowledge and Data Engineering* 35.3 (2023), pp. 2875–2886. DOI: 10.1109/TKDE.2021.3120561.

[5] Nihesh Anderson et al. *Distributional Individual Fairness in Clustering*. 2020. arXiv: 2006.12589 [cs.LG]. URL: https://arxiv.org/abs/2006.12589.

[6] Giuseppe Attanasio et al. "Entropy-based attention regularization frees unintended bias mitigation from lists". In: *arXiv preprint arXiv:2203.09192* (2022).

[7] Arturs Backurs et al. "Scalable Fair Clustering". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 405–413. URL: https://proceedings.mlr.press/v97/backurs19a.html.

[8] Sayan Bandyapadhyay, Fedor V. Fomin, and Kirill Simonov. *On Coresets for Fair Clustering in Metric and Euclidean Spaces and Their Applications*. 2020. arXiv: 2007.10137 [cs.DS]. URL: https://arxiv.org/abs/2007.10137.

[9] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press, 2023.

[10] Suman Bera et al. "Fair Algorithms for Clustering". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/fc192b0c0d270dbf41870a63a8c76c2f-Paper.pdf.

[11] Conrad Borchers et al. "Looking for a handsome carpenter! debiasing GPT-3 job advertisements". In: *arXiv preprint arXiv:2205.11374* (2022).

[12] Brian Brubach et al. "A pairwise fair and community-preserving approach to k-center clustering". In: *International conference on machine learning*. PMLR. 2020, pp. 1178–1189.

[13] Flavio Calmon et al. "Optimized pre-processing for discrimination prevention". In: *Advances in neural information processing systems* 30 (2017).

[14] L Elisa Celis et al. "Classification with fairness constraints: A meta-algorithm with provable guarantees". In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 319–328.

[15] Junyi Chai and Xiaoqian Wang. "Fairness with adaptive weights". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 2853–2866.

[16] Darshan Chakrabarti et al. *A New Notion of Individually Fair Clustering: α-Equitable k-Center*. 2022. arXiv: 2106.05423 [cs.LG]. URL: https://arxiv.org/abs/2106.05423.

[17] Deeparnab Chakrabarty and Maryam Negahbani. *Better Algorithms for Individually Fair k-Clustering*. 2021. arXiv: 2106.12150 [cs.DS]. URL: https://arxiv.org/abs/2106.12150.

[18] Pengyu Cheng et al. "Fairfil: Contrastive neural debiasing method for pretrained text encoders". In: *arXiv preprint arXiv:2103.06413* (2021).

[19] Flavio Chierichetti et al. "Fair clustering through fairlets". In: *Advances in neural information processing systems* 30 (2017).

[20] John Joon Young Chung, Ece Kamar, and Saleema Amershi. "Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions". In: *arXiv preprint arXiv:2306.04140* (2023).

[21] Harnoor Dhingra et al. "Queer people are people first: Deconstructing sexual identity stereotypes in large language models". In: *arXiv preprint arXiv:2307.00101* (2023).

[22] Emily Dinan et al. "Queens are powerful too: Mitigating gender bias in dialogue generation". In: *arXiv preprint arXiv:1911.03842* (2019).

[23] Hyungrok Do et al. "Fair generalized linear models with a convex penalty". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 5286–5308.

[24] Cynthia Dwork et al. *Fairness Through Awareness*. 2011. arXiv: 1104.3913 [cs.CC]. URL: https://arxiv.org/abs/1104.3913.

[25] Seyed Esmaeili et al. "Probabilistic Fair Clustering". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 12743–12755. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/95f2b84de5660ddf45c8a34933a2e66f-Paper.pdf.

[26] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. "A confidence-based approach for balancing fairness and accuracy". In: *Proceedings of the 2016 SIAM international conference on data mining*. SIAM. 2016, pp. 144–152.

[27] Benjamin Fish et al. "Gaps in Information Access in Social Networks?" In: *The World Wide Web Conference*. WWW '19. San Francisco, CA, USA: Association for Computing Machinery, 2019, pp. 480–490. ISBN: 9781450366748. DOI: 10.1145/3308558.3313680. URL: https://doi.org/10.1145/3308558.3313680.

[28] Yacine Gaci et al. "Debiasing pretrained text encoders by paying attention to paying attention". In: *2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics. 2022, pp. 9582–9602.

[29] Isabel O Gallegos et al. "Bias and fairness in large language models: A survey". In: *arXiv preprint arXiv:2309.00770* (2023).

