Assessing Demographic Bias Transfer from Dataset to Model: A Case Study in Facial Expression Recognition

Iris Dominguez-Catena*, Daniel Paternain and Mikel Galar

Institute of Smart Cities (ISC), Department of Statistics, Computer Science and Mathematics
Public University of Navarre (UPNA)
Arrosadia Campus, 31006, Pamplona, Spain

Abstract

The increasing amount of applications of Artificial Intelligence (AI) has led researchers to study the social impact of these technologies and evaluate their fairness. Unfortunately, current fairness metrics are hard to apply in multi-class multi-demographic classification problems, such as Facial Expression Recognition (FER). We propose a new set of metrics to approach these problems. Of the three metrics proposed, two focus on the representational and stereotypical bias of the dataset, and the third one on the residual bias of the trained model. These metrics combined can potentially be used to study and compare diverse bias mitigation methods. We demonstrate the usefulness of the metrics by applying them to a FER problem based on the popular Affectnet dataset. Like many other datasets for FER, Affectnet is a large Internet-sourced dataset with 291,651 labeled images. Obtaining images from the Internet raises some concerns over the fairness of any system trained on this data and its ability to generalize properly to diverse populations. We first analyze the dataset and some variants, finding substantial racial bias and gender stereotypes. We then extract several subsets with different demographic properties and train a model on each one, observing the amount of residual bias in the different setups. We also provide a second analysis on a different dataset, FER+.

1. Introduction

When algorithms and automated systems interact with users, they can often cause harm in many unintentional ways. This effect is multiplied when the system is a machine learning system trained to imitate human behavior, which is inherently conditioned by prejudice and cognitive biases. In machine learning, the biases gathered in the training information can leak to the trained models, which are otherwise expected to be fair. When these systems are deployed to the real world they have been shown to exhibit gender, racial and other demographic biases [1, 2, 3].

The current state-of-the-art systems employ very large datasets, with the most renowned example being Imagenet, with over 14 million images. Further analysis on this dataset has shown critical biases and problematic data [2, 4]. The amount of models trained on this dataset

[5], which is already a standard for model pretraining, means that any transference of bias from the dataset to the trained models could impact very large populations in unpredictable ways.

This work explores a new metric-based methodology for the analysis of bias in machine learning problems, focusing on the measure of bias transfer between the dataset and the trained model. Despite the definition of multiple fairness metrics [6], to the best of our knowledge there is no bias metric supporting multi-class classification problems studied for multiple potentially protected groups, capable of isolating dataset bias from model bias. In particular, our methodology is oriented to the bias transfer from the dataset to the model, where most mitigation systems can be implemented (given a fixed dataset with its own inherent bias).

Three metrics are proposed in this work. The first two metrics measure representational bias in the dataset. One of them is dedicated to quantifying the representational imbalance, where some demographic groups are over or under-represented in the dataset. The other one is a novel usage of the Normalized Mutual Information (NMI) and Normalized Pointwise Mutual Information (NPMI) metrics [7]. We propose using these metrics for stereotype measurement, a type of bias where demographic group representations differ among classes. These two metrics serve as a baseline for measuring the bias present in the input data. Over that baseline, we can then employ a third proposed metric to measure the bias present in the trained model. This metric measures the variation of recall per class among multiple simultaneous demographic groups, giving a single output value quantifying the amount of

bias with respect to the different demographic groups. The three metrics combined enable the study of the bias transfer from the dataset to the trained model.

In particular, we apply these metrics to analyze the bias transfer in a Facial Expression Recognition (FER) problem. The objective of FER systems is to identify the facial expression of people in either video or images, in an attempt to detect the underlying emotion. The nature of the data (human faces) in these problems make them prone to diverse biases and misrepresentations. Human biases in this context have already been studied [8].

Although the application of these metrics requires demographic information of the subjects in the dataset, the usage of existing demographic models makes recovering some of this information possible for unlabeled datasets. In the context of FER problems we employ a pretrained model, Fairface [9] to obtain some demographic descriptors, namely apparent race and gender, from the face images. Although these demographic predictions are only an approximation of the real attributes, in the absence of more accurate data they can be used to perform a general analysis.

This work is aimed at helping reduce the racial, ethnic and gender inequalities that can arise in the development and deployment of AI systems. Specifically, the proposed study of demographic bias transfer can be used to both detect demographic bias in machine learning systems and assess the impact of mitigation methods, guiding the efforts in the implementation of fairer systems in general.

