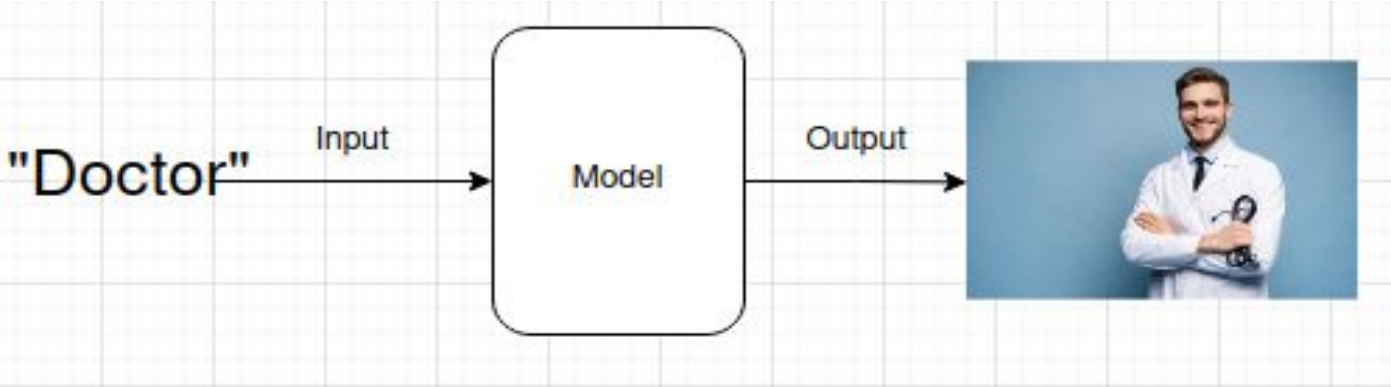
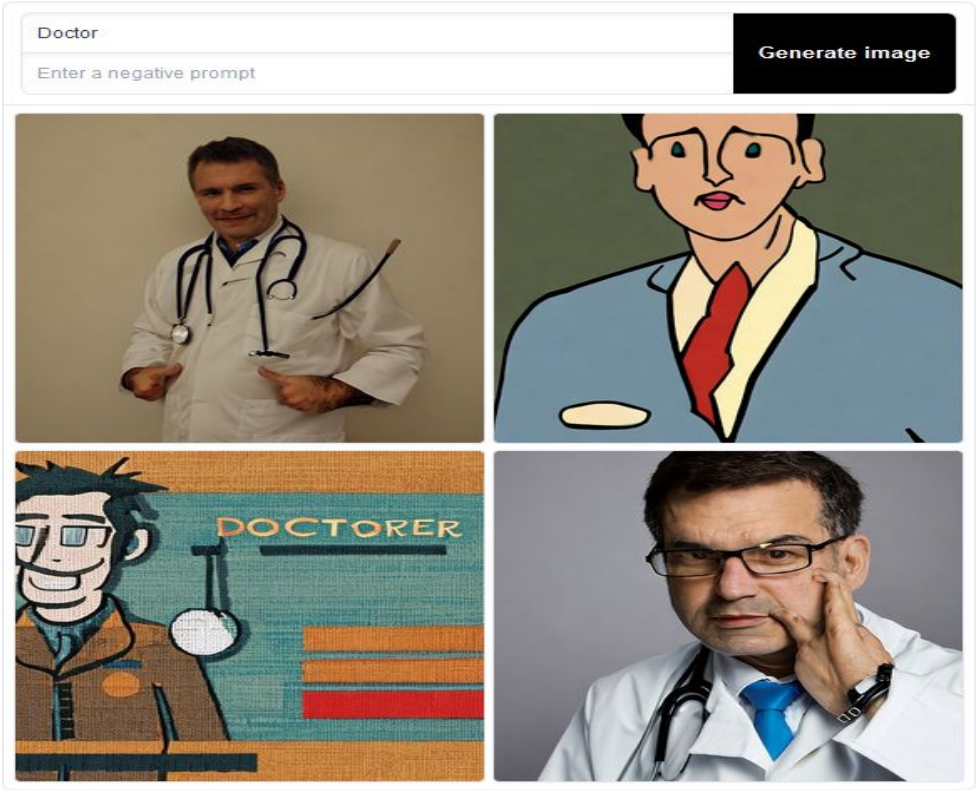
Bias Analysis In Large ML Training Datasets