[30] Aparna Garimella, Rada Mihalcea, and Akhash Amarnath. "Demographic-aware language model fine-tuning as a bias mitigation technique". In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 2022, pp. 311–319.

[31] Aparna Garimella et al. "He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 2021, pp. 4534–4545.

[32] Samuel Gehman et al. "Realtoxicityprompts: Evaluating neural toxic degeneration in language models". In: *arXiv preprint arXiv:2009.11462* (2020).

[33] Mehrdad Ghadiri, Samira Samadi, and Santosh Vempala. *Socially Fair k-Means Clustering*. 2020. arXiv: 2006.10085 [cs.LG]. URL: https://arxiv.org/abs/2006.10085.

[34] Michael Gira, Ruisu Zhang, and Kangwook Lee. "Debiasing pre-trained language models via efficient fine-tuning". In: *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. 2022, pp. 59–69.

[35] Aditya Grover and Jure Leskovec. "node2vec: Scalable Feature Learning for Networks". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 855–864. ISBN: 9781450342322. DOI: 10.1145/2939672.2939754. URL: https://doi.org/10.1145/2939672.2939754.

[36] Xudong Han, Timothy Baldwin, and Trevor Cohn. "Balancing out bias: Achieving fairness through balanced training". In: *arXiv preprint arXiv:2109.08253* (2021).

[37] Xudong Han, Timothy Baldwin, and Trevor Cohn. "Fair enough: Standardizing evaluation and model selection for fairness research in NLP". In: *arXiv preprint arXiv:2302.05711* (2023).

[38] Moritz Hardt, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning". In: *Advances in neural information processing systems* 29 (2016).

[39] Lukas Hauzenberger et al. "Modular and on-demand bias mitigation with attribute-removal subnetworks". In: *arXiv preprint arXiv:2205.15171* (2022).

[40] Max Hort et al. "Bias mitigation for machine learning classifiers: A comprehensive survey". In: *ACM Journal on Responsible Computing* 1.2 (2024), pp. 1–52.

[41] Lingxiao Huang, Shaofeng Jiang, and Nisheeth Vishnoi. "Coresets for Clustering with Fairness Constraints". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach et al. Vol. 32. Curran Associates, Inc., 2019. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/810dfbbebb17302018ae903e9cb7a483-Paper.pdf.

[42] Vasileios Iosifidis and Eirini Ntoutsi. "Dealing with bias via data augmentation in supervised learning scenarios". In: *Jo Bates Paul D. Clough Robert Jäschke* 24.11 (2018).

[43] Nishtha Jain et al. "Generating gender augmented data for NLP". In: *arXiv preprint arXiv:2107.05987* (2021).

[44] Przemyslaw Joniak and Akiko Aizawa. “Gender biases and where to find them: Exploring gender bias in pre-trained transformer-based language models using movement pruning”. In: *arXiv preprint arXiv:2207.02463* (2022).

[45] Christopher Jung, Sampath Kannan, and Neil Lutz. *A Center in Your Neighborhood: Fairness in Facility Location*. 2019. arXiv: 1908.09041 [cs.DS]. URL: https://arxiv.org/abs/1908.09041.

[46] Faisal Kamiran and Toon Calders. “Classification with no discrimination by preferential sampling”. In: *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*. Vol. 1. 6. Citeseer. 2010.

[47] Faisal Kamiran and Toon Calders. “Classifying without discriminating”. In: *2009 2nd international conference on computer, control and communication*. IEEE. 2009, pp. 1–6.

[48] Faisal Kamiran and Toon Calders. “Data preprocessing techniques for classification without discrimination”. In: *Knowledge and information systems* 33.1 (2012), pp. 1–33.

[49] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. “Discrimination aware decision tree learning”. In: *2010 IEEE international conference on data mining*. IEEE. 2010, pp. 869–874.

[50] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. “Decision theory for discrimination-aware classification”. In: *2012 IEEE 12th international conference on data mining*. IEEE. 2012, pp. 924–929.

[51] Faisal Kamiran et al. “Exploiting reject option in classification for social discrimination control”. In: *Information Sciences* 425 (2018), pp. 18–33.

[52] Toshihiro Kamishima et al. “Fairness-aware classifier with prejudice remover regularizer”. In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*. Springer. 2012, pp. 35–50.

[53] Ahmad Khajehnejad et al. “CrossWalk: Fairness-Enhanced Node Representation Learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 36.11 (June 2022), pp. 11963–11970. DOI: 10.1609/aaai.v36i11.21454. URL: https://ojs.aaai.org/index.php/AAAI/article/view/21454.