Code to replicate the reported results is available at GitHub[1]. An additional Appendix to this work with the results for a second dataset, FER+ [10] is also available at the same repository.

2. Related Work

2.1. Algorithmic Fairness

The concept of algorithmic fairness is usually built around the absence of bias or harm. In particular, several independent bias taxonomies have been proposed, focused on different definitions and aspects of bias. Taxonomies like that of [11] consider up to 7 sources of bias in a machine learning pipeline, split in two subgroups:

- Biases in the data generation, comprising historical bias, representational bias and measurement bias.
- Biases in the building and implementation of the system, comprising learning bias, aggregation bias, evaluation bias and deployment bias.

This work focuses on the measurement of bias in both subgroups independently, and more specifically, in the

source dataset and the trained model. The measurement of bias in the source dataset aggregates historical and representational biases, and is common to any model and deployment that employs the same dataset. The measurement of model bias includes learning and aggregation biases. Jointly measuring both biases will give us an estimation of the amount of bias that leaks from the dataset to the model. This can help guide our efforts to the most critical parts of the system, easing the study of mitigation strategies. The remaining forms of bias are not directly applicable to our analysis.

2.1.1. Bias metrics

There are multiple fairness and bias metrics defined through the literature [6]. They are commonly defined for binary classification problems and two populations, a general population and a protected group, with one or several protected attributes indicating the membership to these populations.

However, in many applications there are multiple demographic groups that could require protection at the same time, and more than two target classes, with none of them being clearly advantageous. To the best of our knowledge, there is no bias metric covering this casuistry that is also able to isolate true model bias from validation dataset imbalance. Therefore, we focus on multi-class systems, studying bias across multiple demographic groups. Our bias definition considers any differential treatment suffered by any of the demographic groups in any of the problem classes. To design our metrics, we take the following as reference.

Disparate impact [12] asserts that the proportion of predictions of the positive class is similar across groups:

$$\frac{P(\hat{y} = 1 | s \neq 1)}{P(\hat{y} = 1 | s = 1)} \geq 1 - \varepsilon \,, \qquad (1)$$

where s is the protected attribute (1 for the privileged group and $\neq 1$ otherwise) and $\hat{y}$ the class prediction (1 for the positive class and $\neq 1$ otherwise).

Disparate Impact is arguably the most common fairness metric. However, it requires a defined positive outcome and privileged and protected groups, making it unfit for our application.

Overall accuracy equality [13] requires a similar accuracy across groups:

$$|P(y = \hat{y} | s = 1) - P(y = \hat{y} | s \neq 1)| \leq \varepsilon \,, \quad (2)$$

where s is the protected attribute (1 for the privileged group and $\neq 1$ otherwise), $\hat{y}$ the predicted class, and y the real class.

As it focuses on the accuracy, the Overall accuracy equality metric is applicable in both binary and multi-class problems. It also treats all target classes equally,

without requiring one of them to be defined as positive or advantageous. Unfortunately, it still requires defined privileged and protected groups, making it suitable only for the analysis of individual demographic groups.

Mutual Information [14] measures the statistical dependence between two attributes:

$$\sum_{\hat{y} \in \hat{Y}} \sum_{s \in S} P(\hat{y}, s) \log \frac{P(\hat{y}, s)}{P(\hat{y})P(s)} \leq \varepsilon , \quad (3)$$

where s denotes the protected attribute (from a set S) and $\hat{y}$ denotes the prediction (from a set $\hat{Y}$).

Mutual Information is one of the few fairness metrics that can be directly applied to multi-class classification problems, even when considering several potentially protected groups. Despite this, the metric only measures the dependence between the protected attributes and either the predicted or real class. In problems where the validation partition is not balanced (the real class and protected attributes are dependent), the Mutual Information measured over the trained model prediction becomes unable to disentangle that imbalance from the potential model bias.

Nonetheless, for our analysis we propose the employment of two variants of the Mutual Information metric to measure the bias not in the final model, but in the source dataset, where they only consider the real class and the protected attributes. These variants are the Normalized Mutual Information (NMI) and the Normalized Pointwise Mutual Information (NPMI) proposed by [7] in the context of collocation extraction. They are both bounded and easier to interpret than the classical Mutual Information, and in the case of the NPMI, it can also detect specific stereotypes on top of the general dataset bias. The mathematical definitions are given in Section 3.2.

