ASSESSING DEMOGRAPHIC BIAS TRANSFER FROM DATASET TO MODEL: A CASE STUDY IN FACIAL EXPRESSION RECOGNITION

Iris Dominguez-Catena[a] Daniel Paternain Mikel Galar

July 2022

Institute of Smart Cities (ISC), Department of Statistics, Computer Science and Mathematics
Public University of Navarre (UPNA)

upna
Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

[a] iris.dominguez@unavarra.es

OUTLINE

Introduction

Proposal

Experiments and results

Conclusion

CONTEXT: FACIAL EXPRESSION RECOGNITION

Figure 1: Sample from FER+ dataset[1].

[1] Barsoum et al., "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution".

CONTEXT: DATASET BIAS

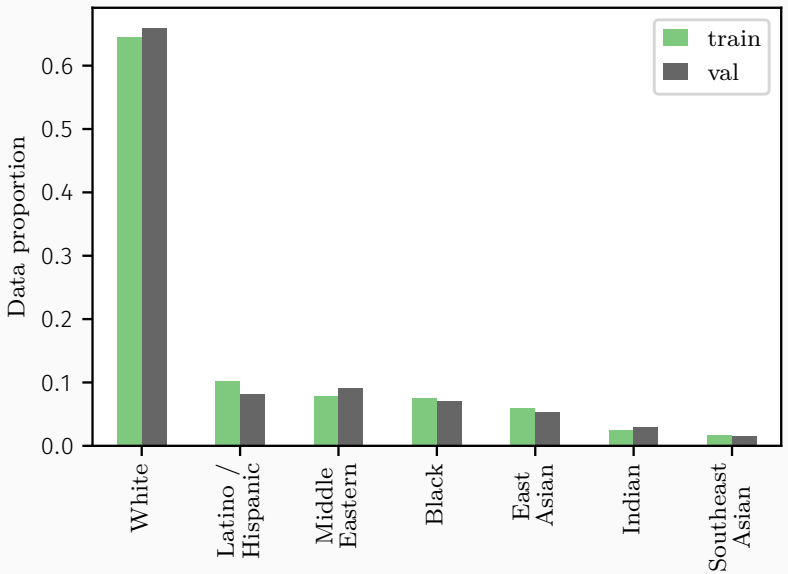


Figure 2: Apparent race distribution of Affectnet.

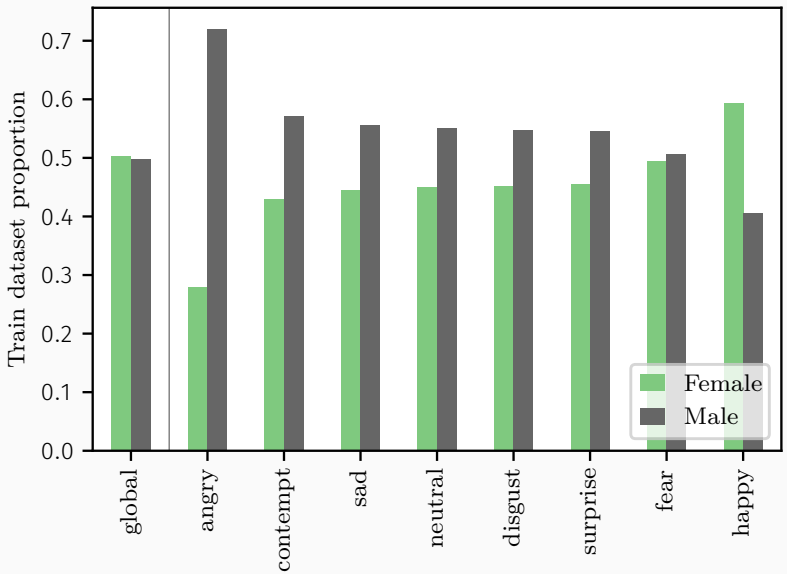


Figure 3: Apparent gender distribution of Affectnet for each label (train partition).

OBJECTIVE AND PROPOSAL

Our objective: to measure bias in the datasets, and observe if it transfers to the model.

Our proposal:

- Three metrics for **dataset bias**.
  - Representational bias.
  - Stereotypical bias.
    - Local.
    - Global.
- One metric for **model bias**.

BACKGROUND: PREVIOUS METRICS

Fairness: absence of **unwanted bias**. Equal treatment of the users independent from their protected attributes.

- **Disparate impact**[2].
  - **Positive (advantageous) class**.
- **Overall accuracy equality**[3].
  - Multi-class.
  - **Single protected group**.
- **Mutual Information**[4].
  - Class independent.
  - Multiple protected groups.
  - **Balanced test dataset**.

[2] Feldman et al., "Certifying and Removing Disparate Impact".

[3] Berk et al., "Fairness in Criminal Justice Risk Assessments".

[4] Kamishima et al., "Enhancement of the Neutrality in Recommendation".

Representational bias: Under or overrepresentation of a demographic group in the **global** dataset.

- **Normalized Standard Deviation** (NSD):

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

  where x is the normalized vector of population distribution.
- Bounds:
  - 0: no bias.
  - 1: total bias.

Example

W	LH	ME	B	EA	I	SA	NSD
1/7	1/7	1/7	1/7	1/7	1/7	1/7	0
1/2	1/2	0	0	0	0	0	0.65
1	0	0	0	0	0	0	1

Example

Source data

Race	angry	happy	Total
ME	10	0	10
EA	0	10	10
W	10	10	20
Total	20	20	40

Race	angry	happy	Total
ME	5	5	10
EA	5	5	10
W	10	10	20
Total	20	20	40

Stereotypical bias: under or overrepresentation of a demographic group in a **specific** class.