[54] James Kirkpatrick et al. “Overcoming catastrophic forgetting in neural networks”. In: *Proceedings of the national academy of sciences* 114.13 (2017), pp. 3521–3526.

[55] Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. “Fair k-Center Clustering for Data Summarization”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, Sept. 2019, pp. 3448–3457. URL: https://proceedings.mlr.press/v97/kleindessner19a.html.

[56] Deepak Kumar et al. “Parameter-efficient modularised bias mitigation via AdapterFusion”. In: *arXiv preprint arXiv:2302.06321* (2023).

[57] Preethi Lahoti et al. “Fairness without demographics through adversarially reweighted learning”. In: *Advances in neural information processing systems* 33 (2020), pp. 728–740.

[58] Anne Lauscher, Tobias Lueken, and Goran Glavaš. “Sustainable modular debiasing of language models”. In: *arXiv preprint arXiv:2109.03646* (2021).

[59] Brian Lester, Rami Al-Rfou, and Noah Constant. “The power of scale for parameter-efficient prompt tuning”. In: *arXiv preprint arXiv:2104.08691* (2021).

[60] Tianyi Li et al. “’Propose and Review’: Interactive Bias Mitigation for Machine Classifiers”. In: *Available at SSRN 4139244* (2022).

[61] Yanhui Li et al. “Training data debugging for the fairness of machine learning software”. In: *Proceedings of the 44th International Conference on Software Engineering*. 2022, pp. 2215–2227.

[62] Yunqi Li et al. “Fairness in recommendation: Foundations, methods, and applications”. In: *ACM Transactions on Intelligent Systems and Technology* 14.5 (2023), pp. 1–48.

[63] Paul Pu Liang et al. “Towards understanding and mitigating social biases in language models”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 6565–6576.

[64] Tomasz Limisiewicz and David Mareček. “Don’t Forget About Pronouns: Removing Gender Bias in Language Models Without Losing Factual Gender Information”. In: *arXiv preprint arXiv:2206.10744* (2022).

[65] Alisa Liu et al. “DExperts: Decoding-time controlled text generation with experts and anti-experts”. In: *arXiv preprint arXiv:2105.03023* (2021).

[66] Haochen Liu et al. “Does gender matter? towards fairness in dialogue systems”. In: *arXiv preprint arXiv:1910.10486* (2019).

[67] Pranay K Lohia et al. “Bias mitigation post-processing for individual and group fairness”. In: *Icassp 2019-2019 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE. 2019, pp. 2847–2851.

[68] Ximing Lu et al. “Neurologic decoding:(un) supervised neural text generation with predicate logic constraints”. In: *arXiv preprint arXiv:2010.12884* (2020).

[69] Ximing Lu et al. “Quark: Controllable text generation with reinforced unlearning”. In: *Advances in neural information processing systems* 35 (2022), pp. 27591–27609.

[70] Xinyao Ma et al. “PowerTransformer: Unsupervised controllable revision for biased language correction”. In: *arXiv preprint arXiv:2010.13816* (2020).

[71] Sepideh Mahabadi and Ali Vakilian. “Individual Fairness for k-Clustering”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 6586–6596. URL: https://proceedings.mlr.press/v119/mahabadi20a.html.

[72] Justus Mattern et al. “Understanding stereotypes in language models: Towards robust measurement and zero-shot debiasing”. In: *arXiv preprint arXiv:2212.10678* (2022).

[73] Aditya Krishna Menon and Robert C Williamson. “The cost of fairness in binary classification”. In: *Conference on Fairness, accountability and transparency*. PMLR. 2018, pp. 107–118.

[74] Ali Omrani et al. “Social-group-agnostic bias mitigation via the stereotype content model”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 4123–4139.

[75] Xiangyu Peng et al. “Reducing non-normative text generation from language models”. In: *arXiv preprint arXiv:2001.08764* (2020).

[76] Andrija Petrović et al. “FAIR: Fair adversarial instance re-weighting”. In: *Neurocomputing* 476 (2022), pp. 14–37.

[77] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. “Fairness in rankings and recommendations: an overview”. In: *The VLDB Journal* (2022), pp. 1–28.

[78] Rebecca Qian et al. “Perturbation augmentation for fairer nlp”. In: *arXiv preprint arXiv:2205.12586* (2022).

