A Study on Gender Bias in Image Datasets and Language Processing

Aparna Nair
University of Southern California, Los Angeles, California 90089, USA
(Dated: May 3, 2021)

In this paper we discuss how machine learning models can be biased towards certain minorities. We conduct two studies, In the first study we investigate the effectiveness of gender identification models and their implications in today's world. In the second study we demonstrate how word embeddings can be biased against women and discuss on ways to mitigate the same.

I. INTRODUCTION

The surge of Data and Deep Learning Techniques have lead to proliferation of Image recognition systems and Language models being used for a multitude of applications. Image recognition techniques have improved dramatically and as of April 2020, the best image identification model only has an error rate of 0.08%. [8] These performance improvements create incentives for widespread deployment of these algorithms.

Nowadays Facial Recognition systems are being used for high stakes tasks such as criminal profiling, security screening etc. The high accuracy achieved by these models are being used to justify it's widespread deployment, but it is essential to note that this accuracy is only achieved under ideal conditions.The model and dataset being used often present it's own set of biases [23] [17].

The first study being conducted in this paper involves navigation of different image datasets and analyzing corresponding gender bias present. Image Recognition systems are now commercialized and being deployed for a range of different applications. Research conducted shows that the algorithmic classification systems often reproduce and amplify the existing societal bias present [12]. In healthcare systems, due to inadequacies of medical imaging datasets, women are often miss classified[15]. Hence image datasets and classification systems need to be examined for biases.

The Second Study conducted in this paper explores gender bias present in word embeddings. Word Embeddings are trained on Natural Language Processing data and they represent text as vectors. A study published by researchers from Princeton University revealed many instances of discriminatory associations between words, predominantly female names were more closely associated to familial terminology while male names were more closely associated with careers [7]. NLP models are increasingly being used for applications like Resume Screening and such biases in model leads to minority groups being marginalized and discriminated against.An example for this was found in the software employed by Amazon for recruitment and resume screening. Since historically, men occupied higher number of positions in technological roles, the screening process was biased against women. The software tool learnt to penalize applications containing the words 'women' and different iterations of it[9]. Hence it is essential to evaluate how the existing word embeddings system's operate and when they should be deployed.

II. MODEL FAIRNESS

In this section we define what factors contribute to an algorithm producing discriminatory results.
The model results can be skewed and discriminatory against minority groups due to the presence of diverse types of bias. In the coming section we discuss types of bias that can affect the performance of the model.

Bias raises concerns over efficiency and optimization of the model's performance but learned bias can cause greater harm when it's results involve humans and adversely affect them. Biases formed by the model often resemble human like biases towards race, sex, religion, and many other common forms of discrimination.

A. Data Bias

1. *Selection Bias*

The duration and role of training varies across different machine learning applications and during this phase the model forms learned biases. One of the prime reasons behind this occurrence is due to the presence of inadequate data and data not being adequately prepared for the algorithm. If the data is not representative of the target population, training can directly create harmful learned biases biases[19]. Hence in this case the dataset might over-represent or under-represent a certain group. Microsoft's emotion recognition technology faced similar issues when it failed to accurately detect emotions in children, the elderly and minorities due to lack of representative training data [13].

2. *Latent Bias*

When an algorithm missclassifies or incorrectly identifies instances based on biased historical data, it exhibits latent bias. Hence in this case the algorithm acts to exacerbate existing preconceived notions and discriminates historically under-represented groups[5].
A study published by researchers from Princeton University revealed many instances of discriminatory associations between words. Female names were more closely associated to familial terminology while male names were more closely associated with careers [7]. Hence in this case the algorithm elevates societal gender stereotypes and essentially automates discrimination.

B. Algorithmic Bias

Other than the dataset being biased, the algorithm used to model the data can also be biased. The concept of Jim Code introduced by Ruha Benjamin can be used to describe the discriminatory behaviour of the algorithm :

- Engineered Equity- The phenomenon that for some people to be elevated others should be contained
- Default Discrimination – They tend to reproduce the default settings of racism or discrimination present in the society onto the algorithm.
- Coded Exposure – Technologies that fair to see racial different or leave them over exposed to surveillance.
- Techno Benevolence – Technologies that claim to address bias but in reality, contribute to or deepen discrimination[2].

III. RELATED WORK

Explain Images with Multimodal Recurrent Neural Networks This paper uses a multimodal RNN model to draw out descriptions from images.The paper is relevant as our proposed system also plans to adopt a CNN model to retrieve image features and connects it to a RNN model to produce text describing the content of the image.