- **Normalized Pointwise Mutual Information** (NPMI)[5]:

$$\text{NPMI}(s, y) = -\frac{\ln \frac{P(s,y)}{P(s)P(y)}}{\ln P(s, y)}\ , \tag{2}$$

  where s denotes a protected attribute and y a class.

- Bounds:
  - -1: total underrepresentation.
  - 0: no bias.
  - 1: total overrepresentation.

[5] Bouma, "Normalized (Pointwise) Mutual Information in Collocation Extraction".

Example

Source data

Race	angry	happy	Total
ME	10	0	10
EA	0	10	10
W	10	10	20
Total	20	20	40

Race	angry	happy	Total
ME	5	5	10
EA	5	5	10
W	10	10	20
Total	20	20	40

NPMI

Race	angry	happy
ME	0.5	-1
EA	-1	0.5
W	0	0

Race	angry	happy
ME	0	0
EA	0	0
W	0	0

Stereotypical bias: under or overrepresentation of a demographic group in a **specific** class.

- **Normalized Mutual Information** (NMI)[6]:

$$\mathrm{NMI}(S, Y) = -\frac{\sum_{y \in Y} \sum_{s \in S} P(s, y) \ln \frac{P(s,y)}{P(s)P(y)}}{\sum_{y \in Y} \sum_{s \in S} P(s, y) \ln P(s, y)} , \tag{3}$$

  where S denotes the demographic groups and Y the labels.

- Bounds:
  - 0: no bias.
  - 1: total bias.

[6] Bouma, "Normalized (Pointwise) Mutual Information in Collocation Extraction".

Example

Source data

Race	angry	happy	Total
ME	10	0	10
EA	0	10	10
W	10	10	20
Total	20	20	40

Race	angry	happy	Total
ME	5	5	10
EA	5	5	10
W	10	10	20
Total	20	20	40

NPMI

Race	angry	happy
ME	0.5	-1
EA	-1	0.5
W	0	0

Race	angry	happy
ME	0	0
EA	0	0
W	0	0

NMI

0.25

0

Model bias: for every label, how different is the accuracy between groups.

- **Recall** (R):

$$\mathrm{R}(y, s) = P(\hat{y} = y | y, s) \; . \tag{4}$$

- **Intraclass Disparity** (ID) (for each class y):

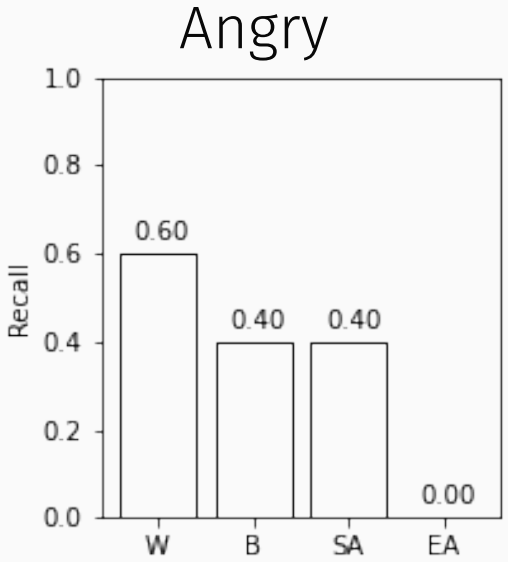
$$\mathrm{ID}(y) = \frac{1}{n-1} \sum_{s \in S} \left(1 - \frac{\mathrm{R}(y, s)}{\max\limits_{s' \in S} \mathrm{R}(y, s')} \right) \; , \tag{5}$$

- **Overall Disparity** (OD):

$$\mathrm{OD} = \frac{1}{|C|} \sum_{c \in C} \mathrm{ID}(c) \; . \tag{6}$$

- Bounds:
  - 0: no bias.
  - 1: total bias.

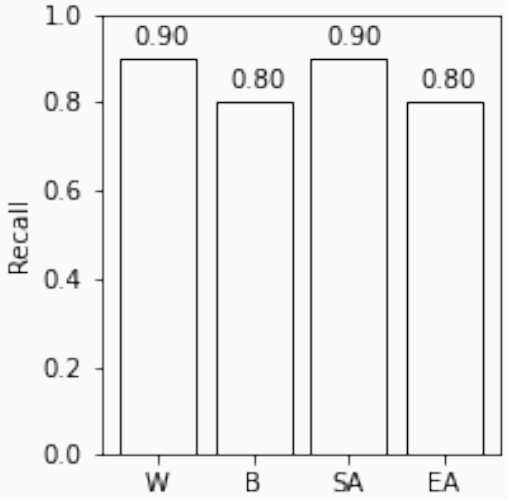
PROPOSAL: MODEL BIAS (EXAMPLE)