Multi-model



Bias example:

https://huggingface.co/spaces/stabilityai/stable-diffusion



Bais Danger

- Amazon’s scrapped secret AI recruiting tool that showed bias against women.
- https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Laion high resolution dataset

- Freely available image-text dataset.
- Total number of rows: 174,704,931
- Number of columns: 9

	URL	TEXT	WIDTH	HEIGHT	similarity	LANGUAGE	hash	pwatermark	punsafe
0	https://mmedia.ozone.ru/multimedia/1018085734.JPG	Штора рулонная Эскар, цвет: абрикосовый, ширин...	1024	1200	0.275895	ru	-3448109611037364457	0.052757	0.001292

Штора рулонная Эскар, цвет: абрикосовый, ширина 140 см, высота 170 см81012140170Рулонными шторами Эскар можно оформлять окна как самостоятельно, так и использовать в комбинации с портьерами. Это поможет предотвратить выгорание дорогой ткани на солнце и соединит функционал рулонных с красотой навесных. Преимущества применения рулонных штор для пластиковых окон: - имеют прекрасный внешний вид: многообразие и фактурность материала изделия отлично смотрятся в любом интерьере;- многофункциональны: есть возможность подобрать шторы способные эффективно защитить комнату от солнца, при этом она не будет слишком темной;- есть возможность осуществить быстрый монтаж.ВНИМАНИЕ! Размеры ширины изделия указаны по ширине ткани! Во время эксплуатации не рекомендуется полностью разматывать рулон, чтобы не оторвать ткань от намоточного вала. В случае загрязнения поверхности ткани, чистку шторы проводят одним из способов, в зависимости от типа загрязнения:легкое поверхностное загрязнение можно удалить при помощи канцелярского ластика;чистка от пыли производится сухим методом при помощи пылесоса с мягкой щеткой-насадкой;для удаления пятна используйте мягкую губку с пенообразующим неагрессивным моющим средством или пятновыводитель на натуральной основе (нельзя применять растворители).

Problem

- Does image-text pair dataset with image URLs and short textual descriptions provides equal representation across age, gender, race, and emotions.

Frameworks for age, gender, race and emotion

- Framework Deepface
- Framework Democlassi

- Race: White, Black, Asian, Indian, Unknown-race.
- Emotion: Anger, Disgust, Fear, Happy, Sad, Surprise and Neutral.

Framework Deepface

- Based on Keras and TensorFlow.
- Uses VGG-Face model for age, gender, race and emotions.
- Last layer of the model replaced with a new layer having seven output nodes, one for each emotion: anger, disgust, fear, happy, sad, surprise and neutral.
- Age and gender classification trained using Adience dataset.
- Race classification trained using RaceFace dataset.
- Emotion classification trained using FER-2013 dataset.


{
"age": 28.66,
"emotion": "neutral",
"gender": "Woman",
"race": "latino hispanic"
}
{
"age": 29.27,
"emotion": "happy",
"gender": "Woman",
"race": "white"
}
{
"age": 29.27,
"emotion": "surprise",
"gender": "Woman",
"race": "white"
}
{
"age": 29.74,
"emotion": "neutral",
"gender": "Woman",
"race": "white"
}

