Observando o viés no reconhecimento de emoções

Resumo do Projeto Final - Human-Centered Machine Learning

Breno Tanure
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, MG - Brasil
tanure@dcc.ufmg.br

Ewerton Santos
Departamento de Ciência da Computação
Universidade Federal de Minas Gerais
Belo Horizonte, MG - Brasil
ewerton_dc@hotmail.com

1. Introdução e Motivação

A motivação para este trabalho é decorrente de estudos recentes acerca de uma área denominada Human-Centered Machine Learning (HCML), que de acordo com Stevie Chancellor et.al. (em [1]), “HCML foca nos impactos em indivíduos, comunidades e na sociedade, tornado explícito pelas suas contribuições para domínios centrados no humano, além de desafios e objetivos auto-declarados dentro desses artigos”. Algumas pesquisas em HCML buscam entender o efeito de viés em bancos de dados e em algoritmos, suas consequências para o resultado de estudos e o impacto deles na sociedade, além de entender como alguns algoritmos de aprendizado de máquina podem, por design, contribuir para a perpetuação de estigmas sociais.

Quando algoritmos são utilizados para auxiliar (ou como única fonte de decisão) para a tomada de decisões importantes na sociedade, como decidir quem deve continuar em cárcere ou não, a quem uma instituição financeira deve ou não emprestar dinheiro, quem é uma ameaça em potencial à segurança alheia, etc., é necessário avaliar se o código é justo, se ele possui algum viés e se ele é moral e eticamente confiável para ser usado nessas situações. Com o objetivo de auxiliar na construção dessa justiça, Kärkkäinen et.al. criaram o FairFace ([6]), um dataset de faces humanas que busca atingir maior igualdade entre gênero e raças, na tentativa de reduzir a possibilidade de um eventual viés nos resultados de uma classificação vir dos dados.

Isso nos motivou a buscar compreender como um viés racial ou de gênero pode se manifestar em um algoritmo de classificação. Para isso, usamos uma rede neural pré-treinada com 20000 fotos do conjunto de dados *FERC2013* para reconhecer emoções nas imagens do FairFace. Analisamos os resultados obtidos graficamente e com intervalos de confiança, a fim de buscar responder: existe algum viés relacionado a essa classificação de emoções? Se sim, como ele ocorre?

2. Trabalhos relacionados

Correa et.al em *Emotion Recognition Using Deep Convolutional Neural Networks* ([7]) treinaram uma rede neural convolucional para reconhecimento de emoções em imagens de faces humanas para compreender como essa ferramenta da computação pode ser usada no papel por eles aplicado. Nós partimos dessa solução e do entendimento da qualidade dos resultados por eles obtidos para usar a mesma rede neural, com os mesmos pesos (não houve treinamento posterior), para reconhecer emoções no dataset do FairFace.

Outro trabalho que podemos mencionar é o artigo que propõe o FairFace. Nele, os autores realizam um estudo de balanceamento de datasets, analisam os problemas presentes em alguns bancos de dados comumente utilizados e sugerem um dataset novo, que visa a mitigar o viés racial de datasets. Além disso, eles realizaram experimentos para avaliar a performance de generalização de seus dados. Como resultado (além da imensa contribuição do FairFace em si), perceberam que o modelo que treinaram em seus dados novos possuía acurácia substancialmente maior e que esse resultado era consistente entre os grupos de gênero e de raça.

3. Problema

Com os avanços no nível de acurácia e utilização mais frequente de algoritmos de reconhecimento facial, também cresce a preocupação com os impactos éticos e sociais de tais práticas. Uma das grandes questões problemas relacionadas ao reconhecimento facial está em torno do viés das previsões, o qual pode prover de duas fontes: dos dados ou dos modelos.

Há diversas questões éticas e morais acerca do uso de algoritmos de aprendizado de máquina para determinadas tarefas de classificação, e a presença de um viés de gênero ou de etnia nesse código pode perpetuar um estima social, além de ser reflexo desse.

