Developing AI Data: Minimizing AI Algorithm Bias for Facial Analytics

---

A Capstone Project
Presented to
the Graduate School of
The University of North Carolina Wilmington

In Partial Fulfillment
of the Requirements for the Degree
Masters of Science in Computer Science and Information Systems
Computer Science

---

by
Lindsay Kness
May 2023

Proposed to:
Dr. Karl Ricanek, Committee Chair
Dr. Minoo Modaresnezhad
Dr. Laavanya Rachakonda

Table of Contents

Chapter 1

Introduction

The rise of data drives the rise of Artificial Intelligence (AI). In today's world, data is ubiquitous; however, there are gaps in the data that provide potentially significant challenges in building AI models. This work examines the development of a face database with data labeled for twelve facial attributes. This database is unique; it provides facial attributes labeled via crowdsourcing for a set of faces validated by country of origin. This effort is the first open-source dataset that observes the effect bias has in data labeling and attempts to reduce or offset the bias to provide a larger volume of accurate data for machine learning. The dataset captures images of faces from industrial and emerging countries whose facial attribute labels are validated and then separately organized by country of origin.

This capstone project aims to create a process for labeling data to produce a usable country of origin image dataset for machine learning. The goal is to implement this procedure using a crowdsourcing website, The Hive AI (Hive), to accurately label thousands of images. Using this platform will improve the efficiency of this process while decreasing discrepancies in data.

Chapter 2

Background Research

"Gender Shades," a paper by Joy Buolamwini and Timnit Gebru, investigates the inconsistencies in data regarding gender and skin tone [1]. The authors created a dataset from three African and three European countries, grouping the subjects by gender and skin tone. They then evaluated three commercial gender classification systems using the dataset. The paper states, "While the companies appear to have relatively high accuracy overall, there are notable differences in the error rates between different groups [2]." The companies in question are Microsoft, IBM, and Face++. Each performed best on lighter males and had the highest error rate on darker females. They state, "Automated systems are not inherently neutral [2]" It is vital to standardize labeling and ensure quality control for companies building AI that research or analysis of humanity can use.

There is a need for a dataset with a range of regional demographic data to encourage others to build fair and consistent facial recognition algorithms. Relying on a platform to help label images from different demographics, regions, and cultural backgrounds can contribute to the movement of minimizing algorithm bias and creating a more expansive global dataset. This work aims to help others by providing

a robust process for using crowdsourcing to label data attributes. **In this work, I will identify the challenges and gaps in performing crowdsourcing for the purpose of generating face attribute labels.**

Additionally, this dataset created can be used to provide machines with the data necessary to enhance their face-based machine learning algorithms. The goal is to contribute to the movement of collecting more precise data, which is instrumental in increasing AI model performance while reducing bias. Data labeling is a multi-step process. The first step is collecting raw data that needs to be labeled. In this instance, the raw data is facial images from specific West Asian and North African countries. The next step is to take the raw data and develop a process to apply labels to the images. Data labeling is "a process of identifying raw data (images, text files, videos, etc.) and adding one or more meaningful and informative labels to provide context so that a machine learning model can learn from it [3]." Outsourcing the data labeling to an established platform eliminates the need to hire temporary employees, reduces workload, and eliminates the need for an in-house data labeling team that could be costly. It allows data scientists to channel their efforts to tasks requiring specific skill sets [4].

Though it is possible to label each of these images manually, it would not be efficient on a dataset that contains a large number of records; It would be "expensive" in time and resources. Additionally, utilizing a single person or a team of labelers that are assigned their own set of images would result in errors to the labels.

For example, it would take four eight-hour days, 33 hours, to label 1,000 images, with twelve different labels taking ten seconds for each label. The objective is to create processes using a crowdsourcing platform to improve efficiency while maintaining accuracy.

When validating images, a standard method must be developed and upheld.

If only one person labels an image, then there would be no way to validate that they labeled it correctly. With only one data point, there would need to be a high level of confidence in the work produced by the individual that they would label every image correctly. Alternatively, a secondary reviewer must be part of the validation step. If person 'A' labels and person 'B' validates the labels, then the image is independently verified, and said to be correctly labeled. However, if person 'B' disagrees with person 'A', there is a discrepancy. It would then be considered inconclusive, "not leading to a firm conclusion; not ending doubt or dispute [5]" or not having an answer based on the selections. It beckons the question of how we treat that data. Where will it go? Should there be a third validator, or should it be thrown away?