2.2. Facial Expression Recognition

The problem of automatic FER is commonly used as a proxy to the more general emotion recognition. Although many works raise questions about the universality of facial expressions in conveying emotions across cultures [8, 15], and despite developments in other emotion measurement modalities [16], FER is still one of the most widely used methods. The applications of these systems are multiple, ranging from robotics [17] to assistive technology [18].

Most works deal with either a continuous emotion codification, such as the Pleasure-Arousal-Dominance model [19] or a discrete codification, such as the six/seven basic emotions proposed by [20]. This work focuses on the second approach, the most used in modern machine learning.

3. Proposal

3.1. Representational Bias Metric

The most basic level of demographic analysis that can be performed in a dataset is that of representational bias, where there is unequal representation of different demographic groups in the overall dataset. A clear predominance of a demographic group in the dataset can hint at a potential bias in favor of that group, generating a differentiated and potentially harmful behavior of the resulting model. Previous bias metrics (presented in Section 2.1.1) only measure the bias for the final model, so they cannot detect this specific bias. As a way to measure the representational bias of the dataset, we propose using a metric based on the standard deviation of the normalized demographic distribution. This **Normalized Standard Deviation** (NSD) adjusts the standard deviation of a normalized vector:

$$\mathrm{NSD}(x) = \frac{n}{\sqrt{n-1}} \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} , \quad (4)$$

where n is the number of elements of the vector x and $\bar{x}$ stands for the arithmetic mean.

The NSD calculated over a normalized demographic distribution is bounded in the interval $[0, 1]$, where 0 is no bias (uniform distribution) and 1 is total bias (the entire population belongs to a single group).

3.2. Stereotypical Bias Metric

A second level of analysis of the dataset not covered by previous metrics is that of the presence of stereotypes, understood as a variation in the demographic profile of each target class in the dataset. In the context of FER, this can result in a different prior for an emotion label depending on the perceived demographic group of the sample.

For most FER datasets, and in general for most multi-class multi-demographic datasets, the analysis and quantification of stereotypical bias is complex due to the double class and demographic imbalance usually accepted in these datasets. To decouple this secondary bias from the main representational bias, we employ the NPMI metric proposed by [7].

The **Normalized Pointwise Mutual Information** (NPMI) measures the statistical dependence between two attributes for a specific pair of values:

$$\mathrm{NPMI}(s, y) = -\frac{\ln \frac{P(s,y)}{P(s)P(y)}}{\ln P(s, y)} , \quad (5)$$

where s denotes the protected attribute and y denotes the class. NPMI values lay in the range $[-1, 1]$, with 1 being total correlation (overrepresentation), 0 being no

correlation, and -1 being inverse correlation (underrepresentation).

For an aggregated value representing the summary of the NPMI biases, we employ the NMI metric, also proposed by [7].

The **Normalized Mutual Information** (NMI) measures the statistical dependence between two attributes:

$$\text{NMI}(S, Y) = -\frac{\sum_{y \in Y} \sum_{s \in S} P(s, y) \ln \frac{P(s, y)}{P(s)P(y)}}{\sum_{y \in Y} \sum_{s \in S} P(s, y) \ln P(s, y)},$$

(6)

where S denotes all the demographic groups studied and Y the set of classes of the problem. The NMI value lies in the range $[0, 1]$, with 0 being no bias and 1 being total bias.

3.3. Model Bias Metric

Based on the metrics presented in Section 2.1.1, we require a new metric to measure the model bias in problems like FER, where we operate over multiple demographic groups and multiple classes, with an unbalanced test dataset. For the calculation of the model bias metric, we expect the model to perform with a similar recall for each demographic group and for each target class to be considered fair. The **Recall** $R(y, s)$ for a class y and a demographic group s is defined as

$$R(y, s) = P(\hat{y} = y | y, s) .$$

(7)

It is important to note that each of the classes in the problem can have different inherent difficulties. Therefore, we first calculate the intraclass disparity for each class by aggregating the recalls of that class for all demographic groups. Later, we aggregate those intraclass disparities into a final dataset metric. The **Intraclass Disparity** (ID) for each class y, aggregated for S demographic groups, is defined as:

$$\text{ID}(y) = \frac{1}{n-1} \sum_{s \in S} \left(1 - \frac{R(y, s)}{\max_{s' \in S} R(y, s')} \right) ,$$

(8)

with the convention that $\text{ID}(y) = 0$ if $\max_{s' \in S} R(y, s') = 0$.

