Race, Faces, Fairness:
Gender & Race Disparity in Commercial Face Recognition

Adam Munawar Rahman.
Wesleyan University
16th May 2019
QAC 239 Final Report

Abstract

Recent literature concerning artificial intelligence demonstrate that machine learning algorithms have race and gender discrimination. In this work, we conduct a status quo check by evaluating two commercial gender classification systems - Amazon Rekognition and Sightengine - using the large-scale UTKFace Dataset to ascertain intersectional gender and race disparities in these systems, i.e. if these systems can accurately identify the perceived gender of a labelled Asian, Black, or White face. Our results show that both of these classification systems exhibit bias towards non-white female faces, assigning disproportionately higher male confidence values. We also observe frequent errors in both classifiers identifying non-white male faces, with either higher-than-normal female confidence values assigned or the classifier not even being able to identify a face. By auditing these two classifiers with an intersectional approach, we hope to pinpoint precise flaws and identify what ought to be corrected in machine learning algorithms in the future.

1. Introduction

Artificial intelligence is scaling with society's demands at a rapid pace. Beyond the mainstream and widely recognized implementations of AI in self-driving vehicles and speech recognition, machine learning algorithms are also deployed in sustaining the stock market, generating written content for publications, and are even incorporated in criminal justice pipelines (Citron and Pascale, 2014) [2]. AI services such as Amazon Rekognition, for example, are being sold to police departments to assist in identifying suspects. Another service, Sightengine, brands itself as ”The Leading Image and Video Moderation API” - is used to detect and filter undesirable content in photos, videos, and livestreams.

However, these and other third-party services - such as Microsoft Face API and IBM Watson - have demonstrated a degree of race and gender bias. The American Civil Liberties Union recently conducted a test where they found Amazon's Rekognition to misidentify 28 members of Congress as criminals, with most of the false matches being disproportionately people of color (Snow, 2018) [3]. With Amazon having made a sales pitch to ICE with their Rekognition software, the misuse of artificial intelligence is more profoundly relevant than ever, as the harmful . For scientists who are cognizant of social injustices, testing and auditing these classification systems becomes a moral obligation to ensure that as we move towards an AI-driven future, it does not further reinforce pre-existing structural oppression. By analyzing popular services Rekognition and Sightengine,

2. Related Work

Previous research has demonstrated how machine learning models have encoded social bias - a study by Bolukbasi et al.[4] how Word2vec, a group of models used to produce word embeddings and are trained to determined linguistic context, displays gender bias when they trained an analogy generator. The model output an analogy associating men with programming and woman with homemaking - demonstrating the social biases that have been integrated into Word2vec as a result of poor data curation. As more and more software is built upon these models, these biases may propagate, resulting in these pre-existing biases being reinforced. This research specifically shows bias in natural language processing algorithms, but less work has been performed on computer vision APIs in particular - and the implications of prejudiced computer vision algorithms are particularly disastrous, as previously discussed.

This work is inspired by and serves to extend the research conducted by Buolamwini et al. [1], whose study ”Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification” brought to light the severe racial disparities in three commercial classification systems - Microsoft Face API, Face++, and IBM Watson. They constructed a new dataset titled the ”Pilot Parliaments Benchmark” dataset which incorporated faces of representatives from European and African countries and was labelled based on skin color using the Fitzpatrick dermatologist developed system. The results of their study determined that among the three classification systems, darker-skinned female faces were the most misclassification - with error rates up to 34.7%, while the maximum error rate for lighter-skinned males was up to 0.8%. Buolamwini et al.'s work has strong social justice implications, as it fuses engineering and data science with a call for accountability and transparency in the services that drive tech industries.

3. Method

3.1. Dataset

We began by retrieving the UTKFace Dataset, the most relevant highlights of which include:

- 20k+ face images in the wild (only single face in one image)
- Provides the corresponding aligned, cropped faces
- Images are labelled by age, gender, and ethnicity

Figure 1. Samples from the UTKFace Dataset

For the purposes of this work, we refer to the labels used by this dataset i.e. a face's gender is labelled as either male or female, and we select the race labels of White, Black, or Asian. We recognize that race and gender are complex spectrums, and understand that the labels used in this work are that of *percevied* race and gender. Adhering to these labels allows us to isolate the biases present in the commercial classifiers, and this dataset in particular not only provided conveniently prelabeled images, but ones that were aligned and cropped so that the facial features are centered in the frame, allowing for minimal noise to be fed into the classifiers. Our selection of the UTKFace Dataset allows us to test faces distinct from that of the Pilot Parliaments Benchmark Dataset from Buolamwini et al., as UTKFace consists of 'faces in the wild' i.e. images of common folk obtained from a variety of unconstained sources, and not necessarily from a preexisting database. This means we can audit the classifiers for more general use cases, and evaluate how they perform.