Neste trabalho queremos entender essas fontes de viés e tentar descobrir, no caso de reconhecimento de emoções, se ele está presente e de onde ele é proveniente.

4. Metodologia

O primeiro passo foi realizar uma análise exploratória dos dados do FairFace para assegurar que eles são bem distribuídos entre suas classes. De fato, o trabalho realizado pelos autores deste projeto resultou em um banco de dados muito equilibrado em relação a gênero e com bom equilíbrio em relação a raça, especialmente entre negros, asiáticos do Leste e Indianos. O dataset apresenta boa variedade de idades, o que pode ser visto como uma vantagem em nossa tarefa, pois permite que o reconhecimento de imagens seja testado em humanos diversos e com uma ampla gama de idades.

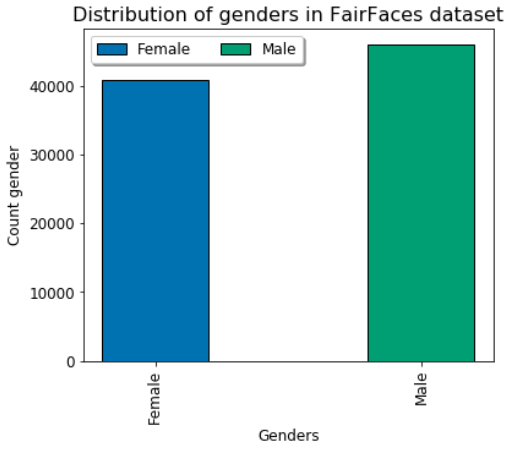

Figura 1: distribuição de gênero entre as imagens do FairFace.

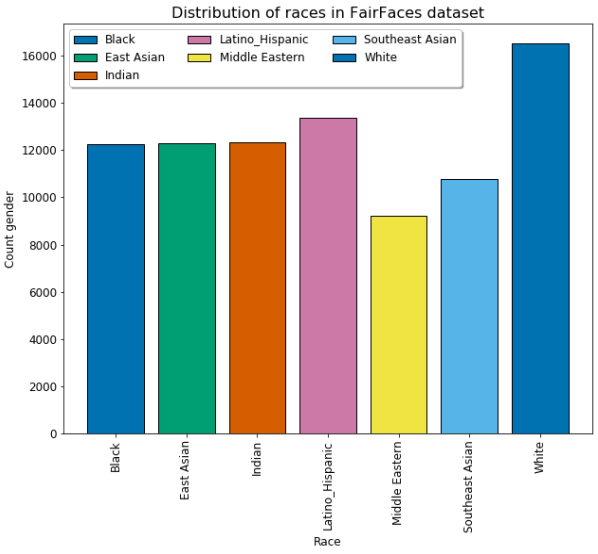

Figura 2: distribuição de etnia entre as imagens do FairFace.

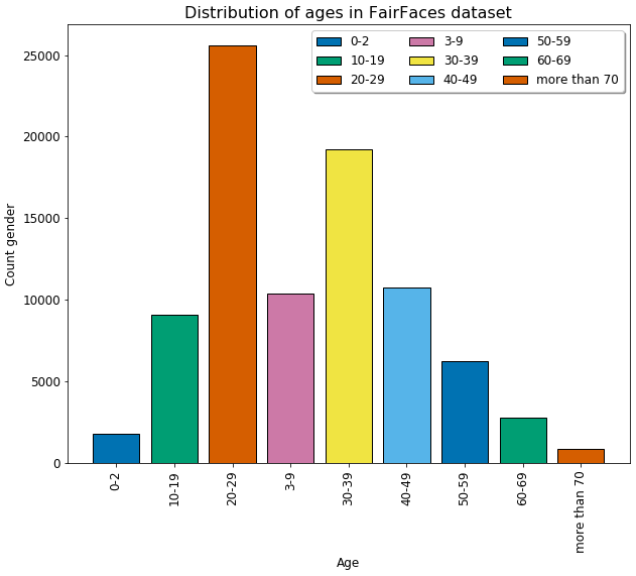

Figura 3: distribuição de idades entre as imagens do FairFace.