This metric uses the maximum recall $R(y, s)$ of any demographic group (s) for the class (y) as the baseline to obtain a relative value between 0 (same performance as the maximum recall group) and 1 (recall 0 relative to the maximum recall group) for each demographic group. These values are then aggregated, obtaining the final metric. This measure considers all the groups, maximizing the ID metric to 1 when all groups except the one with the highest recall have an accuracy of 0 (situation of maximum privilege or bias).

Finally, we can study the class disparities by themselves or aggregate them with a simple mean over the set of

classes C, giving us an **Overall Disparity** OD for the model, defined as:

$$\text{OD} = \frac{1}{|C|} \sum_{c \in C} \text{ID}(c) .$$

(9)

The OD value has a lower bound of 0 (no bias) and an upper bound of 1 (maximum bias). The same bounds apply to the individual ID scores.

4. Case Study

4.1. Dataset

Affectnet [21] is a large FER dataset, composed of 420,299 images of facial expressions. $291,651$ of these images are classified in basic emotions, namely *neutral*, *happy*, *sad*, *surprise*, *fear*, *disgust*, *angry* and *contempt*. The dataset is divided into two partitions, one for training and one for validation, with $287,651$ and $4,000$ images respectively.

The dataset was created from Internet image searches, based on queries related to both the emotional labels considered and gender, age, and ethnicity descriptors. The search query strings were also translated to 6 different languages.

The final images have 425 by 425 pixels of resolution.

An additional Appendix to this work with the results for a second dataset, FER+ [10] is available[2].

4.2. Determination of Demographic Labels

Most large FER datasets [22, 21] are gathered from regular Internet image searches, with relatively low data curation. Outside FER problems, many datasets have already undergone heavy bias analysis [4], but the lack of meta-information and demographic labels on subjects for in the wild (ITW) FER datasets has hindered the bias analysis in this context. Affectnet is not an exception and, despite the diversity considerations during its generation, demographics labels are not available.

In recent years, the development of new datasets for demographic annotation, such as Fairface [9], has enabled the demographic relabeling of existing datasets. It is important to note that this kind of annotation is highly subjective and imperfect, and any demographic label obtained constitutes only a proxy measure for the real demographic characteristics of the subjects, which is already a subjective and complex concept. For example, the gender classification of Fairface is binary (Male, Female), which already constitutes a bias against nonbinary people [1]. The race classification only uses the seven most common

---

descriptors (White, Latino/Hispanic, East Asian, Black, Middle Eastern, Indian, and Southeast Asian). Furthermore, both categories are treated as single-label classifications. Despite these issues, the Fairface model gives us an approximation of the demographic profile of the dataset, enough for a general bias analysis in the absence of more accurate data.

4.3. Experimental Setup

We employ a simple VGG11 [23] network with no pre-training as the base test model. This is a classical convolutional architecture that is often used as a baseline for machine learning applications.

The experiments are developed in PyTorch 1.8.0 and Fastai 2.3.1. The hardware used is a machine equipped with a GeForce RTX 2080 Ti GPU, 128 GB of RAM, an Intel® Xeon® Silver 4214 CPU, and running CentOS Linux 7.7.

All the models are trained under the same conditions and hyperparameters, namely, a maximum learning rate of $1e^{-2}$ with a 1cycle policy (as described in [24] and implemented in Fastai) for 100 iterations. This parameter was decided using the *lr_finder* tool in Fastai. The batch size is set to 256, the maximum allowed by the hardware setup. For each dataset, we train the model 10 times and average the results over them.

We have also applied the basic data augmentation provided by Fastai through the *aug_transforms* method, including left-right flipping, warping, rotation, zoom, brightness, and contrast alterations.

4.4. Experiments

4.4.1. Dataset bias

To analyze the dataset, first we process it using the Fairface model to obtain a demographic estimation for each image. We then proceed to calculate the demographic representation profile of the dataset, and compute the NSD bias metric presented in Section 3.1. After that, we can also use the NPMI metric to highlight any stereotypical bias inherent in the distribution of labels for each demographic group, as explained in Section 3.2.

The demographic information added to the dataset through this relabeling process also enables the creation of derived datasets that can be used to simulate different bias situations. In particular, we generate:

- Two *balanced subsets* for the racial and gender demographics. These datasets have the same representation of each demographic group considered for each of the target labels. This balancing removes both representational and stereotypical biases.

- Two *artificially biased gender subsets*, that contain examples of only one gender.

The two gender biased and the gender balanced subsets are generated with exactly the same number of examples for each class to enable the comparison between the three of them.

