Reducing Bias in Emotion Labeling

Abigail Nay
abna2071@colorado.edu

Jake Swartwout
jacob.swartwout@gmail.com

Mari Griego
magr9892@colorado.edu

Cathy Yang
cathy.yang@colorado.edu

Dominic Fikany
dofi1111@colorado.edu

ABSTRACT

UPDATED—May 5, 2020. This research paper aims to identify the primary factors that lead to emotional labeling bias in facial recognition. The negative impact that biased datasets and labeling can have on machine learning and facial recognition algorithms can produce inaccurate and even discriminatory results. Through research and a survey study, our group collected data on the responses of surveyors to come to conclusions about the impact of four major identifying factors; age, gender, race, and income. After finding significant relevance in research that these identifying factors can have on emotional labeling, the goal of this research project is to further recognize their influence, and how to reduce the impact it has on emotional labeling. Ten different surveys were deployed through three methods of sampling over the course of three days allowing us to obtain data to make inferences regarding the four identifying factors.

Author Keywords

Emotion; labeling; facial recognition; AI; machine learning; survey

INTRODUCTION

Machine learning is currently one of the biggest buzzwords of the tech industry. With more data being collected and made available to computer and data scientists everyday, more problems are being solved with complex algorithms and other tech-based solutions. Facial recognition is a sphere of machine learning that has promising results when conducted correctly, but can have undesirable consequences when not thoroughly vetted for bias. One example of these negative consequences can be seen with Google Photos facial recognition software classifying black people as Gorillas, an unfortunate result of possible biases in the training data fed to the original model (Guynn, 2015). With facial recognition being used to classify more fluid ideas like gender and emotion, it is also very important to reduce the biases found within training datasets if we are to create useful facial recognition algorithms and avoid problematic outcomes.

The quality of the data fed into any algorithm will determine how effective and correct that program will be at solving real-world problems. Statisticians are aware of this problem and recognize the need to pull from clean and useful data. Dirty data, data not representing the true population, and data that is heavily biased will lead to the creation of models that are utterly useless, and this process has become so infamous the saying 'garbage in, garbage out' has come to intuitively mean feeding a model bad data will produce a bad model (Tweedie et al. 1994).

In machine learning, there are many data cleansing checks to be made before feeding a model a training dataset. One of these involves collecting quality data and making sure it is representative of the population. In emotion recognition this means collecting a wide range of emotional displays from a diverse group of people in order to collect images showcasing a spectrum of emotions in a representative population sample. The second data quality check, and the one we are most interested in addressing in this paper, centers around the bias in labeling different images. If we feed a flower database into a machine learning algorithm with some daisies incorrectly labeled as roses, then our algorithm will reflect this labeling bias and will be less useful and accurate in the long run.

When it comes to emotion recognition, it is especially important to have an accurate model for certain proposed use cases. One suggested use for an emotion recognition algorithm is in job relocation and assignments, where employees would be assigned to tasks that gave them the most satisfaction and joy based on their facial expressions during certain tasks (Purdy et al. 2019). Situations like these are where we could possibly introduce a lot of bias in certain professions dominated by men, such as in civil engineering jobs where 89% of workers are male (Purdy et al. 2019). If this algorithm were trained to assign civil engineers to certain tasks, and men were used to label the emotions of people in a training dataset we may find that the model is only accurate when predicting the emotional state of men. Thus, if women in civil engineering were monitored by that algorithm, they may find their emotional states consistently mislabeled and could see themselves assigned to jobs they would not enjoy or being passed up for opportunities they would otherwise like.

In our research we aim to address some of these concerns by studying the effects of gender, age, race, and income on labeling bias in machine learning contexts. By understanding what factors influence a person labeling emotion, we can provide

feedback and suggestions for reducing labeling bias in future machine learning databases.

Our research was based on two parts, where we first read through and understood related studies and research to ground our understanding of different labeling biases. This is what informed our hypothesis when creating our own surveys and research methods. Through our preliminary findings we understood that there were many different factors that could potentially influence how someone labels emotions.

We decided to study the effects of gender on emotion labeling based on a study of gendered perceptions of subtle facial expressions. In this study researchers studied over 400 different subjects and examined their different recognitions of exaggerated and subtle facial expressions. While the researchers concluded there was no difference in recognition between men and women when perceiving the exaggerated facial expressions, they were able to conclude that women were significantly better at perceiving subtle facial expressions than men were (Hoffmann et al. 2010). This paper informed our decision to study effects of gender differences on emotion labeling.