Posteriormente montamos uma rede neural convolucional com a mesma arquitetura utilizada por Enrique Correa et.al. e acessamos o repositório de seu estudo [7] e baixamos os arquivos com os pesos da rede obtidos durante o treinamento por eles realizado. Definimos esses pesos em nossa rede, de forma a termos em mãos exatamente a mesma ferramenta utilizada no estudo acima citado, que já sabemos ser eficaz na tarefa de reconhecimento de emoções em imagens.

Após montar a rede, fornecemos as 86,744 imagens do conjunto de treino FairFace para classificação entre 7 emoções possíveis: Bravo(a), Triste, Feliz, com Medo, com Nojo, Neutro(a), Surpreso(a) (respectivamente, *Angry, Sad, Happy, Fear, Disgust, Neutral, Surprise*, nos termos

originais de classificação do estudo [7]). Analisamos os resultados obtidos graficamente para tentar responder às perguntas feitas na Introdução deste trabalho e, além disso, calculamos os intervalos de confiança para cada emoção, de forma a entender como alterar a etnia altera o intervalo para essa emoção.

Intervalos de confiança são intervalos estimados de um parâmetro de interesse de uma população, bastante usados na Estatística para indicar a confiabilidade de uma estimativa. O intervalo determina a probabilidade de obtermos o resultado indicado para dada população. Neste trabalho calculamos os intervalos de confiança das emoções para analisar a probabilidade de uma classe de etnia receber a classificação de uma dada emoção. Para o cálculo dessa métrica usamos a técnica *Adjusted Wald*, que permitiu obter um intervalo que contém a proporção em média com 95% de confiança.

A fórmula usada para o cálculo descrito acima encontra-se na figura abaixo:

$$\hat{p} \pm 1.96 * \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Calcular intervalos de confiança também permite perceber se a rede possui alguma tendência de gênero ou de etnia para classificação de dadas emoções. Os resultados obtidos e sua análise encontram-se na seção abaixo.

5. Resultados e análises

5.1. Análise por gênero

Primeiramente, analisamos a classificação de emoção por gênero:

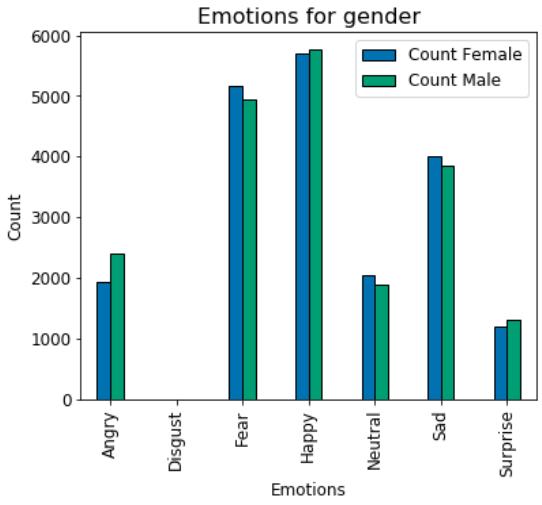

Figura 4: distribuição das classificações de emoções por gênero.

É possível observar grande equilíbrio entre as classificações, possível reflexo do balanceamento de gênero do dataset. Isso é mais uma evidência do bom trabalho que os autores fizeram ao criar o FairFace.

Há uma leve disparidade em relação a emoção *Angry*, mas ela é comparável à disparidade entre o gênero dos indivíduos já presente no dataset. Dessa forma, consideramos os resultados balanceados em termos de gênero.