From the dataset, we extract all images with ages labelled between 18 - 36 - the range of 'young adult' - we do this to trim the dataset to a reasonable degree while having a collection of faces that display a large amount of sexual dimorphism i.e. it is easier to perceive the gender differences between young adults than between children. As we are restricted by the total number of queries we can make to each commercial classifier before hitting the API usage limit, we collected 83 female and 83 male labelled faces for Asian, Black, and White labelled faces, for a count of 166 faces for each race and 498 faces total in our dataset. Figures 1, 2, and 3 show samples of the faces collected; each image is 200 by 200 pixels and the faces are centered in each frame, allowing a degree of uniformity throughout our dataset. Now that we have successfully compiled a reasonably diverse set of images as lists of paths, we are prepared to feed them into the classifiers we are evaluating.

Figure 2. Selection of Asian labelled faces

Figure 3. Selection of Black labelled faces

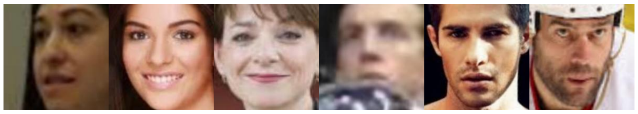

Figure 4. Selection of White labelled faces

3.2. Classification & Processing

After setting up credentials with both Sightengine and Amazon Rekognition APIs, we are able to create client objects for each API and feed it the paths to the images. We initialize three empty dataframes for each racial group we will be analyzing - Asian, Black, and White. For each face in the collection, we:

- Tag it with its original UTKFace gender label
- Send a request to the API and retrieve the response
- Extract the gender label and confidence values from the response

- Convert the gender label to 0 (Male), 1 (Female), or 2 (undetermined), to match the UTKFace label (we add the additional 'undetermined' label to account for classification failure, since there are no unlabelled faces in the dataset)
- Scale the confidence value by 100 (e.g. Sightengine may return 0.9, which becomes 90.0)
- Append the label and confidence values to the entry for that face in the dataframe

The Rekognition classifier returns a gender label (either 'Male' or 'Female') and confidence probabilities for either one, in the form of high-precision floating-point values. The Sightengine classifier only returns gender confidence values as low-precision floating-point numbers, so the gender label is determined by computing whichever confidence value is greater than 50%. Even if the request made to the API is a success, if the classifier is unable to detect a face, the response will not return any confidence values - in this case the data is read as an 'undetermined' label and the confidence values are set to the floating-point not-a-number value in Python so they do not contribute to the mean confidence values. These events count towards the classifier 'failure' rate, while a mismatch between the tagged gender and the classified gender is counted towards the classifier 'error' rate. This procedure is repeated for all three racial groups with both classifiers, and dataframes for each racial group are constructed for processing.

After building dataframes for each racial group, we manipulate each dataframe to obtain the following datapoints per race:

- Rekognition error count - the number of misidentifications made by the Rekognition classifier
- Sightengine error count - the number of misidentifications made by the Sightengine classifier
- Rekognition failure count - the number of times the Rekognition classifier is unable to recognize a face
- Sightengine failure count - the number of times the Sightengine classifier is unable to recognize a face
- Mean Rekognition Male Confidence - the mean of all male confidence values assigned by Rekognition
- Mean Rekognition Female Confidence - the mean of all female confidence values assigned by Rekognition
- Mean Sightengine Male Confidence - the mean of all male confidence values assigned by Sightengine
- Mean Sightengine Female Confidence - the mean of all female confidence values assigned by Sightengine

To emphasize our intersectional approach, we subdivide the error/failure count data into male and female error/failure count data. This allows us to more precisely identify flaws in the classification - for example, if a classifier has many mismatches with a particular race, it could be that it is having trouble identifying female faces of that race in particular, while successfully identifying male faces. The male/female confidence data is also subdivided by gender, so that we may identify a classifier's tendency to assign 'male' attributes to female faces of a specific race, or vice versa. For example, a classifier may successfully label a collection of black female faces, but the male confidence value assigned to these faces may be higher than those assigned to white female faces, which indicates that the classifier lacks appropriate training data for black women.

4. Results

Our computed results are shown in the corresponding tables on the last page. Table 1 displays the classification results from Amazon Rekognition, broken down by race and gender. Table 2 displays the classification results from Sightengine, broken down by race and gender. Table 3 displays a comparison between the two classification systems, broken down by Asian and Black faces. Table 4 displays a comparison between the two classification systems, broken down by White vs. non-white faces. Table 5 displays a comparison between the two classification systems, broken down by gender.

4.1. Discussion

We observe from Table 1 that despite Amazon Rekognition successfully being able to identify a face from all images in our dataset, it has a highest error percentage when labelling black male faces - at 13.6%. We also observe a high error percentage with Asian male faces. Between Asian male and female faces, we observe a relatively high female confidence value - around 16% - which is comparable to the 15.1% male confidence value for black female faces. The 0% error rate for white female versus the 12% error rate for white male faces is curious, and we acknowledge this as not necessarily Amazon Rekognition's algorithm being biased *against* white male faces - which is dubious - but rather the particularities of our dataset, which may be adjusted for if the sample size of our dataset were increased.