Age and income were based on two studies examining the differences in emotion labeling of children with different backgrounds. Our decision to examine race comes from numerous studies examining different cultural and societal influences on emotion labeling. With race, we wanted to see if there would be a cultural difference in emotion labeling within the United States and wanted to concentrate on finding racial differences in a culturally diverse nation. We decided to test the age and income hypothesis as well because we wanted to expand the research from children and see if those same findings would hold with adults. Would age impact emotion recognition at the same level as differences between children in different development stages? And would living in poverty in adulthood affect emotion perception in ways like children living in poverty?

In order to answer these questions and test our hypotheses we moved into the second stage of our research, where we gathered pictures off Instagram and created a labeling survey to test differences between people of diverse backgrounds in the United States. After posting our survey to numerous Reddit boards dedicated to data collection and sending out surveys to family and friends, we were able to receive 150 responses over the course of three days. After doing a thorough analytical review of the data we were able to draw conclusions about the different influences on an individual's emotion recognition tendencies. Through this paper we set forth our findings and give recommendations on reducing labeling bias in emotion datasets for machine learning algorithms.

RELATED WORKS

Sources of Bias:

Many other papers have attempted to find sources of bias in how we label emotions. There has been a large focus on the culture of the labeler, as it is clear that different cultures interpret emotions in various ways. There have been varying results in this area of research. One study done by Ekman (1973) found that there did not seem to be differences across cultures. While they only surveyed college aged students, their findings showed that there seemed to be a high agreement on the primary emotion, and even agreement on the secondary emotion. However, Matsumo (1993) found that there were several differences within his line of research. In Japan, the cultural norm of hiding emotions may result in more sensitivity to facial emotions and less expression of them, so we must be careful in trying to group together results of this nature. Even among ethnic groups of American culture, they found slight cultural differences among clusters. Another study by Elfenbein (2003) may give us clues as to why this occurs. They found that there was higher accuracy for those attempting to label the culture they live in, but little difference from ethnicity. Perhaps the construction of these surveys caused a divide in this regard. Differing cultural backgrounds is a variable to be accounted for, as it has made significant impact in related studies and research. Looking for these potential factors within our results may compare similarly to Ekman's research.

We found several other papers as well, which inspected various other aspects that could possibly affect how emotion is presented and perceived. In one paper by Fox (2000), presented the idea that angry faces serve as distractors. People were quicker to find the singular angry face in a crowd of happy ones over the single happy face in a crowd of angry ones. By flipping the faces, there is no improved ability to discern the angry face, meaning that the improvement was due to the emotion being conveyed rather than the appearance. This becomes highly relevant towards our results and findings among non-white respondents. Although we did not use face inversions, we did conclude similar results among the 'Angry' emotional label.

Erhart (2019) looked at the emotional recognition abilities of children living in poverty. Through testing on pictures of both low and high intensity emotions, children in poverty had smaller improvements with greater intensity than the other children. While we won't focus on studying children directly, we included an income level question to see if this result persists into adulthood. The connection between income levels and emotional labeling was made apparent with our results, and this tie can lead to potential insights into more socio-cultural findings.

Another area was the division between genders. Hoffman (2010) found that women are able to better identify subtle emotional differences, such as anger, disgust, and fear. Women also tended to be labeled with these emotions more often. At the same time, there was no significant difference in labeling when "full-blown" emotions were shown. The importance this piece has to our research is clearly evident in two ways: first, gender plays a significant role in Hoffman's research, and our research; secondly, his findings show that emotional intensity plays an important part in emotional labelling. So, another way of viewing the data, is that it showed women could label the emotions at lower intensities than men. This will become especially prevalent within our findings. This makes Hoffman an important piece of literature to consider. Brody (2010) is also very important to our work. He found that images of women

expressing negative emotions were more likely to be labeled as sad or angry, in comparison to men who were more likely to be labeled as happy. What is interesting and integral to our better understanding of labeling, however, is that Brody provides a reason as to why women may label more accurately than men from more of a psychological point of view. This comes down to three things in his research: personality variables, culture, and social context. These more specifically come down to pop culture influence, socialization, and general differences in goals and roles these genders serve and perpetuate. In this way, Brody provides an important perspective on the importance of gender labelling, as his main goals were to avoid "stereotype threat" in future labelling. This, meaning doing whatever we can to ensure future machine learning algorithms do not further perpetuate social stereotypes. For these reasons, we made sure to include a gender-related question in our survey to study this result, which provided a multitude of insights regarding these differences in emotion labeling. These small, but significant differences became a major point of interest in our findings. Through our research, and the aforementioned experiments, gender differences appear to be one of the most important forms of bias among labelling responders.

Survey Construction:

We examined and researched many different ways to collect labels from people, as the format of input can influence the results. We originally wanted to have the emotional labels be 'free response', where the respondents could enter whatever word they wished. We learned from Barrett (2007) that giving labelers options to choose from altered their results (at least in the context of emotion labeling). These options made the emotions less ambiguous, resulting in more agreement. While it would be interesting to study linguistic differences, we concluded that consistency would be best in discovering the actual sources of bias for our data.