$\longrightarrow$

$1 - \frac{R(y,s)}{\max_{s' \in S} R(y,s')}$

$\longrightarrow ID = 0.555$



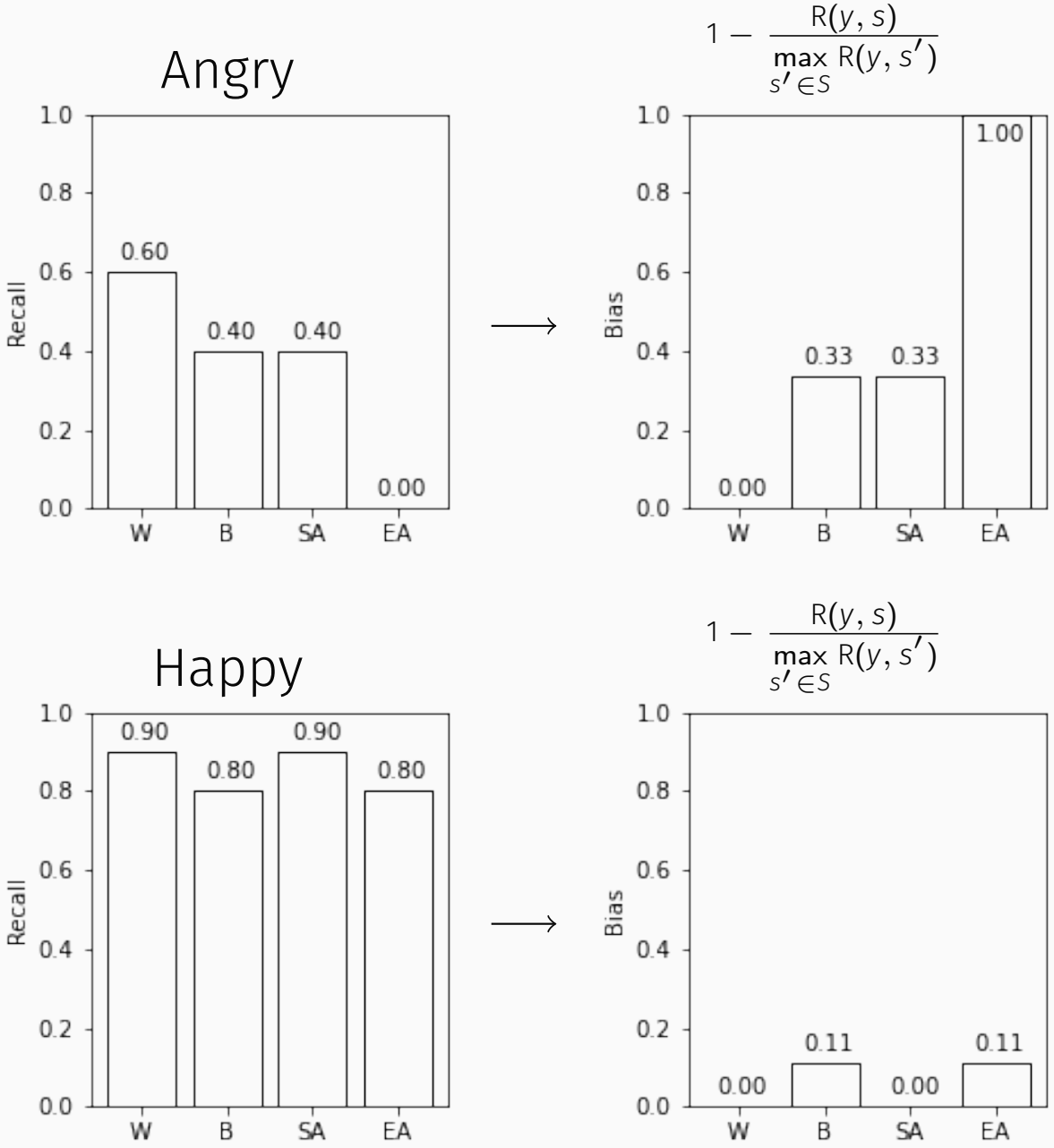
$\longrightarrow$

$1 - \frac{R(y,s)}{\max_{s' \in S} R(y,s')}$

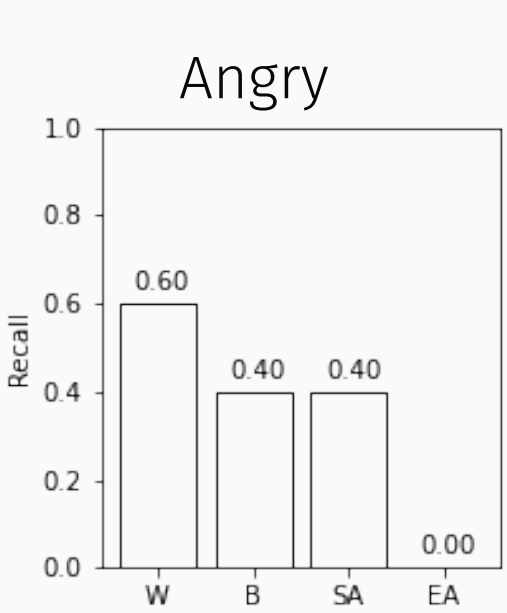
$\longrightarrow ID = 0.074$

$OD = 0.31$

PROPOSAL: MODEL BIAS (EXAMPLE)

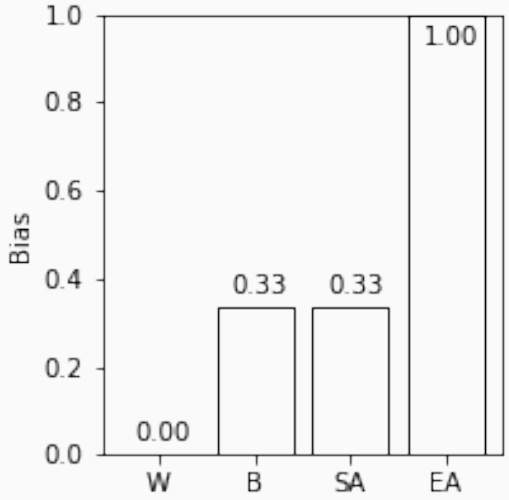


PROPOSAL: MODEL BIAS (EXAMPLE)