Stereotyping and Bias in the Flickr30K Dataset This paper uses Flickr30K dataset to provide evidence of model bias.It focuses on the different types of bias present and demonstrates how a model can produce stereotype driven descriptions of images.

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings This paper uses Google News data produce word embedding vectors. The paper shows how the anologies mapped between words can be problematic and perpetuate further bias and discrimination against certain groups.

Additionally I will also be referring to the papers [3] [6] [10] [11] [14] [16] [18] [20] [21] [22] and [24] mentioned in references.

IV. STUDY 1 - EXAMINING GENDER BIAS IN IMAGE DATASETS

The model will aim to identify the gender based on the image. We analyze how accurately the model can detect gender from an image [4].
The goal behind detecting gender is to identify how significant of a factor gender is and how it affects other applications, for e.g. to detect if image captions are biased against gender.Even when Gender Identification models have high accuracy, they tend to miss classify among groups and this goes unnoticed. A study conducted on IMDB-WiKi dataset achieved a 91.17% test accuracy for gender identification but it was discovered that females had a higher chance of being miss classified. Hence though this study we want to find out how these gender identification models perform and whether the current systems in place should be employed at all.
Gender Identification as a stand alone experiment is not recommended and the purpose of this study is to detect gender biases and not perpetuate binary gender roles. This will be further discussed in the discussion.

Dataset

Two datasets are being used to perform Gender Identification.

1. UTKFace dataset The UTKFace dataset consists of over 200K images with gender, age and ethnicity annotations. The labels used for gender are: 0 - Male and 1 - Female.

FIG. 1. Images from UTKFace Dataset

[**cvpr**]

2. CelebA Image Dataset This dataset is 1.3GB with around 200K images and was specifically designed for gender classification [**zhifei2017cvpr**].

FIG. 2. Images from CelebA Image Dataset

A. Model Explanation

UTKFace dataset
The dataset is distributed into male and female as below:

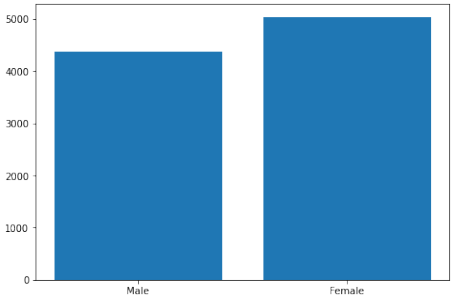

FIG. 3. Distribution of Dataset

A convolutional Neural Network Model is used to identify the Gender from the image. The CNN model consists of an input layer, 4 convolutional layers and an output layer. The model was run for 20 epochs.

Celeb A dataset
A convolutional Neural Network Model is used to identify the Gender from the image. The CNN model consists of an input layer, 5 convolutional layers and an output layer. The model could not be run for 20 epochs due to lack of GPU Power.

B. Experiment and Results

UTKFace dataset The accuracy of the model against UTKFace dataset is 90%. The Model loss is plotted below.
The model accuracy alone is not a good measure to assess the performance of the model. We access the

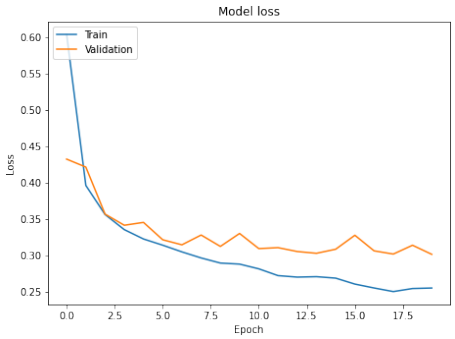

FIG. 4. Model Loss

Precision, Recall and confusion matrix of the model.

	precision	recall	f1-score	support
0	0.91	0.91	0.91	3111
1	0.90	0.89	0.90	2816
accuracy			0.90	5927
macro avg	0.90	0.90	0.90	5927
weighted avg	0.90	0.90	0.90	5927

FIG. 5. Precision and Recall

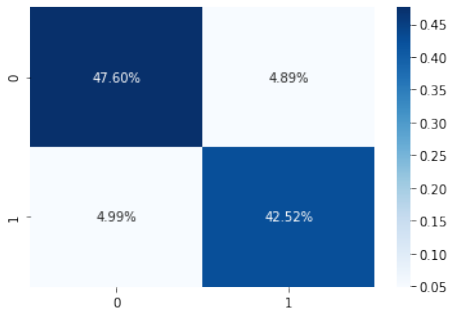

FIG. 6. Confusion Matrix

According to the results presented by the confusion matrix, we can conclude that the model does a good job in classifying image by gender. But from the confusion matrix we know that the False Positive and the False Negative values are 290 and 296 respectively. Hence a portion of images are being miss classified. Hence we need to analyze the model and check which factors are essential while making the classification and how these factors differ in images of men and images of women.