5.2. Análise por idade

A segunda análise realizada foi em termos de idade, cujo resultado pode ser visto na figura abaixo:

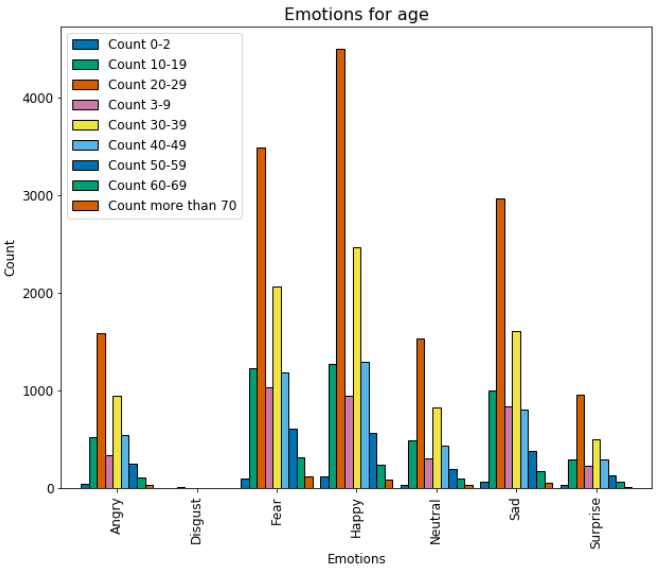

Figura 5: distribuição das classificações por idade.

Nesse resultado é possível notar maior disparidade entre os resultados nessa análise: pessoas de 20 a 29 anos possuíram muito mais classificações que as outras, especialmente para a emoção Feliz. A diferença foi de quase 50% em relação à classe de idade com o segundo maior número de classificações para *Happy*, o que difere significativamente da diferença percebida entre essas classes nos dados do FairFace.

O gráfico dessa emoção especificamente pode ser indicativo de um pequeno viés no algoritmo, dado a maior disparidade presente, embora em geral a proporção entre as classes para as classificações tenda a seguir o padrão da distribuição entre as classes de idade no próprio dataset (imagem 3). Isso contribui para a hipótese de que o algoritmo também não parece apresentar viés significativo em relação a idades.

5.3. Análise por raça

Posteriormente, foi realizada uma análise da distribuição das classificações de emoções em relação a raça, de modo a observar quantas instâncias de cada etnia foram classificadas com cada uma das emoções:

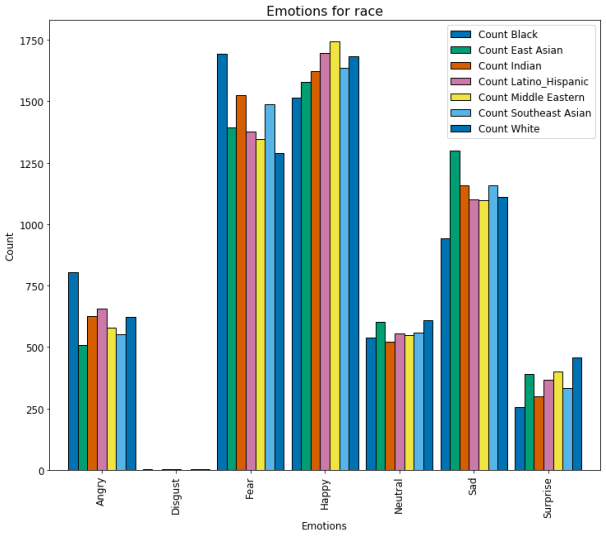

Figura 6: distribuição das classificações por raça.

Com esse resultado é possível notar uma disparidade maior, que parece não seguir o padrão da proporção entre as classes de etnia do FairFace (imagem 2). O algoritmo tende a classificar pessoas negras com Medo e Bravo bem mais frequentemente do que pessoas de outras raças, especialmente brancas. Em contrapartida, pessoas negras foram classificadas bem menos vezes como Feliz, enquanto pessoas do Oriente Médio, minoria no dataset, tiveram o maior número de instâncias classificadas com essa emoção.

Pessoas do Leste Asiático, classe bem equilibrada em relação a negros e indianos, foram mais vezes classificadas como Triste do que qualquer outra raça; inclusive os negros foram classificados o menor número de vezes com essa emoção.