There are also other ways in which to collect the labels. Paiva-Silva (2016) details a few methods of relating emotions. These included matching faces with similar emotions, tracking eye movements, brain scanning, and the forced-choice method mentioned above. For large quantities of data needed to train computer vision systems, many of these methods seemed unfeasible or not useful, so we decided on a multiple choice selection. This same paper also informed our decision for those choices, saying that there is consensus among the six emotional categories and the 'Neutral' category.

A possible issue was brought to our attention by Griffith (1997), who pointed out the problem with the contextuality of emotions. Our recognition of emotions rely heavily on context, and we merely seek out what we expect to see. However, computer vision systems are generally trained on static images, so we are unable to provide this context. This issue relies more on the construction of our systems, rather than any preventable collection bias. During our data labeling, we noticed this issue in context, and evaluated the images on merely the face itself.

Reducing Bias After Collection:

We also looked at a few papers which sought to reduce bias algorithmically. Zhang (2017) looked at ways to reduce bias by detecting discrepancies in the data and altering the labels. For a binary classification, the algorithm looked at finding the peaks in the data and shifting them to become more balanced. Zeng (2018) attempted to solve this problem as well, by using the actual images to construct computer vision models and apply these back to the labels. By comparing the original labels and the constructed labels through the use of a Latent Truth Net, a better set of labels could be produced. We did not alter the discrepancies within the data, but rather excluded the data that upheld little result or significance.

While these papers are useful in understanding the area of research to consider, we focused more on preventing historical and collection bias. This way, we could hopefully solve the root of the problem and then later further reduce the bias in the labels.

METHODS

We collected data on how people labeled emotions in images of people's faces. To do this, we compiled a dataset of face images, analyzed the reliability between ourselves, compiled the best images into a survey, and then collected and analyzed the results.

Dataset creation:

Our initial dataset was created by scraping images off of Instagram. Instagram is a public free to use platform with a large user base, meaning that a large range of demographics are welcome to participate in the platform. The images are also self labeled through hashtags, allowing us to collect images according to a societal collection of interpretations, rather than a single company's decision (which would have occurred if we scraped from say, Google images). A variety of hashtags were used, some being express emotional states, and others with no emotional bias. This was done so there would be a collection of emotions with a focus on those that we are interested in. We scraped from the following tags: #Angry, #Badhaircut, #Beatup, #Candid, #Caught, #Cosplay, #Disgusted, #Disgusting, #Emotion, #Emotions, #Everyday, #Face, #Feeling, #Guys, #Happy, #Homelessfits, #Horrible, #Model, #Mood, #Omg, #Ouch, Photooftheday, #Portrait, #Sad, #Selfie, #Surprised, #Tbt, #Untiltomorrow, #Weddingguest, and #whoops. We originally scraped 200 images for each hashtag, but many returned unusable or irrelevant images, and thus resulted in drastically fewer images for our data set.

We created a set of rough standards for identifying an acceptable image. For each image we required that there being a single, uncovered human face in the image, no words obstructing the subject's face, was reasonably close to the subject and vertical, as well as being generally clear. We also excluded overly sexual images, as well as those which featured famous or recognizable people. We found that having prior knowledge of the subject resulted in a different emotional categorization, as it was based off of their usual state and not their current or apparent state. Otherwise, no modifications were made to the images.

We had many more images for Happy and Neutral than the other categories, so we reduced these randomly to make for a more balanced dataset. After the initial cleaning, we had

500 images total. Happy images were reduced to 147 from 307, Neutral was reduced from 207 to 147, 61 Sad, 45 Other, 45 Anger, 22 Surprise, 21 Fear, and 11 Disgust. Because we found it difficult to find many disgust images (even after specifically scraping for them), we decided that this emotional category was not large enough or common enough to warrant its own category. It was instead merged with the Other category.

After labeling the images, we calculated our Intercoder Reliability to better understand our groups similarities and differences in labeling. We used the sample of 500 images and then labeled them within the same six categories. After individually labeling these images and compiling all of the results, we then re-coded our qualitative responses to numerical form. We were able to use the ReCal3 0.1 program to calculate our Intercoder Reliability Rate. Our average percent of agreement was 54%, with our highest percentage of agreement being 61.2%, and the lowest percent of agreement at 43.3%. This program used the Fleiss Kappa, Cohen's Kappa, and the Krippendorff's Alpha algorithms to calculate IRR, and ultimately based our average on Cohen's Kappa results. Based on these results, our IRR was nearly average, however, we had some discrepancies with our consistency. The group continued to self label images, rather than rely on an individual's single label.