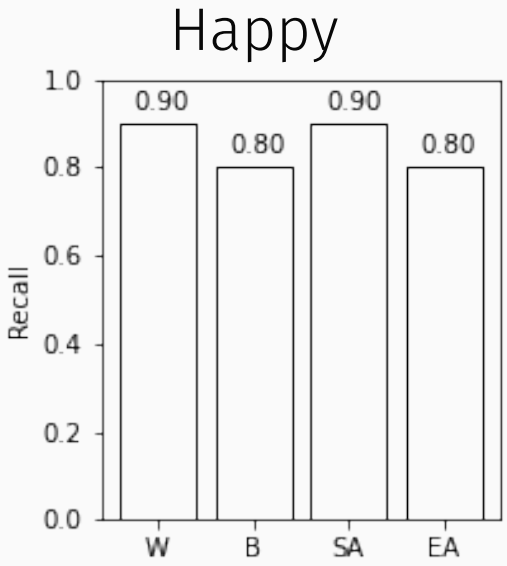
$\longrightarrow$

$$1 - \frac{\mathrm{R}(y, s)}{\max\limits_{s' \in S} \mathrm{R}(y, s')}$$



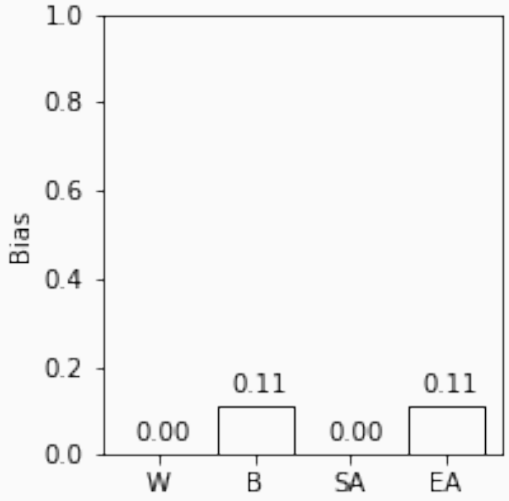
$\longrightarrow ID = 0.555$

$OD = 0.31$



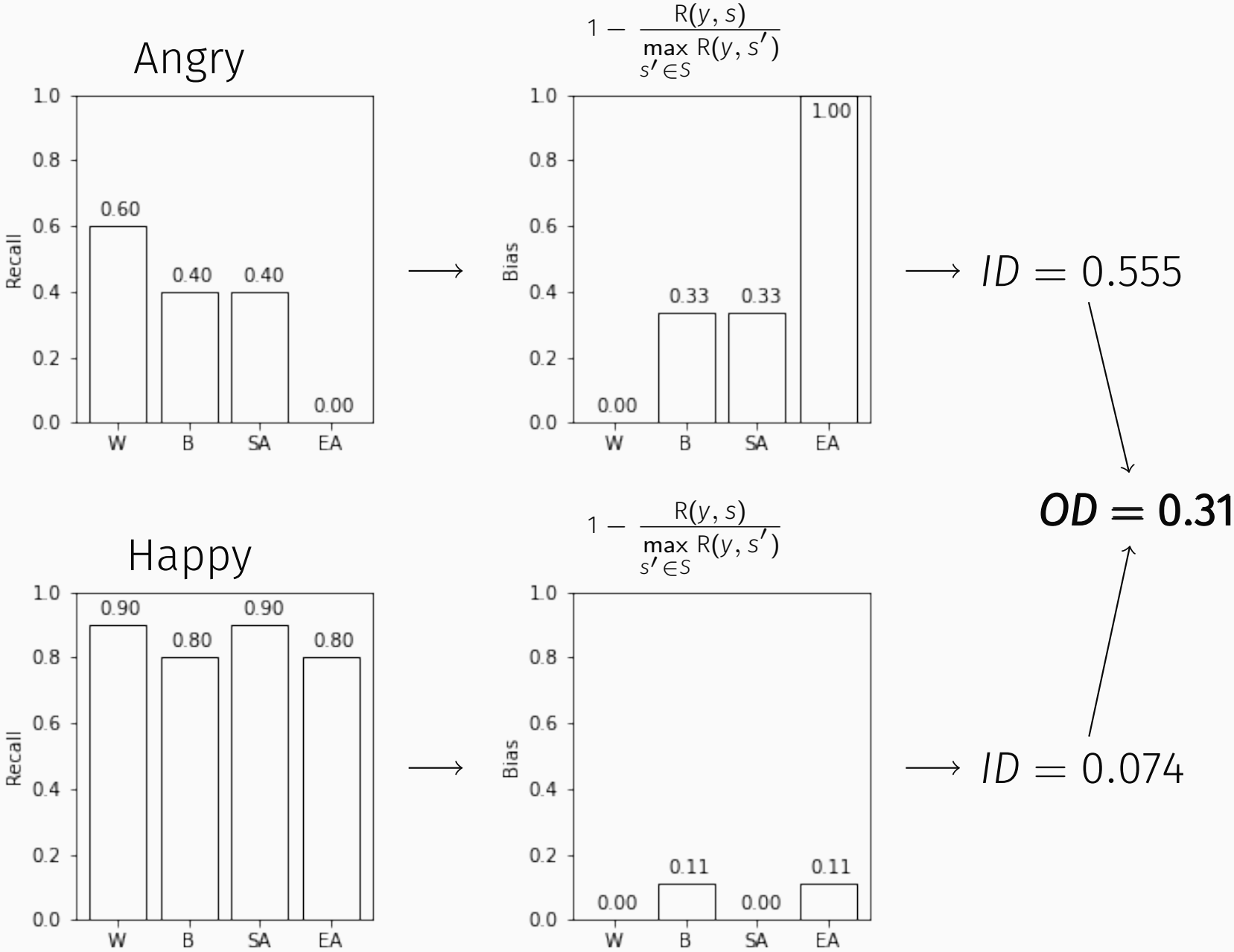
$\longrightarrow$

$$1 - \frac{\mathrm{R}(y, s)}{\max\limits_{s' \in S} \mathrm{R}(y, s')}$$



$\longrightarrow ID = 0.074$

PROPOSAL: MODEL BIAS (EXAMPLE)



- **Affectnet**[7]:
  - Complementary analysis: FER+[8] [9].
- VGG11 model[10], same hyperparameters.

Preprocessing

Demographic relabeling with an auxiliary model, FairFace[a].

- Two *apparent* gender categories.
- Seven *apparent* race categories.

[a] Karkkainen and Joo, "FairFace".

[7] Mollahosseini, Hasani, and Mahoor, "AffectNet".

[8] Barsoum et al., "Training Deep Networks for Facial Expression Recognition with Crowd-Sourced Label Distribution".

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

[10] Simonyan and Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition*.

EXPERIMENTS: DERIVATIVE DATASETS

Original

gender	%	Female	Male
angry	8.7	6,961	17,921
contempt	1.3	1,611	2,139
disgust	1.3	1,720	2,083
fear	2.2	3,149	3,229
happy	46.7	79,830	54,585
neutral	26.0	33,672	41,202
sad	8.9	11,304	14,155
surprise	4.9	6,403	7,687

Balanced

gender	%	Female	Male
angry	8.7	6,961	6,961
contempt	1.3	1,049	1,049
disgust	1.3	1,063	1,063
fear	2.2	1,784	1,784
happy	46.7	37,604	37,604
neutral	26.0	20,946	20,946
sad	8.9	7,122	7,122
surprise	4.9	3,941	3,941

Biased

gender	%	Female	Male
angry	8.7	6,961	0
contempt	1.3	1,049	0
disgust	1.3	1,063	0
fear	2.2	1,784	0
happy	46.7	37,604	0
neutral	26.0	20,946	0
sad	8.9	7,122	0
surprise	4.9	3,941	0

Stratified 0.28

gender	%	Female	Male
angry	8.7	1,949	5,018
contempt	1.3	451	599
disgust	1.3	482	583
fear	2.2	882	904
happy	46.7	22,352	15,284
neutral	26.0	9,428	11,537