5.4. Análise de intervalos de confiança

A análise de intervalos de confiança fornece dados interessantes, que nem sempre podem ser observados com os gráficos anteriores. Os gráficos abaixo mostram os intervalos de confiança para cada etnia com emoções fixadas uma a uma:

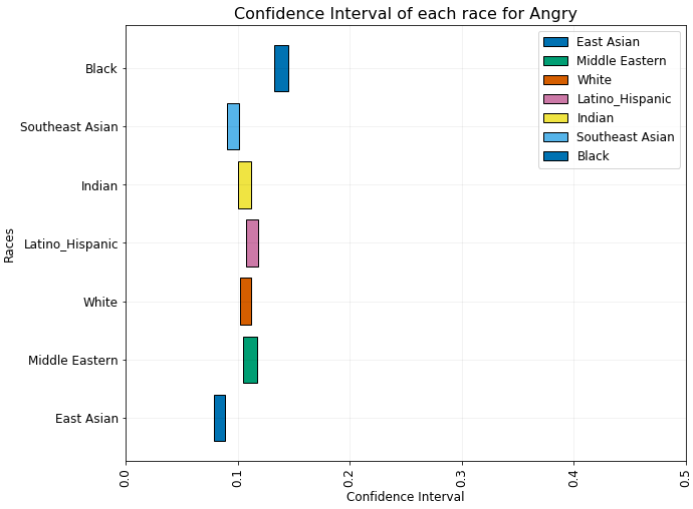

Figura 7: intervalos de confiança para cada etnia com a emoção fixada em *Angry*.

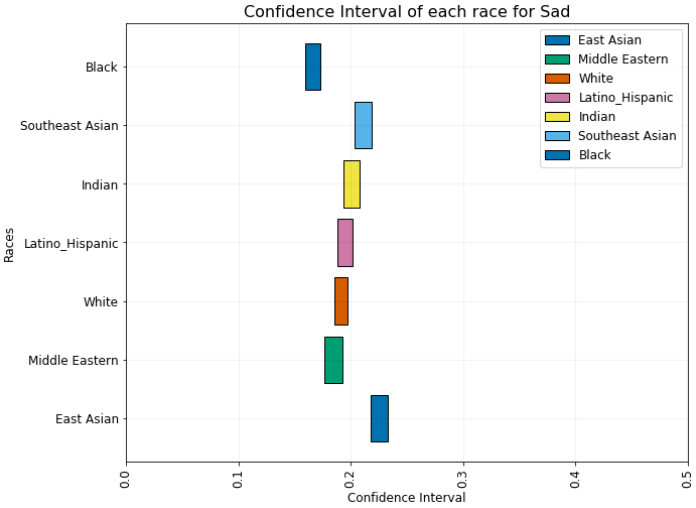

Figura 8: intervalos de confiança para cada etnia com a emoção fixada em *Sad*.

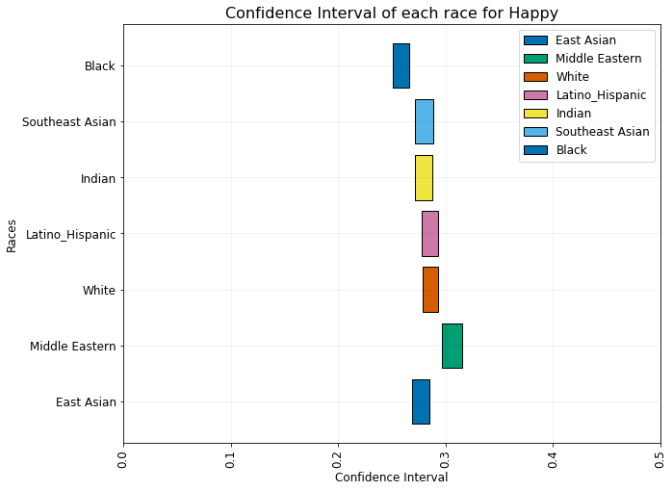

Figura 9: intervalos de confiança para cada etnia com a emoção fixada em *Happy*.