We then continued to label the images as a team. Over the course of a few rounds of data labeling, we discussed the various interpretations of different images. The final round of labeling was done completely individually, and the results were collected. Only those images with an agreement 80% or more, were kept. Doing this allowed us to put more focus on differences between our group and the outside labelers, rather than analyzing difficult or ambiguous images. Our final dataset was reduced to 309 images.

These images were then compiled into a Qualtrics survey for outside users to label. We considered publishing this survey through Amazon's Mechanical Turk in order to collect responses since the properties that Mechanical Turk allows would have been a preferred platform that could have granted quick labeling by a wide range of people. However, using Mechanical Turk proved to be difficult to manage due to time constraints and an unfamiliarity with the system. Instead of utilizing Mechanical Turk we decided to use the Qualtrics survey system.

Created Survey:

After compiling our finalized dataset, we used the Qualtrics survey system to import our images and questions. There were a total of ten surveys, each with 33-42 images in each survey. This was done to encourage a higher rate of participation. We included four demographic questions regarding race, age, gender, and average income. We used the average tax bracket to generate our income questions, as well as used the national Census ethnic questions to determine our ethnic categories. These factors would be used to analyze and determine the impact that they have on people's ability to label emotion. Every survey was programmed to demand a response for all questions, leaving no room for uncompleted surveys. All of the surveys were open for the same duration of three days.

The survey was distributed in three major ways; snowball, convenience, and voluntary response sampling. The group distributed the surveys amongst family, friends, and acquaintances for the convenience sample. We then prompted them to continue to send the surveys to others they know in order to collect more diverse data which was our snowball sample. In addition to this, we distributed our survey on Reddit, on multiple pages for voluntary responses. Some of these pages include r/SampleSize, r/dataset, r/artificial, r/college, r/ArtificialIntelligence, and r/learnmachinelearning. The distribution of the survey URLs to Reddit pages were dispersed equally among these pages within multiple posts. From these different sources of sampling, we gathered the final responses from Qualtrics and analyzed the results. Over the course of the three days we accumulated 150 responses in total across all ten surveys. We utilized the Qualtrics "Data Reports" to examine the survey's initial results and analysis. The goal of these surveys were to obtain personal identifying factors that could lead one to have unintentional bias.

Data Analysis:

After collecting a large sample size with our surveys, we began to do an in-depth data analysis in order to draw conclusions from our results. Since we had sent out ten different surveys, we first compiled the results together in order to aggregate our data. Each row of our datset included the gender, age, income, and race of the person taking our survey. Every column was labeled with a number representing the image being assessed, and in each row held the labels each individual gave an image (Happy, Sad, Angry, Surprised, Fear, Neutral, or Other). Each label was originally encoded in numeric data, but because of the potential to have a 4 (Surprised) weight more in statistical analysis than a 1 (Happy) in traditional hypothesis testing, we decided to convert the numeric data back into categorical data.

Once the values were re-coded the data lent itself well to textual analysis looking at similarity scores among different groups of people. Once the data was compiled and cleaned, we had 150 rows of data with labels for 300 images scraped from Instagram. In order to derive insights between different groups we first separated the dataset into different clusters based on age, gender, income, and race. Once each cluster had its own data, we began to calculate the average response for each image. Because of the potentially skewing properties of the mean with numerically encoded data, we chose the mode so we could do a straight-forward textual analysis.

Going cluster by cluster for each image we determined what the most common label was for each group (Male/ Female, White/Asian, etc.) in order to compare group averages against each other. Using the mode, we then placed those data points into another dataset where each row represented a cluster, and each column had the median response label for that image from each demographic. This allowed us to easily visualize and calculate the differences in labeling among different groups.

In order to test differences among different groups and clusters of data we calculated an agreement score between two different groups. Our agreement score was based on how many images the two groups agreed upon, or more specifically: the number of images that the two groups had the same labels for

divided by the total number of images labeled by the group. For example, when calculating the differences and similarities in labeling between men and women, we would compare the average label for image 236 for women (Happy) against the average label for image 236 for men (Happy). If the labels were in agreement, we would count that in our numerator. Once we had found all the images each group had agreed on together, we were able to calculate the similarity score by taking the number of agreed labels over the total number of labels given.

While we were calculating agreement scores, we also took observations via memoing where we noticed what patterns we observed in disagreements and agreements among different groups. This was also informed by the first phase of our research where patterns we began to notice represented similar patterns in research papers we had read through in our initial topic exploration. Using these observations we were able to derive deeper insights from our data and understand certain labeling biases that may affect machine learning algorithms.