By using a crowdsourcing platform, there is an aim to solve two main problems. **The problems are: 1 the efficiency of labeling the data:to speed up the process of labeling face images for attribute labels, and 2 ensuring that the labels are correct.**

2.0.1 Crowdsourcing

"Crowdsourcing is the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community rather than from traditional employees or suppliers [6]." The word crowdsourcing is a combination of crowd and outsourcing. In contrast to outsourcing, crowdsourcing usually does not require specific skills or backgrounds but involves more public groups. There are many crowdsourcing categories that aim to complete a task. Crowdfunding, opinion seeking, and manual tasks are three common types. Crowdfunding is "the practice of funding a project or venture by raising many small amounts of money from a large number of people, typically via the internet [7]."

Some famous companies are Kickstarter and GoFundMe [8]. Opinion seeking is when crowds provide suggestions to improve products or services. An example is SurveyMonkey or MyStarbucksIdea.

Another kind of crowdsourcing is manual tasks, which are divided into subgroups, micro and macro tasks. Manual tasks are when companies set up manageable or micro tasks and outsource them on a platform for others to complete. These tasks usually do not require a specific skill set. A popular platform that does this is Amazon Mechanical Turk (MTurk) [8]. Since the project requires data labeling, the choice of platforms can be narrowed down to a few key players, Amazon Mechanical Turk (MTurk), a popular crowdsourcing platform, and The Hive AI (Hive), a company that focuses on data labeling, and Isahit, a french company that has an image annotation use-case. After comparing the three platforms, The Hive AI best supports the objectives of this project.

Hive markets its product as an excellent system for "creating training datasets for machine learning models [9]." Each client is assigned a manager to ensure that the process goes smoothly. It is scalable, with an astounding two million workers generating tens of millions of accurate labels daily [9]. Their mandated consensus and high worker accuracy make for excellent data quality. They are cost efficient compared to other platforms or manually performing the labor [9]. Hive also has a focus on deep learning, and "Deep learning enables human-like interpretation of video, image, text, and audio — allowing businesses to automate more 'intelligent' processes and implement new processes not scalable with manual labor [10]."

Hive has a set pay scale, so unlike MTurk, there is less flexibility in the payment model and, therefore, more predictability in the cost to use the platform. Isahit does not list an average cost or payment model on their platform. The website does state that they are the most ethical option and pay. Both Hive and MTurk platforms have

built-in checks to ensure that the workers are doing an excellent job and are still qualified to work [9, 11]. Though both platforms address the issues associated with the inconsistencies of data labeling and the errors involved when relying on human input, Hive works better for this project as these checks are available for every project without additional costs, unlike MTurk. Isahit does not list that they include built-in checks but does state that they help train their workers "more than 3 hours of training per project [12]."

MTurk's pricing is tiered and based on the task's complexity and additional requests from the project owner. With additional payment, the project owner can request a specific demographic of users that they can request to complete their work, task, or survey [11]. With Hive, due to the standard payment, it will be lower cost to use, and workers will be less likely to choose a task based on pay, eliminating the possibility of a competitive workforce that may happen with tiered pricing.

Isahit has a diverse group of women workers stating, "Our workforce is multicultural, coming from 44 different countries, speaking more than 16 languages, with different academic backgrounds and professional experiences [12]." Similarly, Hive has a large population of spanish speaking workers. Though the diversity of the workers could have been a pro for this project it also adds another variable to consider. The fear of things getting lost in translation should be considered, Isahit having a workers from a multitude of languages could increase the translation mistakes. That weighs more heavily with them also being a French company and the possibility of miscommunication could increase. Hive's team is based in California and recommends to use one other language for translation. The decision to use Hive comes down to the simplicity of use, as their company's focus is data labeling, and their overall cost structure, which is lower than competitors for quality of work, and the pricing is more straightforward.

Chapter 3

Methodology

3.1 Methodology and plan

The procedure I am proposing will use a crowdsourcing platform to decrease the time spent labeling images to create the data set. This section discusses the steps taken to collect data that will need to be labeled. Additionally, I will provide a review of the tools Hive utilizes to help verify that the labels are accurate. This section also discusses the twelve labels used for each image and the corresponding definitions.

3.1.1 Creating the Dataset

To replicate the collection of raw data, start by creating user-defined queries to gather celebrity names from a particular country. Query Google Knowledge Graph with the user-defined queries to compile a list of names of famous people for each of the regions being analyzed. These famous people can be actors, celebrities, socialites, politicians, etc.

Using an automated approach, search each famous person's name on Google and determine their gender and country of birth. Then, confirm that the individual's

origin is the same as the country of interest; if it is not, ignore the person and go to the next on the list. When the individual is determined to be an acceptable candidate for the dataset, if there is an available image of that person, it is downloaded from Google Knowledge Graph Panel, and serve as the reference image for the image verification step.