Celeb A dataset The accuracy of the model against UTKFace dataset is 72.47%. The Model loss is plotted below.Hence the accuracy of this model has significantly dropped. This can be attributed to the images being noisy and uncropped.

V. STUDY 2 - EXAMINING GENDER BIAS IN WORD EMBEDDINGS

Word Embedding acts as a basic unit for Natural Language Processing. Word Embeddings are trained on text and represent each word as a vector. Words which are deemed to be similar have vectors which lie closer together as well. The relationship between words can be represented using an expression with its corresponding embedded vectors. For example, 'man is to king as woman is to x' can be denoted as man:king :: woman:x. Hence these vectors are powerful and an arithmetic operation between them helps us detect relationships between different words.
Hence we use these analogies to detect biased representations in text.

A. Dataset

Pre-trained word embeddings were used to analyze and detect gender bias in this experiment. The pre-trained model used id ' GoogleNews-vectors-negative300.bin.gz.'. This is a pre-trained word2vec model published by Google containing Google news data. It is a 1.53 Gigabytes file with 3 million words and phrases[1] .

B. Experiment and Results

In this experiment a function is written to extract the relationship between different terms. The function also returns cosine similarity scores between the words. Cosine Similarity is a popular similarity measure used for text data. A cosine similarity score of 0 implies that the words are the least similar and score of 1 implies that the words are the most similar.
We begin by considering the occupation 'sewing'.Please observe FIG3 and FIG4.

she - sewing = he - ?

```
check_full_list(subj='she', obj='sewing', counterpart='he')
```

```
[('sewing', 0.66992193),
 ('woodworking', 0.5794616937637329),
 ('sew', 0.531487226486206),
 ('carpentry', 0.5230335593223572),
 ('woodcarving', 0.49166661500930786),
 ('wood_carving', 0.47534996271133423),
 ('leatherworking', 0.4700630009174347),
 ('Sewing', 0.4634256958961487),
 ('knitting', 0.4630447924137155),
 ('spinning_weaving', 0.4606916308403015),
 ('woodworking_shop', 0.45343703031539917)]
```

FIG. 7. Male counterpart of sewing

The occupations associated with 'he' include labour intensive and stereotypically male dominated occupations like 'woodworking', 'carpentry' and 'Woodcarving'. The occupations associates with 'she' include 'knitting', 'crocheting' and 'quilting'. Hence we can say that the

he- sewing = she - ?

```
check_full_list(subj='he', obj='sewing', counterpart='she')
```

```
[('sewing', 0.8244272),
 ('knitting', 0.6727133393287659),
 ('quilting', 0.6575378775596619),
 ('Sewing', 0.6378443241119385),
 ('needlework', 0.6374493837356567),
 ('sewing_embroidery', 0.5977171659469604),
 ('crochet', 0.5976918339729309),
 ('crocheting', 0.5905722379684448),
 ('serger', 0.5881949067115784),
 ('needlecraft', 0.5823681354522705),
 ('sewing_quilting', 0.5818668603897095)]
```

FIG. 8. Female counterpart of sewing

examples above exacerbate the preconceived notions of occupations we have in our society.

We move onto observe how adjectives differ for men and women. The adjective most similar to 'he' in terms

she - lovely = he - ?

```
check_full_list(subj='she', obj='lovely', counterpart='he')
```

```
[('lovely', 0.69995606),
 ('magnificent', 0.624822735786438),
 ('marvelous', 0.6054928302764893),
 ('splendid', 0.5995590686798096),
 ('nice', 0.5869458913803101),
 ('fantastic', 0.5587064027786255),
 ('delightful', 0.5561119914054871),
 ('terrific', 0.5524159669876099),
 ('wonderful', 0.5481390953063965),
 ('brilliant', 0.5460424423217773),
 ('beautiful', 0.5450632572174072)]
```

FIG. 9. Male counterpart of lovely

on cosine similarity are 'magnificent', 'marvelous' and 'splendid'. While the adjectives most similar to 'she' describe a woman's physical appearance.