FINDINGS

From our findings we were able to observe noticeable differences among many different groups of people. While our agreement scores gave us broad data about where different groups had the most differences and similarities, our notes from memoing provided more insight into exactly why certain agreement scores came up. In this section we will walk you through all our significant results and explain which hypotheses had the most truth to them.

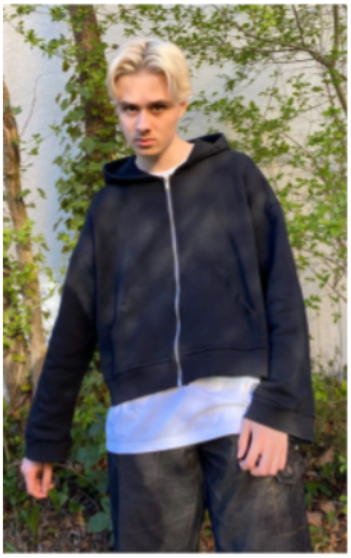

Figure 1 This image was labeled as angry by our Female respondents and neutral by our Male respondents. Our group had labeled this image angry with 80% agreement.

Hypothesis 1:

Women are going to be better at detecting subtle emotions than men.

In our preliminary research we came across a paper detailing how there appears to be no difference in emotion labeling among men and women when the pictures represent exaggerated facial expressions, but women were better at picking up subtle facial expressions (Hoffmann et al. 2010). This hypothesis was tested with over 400 subjects but was the only paper of its kind we found during our preliminary research phase. We decided to test out these findings with our own hypothesis that women would be better at categorizing subtle emotional expressions than men in order to confirm or question the findings we had read about.

Going through our agreement scores we found that men and women had a score of 86.2% between those two groups, irrespective of income, age, and race. Based solely on gender, it would seem men and women tend to have a lot of agreement when labeling emotions. Of all our categories this was the highest agreement rate we saw in our data. Our sample size for this comparison included 74 men and 76 women, so we feel confident that our results are statistically significant and represent a good approximation for the true agreement score between men and women in the United States when labeling emotions.

While there is high agreement between men and women overall in emotion labeling, there are still areas where we noticed patterns of disagreement that support our initial hypothesis. For 66% of the disagreements between men and women we found that women had rated an expression as either angry or sad while men had rated an expression as neutral. In Figure 1 we can see one such image which was rated very differently by men and by women. Women typically labeled Figure 1 angry, while men typically labeled it neutral. Our group had come to an 80% agreement before sending out our survey that this image was portraying anger with one person labeling the emotion neutral as well.

Looking at this image, you can see how some might interpret this as a neutral expression when compared to other facial expressions of anger. Since our group was in agreement on the labeling of this picture, however, we would conclude that our hypothesis seems to be true. Women do tend to label more subtle emotions accurately when compared to men. In this case accurate is measured by how closely the label mirrored our initial labeling agreement among the researchers.

Because of the lowered eyes, slightly knit brows, and the clenched jaw this expression does indeed appear to be more angry than neutral. Because these visual clues were more subtle, we can conclude that women picked up on these smaller expressions which made their labeling more accurate and attuned than men's.

Hypothesis 2:

Minorities in the United States are going to perceive more expressions as negative than White respondents.

In a paper by Matsumoto (1993) there was found to be no significant difference in emotion labeling among different ethnicities. There was, however, evidence showing that Black people tended to rate negative expressions as with a higher intensity than White people (Matsumoto, 1993). While we

did not want to add a quantification of intensity to our survey, this did cause us to wonder if there would be any link between being a minority in America and labeling emotions with more negative terms than White people. In our survey we had Asian, White, Hispanic, Native American, and Black respondents label our images in order to determine if there was a racial difference in emotion labeling within the United States.

When we got our results, we had to omit analysis for Hispanic, Native American, and Black racial categories as we did not have enough respondents to draw any statistically significant conclusions, as we had 1 Native American, 4 Black, and 14 Hispanic respondents. We did, however, have enough Asian respondents to draw some conclusions between differences in emotion labeling when compared to White respondents.

Going through our agreement scores we found that Asian respondents and White respondents had a score of 72.41% between those two groups, irrespective of income, age, and gender. Based solely on race, it would seem Asian respondents and White respondents tend to have a fair bit of agreement when labeling emotions. Since this was not as high as the agreement score between men and women, there does seem to be a noticeable difference in how these two groups label emotions. Our sample size for this comparison included 41 Asian respondents and 90 White respondents, so we can confidently say our results are statistically significant and represent a good approximation for the true agreement score between Asian people and White people in the United States when labeling emotions.