4.4.2. Model bias

All trained models are evaluated on the whole Affectnet validation partition, and the OD metric described in Section 3.3 is used as a measure of bias.

In addition to the subsets proposed in the previous experiment, a series of *stratified subsets* is generated to evaluate the influence of the size of the dataset on the residual bias of the model. Each subset will contain a fixed percentage of the original train partition, maintaining the same demographic distribution. Hence, they will preserve the dataset biases of Affectnet.

In summary, we want to study the behavior of the model bias metric when increasing the number of training examples and when training on balanced subsets. If balancing the dataset is an efficient mitigation strategy, we expect a lower metric with little to no impact on general accuracy, for the same training data size.

5. Results And Discussion

5.1. Dataset bias

Figure 1 shows the apparent race distribution of the dataset in Affectnet. We can observe a strong imbalance in favor of the white race, which comprises 64.4% of the training data. The apparent gender distribution (not shown) is much more balanced, with 49.7% of the training data classified as male and the rest as female. In the case of gender, imbalances arise in the analysis per label, summarized in Figure 2, reaching an imbalance 72% − 28% in the case of the *angry* class.

These intuitive indicators of bias are also reflected in the NSD and NMI metrics results presented in Table 1, where both the original dataset and some variations are analyzed. Representational bias is measured with the NSD metric and stereotypical bias with the NMI. Note that the NSD and NMI metrics do not share the same scales and are not comparable to each other. The number in parentheses denotes the number of demographic groups considered, 2 for gender and 7 for race. These results show a zero representational and stereotypical bias for the artificially balanced datasets in their respective categories, as expected. Note that the gender biased datasets are excluded from the gender bias calculations, as they only include examples of one gender. For the original dataset, we can observe a relatively low representational bias in the

gender category ($NSD = 0.0057$) compared to the race category ($NSD = 0.5902$). On the contrary, the stereotypical bias is higher in the gender category ($NMI = 0.0089$) compared to the race category ($NMI = 0.0021$). This agrees with the imbalance perceived in Figure 2.
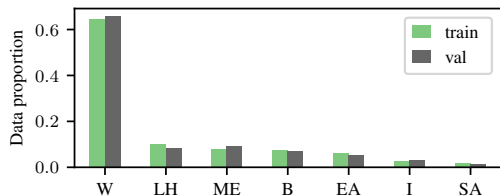

Figure 1: Apparent race distribution of Affectnet (W: White, LH: Latino / Hispanic, ME: Middle Eastern, B: Black, EA: East Asian, I: Indian, SA: Southeast Asian).
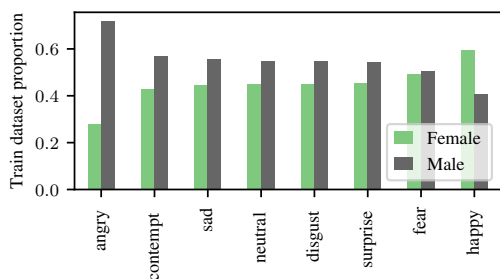

Figure 2: Apparent gender distribution of Affectnet for each label.

Dataset		Representational bias (NSD)		Stereotypical bias (NMI)	
		Race (7)	Gender (2)	Race (7)	Gender (2)
Original		0.5902	0.0057	0.0021	0.0089
Balanced	Race	**0.0000**	0.0159	**0.0000**	0.0091
	Gender	0.5929	**0.0000**	0.0017	**0.0000**
Gender biased	M	0.5702	–	0.0020	–
	F	0.6129	–	0.0020	–

Table 1

Representational (NSD) and stereotypical (NMI) bias metrics for the original dataset and the considered subsets.

The stereotypical bias detected through the NMI can be further analyzed with NPMI. Figure 3 shows the NPMI matrices for both the apparent gender and race analysis. In the race category, the most prominent value indicates an underrepresentation (-0.13) of the East Asian group in the *angry* class. In the gender category two stereotypical biases stand out: for the *angry* class, an underrepresentation of people recognized as female (and a corresponding

overrepresentation of people recognized as male), and the opposite for the *happy* class, with an overrepresentation of people recognized as female and underrepresentation of people recognized as male. The results for the gender category are consistent with the *angry-men-happy-women* social bias already known in the literature [25].
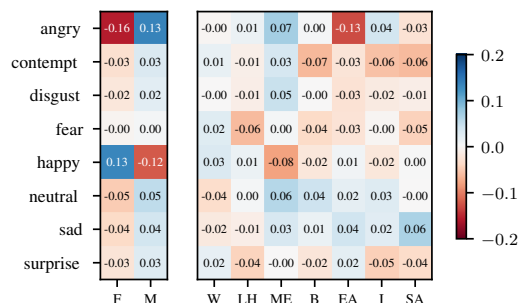