Next, using the list of names of famous people for which you obtained a reference image, use an automated web searching tool to collect images of those famous people from Google search image tab. Run a face detector on each of the collected images to validate whether each image has a face or not. If not, the process will ignore or remove that image. If a face is detected, continue to the next step, checking if the width and height of the bounding box containing the detected face is greater than 64 pixels; this step guarantees a minimum size of the face image. If not, the images should be ignored or removed. If the image passes, it will be compared to the verification image using a face recognition service. The system will then check that the similarity score is greater than an empirically derived threshold. Again, if the image does not pass, it will be removed or ignored. Alternatively, a manual comparison of the collected and verification images can be performed. Finally, download the passed image to form a point in the dataset. Figure 3.1 illustrates this process. The dataset has labels by country, and each image has a name with the country, the ID, and the gender. *Note: A region dataset will have multiple countries' data in it.*

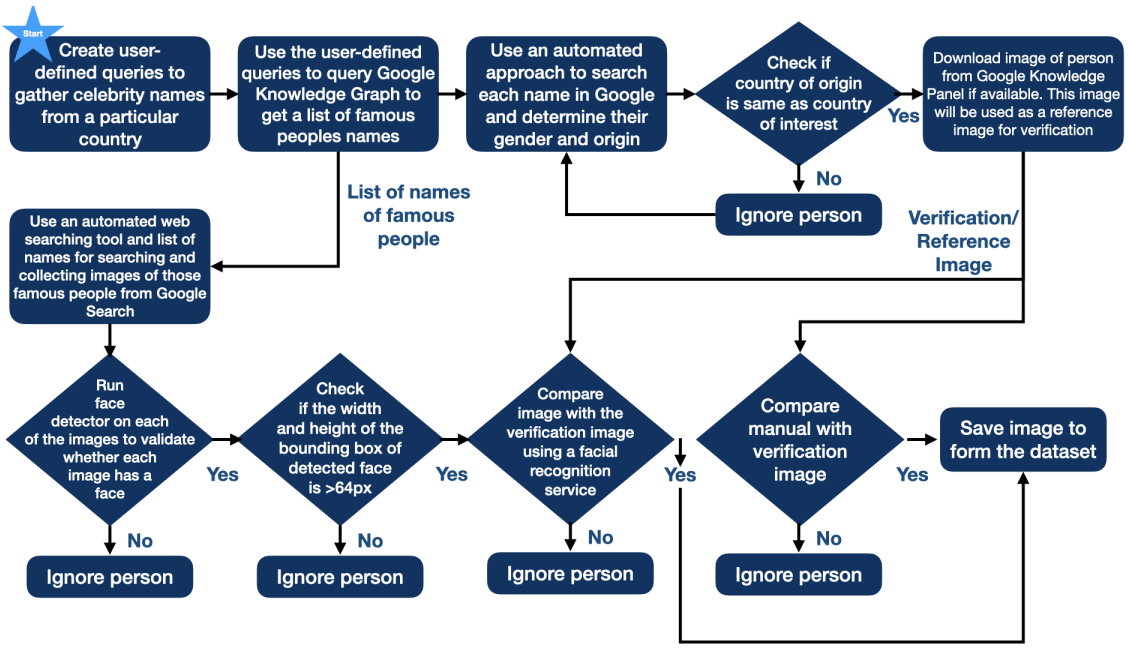

Figure 3.1: Data Collection Overview

3.1.2 The Hive AI Process

Hive has a built-in process to address the issue of inaccuracy and help users get the best data. These tests are there to aid in identifying, and removing a worker who is making too many mistakes when labeling. The Hive has three tasks in their system which are defined below, the qualifier, the real task, and the honeypots. "The qualifier is a set of instructions and qualification tasks that a worker must successfully pass to begin completing work on the job [13]." If they pass, they will start labeling real tasks, the product labels we need. While a worker goes through the tasks, there will be honeypot tasks to regulate workers to ensure they are still performing at a high enough standard to include their responses in the project. "Honeypot tasks are a subset of real tasks that are hand-labeled by the job creator. These tasks are then mixed into the worker's job feed along with real tasks. A worker's performance on the honeypot tasks will dictate whether they are allowed to continue working on your

job. If a worker's accuracy drops below the acceptable threshold, the worker will be temporarily banned from the job and will have to re-qualify to continue working on the job [13]." See Figure 3.2 for illustration of the honeypot examples for this.