A pattern of disagreement we found when examining different emotion labeling occurred when White respondents rated images angrier and sadder than Asian respondents, who labeled those same images as neutral. This seemingly contradicts our initial hypothesis that minorities in America would rate some images as negative more often than White people, who would rate those same images as neutral. It is interesting and important to note that of those disagreements where Asian respondents labeled the image neutral and White respondents labeled the image angry or sad, approximately one-third of those images were of Asian models. Thus, it is important to note that perhaps White people might view images of people of color in a more negative way than other people of color.

Figure 2 shows us one such image which was rated as angry by White respondents and neutral by Asian respondents. When our team originally went through labeling photos in our dataset, we had labeled this image as a neutral expression with 80% agreement. There was one dissenter who had labeled the image as angry, and that person did happen to be White, which supports our findings above. Thus, we can conclude our original hypothesis that minorities would rate more images with negative labels than their White counterparts does not seem to hold true with this data collection. In fact, we can see from our results that it appears White people tend to label expressions as more negative than Asian people.

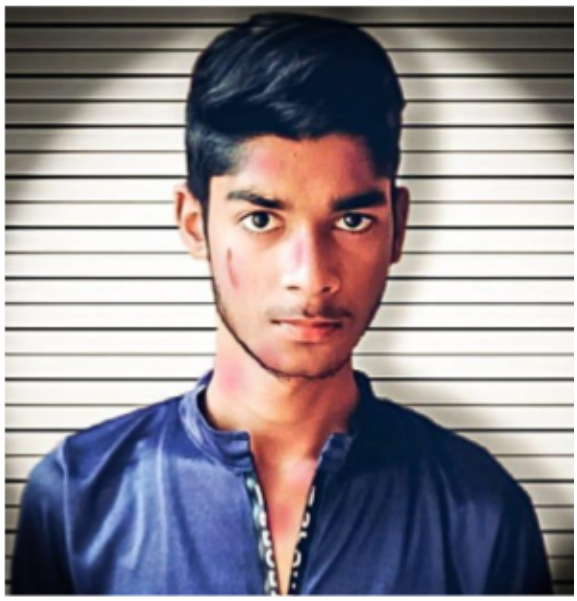

Figure 2 This image was labeled angry by White respondents and neutral by Asian respondents. Our group labeled this image as neutral with 80% agreement.

Hypothesis 3:

Older adults will label things significantly differently from younger adults and teenagers because as you grow older you are more experienced in interacting with others and interpreting their facial expressions.

In our preliminary research we came across a study by Harrigan (1984) testing the effects of task order in labeling accuracy among different age groups of children. While the study did find increased accuracy when adding an emotion recognition task preceding a labeling task, the more interesting findings our group focused on were how accuracy increased with age (Harrigan, 1984). Since this was a study involving children there was no evidence to suggest that adults would also benefit in emotion labeling accuracy with age, so we decided to test if there would be a difference between older adults and younger adults/teenagers in emotion labeling. In our survey we had 18-21 years, 22-27 years, 28-35 years, 36-40 years, and 40+ year old respondents label our images in order to determine if there was an age difference in emotion labeling within the United States.

When analyzing and calculating the agreement scores, we found that agreement among all different age groups was at a low 31% and the only consistently labeled images were images where Happy was the dominant label, which were 77% of the agreed upon labels. Surprisingly, the similarity scores between 18-21 year-olds and people over 41 years old were pretty high, with 72% agreement. We would have expected there to be a higher similarity score between the age groups of 36-40 years old and 41+ years old but at 62.07% we found this score to be lower than it was between the oldest and youngest group we surveyed.

Although there seemed to be a significant difference in labeling overall between every age group, we could not find any patterns that appeared to explain the differences. There

were no two groups with a similarity score below 60% which could have explained the incredibly low overall agreement rate among every age group. We also saw no pattern to suggest, based on our hypothesis, that older adults would have more experience in interpreting facial expressions and thus would have more agreement with the other older age groups. In fact, we found that the oldest age group of 41+ had a larger similarity score with the younger age group of 18-21 than with their closest age group of 36-40. Overall, we feel the other factors explored in our hypotheses (race, gender, income) explain more of the variation in labeling than age seems to.

Hypothesis 4:

People with lower incomes will rate more pictures neutral than people with higher incomes, as people living in poverty tend to need more expressiveness in order to accurately interpret facial expressions.

Based on previous studies which showed a link between poverty and difficulties in emotion recognition tasks, Erhart et al. (2019) wanted to see if there was a link between poverty and recognizing certain emotions, or if emotion intensity was the deciding factor. Based on their results they concluded people living in poverty needed emotion expressions with a higher intensity before they were able to accurately label images with an emotion (Erhart et al., 2019). Erhart et al. (2019) also found that while each child experienced gains in accuracy with higher intensities, children living in poverty had less gains than children who were not living in poverty. From this, we wanted to see if poverty would affect adults in similar ways. We believed that adults with lower incomes would have significantly different labels than adults with higher incomes, and that the lower income respondents would rate more images as neutral than higher earners.