From Table 2, we see that Sightengine's classifier exhibits more apparent bias - with an error rate of 10% for black female faces - closer to Buolamwini et al.'s value of 13%. We also observe similarly to Rekognition that the mean confidence value assigned to black female faces is around 15%, while the male confidence values for Asian and White faces is smaller, around 6% and 11% respectively. Sightengine has the second-highest failure rate with Asian male faces, and has similar failure rates at around 2-4% for all faces except white male faces, which is less than 1%.

Tables 3, 4, and 5 offer a breakdown specifically only by race or gender. We observe from Table 5 that while the male error rate for Rekognition is higher, the female failure rate for Sightengine is greater - this implies that while Rekognition is able to at least classify all faces, its training data assigns disproportionate confidence values. On the other hand, Sightengine is not even able to classify all types of faces, and is disproportionately biased against female faces - and as we observe from Table 2, this is affected mostly by black female faces.

Although we are able to discern algorithmic bias from the limited amount of queries we could make to the API and our dataset of 498 images, more patterns may have emerged if we were able to incorporate more input images. We also relied heavily on the dataset's racial labels, as opposed to Buolamwini et al.'s dataset which was based on a unique skin-color classification, which may have generated more nuanced results with the tradeoff being more time needed to prepare a dataset.

5. Conclusion

Machine learning services are touted for their objective and accurate reliability as components of software engineering pipelines. Entire platform infrastructures are built with Amazon's services, and numerous organizations deploy Sightengine for content moderation. With how ubiquitious these services are, they ought to be more objective and uniform - and yet, as our results demonstrate, they demonstrate consistent bias and - in the case of Sightengine specifically - can even fail to recognize a human face, which is disproportionate across race and gender lines.

The bias exhibited by these classification services is a consequence of systemic prejudice, of a lack of representation in our data, and a neglect towards the ramifications of these flaws. As artifical intelligence becomes more woven into the threads of society, the consequences of poor diversity will become more apparent as systemic prejudice becomes reinforced - whether computer vision algorithms mislabel marginalized persons as criminals, or self-driving cars fail to identify humans on walkways, or content moderation systems disproportionately mark certain content as 'inappropriate' which further reduces diversity. We observe from our results that specific societal prejudices - such as the dehumanization of black men, the effeminization of asian men, and the masculinization of black women - end up being encoded in the behavior of these classifiers due to the type of content that is used to train them.

More research on commercial machine learning systems ought to be conducted in the future, in order to keep large companies in check and maintain the transparency and accountability of technology. It is empowering to know that we as scientists have access to a variety of tools that permit us to scrutinize and isolate the biases inherent in our technologies, and develop the methods to fix it - one might use a variant of the UTKFace Dataset, the Pilot Parliaments Benchmark Dataset, or a brand new set of faces to craft their own just, accurate classifier that exhibits less bias, and use it to compete with large commercial services. Artificial intelligence promises a bright future, and it ought to be one that is welcoming to all shades.

References

[1] J. Buolamwini and T. Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research*, 15(1):1–15, 2018. 1

[2] D. K. Citron and F. A. Pasquale. The scored society: due process for automated predictions. 2014. 1

[3] J. Snow. Amazon's face recognition falsely matched 28 members of congress with mugshots. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28, 2018. 1

[4] J. Y. Z. Tolga Bolukbasi, Kai-Wei Chang. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in Neural Information Processing Systems*, 29(1):4349–4357, 2016. 1

Race	Gender	Error (%)	Failure (%)	Male Confidence	Female Confidence
Asian	Female	1.204819	0.0	5.822314	94.177686
Asian	Male	6.024096	0.0	83.719559	16.280441
Black	Female	0.401606	0.0	15.102459	84.897541
Black	Male	13.654618	0.0	79.848995	20.151005
White	Female	0.000000	0.0	11.669456	88.330544
White	Male	12.048193	0.0	79.605527	20.394473

Table 1. Amazon Rekognition Results

Race	Gender	Error (%)	Failure (%)	Male Confidence	Female Confidence
Asian	Female	3.614458	2.811245	5.822314	94.177686
Asian	Male	7.228916	2.008032	90.860656	9.139344
Black	Female	10.040161	2.409639	15.102459	84.897541
Black	Male	2.008032	2.008032	93.557377	6.442623
White	Female	3.614458	4.417671	11.669456	88.330544
White	Male	6.425703	0.803213	86.579592	13.420408

Table 2. Sightengine Results

Classifier	Asian Error (%)	Asian Failure (%)	Black Error (%)	Black Failure (%)
Rekognition	3.614458	0.000000	7.028112	0.000000
Sightengine	5.421687	2.409639	6.024096	2.208835

Table 3. Classifier Performance, Race (Asian & Black)

Classifier	White Error (%)	White Failure (%)	Non-white Error (%)	Non-white Failure (%)
Rekognition	6.024096	0.000000	5.321285	0.000000
Sightengine	5.020080	2.610442	5.722892	2.309237

Table 4. Classifier Performance, Race (Non-white vs. White)

Classifier	Male Error Rate	Female Error Rate	Male Failure Rate	Female Failure Rate
Rekognition	10.575636	0.535475	0.000000	0.000000
SightEngine	5.220884	5.756359	1.606426	3.212851

Table 5. Classifier Performance, Gender