sad	8.9	3,165	3,963
surprise	4.9	1,793	2,152

Balanced 0.5

gender	%	Female	Male
angry	8.6	3,480	3,480
contempt	1.3	524	524
disgust	1.3	532	532
fear	2.2	892	892
happy	46.7	18,802	18,802
neutral	26.0	10,473	10,473
sad	8.9	3,561	3,561
surprise	4.9	1,970	1,970

EXPERIMENTS: DERIVATIVE DATASETS

Original

gender	%	Female	Male
angry	8.7	6,961	17,921
contempt	1.3	1,611	2,139
disgust	1.3	1,720	2,083
fear	2.2	3,149	3,229
happy	46.7	79,830	54,585
neutral	26.0	33,672	41,202
sad	8.9	11,304	14,155
surprise	4.9	6,403	7,687

Balanced

gender	%	Female	Male
angry	8.7	6,961	6,961
contempt	1.3	1,049	1,049
disgust	1.3	1,063	1,063
fear	2.2	1,784	1,784
happy	46.7	37,604	37,604
neutral	26.0	20,946	20,946
sad	8.9	7,122	7,122
surprise	4.9	3,941	3,941

Biased

gender	%	Female	Male
angry	8.7	6,961	0
contempt	1.3	1,049	0
disgust	1.3	1,063	0
fear	2.2	1,784	0
happy	46.7	37,604	0
neutral	26.0	20,946	0
sad	8.9	7,122	0
surprise	4.9	3,941	0

Stratified 0.28

gender	%	Female	Male
angry	8.7	1,949	5,018
contempt	1.3	451	599
disgust	1.3	482	583
fear	2.2	882	904
happy	46.7	22,352	15,284
neutral	26.0	9,428	11,537
sad	8.9	3,165	3,963
surprise	4.9	1,793	2,152

Balanced 0.5

gender	%	Female	Male
angry	8.6	3,480	3,480
contempt	1.3	524	524
disgust	1.3	532	532
fear	2.2	892	892
happy	46.7	18,802	18,802
neutral	26.0	10,473	10,473
sad	8.9	3,561	3,561
surprise	4.9	1,970	1,970

EXPERIMENTS: DERIVATIVE DATASETS

Original

gender	%	Female	Male
angry	8.7	6,961	17,921
contempt	1.3	1,611	2,139
disgust	1.3	1,720	2,083
fear	2.2	3,149	3,229
happy	46.7	79,830	54,585
neutral	26.0	33,672	41,202
sad	8.9	11,304	14,155
surprise	4.9	6,403	7,687

Balanced

gender	%	Female	Male
angry	8.7	6,961	6,961
contempt	1.3	1,049	1,049
disgust	1.3	1,063	1,063
fear	2.2	1,784	1,784
happy	46.7	37,604	37,604
neutral	26.0	20,946	20,946
sad	8.9	7,122	7,122
surprise	4.9	3,941	3,941