Figure 3: NPMI analysis of Affectnet. (F: Female, M: Male, W: White, LH: Latino / Hispanic, ME: Middle Eastern, B: Black, EA: East Asian, I: Indian, SA: Southeast Asian)

5.2. Model bias

The results regarding the trained model bias metric OD and accuracy are shown in Table 2, and graphically in Figure 4.

Regarding racial bias, across the stratified subsets we can observe consistently high model bias values, clearly decreasing from the highest bias in the smallest dataset (0.422 ± 0.043) to the lowest in the largest dataset (0.268 ± 0.023). The race balanced dataset, when compared to an imbalanced one of similar size (Stratified / 0.13), does not appear to improve the bias metric (0.297 ± 0.016 vs. 0.284 ± 0.017), while decreasing accuracy (45.7 ± 0.3 vs. 48.4 ± 0.5). In this case, increasing the dataset size (Stratified / 0.22 and higher) improves the accuracy without impacting the racial bias metric (stable around 0.29 and decreasing for larger sizes). Note that the accuracy is calculated over the whole test partition of Affectnet, which is not balanced in terms of labels and the demographic groups studied.

In the gender category, we observe comparatively lower bias values overall, from 0.264 ± 0.053 to 0.133 ± 0.015, consistent with the lower gender bias of the original dataset ($NSD = 0.0057$, $NMI = 0.0089$). Balancing the gender does not significantly improve the accuracy results for a similar sized dataset (Stratified / 0.36, accuracy 51.8 ± 0.4 vs. 51.1 ± 0.6), but in this case the gender bias value improves significantly (0.091 ± 0.014, lower than all other models). Additionally, although the artificially gender biased datasets have a similar accuracy to the bal-

anced one (50.7 ± 0.5 and 50.2 ± 0.4 vs. 51.8 ± 0.4), their bias metric is substantially higher (0.185 ± 0.017 and 0.242 ± 0.018 vs. 0.091 ± 0.014).

| Train data | | Size | Accuracy | Bias | |
				Race OD	Gender OD
Original	1%	2,839	33.5 ± 0.9	0.422 ± 0.043	0.264 ± 0.053
	2%	5,678	39.2 ± 0.9	0.362 ± 0.027	0.214 ± 0.013
	3%	8,517	41.5 ± 0.8	0.347 ± 0.019	0.186 ± 0.033
	5%	14,195	44.4 ± 0.4	0.315 ± 0.033	0.192 ± 0.022
	8%	22,712	46.4 ± 0.5	0.288 ± 0.017	0.174 ± 0.029
	13%	36,907	48.4 ± 0.5	0.284 ± 0.028	0.144 ± 0.026
	22%	62,458	49.6 ± 0.6	0.292 ± 0.024	0.166 ± 0.018
	36%	102,204	51.1 ± 0.6	0.289 ± 0.014	0.157 ± 0.025
	60%	170,340	53.6 ± 0.5	0.279 ± 0.030	0.149 ± 0.018
	100%	283,901	$\mathbf{55.8 \pm 0.2}$	$\mathbf{0.268 \pm 0.023}$	0.133 ± 0.015
Balanced	Race	32,452	45.7 ± 0.3	0.297 ± 0.016	0.177 ± 0.017
	Gender	117,790	51.8 ± 0.4	0.273 ± 0.026	$\mathbf{0.091 \pm 0.014}$
Gender biased	M	117,790	50.7 ± 0.5	0.277 ± 0.015	0.185 ± 0.017
	F	117,790	50.2 ± 0.4	0.315 ± 0.022	0.242 ± 0.018

Table 2
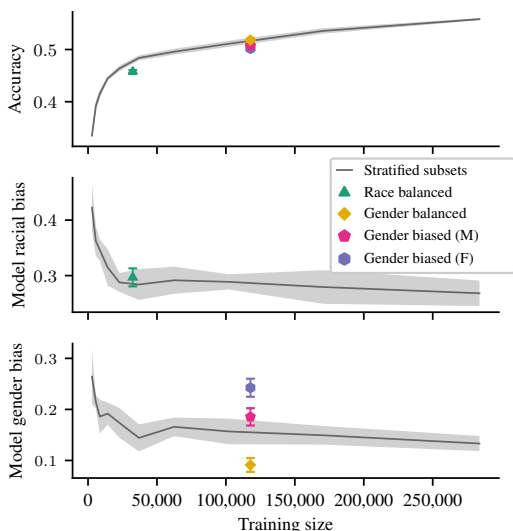Bias metric summary for the model when trained on dataset variations.