Detectors in Deepface

Original Image

RetinaFace

MtCnn

Dlib

MediaPipe

Ssd

OpenCv

Framework Democlassi

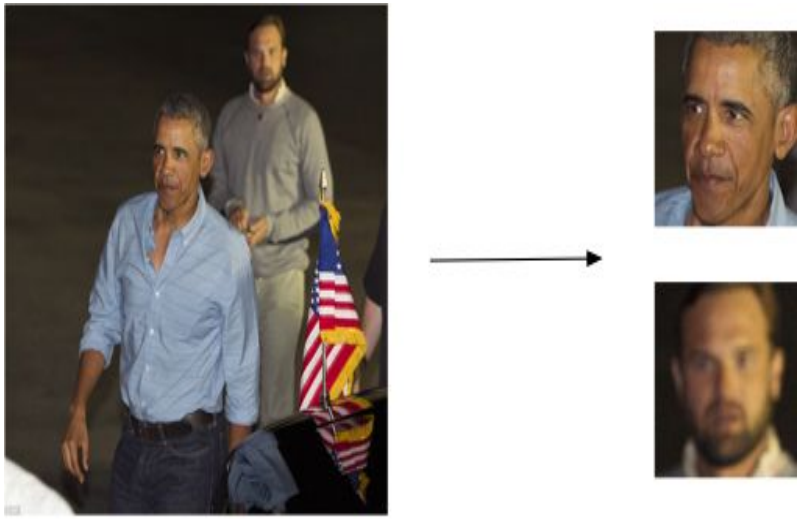
- Based on PyTorch and PyTorch-Ignite.
- This framework model provides three pre-trained models for age, gender, and race
  - Visual Geometry Group (VGGNet-19)
  - Separable convolution(SEPCONV)
  - Residual Network(ResNet-50)
- Only a single pre-trained model SEPCONV for emotion detection.
- All the models for age, gender and race were trained using the UTK Face dataset.
- Emotion model was trained using a FER-2013 dataset.

Limitation of the frameworks

- These frameworks have multi-face detection problem.

Framework Face_recognition

- Built using Dlib.
- Dlib open-source library written in C++, which provides a wide range of machine learning tools.



Input

Output

Questions

- Can face_recognition detect all people in different images?
- Which detector is best suited for Deepface ?
- Which pretrained model gives the best output for Democlassi?
- Are the framework biased ?

Benchmarking of frameworks

Benchmarking face_recognition

- A total of 170 images were downloaded from open-source websites such as Getty Images, Adobe Stocks and iStocks.
- The images that were used images contained a single face per image.
- These images included people from all the different races.

Face_recognition evaluation

Status	Number
Total images	170
Faces found	164
Double detected	12
Faces not detected	12

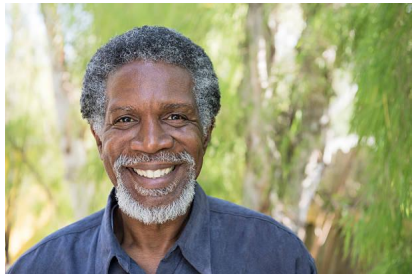
Figure 4.1: Double face detected. Left original image, right faces found

Benchmarking Deepface and Democlassi

- A total of 70 images were distributed among five races: white, black, Asian, Indian, and unknown-race (Latino and Middle Eastern).
- OpenCV detector was used for nd ResNet-50 was used for deepface and democlassi. (evaluation backup slides)

Benchmarking Deepface and Democlassi : Example

- 14 images in the 70-image set that belong to the black race.
- 7 images for black males.
- 7 images for black females.
- There are 7 images for black males and black females depicting the emotions: anger, disgust, fear, happiness, sadness, surprise, and neutrality.



Age

- The models produced consistent output ranging from 24 to 45, regardless if the image belonged to a child or an older person.

Gender

Gender	Number	Percentage (%)
Male (years)	35	50
Female (years)	35	50

Table 5.2: Gender estimation after manual analysis

Gender	Democlassi(ResNet-50)	Deepface(OpenCV)
Male (%)	57.14	72.86
Female (%)	42.86	21.14
Time taken (sec)	4	64

Table 5.18: Automatic analysis on Gender

Race

Race	Number	Percentage (%)
White	14	20
Black	14	20
Asian	14	20
Indian	14	20
Unknown	14	20

Table 5.3: Race estimation after manual analysis

Race	Democlassi(ResNet-50)	Deepface(OpenCV)
White (%)	21.43	28.57
Black (%)	40.00	15.71
Asian (%)	38.57	25.71
Indian (%)	0.00	2.86
Unknown (%)	0.00	27.14

Table 5.19: Automatic analysis on Race

Emotion

Emotion	Expected	Democlassi(ResNet-50)	Deepface(OpenCV)
Neutral (%)	14.29	17.14	41.43
Disgusted (%)	14.29	0.00	4.29
Fear (%)	14.29	20.00	12.86
Happy (%)	14.29	10.00	15.71
Sad (%)	14.29	32.86	14.28
Surprise (%)	14.29	0.00	0.00
Neutral (%)	14.29	20.00	11.43

Table 5.19: Automatic analysis on Emotion

Benchmarking Conclusion

- Both frameworks are baised.