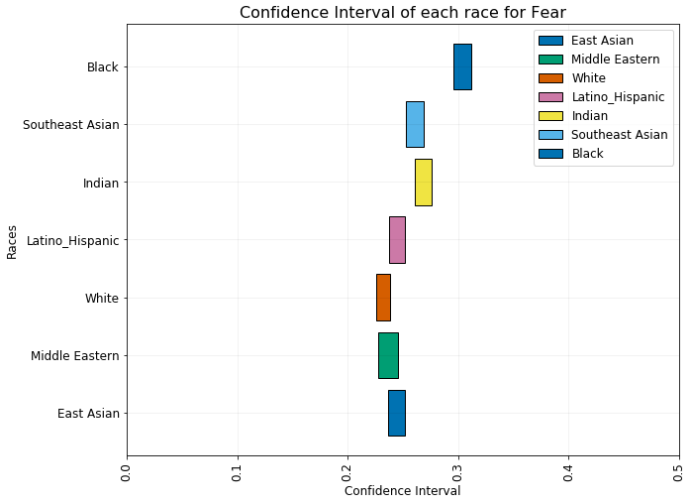

Figura 10: intervalos de confiança para cada etnia com a emoção fixada em *Fear*.

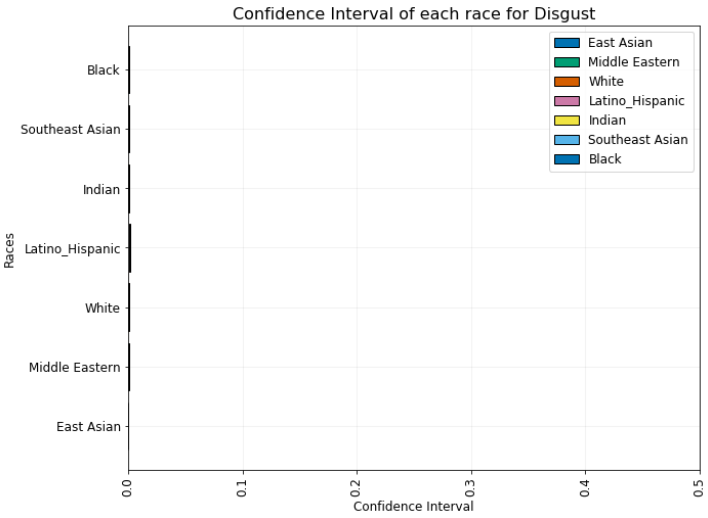

Figura 11: intervalos de confiança para cada etnia com a emoção fixada em *Disgust*.

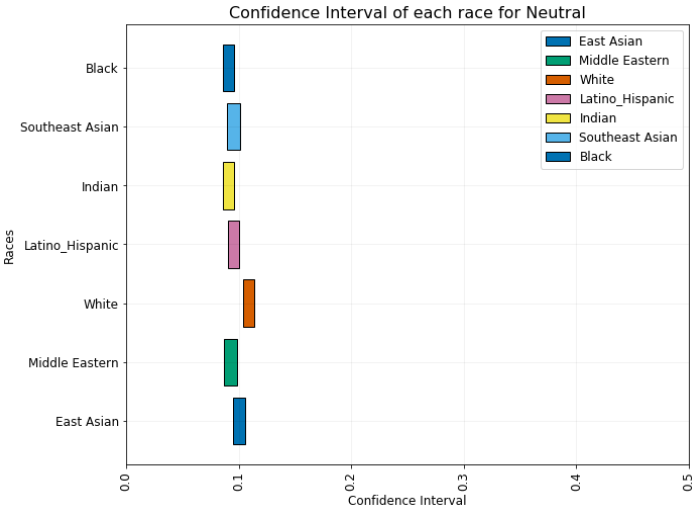

Figura 12: intervalos de confiança para cada etnia com a emoção fixada em *Neutral*.