Figure 4: Accuracy and apparent race and gender biases for models.

5.3. Bias transfer

Regarding the apparent race analysis of the system, the proposed metrics reveal an important representational imbalance of the dataset coupled with some stereotypical bias. Despite this, when comparing a stratified subset to a dataset of the same size but balanced by race, we observe no improvement in the trained model bias. This suggests that either the source of the model bias is not the dataset bias (inherent differences between racial expressions, for example) or that directly balancing the dataset is not a

successful mitigation approach in this context (because of quality differences between the images associated with each race, for example).

Regarding the gender bias, the metrics reveal a comparatively lower representational bias, but a much more pronounced stereotypical bias. In this case, balancing the dataset seems to substantially improve the trained model bias, mitigating the bias transfer. This suggests that the stereotypical bias detected in the dataset has a large impact on the trained model, but the balancing of the dataset corrects it properly.

Additionally, we observe a strong tendency to reduce the bias scores as the size of the training dataset increases, even when the datasets have the same representational and stereotypical biases.

Although the metrics have unveiled both gender and racial bias in the source dataset, these bias transference results suggest that dataset gender bias has a greater impact in the final model. Thus, dataset gender bias seems more susceptible to mitigation measures in the early stages of the AI life-cycle, whereas racial bias may require different mitigation measures in later stages. Further studies would be required to evaluate the impact of different mitigation techniques in this case, but are out of the scope of this paper.

6. Conclusion

The metrics presented have been shown to be useful, reflecting some of the biases present in both real and manipulated datasets through easily interpretable values. The analysis of these metrics allows the study of the bias transfer from dataset to trained model, which can be useful for understanding the bias in different stages of a machine learning pipeline, and consequently in the study of mitigation strategies.

In our case study, we have revealed the heavy racial representational bias of a popular FER dataset, Affectnet, and the presence of stereotypical gender biases. The experiments also show how the resulting model seems almost invariant to the removal of the racial bias, while being severely impacted by any gender bias, either induced or corrected, in the source dataset. In the Appendix to this document[3] the results for a second dataset, FER+, are provided, showing similar tendencies to the ones found for Affectnet. These results, while specific to this model and training setup, expose the complexity of the bias analysis and its impact in real world problems.

As future work lines, the same analysis could be performed for more datasets, models, and training setups. Further work in the development of new demographic datasets and models could also improve the accuracy and

---

[3]https://github.com/irisdominguez/Dataset-Bias-Metrics

detail of this bias analysis and extend it to other problems. Different mitigation techniques, especially in the dataset preprocessing stage, could differ wildly in their impact. Finally, the metrics still require further analysis of their properties and potential application to other contexts. For example, although our proposed OD reflects both representational and stereotypical biases, having metrics capable of decoupling them could enable a more in-depth bias analysis.

Furthermore, new application areas could be researched in other multi-class multi-demographic AI systems, such as age, gender and race recognition, AI-based medical diagnosis and sign language gesture recognition, to name a few.

Acknowledgments

References

[1] O. Keyes, The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition, Proceedings of the ACM on Human-Computer Interaction 2 (2018) 1–22. doi:10.1145/3274357.

[2] V. U. Prabhu, A. Birhane, Large image datasets: A pyrrhic win for computer vision?, arXiv:2006.16923 [cs, stat] (2020). arXiv:2006.16923.

[3] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: S. A. Friedler, C. Wilson (Eds.), Proceedings of the 1st Conference on Fairness, Accountability and Transparency, volume 81 of *Proceedings of Machine Learning Research*, PMLR, 2018, pp. 77–91.

[4] C. Dulhanty, A. Wong, Auditing ImageNet: Towards a Model-driven Framework for Annotating Demographic Attributes of Large-Scale Image Datasets, arXiv:1905.01347 [cs] (2019). arXiv:1905.01347.

[5] E. Denton, A. Hanna, R. Amironesei, A. Smart, H. Nicole, On the genealogy of machine learning datasets: A critical history of ImageNet, Big Data & Society 8 (2021) 205395172110359. doi:10.1177/20539517211035955.