Image extraction problem

- Around 175 million images and limited resources.

Solution: Joblist

- List of 60 occupation.
- Example: kindergarten teacher, dental hygienist, speech-language pathologist, dental assistant, childcare worker, medical records , technician, secretary

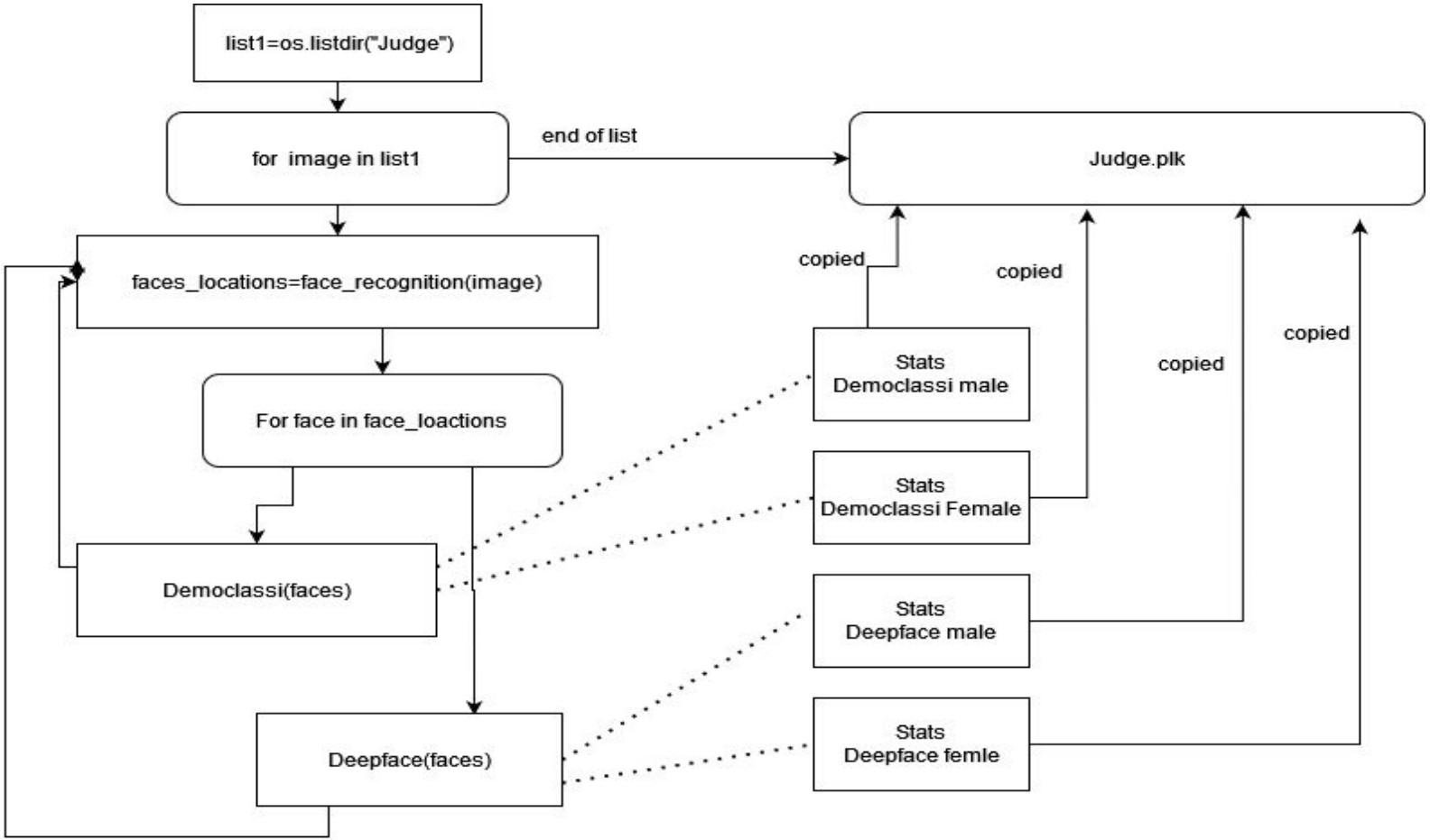
https://github.com/marionbartl/gender-bias-BERT/blob/master/data/Professions%20US%2BDE.tsv