[79] Tahleen Rahman et al. “Fairwalk: Towards Fair Graph Embedding”. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, July 2019, pp. 3289–3295. DOI: 10.24963/ijcai.2019/456. URL: https://doi.org/10.24963/ijcai.2019/456.

[80] Shauli Ravfogel et al. “Null it out: Guarding protected attributes by iterative nullspace projection”. In: *arXiv preprint arXiv:2004.07667* (2020).

[81] Clemens Rösner and Melanie Schmidt. *Privacy preserving clustering with constraints*. 2018. arXiv: 1802.02497 [cs.CC]. URL: https://arxiv.org/abs/1802.02497.

[82] Melanie Schmidt, Chris Schwiegelshohn, and Christian Sohler. “Fair Coresets and Streaming Algorithms for Fair k-means”. In: *Approximation and Online Algorithms*. Ed. by Evripidis Bampis and Nicole Megow. Cham: Springer International Publishing, 2020, pp. 232–251. ISBN: 978-3-030-39479-0.

[83] Irene Solaiman and Christy Dennison. "Process for adapting language models to society (palms) with values-targeted datasets". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 5861–5873.

[84] Ewoenam Kwaku Tokpo and Toon Calders. "Text style transfer for bias mitigation using masked language modeling". In: *arXiv preprint arXiv:2201.08643* (2022).

[85] Alan Tsang et al. "Group-Fairness in Influence Maximization". In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, July 2019, pp. 5997–6005. DOI: 10.24963/ijcai.2019/831. URL: https://doi.org/10.24963/ijcai.2019/831.

[86] Sotiris Tsioutsiouliklis et al. "Fairness-Aware PageRank". In: *Proceedings of the Web Conference 2021*. WWW '21. Ljubljana, Slovenia: Association for Computing Machinery, 2021, pp. 3815–3826. ISBN: 9781450383127. DOI: 10.1145/3442381.3450065. URL: https://doi.org/10.1145/3442381.3450065.

[87] Sotiris Tsioutsiouliklis et al. "Link Recommendations for PageRank Fairness". In: *Proceedings of the ACM Web Conference 2022*. WWW '22. Virtual Event, Lyon, France: Association for Computing Machinery, 2022, pp. 3541–3551. ISBN: 9781450390965. DOI: 10.1145/3485447.3512249. URL: https://doi.org/10.1145/3485447.3512249.

[88] Xun Wang et al. "Pay attention to your tone: Introducing a new dataset for polite language rewrite". In: *arXiv preprint arXiv:2212.10190* (2022).

[89] Kellie Webster et al. "Measuring and reducing gendered correlations in pre-trained models". In: *arXiv preprint arXiv:2010.06032* (2020).

[90] Depeng Xu et al. "Fairgan+: Achieving fair data generation and classification through generative adversarial nets". In: *2019 IEEE international conference on big data (Big Data)*. IEEE. 2019, pp. 1401–1406.

[91] Ke Yang et al. "Adept: A debiasing prompt framework". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 9. 2023, pp. 10780–10788.

[92] Muhammad Bilal Zafar et al. "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment". In: *Proceedings of the 26th international conference on world wide web*. 2017, pp. 1171–1180.

[93] Muhammad Bilal Zafar et al. "Fairness constraints: A flexible approach for fair classification". In: *Journal of Machine Learning Research* 20.75 (2019), pp. 1–42.

[94] Abdelrahman Zayed et al. "Deep learning on a healthy data diet: Finding important examples for fairness". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 12. 2023, pp. 14593–14601.

[95] Abdelrahman Zayed et al. "Should we attend more or less? modulating attention for fairness". In: *arXiv preprint arXiv:2305.13088* (2023).

[96] Meike Zehlike, Ke Yang, and Julia Stoyanovich. "Fairness in ranking: A survey". In: *arXiv preprint arXiv:2103.14000* (2021).

[97] Rich Zemel et al. "Learning fair representations". In: *International conference on machine learning*. PMLR. 2013, pp. 325–333.

[98] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. "Mitigating unwanted biases with adversarial learning". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 335–340.

[99] Lu Zhang, Yongkai Wu, and Xintao Wu. "Achieving non-discrimination in prediction". In: *arXiv preprint arXiv:1703.00060* (2017).

[100] Fan Zhou et al. "Causal-debias: Unifying debiasing in pretrained language models and fine-tuning via causal invariant learning". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 4227–4241.

[101] Indre Žliobaite, Faisal Kamiran, and Toon Calders. "Handling conditional discrimination". In: *2011 IEEE 11th international conference on data mining*. IEEE. 2011, pp. 992–1001.