[6] D. Pessach, E. Shmueli, Algorithmic Fairness, arXiv:2001.09784 [cs, stat] (2020). arXiv:2001.09784.

[7] G. Bouma, Normalized (pointwise) mutual information in collocation extraction, in: Proceedings of GSCL, volume 30, GSCL, 2009, pp. 31–40.

[8] H. A. Elfenbein, N. Ambady, On the universality and cultural specificity of emotion recognition: A meta-analysis, Psychological Bulletin 128 (2002) 203–235. doi:10.1037/0033-2909.128.2.203.

[9] K. Karkkainen, J. Joo, FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation, in: 2021 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, Waikoloa, HI, USA, 2021, pp. 1547–1557. doi:10.1109/WACV48630.2021.00159.

[10] E. Barsoum, C. Zhang, C. C. Ferrer, Z. Zhang, Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution, arXiv:1608.01041 [cs] (2016). arXiv:1608.01041.

[11] H. Suresh, J. V. Guttag, A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle, arXiv:1901.10002 [cs, stat] (2021). arXiv:1901.10002.

[12] M. Feldman, S. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, Certifying and removing disparate impact, arXiv:1412.3756 [cs, stat] (2015). arXiv:1412.3756.

[13] R. Berk, H. Heidari, S. Jabbari, M. Kearns, A. Roth, Fairness in Criminal Justice Risk Assessments: The State of the Art, Sociological Methods & Research 50 (2018) 3–44. doi:10.1177/0049124118782533.

[14] T. Kamishima, S. Akaho, H. Asoh, J. Sakuma, Enhancement of the Neutrality in Recommendation, Proc. of the 2nd Workshop on Human Decision Making in Recommender Systems 893 (2012) 8–14.

[15] R. E. Jack, C. Blais, C. Scheepers, P. G. Schyns, R. Caldara, Cultural Confusions Show that Facial Expressions Are Not Universal, Current Biology 19 (2009) 1543–1548. doi:10.1016/j.cub.2009.07.051.

[16] J. Gonzalez-Sanchez, M. Baydogan, M. E. Chavez-Echeagaray, R. K. Atkinson, W. Burleson, Chapter 11 - Affect Measurement: A Roadmap Through Approaches, Technologies, and Data Analysis, in: M. Jeon (Ed.), Emotions and Affect in Human Factors and Human-Computer Interaction, Academic Press, San Diego, 2017, pp. 255–288. doi:10.1016/B978-0-12-801851-4.00011-2.

[17] E. B. Sonmez, H. Han, O. Karadeniz, T. Dalyan, B. Sarioglu, EMRES: A new EMotional RESpondent robot, IEEE Transactions on Cognitive and Developmental Systems (2021) 1–1. doi:10.1109/

`TCDS.2021.3120562`.

[18] J. L. Joseph, S. P. Mathew, Facial Expression Recognition for the Blind Using Deep Learning, in: 2021 IEEE 4th International Conference on Computing, Power and Communication Technologies (GUCON), IEEE, Kuala Lumpur, Malaysia, 2021, pp. 1–5. doi:`10.1109/GUCON50781.2021.9574035`.

[19] A. Mehrabian, Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament, Current Psychology 14 (1996) 261–292. doi:`10.1007/BF02686918`.

[20] P. Ekman, W. V. Friesen, Constants across cultures in the face and emotion., Journal of Personality and Social Psychology 17 (1971) 124–129. doi:`10.1037/h0030377`.

[21] A. Mollahosseini, B. Hasani, M. H. Mahoor, AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild, IEEE Transactions on Affective Computing 10 (2019) 18–31. doi:`10.1109/TAFFC.2017.2740923`. `arXiv:1708.03985`.

[22] C. F. Benitez-Quiroz, R. Srinivasan, A. M. Martinez, EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 2016, pp. 5562–5570. doi:`10.1109/CVPR.2016.600`.

[23] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, 2015. `arXiv:1409.1556`.

[24] L. N. Smith, A disciplined approach to neural network hyper-parameters: Part 1 – learning rate, batch size, momentum, and weight decay, arXiv:1803.09820 [cs, stat] (2018). `arXiv:1803.09820`.

[25] A. P. Atkinson, J. Tipples, D. M. Burt, A. W. Young, Asymmetric interference between sex and emotion in face perception, Perception & Psychophysics 67 (2005) 1199–1213. doi:`10.3758/BF03193553`.