Image extraction with joblist

- Check each row for the presence of words from the list in the text and extract the corresponding image if a match is found. (Backup slide)

Image analysis

- Both frameworks evaluates the images sequentially to compute the runtime.


list1=os.listdir("Judge")
for image in list1
end of list
Judge.plk
faces_locations=face_recognition(image)
copied
copied
copied
copied
Stats
Democlassi male
For face in face_loactions
Stats
Democlassi Female
Democlassi(faces)
Stats
Deepface male
Stats
Deepface femle
Deepface(faces)

Joblist on Laion high resolution

Jobs from Joblist

- 60 occupation in Joblist.
- During analysis, it was discovered that 18 professions had no relevant images for analysis.
- In some cases, images were extracted for these professions but these images only contained text.
- Occupations with less than 200 relevant images were excluded from further analysis, as they did not have enough data to draw meaningful conclusions.

Jobs from Joblist

- 121K images were analyzed by both the frameworks.
- 20 occupation from Joblist.

Jobs	Total	Democlassi evaluation	Democlassi Male (in%)	Democlassi Female (in %)	Deepface evaluation	Deepface Male (in%)	Deepface Female (in %)
Bartender	5,801	2,699	68.06	31.94	2,699	83.18	16.82
Book-keeper	643	200	42.5	57.5	200	70.00	30.00
Carpenter	15,040	7,800	56.99	43.01	7,800	75.58	24.42
Conductor	12,522	4,446	68.26	31.74	4,446	86.59	13.41
Dietitian	1,342	544	21.86	78.12	543	59.11	40.88
Dispatcher	678	277	58.48	41.52	277	78.34	21.66
Electrician	5,742	829	79.73	20.27	829	89.62	10.37
Firefighter	23,504	10,100	76.96	23.03	10,100	90.67	9.33
Hairdresser	3,953	1,594	36.83	63.17	1,594	60.85	39.15
Housekeeper	736	285	51.93	48.07	285	77.54	22.46
Judge	27,753	30,014	65.25	34.75	30,014	83.26	16.74
Lifeguard	1,806	1,235	62.83	37.17	1,235	82.83	17.16
Mason	62,330	18,223	59.47	40.53	18,223	81.93	18.07
Paralegal	871	264	35.98	64.01	264	56.44	43.56
Plumber	4,426	847	81.22	18.77	847	92.44	7.55
Receptionist	2,331	837	35.01	64.99	837	59.02	40.98
Registered nurse	618	222	41.44	58.56	222	76.58	23.42
Secretary	27,383	37,230	71.74	28.25	37,230	87.33	12.67
Taper	44,458	2,788	54.81	45.19	2,787	77.04	22.96
Kindergarten teacher	438	204	36.27	63.73	204	72.55	27.45