We now find multiple analogies with respect to the model characteristics. Refer to FIG6

```
Subtracting man from male-dominated roles

manager is to man as vice_president is to woman
executive is to man as chairwoman is to woman
doctor is to man as gynecologist is to woman
lawyer is to man as attorney is to woman
programmer is to man as programmers is to woman
professor is to man as associate_professor is to woman
soldier is to man as solider is to woman
officer is to man as Officer is to woman
janitor is to man as receptionist is to woman
rockstar is to man as rocker_chick is to woman

Subtracting woman from female-dominated roles

nurse is to woman as medic is to man
teacher is to woman as headmaster is to man
homemaker is to woman as machinist is to man
housewife is to woman as schoolteacher is to man
midwife is to woman as midwives is to man
secretary is to woman as sectary is to man
maid is to woman as housekeeper is to man
dancer is to woman as magician is to man
receptionist is to woman as receptionists is to man
artist is to woman as painter is to man
```

FIG. 10. Analogies between different gender occupations

The analogies represented include -

1. nurse is to woman as medic is to man
2. teacher is to woman as headmaster is to man
3. housewife is to woman as schoolteacher is to man

The above analogies are discriminatory towards woman and aggravate the existing social biases present.
The professions are clustered together. Words which have similar biases are grouped together.

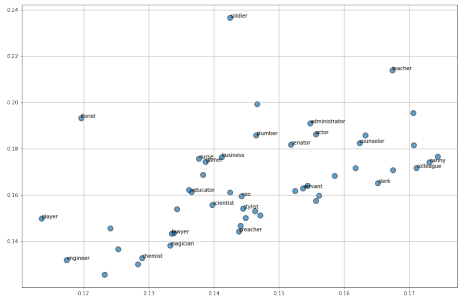

FIG. 11. Cluster of Different Professions

VI. DISCUSSION

In Study 1 we analyze the effectiveness of Gender Identification model. Even though the model performed well with a high accuracy, it is significant to note the large scale implications of this model. From the confusion matrix we know that the model does miss classify images.
The assessment of an algorithm should look beyond computational depth and focus on it's the broader application and impact on humans.
All the Gender Identification models deal with datasets that contain binary gender classification. These models are not representative of all genders. Hence Gender models which categorize images only as Male or Female are discriminatory and discussions should be made on how to work towards including all gender identities.
In Study 2 we analyze the word mapping associated with each gender and it's implications. The words connected to both genders were stereotype driven.
Hence before Natural Language systems are deployed the embedded word vectors need to be neutralized. The vectors should be represented in such a way that they remain equidistant from equality pairs like "he-she". AI algorithms should aim to uplift marginalized societies, instead these algorithms if deployed as is will only widen the gap between different groups. Hence bias needs to be removed from word embeddings and they decisions made by these algorithms should be unbiased and fair.

VII. FUTURE WORK

This study would be extended to include an image captioning model. An image captioning model should successfully describe the image. Once we have the image captions, we will conduct a study to analyze whether the image captions defer for different genders. A comparative study will be conducted. Analysis will be conducted to investigate how image captions will vary from male to female even when the picture look similar.
Study 2 will also be extended to include debiasing models so that the word embedding are neutral and unbiased. These task is critical as language models these days are predominantly being used to screen documents. The language models being deployed have a strong impact on people's lives and their choices. AI systems need to be investigated thoroughly to make sure they aren't introducing any bias knowingly or unknowingly.

VIII. CONCLUSION

Hence we conduct two studies to investigate the ways in which gender bias are present in different systems and how they can be detected. In study 1, we analyze different face image datasets to identify the gender. Even though the accuracy was good, it was seen that the model missclassified instances. Hence when such models are being implemented in different applications, the results of should not be merely assessed using performance metrics. In study 2 we demonstrate how word embedding can be biased against women and can facilitate further societal biases. Hence the vectors need to be neutralized so that they are not biased towards a particular gender.

DATA AVAILABILITY

UTKFace Dataset - https://www.kaggle.com/jangedoo/utkface-new

Celeb A Dataset - https://www.kaggle.com/ashishjangra27/gender-recognition-200k-images-celeba

CODE AVAILABILITY

https://github.com/Aparna1111/A-Study-on-Gender-Bias-in-Image-Datasets-and-Language-Processing—552-Project-

ACKNOWLEDGMENTS

I would like to express my profound gratitude and heartfelt regards to my professor, Dr.Marcin Abram for his exemplary guidance, monitoring and constant encouragement throughout the course of the assignment. I would also like to acknowledge with much appreciation the crucial role of Ninareh Mehrabi for her earnest cooperation and assistance.I am thankful and fortunate enough to get constant support and encouragement from Supriya Devalla and Pratik Singhavi. I am thankful to the comments and suggestions given by my peer reviewers. Their recommendations steered me in the right direction and am thankful for their help.I express my deep sense of indebtedness to my peers whose contribution in

constant suggestions, inspiring discussions and encouragement helped me through the course of this assignment.

REFERENCES

[1] Gunjan Agicha. "Word embeddings in NLP". In: (2019). URL: https://gunjanagicha.medium.com/word-embeddings-ee718cd2b8b5.