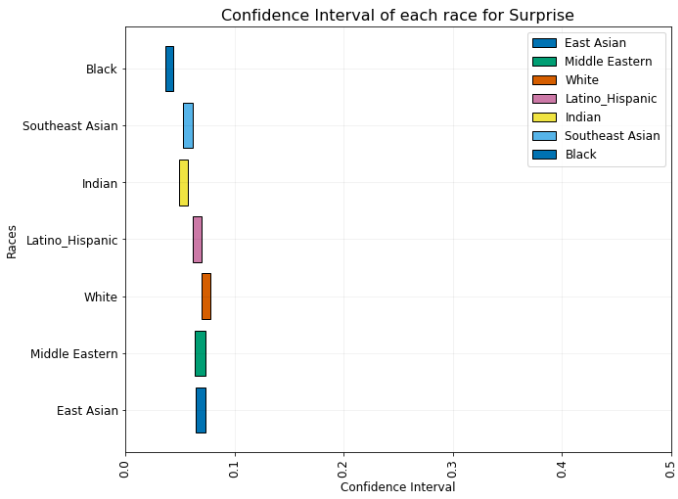

Figura 13: intervalos de confiança para cada etnia com a emoção fixada em *Surprise*.

Como é possível perceber, quanto menor os intervalos de confiança mais difícil (menos provável) é para uma classe (de raça) receber a classificação com a emoção fixada. No caso de *Angry*, fica mais claro que, mesmo com um número menor de instâncias para negros, essa etnia tem uma probabilidade significativamente maior de ser classificada como *Angry*.

Fica evidente uma maior tendência do algoritmo em determinar que pessoas do Leste Asiático possuem a emoção *Sad*. Outro caso em que houve disparidade significativa foi o da emoção *Fear*, cuja maior probabilidade é para negros.

Dois casos interessantes, porém que não trouxeram disparidade significativa foram os das emoções *Neutral* e *Disgust*. Ambas apresentaram bom balanceamento entre as raças, mas a segunda teve poucas instâncias para cada uma.

6. Conclusões e observações finais

Com os resultados acima, é possível perceber que, mesmo com o uso de um dataset feito para ser balanceado e justo, é possível obter viés em classificações, provido pelo algoritmo. A rede neural utilizada neste trabalho obteve bons resultados no estudo em que ela foi proposta, mas algumas evidências de nossos resultados mostram que há certo desbalanceamento das classificações para algumas etnias, especialmente negros.

Isso reitera a importância de se buscar algoritmos e datasets mais justos e balanceados, de forma a evitar que viés possa ser um problema não apenas para o algoritmo em si (dado que espera-se que construir um modelo justo, robusto e com boa performance seja um objetivo em comum de todos os programadores), mas para impedir que esses estudos tragam consequências negativas para a sociedade e para os indivíduos neles envolvidos.

REFERÊNCIAS

[1] *Stevie Chancellor, Eric P. S. Baumer, and Munmun De Choudhury. 2019. Who is the "Human" in Human-Centered Machine Learning: The Case of Predicting Mental Health from Social Media. Proc. ACM Hum.-Comput. Interact. 3, CSCW, Article 147 (November 2019).*

[2] *Catherine D'Ignazio and Lauren Klein. 2019. Chapter One: Bring Back the Bodies. MIT Open Press.*

[3] *Anna Lauren Hoffmann. 2019. Where fairness fails: Data, algorithms, and the limits of anti discrimination discourse. Information, Communication, and Society (2019).*

[4] *Anna Lauren Hoffmann, Nicholas Proferes, and Michael Zimmer. 2018. "Making the world more open and connected": Mark Zuckerberg and the discursive construction of Facebook and its users. New Media and Society 20, 1 (2018), 199–218.*

[5] *Importance of emotions in learning: a neuropsycopedagogical approach*

[6] *Kärkkäinen, Kimmo, and Jungseock Joo. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age." arXiv preprint arXiv:1908.04913 (2019).*

[7] *Correa, Enrique; Jonker, Arnoud; Ozo, Michael; Stolk, Rob. Emotion Recognition using Deep Convolutional Neural Networks June 30, 2016 Link:https://github.com/atulapra/Emotion-detection/blob/master/ResearchPaper.pdf*

..