Biased

gender	%	Female	Male
angry	8.7	6,961	0
contempt	1.3	1,049	0
disgust	1.3	1,063	0
fear	2.2	1,784	0
happy	46.7	37,604	0
neutral	26.0	20,946	0
sad	8.9	7,122	0
surprise	4.9	3,941	0

Stratified 0.28

gender	%	Female	Male
angry	8.7	1,949	5,018
contempt	1.3	451	599
disgust	1.3	482	583
fear	2.2	882	904
happy	46.7	22,352	15,284
neutral	26.0	9,428	11,537
sad	8.9	3,165	3,963
surprise	4.9	1,793	2,152

Balanced 0.5

gender	%	Female	Male
angry	8.6	3,480	3,480
contempt	1.3	524	524
disgust	1.3	532	532
fear	2.2	892	892
happy	46.7	18,802	18,802
neutral	26.0	10,473	10,473
sad	8.9	3,561	3,561
surprise	4.9	1,970	1,970

EXPERIMENTS: DERIVATIVE DATASETS

Original

gender	%	Female	Male
angry	8.7	6,961	17,921
contempt	1.3	1,611	2,139
disgust	1.3	1,720	2,083
fear	2.2	3,149	3,229
happy	46.7	79,830	54,585
neutral	26.0	33,672	41,202
sad	8.9	11,304	14,155
surprise	4.9	6,403	7,687

→

Balanced

gender	%	Female	Male
angry	8.7	6,961	6,961
contempt	1.3	1,049	1,049
disgust	1.3	1,063	1,063
fear	2.2	1,784	1,784
happy	46.7	37,604	37,604
neutral	26.0	20,946	20,946
sad	8.9	7,122	7,122
surprise	4.9	3,941	3,941

→

Biased

gender	%	Female	Male
angry	8.7	6,961	0
contempt	1.3	1,049	0
disgust	1.3	1,063	0
fear	2.2	1,784	0
happy	46.7	37,604	0
neutral	26.0	20,946	0
sad	8.9	7,122	0
surprise	4.9	3,941	0

↓

Stratified 0.28

gender	%	Female	Male
angry	8.7	1,949	5,018
contempt	1.3	451	599
disgust	1.3	482	583
fear	2.2	882	904
happy	46.7	22,352	15,284
neutral	26.0	9,428	11,537
sad	8.9	3,165	3,963
surprise	4.9	1,793	2,152

↘

Balanced 0.5

gender	%	Female	Male
angry	8.6	3,480	3,480
contempt	1.3	524	524
disgust	1.3	532	532
fear	2.2	892	892
happy	46.7	18,802	18,802
neutral	26.0	10,473	10,473
sad	8.9	3,561	3,561
surprise	4.9	1,970	1,970

EXPERIMENTS: DERIVATIVE DATASETS

Original

gender	%	Female	Male
angry	8.7	6,961	17,921
contempt	1.3	1,611	2,139
disgust	1.3	1,720	2,083
fear	2.2	3,149	3,229
happy	46.7	79,830	54,585
neutral	26.0	33,672	41,202
sad	8.9	11,304	14,155
surprise	4.9	6,403	7,687

→

Balanced

gender	%	Female	Male
angry	8.7	6,961	6,961
contempt	1.3	1,049	1,049
disgust	1.3	1,063	1,063
fear	2.2	1,784	1,784
happy	46.7	37,604	37,604
neutral	26.0	20,946	20,946
sad	8.9	7,122	7,122
surprise	4.9	3,941	3,941

→

Biased

gender	%	Female	Male
angry	8.7	6,961	0
contempt	1.3	1,049	0
disgust	1.3	1,063	0
fear	2.2	1,784	0
happy	46.7	37,604	0
neutral	26.0	20,946	0
sad	8.9	7,122	0
surprise	4.9	3,941	0

↓

Stratified 0.28

gender	%	Female	Male
angry	8.7	1,949	5,018
contempt	1.3	451	599
disgust	1.3	482	583
fear	2.2	882	904
happy	46.7	22,352	15,284
neutral	26.0	9,428	11,537
sad	8.9	3,165	3,963
surprise	4.9	1,793	2,152

↘

Balanced 0.5

gender	%	Female	Male
angry	8.6	3,480	3,480
contempt	1.3	524	524
disgust	1.3	532	532
fear	2.2	892	892
happy	46.7	18,802	18,802
neutral	26.0	10,473	10,473
sad	8.9	3,561	3,561
surprise	4.9	1,970	1,970

EXPERIMENTS: DERIVATIVE DATASETS

Original

gender	%	Female	Male
angry	8.7	6,961	17,921
contempt	1.3	1,611	2,139
disgust	1.3	1,720	2,083
fear	2.2	3,149	3,229
happy	46.7	79,830	54,585
neutral	26.0	33,672	41,202
sad	8.9	11,304	14,155
surprise	4.9	6,403	7,687

Balanced

gender	%	Female	Male
angry	8.7	6,961	6,961
contempt	1.3	1,049	1,049
disgust	1.3	1,063	1,063
fear	2.2	1,784	1,784
happy	46.7	37,604	37,604
neutral	26.0	20,946	20,946
sad	8.9	7,122	7,122
surprise	4.9	3,941	3,941