[2] Ruha Benjamin. "A New Jim Code?" In: (2019). URL: https://www.edsurge.com/news/2019-08-20-the-new-jim-code-race-and-discriminatory-design.

[3] C.J. Beukeboom. "Mechanisms of linguistic bias: How words reflect and maintain stereotypic expectancies". English. In: *Social Cognition and Communication*. Ed. by J.P. Forgas, O. Vincze, and J. Laszlo. Psychology Press, 2014, pp. 313–330. ISBN: 9781848726642.

[4] Shruti Bhargava and David Forsyth. "Exposing and Correcting the Gender Bias in Image Captioning Datasets and Models". In: (Dec. 2019).

[5] Annie Brown. "Biased Algorithms Learn From Biased Data: 3 Kinds Biases Found In AI Datasets". In: (2020). URL: https://www.forbes.com/sites/cognitiveworld/2020/02/07/biased-algorithms/?sh=6a38f68176fc.

[6] Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification". In: *FAT*. 2018.

[7] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334 (2017), pp. 183–186. ISSN: 0036-8075. DOI: 10.1126/science.aal4230. eprint: https://science.sciencemag.org/content/356/6334/183.full.pdf. URL: https://science.sciencemag.org/content/356/6334/183.

[8] William Crumpler. "How Accurate are Facial Recognition Systems – and Why Does It Matter?" In: *Center for Strategic and International Studies* (2020).

[9] Jeffrey Dastin. "Amazon scraps secret AI recruiting tool that showed bias against women". In: (Oct. 2018). URL: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.

[10] S. Hajian, F. Bonchi, and Carlos Castillo. "Algorithmic Bias: From Discrimination Discovery to Fairness-aware Data Mining". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016).

[11] Sara Hajian, Francesco Bonchi, and Carlos Castillo. "Algorithmic Bias: From Discrimination Discovery to Fairness-Aware Data Mining". In: KDD '16. San Francisco, California, USA: Association for Computing Machinery, 2016. ISBN: 9781450342322. DOI: 10.1145/2939672.2945386. URL: https://doi.org/10.1145/2939672.2945386.

[12] Sean Illing. "How search engines are making us more racist". In: (May 2018). URL: https://www.vox.com/2018/4/3/17168256/google-racism-algorithms-technology.

[13] Atakan Kantarci. "Bias in AI: What it is, Types Examples, How Tools to fix it". In: *AIMultiple: AI Use cases Tools to Grow Your Business* (2021).

[14] Ryan Kiros, R. Salakhutdinov, and R. Zemel. "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models". In: *ArXiv* abs/1411.2539 (2014).

[15] Agostina J. Larrazabal et al. "Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis". In: *Proceedings of the National Academy of Sciences* 117.23 (2020), pp. 12592–12594. ISSN: 0027-8424. DOI: 10.1073/pnas.1919012117. eprint: https://www.pnas.org/content/117/23/12592.full.pdf. URL: https://www.pnas.org/content/117/23/12592.

[16] Junhua Mao et al. "Explain Images with Multimodal Recurrent Neural Networks". In: *ArXiv* abs/1410.1090 (2014).

[17] Rachel Meade. "Bias in Machine Learning: How Facial Recognition Models Show Signs of Racism, Sexism and Ageism". In: *towards data science* (2019).

[18] Emiel van Miltenburg. "Stereotyping and Bias in the Flickr30K Dataset". In: Portoroz, Slovenia, 2016, pp. 1–4.

[19] Osonde A. Osoba and William Welser IV. *An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence*. Santa Monica, CA: RAND Corporation, 2017. DOI: 10.7249/RR1744.

[20] Jahna Otterbacher et al. "How Do We Talk about Other People? Group (Un)Fairness in Natural Language Image Descriptions". In: *AAAI 2019*. 2019.

[21] Rachel Rudinger, Chandler May, and Benjamin Van Durme. "Social Bias in Elicited Natural Language Inferences". In: *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 74–79. DOI: 10.18653/v1/W17-1609. URL: https://www.aclweb.org/anthology/W17-1609.

[22] Denise Sekaquaptewa et al. "Stereotypic Explanatory Bias: Implicit Stereotyping as a Predictor of Discrimination". In: *Journal of Experimental Social Psychology* 39 (Jan. 2003). DOI: 10.1016/S0022-1031(02)00512-7.

[23] Natasha Singer and Cade Metz. "Many Facial-Recognition Systems Are Biased, Says U.S. Study". In: *The New York Times* (2019).

[24] Oriol Vinyals et al. "Show and tell: A neural image caption generator". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 3156–3164.