Table 1: Jobs with more than 200 images

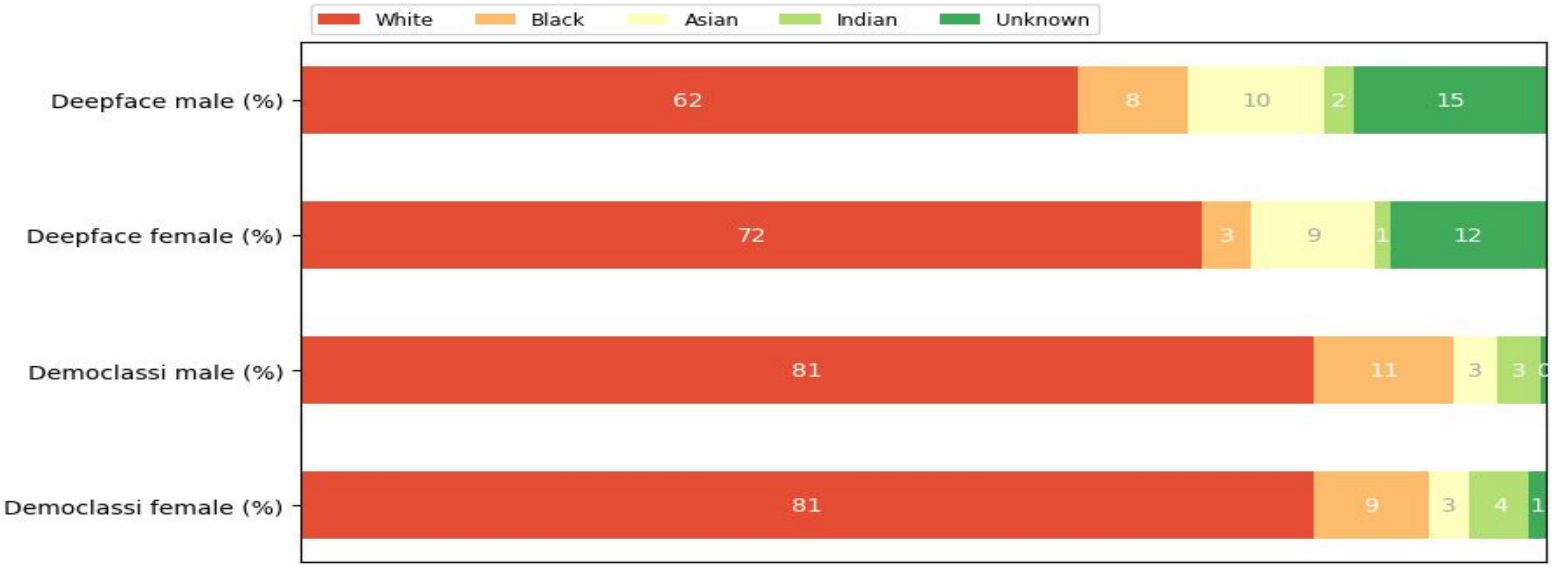
Male dominated jobs

- Plumber (Democlassi: 81.22% male, 18.77% female, Deepface: 92.44% male, 7.55% female)
- Firefighter (Democlassi: 76.96% male, 23.03% female, Deepface: 90.67% male, 9.33% female)
- Electrician (Democlassi: 79.73% male, 20.27% female, Deepface: 89.62% male, 10.37% female)

Female dominated jobs

- Dietitian (Democlassi: 21.86% male, 78.12% female,
  Deepface: 59.11% male, 40.88% female)
- Receptionist (Democlassi: 35.01% male, 64.99% female,
  Deepface: 59.02% male, 40.98% female)
- Paralegal (Democlassi: 35.98% male, 64.01% female,
  Deepface: 56.44% male, 43.56% female)

Race


White
Black
Asian
Indian
Unknown
Deepface male (%)
62
8
10
2
15
Deepface female (%)
72
3
9
1
12
Democlassi male (%)
81
11
3
3
Democlassi female (%)
81
9
3
4
1

Runtime

- Democlassi took 262996 seconds (approximately 73 hours) to complete.
- Deepface took 346837 seconds (approximately 96 hours) to complete the same task.

Conclusion

- Frameworks were biased.
- Joblist contains more male dominated occupation than females.
- Dataset was also found to have bias.

THANK YOU

Backup Slides:

Democlassi evaluation

Democlassi evaluation: Age

Gender	Minimum age	Maximum age
Male (years)	23.9	35.40
Female (years)	26.23	39.51

Table 5.1: Estimated age after manual analysis

	ResNet-50	VGGNET-19	SEPCONV
Male	31.5	28.98	28.7
Female	33.7	29	27.72

Table 5.6: Age distribution by different model in Democlassi

Age Problem

- The models produced consistent output ranging from 24 to 35, regardless if the image belonged to a child or an older person.

Democlassi evaluation: Gender

Gender	Number	Percentage (%)
Male (years)	35	50
Female (years)	35	50

Table 5.2: Gender estimation after manual analysis

Model	ResNet-50	VGGNET-19	SEPCONV
No. of Male (%)	57.14	42.86	81.43
No. of Female (%)	42.86	57.14	18.27
Time taken (sec)	4	5	3

Table 5.7: Gender distribution automatic analysis Democlassi.