Biased

gender	%	Female	Male
angry	8.7	6,961	0
contempt	1.3	1,049	0
disgust	1.3	1,063	0
fear	2.2	1,784	0
happy	46.7	37,604	0
neutral	26.0	20,946	0
sad	8.9	7,122	0
surprise	4.9	3,941	0

Stratified 0.28

gender	%	Female	Male
angry	8.7	1,949	5,018
contempt	1.3	451	599
disgust	1.3	482	583
fear	2.2	882	904
happy	46.7	22,352	15,284
neutral	26.0	9,428	11,537
sad	8.9	3,165	3,963
surprise	4.9	1,793	2,152

Balanced 0.5

gender	%	Female	Male
angry	8.6	3,480	3,480
contempt	1.3	524	524
disgust	1.3	532	532
fear	2.2	892	892
happy	46.7	18,802	18,802
neutral	26.0	10,473	10,473
sad	8.9	3,561	3,561
surprise	4.9	1,970	1,970

RESULTS: INTUITIVE DATASET BIAS

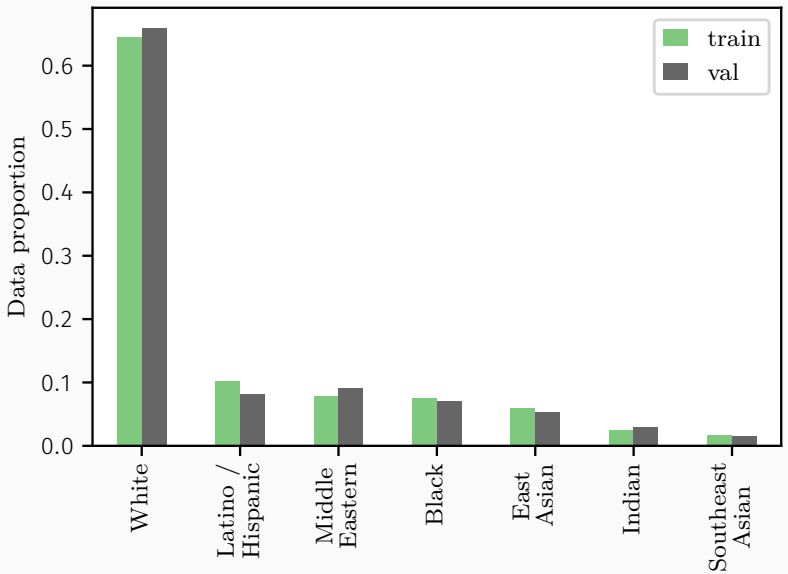


Figure 4: Apparent race distribution of Affectnet.

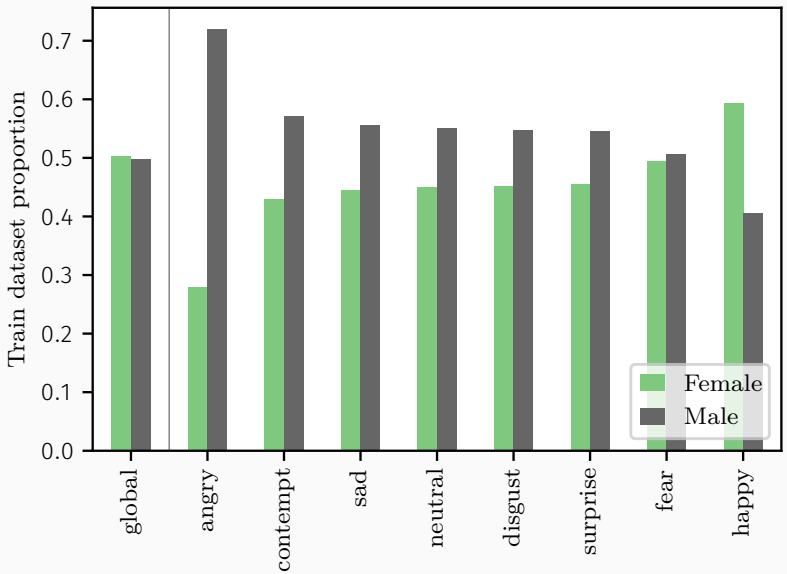


Figure 5: Apparent gender distribution of Affectnet for each label (train partition).

RESULTS: DATASET REPRESENTATIONAL BIAS

		Representational bias (NSD)	
Dataset		Race (7)	Gender (2)
Original		0.5902	0.0057
Balanced	Race	**0.0000**	0.0159
	Gender	0.5929	**0.0000**
Gender biased	M	0.5702	–
	F	0.6129	–

Table 1: Representational (NSD) bias metrics for the original dataset and the considered subsets.