Upon examining our calculated agreement scores we found that overall income and age had similar agreement scores, with all incomes agreeing only 31.04% of the time. Similarly to age, the Happy images tended to have the most agreement, with 72% of the agreed upon labels being linked to happiness. Unlike age, however, there were significant patterns indicating specific differences in emotion labeling based on income levels. We found that between the respondents with the lowest incomes of under $25,000 a year and the respondents with the highest incomes of over $100,000 there was an agreement score of only 41.38% indicating a big difference in the two income levels. When comparing between less than $25,000 and $75,000 - $99,000 we found a similarity score of only 48.28% which indicates there is a big difference between income levels and emotion labeling. Further, because the similarity score decreases with an increased wage gap, we can say that not only are there significant differences among income levels, there is a clear pattern where the greater the income disparity between two different groups, the less similarly they will label emotion images. Our sample was roughly normally distributed by income, so we can confidently say our results are statistically significant and represent a good approximation for the true agreement score between high earning people and low earning people in the United States when labeling emotions.

We found when reading over our observations of data patterns that our hypothesis regarding more neutral labels in low earners was correct. Of the labels which the lowest and highest earners disagreed on, the lower earners tended to label an image neutral where higher earners would label an image happy or surprised. We also found that low earners tended to rate more images as neutral than higher earners, with respondents whose incomes were under $25,000 labeling 34.5% of their images neutral and respondents with incomes over $100,000 rating only 10.34% of their images neutral. We also noticed that of the few images rated at neutral by the top earners, the low earners were more likely to rate that neutral image as angry. This might lend some support to the theory that those who are exposed to more negative emotions on a daily basis are more likely to perceive negativity in facial expressions. Since those earning lower wages are normally in tougher conditions and are looked down upon by others they may be more likely to perceive negative emotions in the faces they see every day. This would require further research in order to conclusively determine people living in poverty see expressions as more negative, but we can see from our results that there is significant evidence to support our original hypothesis.

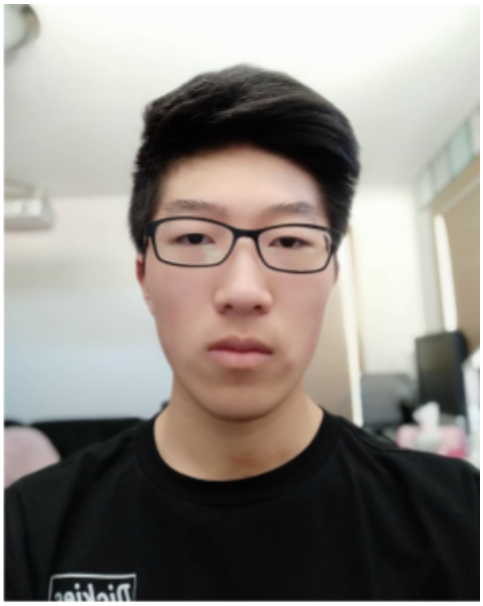

Figure 3 This image was labeled as angry by respondents earning under $25,000 a year and neutral by respondents earning over $100,000 a year. Our group labeled this image neutral with 80% agreement.

An example of a differently labeled image can be seen in Figure 3. This image was labeled as neutral by respondents making over $100,000 a year and was labeled as angry by respondents making under $25,000 a year. Our team labeled this image as neutral with 80% agreement, and no one in our group had felt this image represented anger. Using our label as a base for accuracy we could conclude that higher income earners are more accurate in emotion labeling as they most closely labeled the images like the labels our group had

agreed to. Regardless of accuracy it is clear that higher income earners perceive emotions very differently from lower income earners and it appears very likely that less income lends to more perceived negativity in facial expressions.

CONCLUSIONS

Through our research and the presentation of our results, we hope to provide more information and evidence on the effects of specific identifying factors and their influence on emotional labeling. Machine learning is becoming an increasingly large field which means there will be a large influx of uninformed users, however, we aim to help to bridge the gap through our research. By promoting more academic research and work on machine learning topics such as facial recognition and the data labeling involved, academic researchers as well as lay people can be more aware of these potential biases. Information scientists and researchers can make better claims to accurate and unbiased data with these influences of age, gender, income, and race in mind.

This research paper hopes to create more dialogue surrounding emotional labeling and potential biases surrounding data labeling and machine learning. While we can never perfectly represent the human experience of emotion in digital form, the more comprehension of the dangers that come with biased datasets and labeling will help present and future facial recognition algorithms to more accurately represent emotions. The continuing evolution of artificial intelligence and facial recognition will forever be tied with cultural and societal influences. Labeling data for algorithms in an unbiased way will be a persistent challenge as external identifying factors will generate different responses to emotions. Understanding our own underlying biases allow for more dataset models to be monitored for those individualistic influences more closely.