Democlassi evaluation: Race

Race	Number	Percentage (%)
White	14	20
Black	14	20
Asian	14	20
Indian	14	20
Unknown	14	20

Table 5.3: Race estimation after manual analysis

Races	ResNet-50	VGGNET-19	SEPCONV
White in (%)	21.43	87.14	55.71
Black in (%)	40.0	12.86	12.86
Asian in (%)	38.57	0.0	31.43
Indian in (%)	0.0	0.0	0.0
Unknown in (%)	0.0	0.0	0.0

Table 5.8: Automatic race prediction by different models in Democlassi

Democlassi evaluation: Emotion

- Only one model (SEPCONV) available for emotion

Emotion	Expected Value	Actual Value
Neutral	14.29%	17.14%
Disgusted	14.29%	0.00%
Fear	14.29%	20.00%
Happy	14.29%	10.00%
Sad	14.29%	32.86%
Surprise	14.29%	0.00%
Angry	14.29%	20.00%

Table 5.9: Automatic emotion distribution in Democlassi

Choosing best Democlassi model

- ResNet-50 was chosen for age, gender and race.
- Better performance in classifying race and gender.

Deepface evaluation

Deepface evaluation: Age

Gender	Minimum age	Maximum age
Male (years)	23.9	35.40
Female (years)	26.23	39.51

Table 5.1: Estimated age after manual analysis

Gender	OpenCV	SSD	Dlib	MCTNN	RatinaFace
Males in yrs	30.59	34.42	30.61	34.68	34.52
Females in yrs	31.32	32.5	30.25	31.25	31.8

Table 5.10: Age distribution using different detectors in Deepface

Deepface age problem

- The models generated age estimates for the majority of the images, which fell between the ages of 24 and 45.

Deepface evaluation: Gender

Gender	Number	Percentage (%)
Male (years)	35	50
Female (years)	35	50

Table 5.2: Gender estimation after manual analysis

	OpenCV	SSD	Dlib	MCTNN	RatinaFace
Male (in %)	72.86	74.29	77.14	71.43	71.43
Female (in %)	27.14	25.71	22.85	28.57	28.57
Time (sec)	64	90	68	226	186

Table 5.11: Gender distribution using different detectors in Deepface

Deepface evaluation: Race

Race	Number	Percentage (%)
White	14	20
Black	14	20
Asian	14	20
Indian	14	20
Unknown	14	20

Table 5.3: Race estimation after manual analysis

Race	OpenCV	SSD	Dlib	MCTNN	RatinaFace
White in (%)	28.57	37.14	28.57	25.71	28.57
Black in (%)	15.71	17.14	17.14	17.14	14.28
Asian in (%)	25.71	25.71	22.86	28.57	27.14
Indian in (%)	2.86	0.0	7.14	4.29	1.43
Unknown in (%)	27.14	20.00	24.29	24.29	28.57

Table 5.12: Race prediction using different detectors in Deepface

Deepface evaluation: Emotion

Emotion	Number	Percentage (%)
Angry	10	14.29
Fear	10	14.29
Neutral	10	14.29
Happy	10	14.29
Sad	10	14.29
Surprise	10	14.29
Neutral	10	14.29