RESULTS: DATASET STEREOTYPICAL BIAS

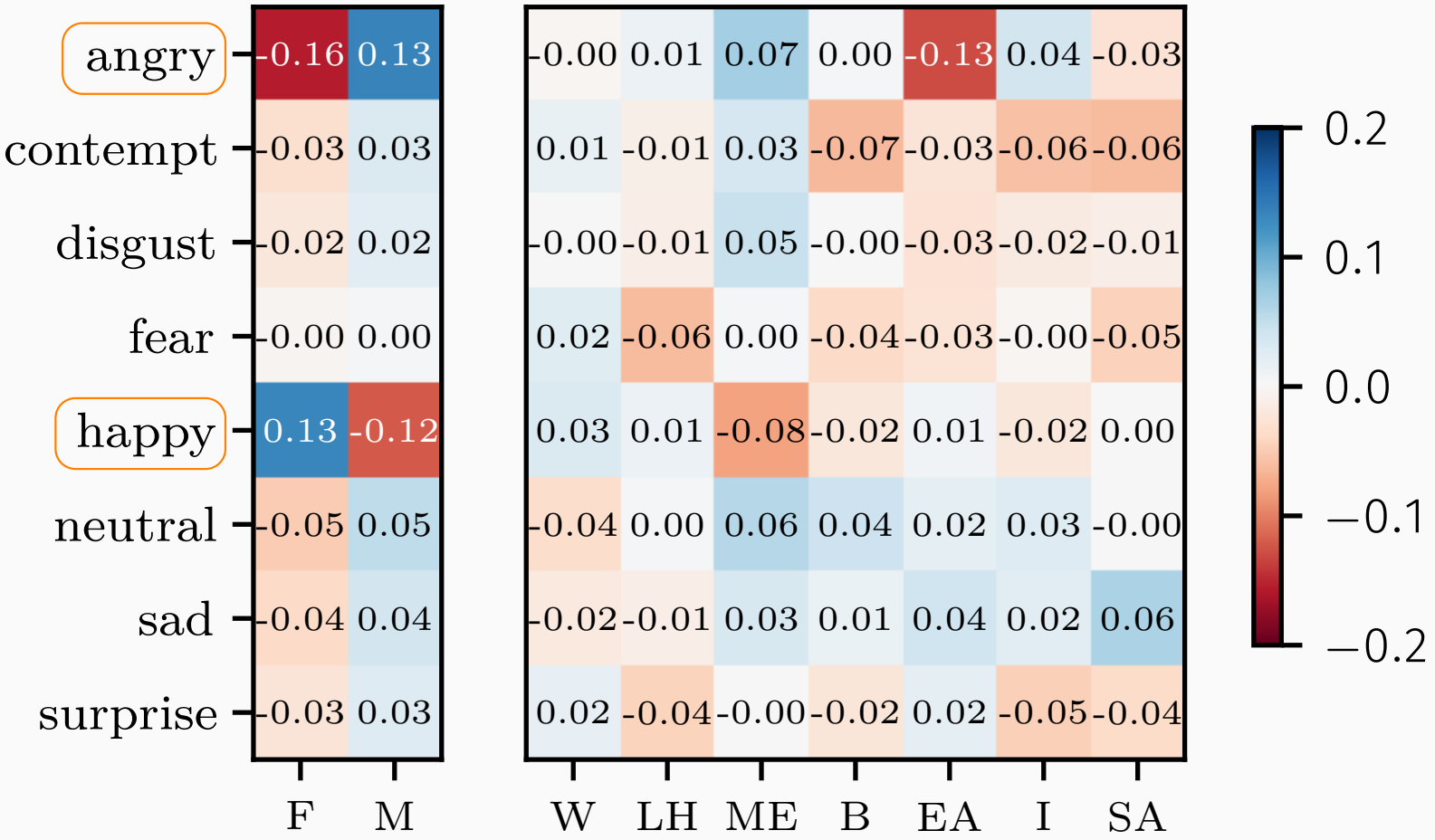


Figure 6: NPMI analysis of Affectnet. (F: Female, M: Male, W: White, LH: Latino / Hispanic, ME: Middle Eastern, B: Black, EA: East Asian, I: Indian, SA: Southeast Asian)

RESULTS: DATASET STEREOTYPICAL BIAS (II)

		Stereotypical bias (NMI)	
Dataset		Race (7)	Gender (2)
Original		0.0021	0.0089
Balanced	Race	**0.0000**	0.0091
	Gender	0.0017	**0.0000**
Gender biased	M	0.0020	–
	F	0.0020	–

Table 2: Stereotypical (NMI) bias metrics for the original dataset and the considered subsets.

RESULTS: MODEL BIAS

				Bias	
Train data		Size	Accuracy	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 3: Bias metric summary for the model when trained on dataset variations.

RESULTS: MODEL BIAS

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 3: Bias metric summary for the model when trained on dataset variations.

RESULTS: MODEL BIAS

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 3: Bias metric summary for the model when trained on dataset variations.

RESULTS: MODEL BIAS

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	**55.8 ± 0.2**	**0.268 ± 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	**0.091 ± 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 3: Bias metric summary for the model when trained on dataset variations.

CONCLUSION

- The metrics presented correlate well with the **intuitive bias**.
- The metrics allow the **study of the bias transfer** from dataset to trained model.
- Affectnet: **racial representational** bias and **gender stereotypical** bias.
  - Balancing the dataset only works sometimes.
- Future work:
  - Different **contexts**.
  - Different **mitigation** techniques.
  - Different **application areas** (multi-class, multi-demographic).

CONCLUSION

- The metrics presented correlate well with the **intuitive bias**.
- The metrics allow the **study of the bias transfer** from dataset to trained model.
- Affectnet: **racial representational** bias and **gender stereotypical** bias.
  - Balancing the dataset only works sometimes.
- Future work:
  - Different **contexts**.
  - Different **mitigation** techniques.
  - Different **application areas** (multi-class, multi-demographic).

CONCLUSION

- The metrics presented correlate well with the **intuitive bias**.
- The metrics allow the **study of the bias transfer** from dataset to trained model.
- Affectnet: **racial representational** bias and **gender stereotypical** bias.
  - Balancing the dataset only works sometimes.
- Future work:
  - Different **contexts**.
  - Different **mitigation** techniques.
  - Different **application areas** (multi-class, multi-demographic).

Questions?

✉ iris.dominguez@unavarra.es

https://github.com/irisdominguez/Dataset-Bias-Metrics