With the knowledge presented in this paper, an effort can now be made towards a decision on what the 'true' label for these should be. Seeing the distinctions, further studies can be done to inspect the causes of these, be it a better ability to discern subtle emotions or rather two differing and correct opinions. Perhaps along with these, the dataset could be constructed through donation and collect the true emotion represented in the image. This could vastly improve the studies, giving us the actually correct emotional label.

Beyond the labeling of data, we hope future work is also put towards the dataset creation. In our work, we encountered many issues in even generating the images. With more personal information infiltrating social media and other platforms, it is important to emphasize the importance of personal data privacy. In a recent AI scandal involving Flickr, millions of people's images were used to train AI facial recognition algorithms for security purposes (Hill and Krolik, 2019). The lack of consent in this process makes for a dangerous situation. Our project and research has been aimed to gather ethically sourced data, with using public images for non-profit research, but not all projects can say the same. For this reason, we believe that the future of AI and facial or emotional labeling will become more reliant on authorized image donation, or some shared database of public images.

REFERENCES

Barrett, L. F., Lindquist, K. A., Gendron, M. (2007). Language as context for the perception of emotion. Trends in cognitive sciences, 11(8), 327-332.

Brody, L. R., Hall, J. A. (2010). Gender, emotion, and socialization. In Handbook of gender research in psychology (pp. 429-454). Springer, New York, NY.

Ekman, P. (1973) Cross-cultural studies of facial expressions. In P. Ekman (Ed.), Darwin and facial expression: A century of research in review (pp. 169-229).

Elfenbein, H. A., Ambady, N. (2003). When familiarity breeds accuracy: cultural exposure and facial emotion recognition. Journal of personality and social psychology, 85(2), 276.

Erhart, A., Dmitrieva, J., Blair, R. J., Kim, P. (2019). Intensity, not emotion: The role of poverty in emotion labeling ability in middle childhood. Journal of experimental child psychology, 180, 131-140.

Fox, E., Lester, V., Russo, R., Bowles, R. J., Pichler, A., Dutton, K. (2000). Facial expressions of emotion: Are angry faces detected more efficiently?. Cognition emotion, 14(1), 61-92.

Griffiths, P. E. (1997). What emotions really are: The problem of psychological categories. Chicago: University of Chicago Press. 137-169.

Guynn, J. (2015, July 1). Google Photos labeled black people 'gorillas'. Retrieved from https://www.usatoday.com/story/tech/2015/07/01/google-apologizes-after-photos-identify-black-people-as-gorillas/29567465/

Harrigan, J. A. (1984). The effects of task order on children's identification of facial expressions. Motivation and emotion, 8(2), 157-169. https://doi.org/10.1007/BF00993071

Hoffmann, H., Kessler, H., Eppel, T., Rukavina, S., Traue, H. C. (2010). Expression intensity, gender and facial emotion recognition: Women recognize only subtle facial emotions better than men. Acta psychologica, 135(3), 278-283. https://doi.org/10.1016/j.actpsy.2010.07.012

Kashmir Hill and Aaron Krolik. 2019. How Photos of Your Kids Are Powering Surveillance Technology. (October 2019). Retrieved May 2020 from https://www.nytimes.com/interactive/2019/10/11/technology/flickr-facial-recognition.html

Matsumoto, D. (1993). Ethnic differences in affect intensity, emotion judgments, display rule attitudes, and self-reported emotional expression in an American sample. Motivation and emotion, 17(2), 107-123. https://doi.org/10.1007/BF00995188

Paiva-Silva, A. I. d., Pontes, M. K., Aguiar, J. S. R., de Souza, W. C. (2016). How do we evaluate facial emotion recognition? Psychology Neuroscience, 9(2), 153–175.

Purdy, M., Zealley, J., Maseli, O. (2019, November 18). The Risks of Using AI to Interpret Human

Emotions. https://hbr.org/2019/11/the-risks-of-using-ai-to-interpret-human-emotions

Tweedie, R. L., Mengersen, K. L., Eccleston, J. A. (1994). Garbage in, garbage out: can statisticians quantify the effects of poor data?. Chance, 7(2), 20-27.

Zeng, Jiabei. "Facial Expression Recognition with Inconsistently Annotated Datasets." The European Conference on Computer Vision (ECCV), 2018, pp.

Zhang, J., Sheng, V. S., Li, Q., Wu, J., Wu, X. (2017). Consensus algorithms for biased labeling in crowdsourcing. Information Sciences, 382–383, 254–273. https://doi.org/10.1016/j.ins.2016.12.026