Table 5.4: Emotion estimation after manual analysis

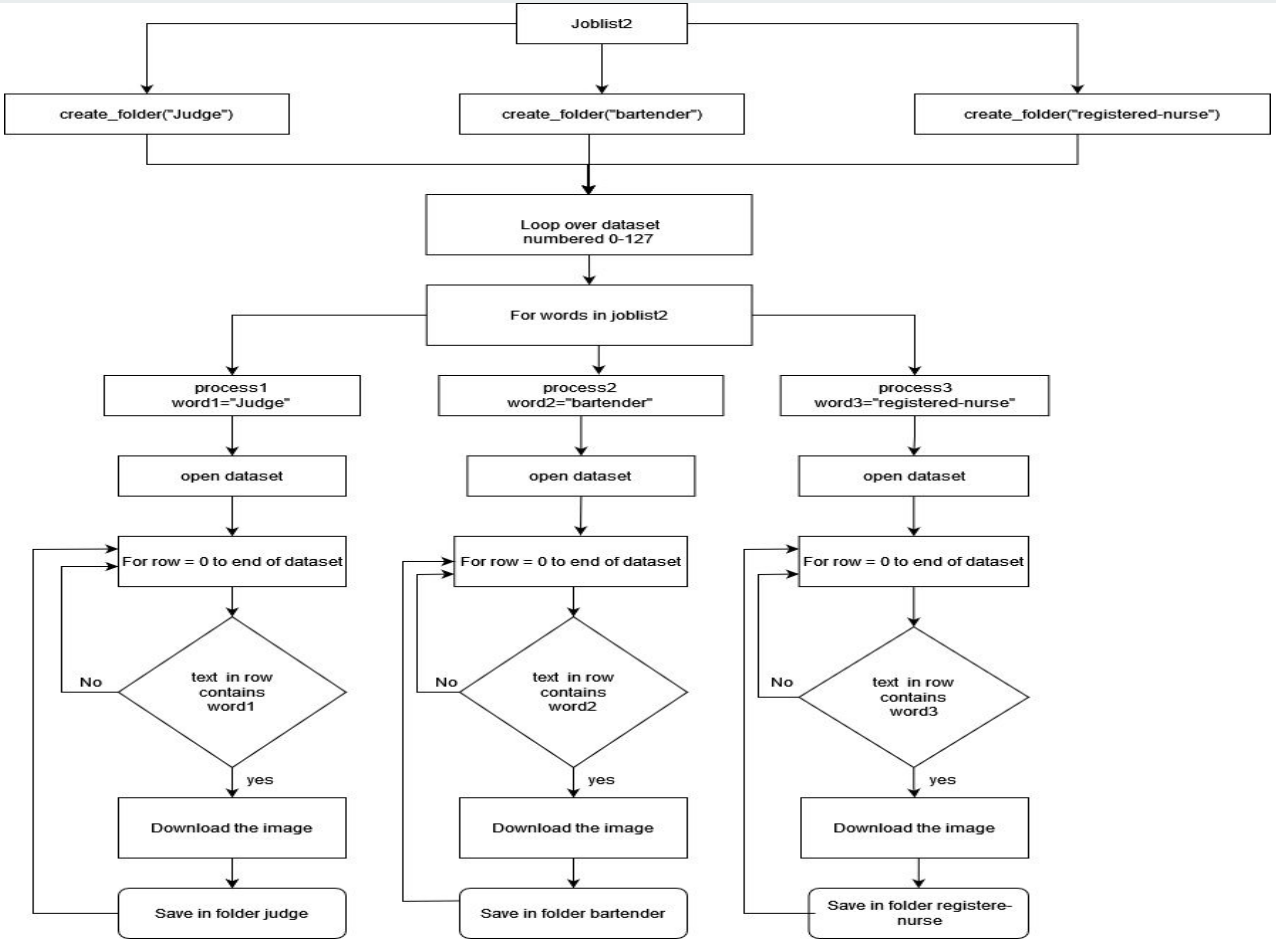
Emotion	OpenCV	SSD	Dlib	MCTNN	RatinaFace
Neutral in (%)	41.43	40.00	47.14	45.71	41.43
Disgusted in (%)	4.29	1.43	0.00	0.00	1.43
Fear in (%)	12.86	7.14	5.71	11.43	8.57
Happy in (%)	15.71	20.00	18.57	22.86	20.00
Sad in (%)	14.28	8.57	10.0	1.43	7.14
Surprise in (%)	0.00	0.00	0.00	0.00	0.00
Angry in (%)	11.43	22.86	18.57	18.57	21.43

Table 5.13: Automatic emotion analysis by Deepface

Choosing the Deepface detector

- Due to its small runtime, the OpenCV detector was utilized for further analysis.

Image Extraction flowchart


Joblist2
create_folder("Judge")
create_folder("bartender")
create_folder("registered-nurse")
Loop over dataset numbered 0-127
For words in joblist2
process1 word1="Judge"
process2 word2="bartender"
process3 word3="registered-nurse"
open dataset
open dataset
open dataset
For row = 0 to end of dataset
For row = 0 to end of dataset
For row = 0 to end of dataset
No
text in row contains word1
No
text in row contains word2
No
text in row contains word3
yes
yes
yes
Download the image
Download the image
Download the image
Save in folder judge
Save in folder bartender
Save in folder registere-nurse

Future works

- To improve the accuracy of the results, a new joblist that provides equal representation of images for males and females should be used.
- It is recommended to either analyze all the images in the dataset or use frameworks with low biases to better understand the overrepresentation of white males.
- Researching new alternative methods to detect emotions that are less biased would be beneficial.