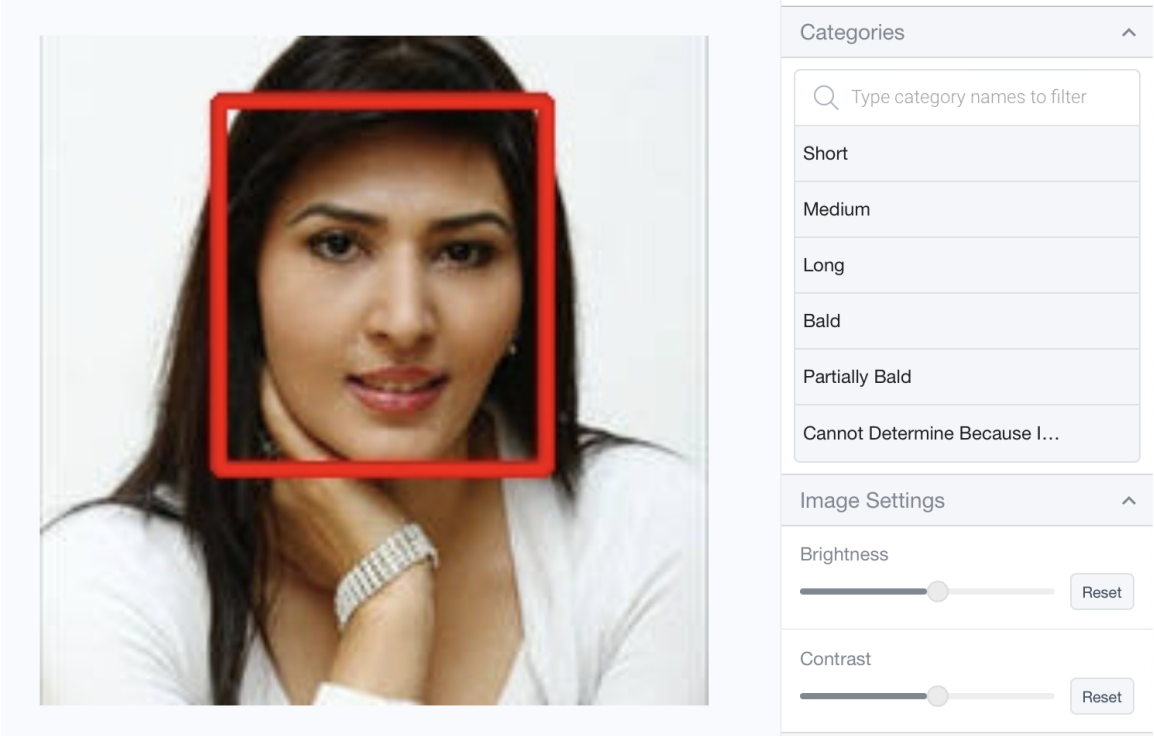

Figure 3.2: Hive honeypot task example

An important consideration when using Hive is addressing encounters with language differences. Hive has a large community of Spanish-speaking workers. Because of this, they provide an area to give instructions, and labels, in other languages. The Hive AI recommends that task instructions and information are provided in dual languages (English and Spanish in our case). Diminishing language barriers assists in reducing confusion, increasing effectiveness, and decreasing mistakes, as unequal information can lead to unequal outcomes.

3.1.3 Labels

Twelve predefined physical attributes were used for labeling the images. They are age, hair color, hair length, mouth open, glasses, expression, eyebrow thickness makeup, eye makeup, lipstick, beard, mustache. The twelve labels, derived by the face aging lab team, are theoretically created by an earlier research paper (not identified) and narrowed down through agreements between the team and the ORNL team they were working with. Each label listed is also defined in the chapter Hive Labels, and contains a reference for the figure. These labels are for the project; every image will have a descriptive label when applicable.

Table 3.1: Hive Labels/Categories.

Begin of Table		
Category	Definition/description of the task	Category Sub-labels
Age	Determine the age category of the person. If the person in the photo appears to be between 0 and 12 years old, select Child. If the person in the photo appears to be between 13 and 19 years old, select Teen. If the person in the photo appears to be between 20 and 29 years old, select Young Adult. If the person in the photo appears to be between 30 and 50 years old, select Middle Aged. If the person in the photo appears to be older than 50 years old, select Older.	Child, teen, Young Adult, Middle Age, Older

Continuation of Table 3.1		
Category	Definition/description of the task	Category Sub-labels
Expression	If the person in the photo is smiling, select Smiling.If the person in the photo appears to be frowning or angry, select Frowning/Angry. If the person in the photo has a blank expression (the facial muscles are relaxed), select Blank Expression. If the person in the photo is making a sad expression, select Sad.	Smiling, Frowning/Angry, Blank Expression, Sad
Hair Color	Look only at the hair on the head. If the person in the photo has dark brown or black hair, select Dark Brown/Black. Look only at the hair on the head. If the person in the photo has light brown or blonde hair, select Light Brown/Blonde.Look only at the hair on the head. If the person in the photo has gray or white hair, select Gray/White. Look only at the hair on the head. If the person in the photo is bald or if the hair is not visible, select Not Visible/Bald.	Dark Brown/Black, Light Brown/Blonde, Gray/White, Not Visible, Bald

Continuation of Table 3.1		
Category	Definition/description of the task	Category Sub-labels
Hair length	If the person's hair does not touch the collar or the shoulders, select Short.If the hair is long enough to touch the collar or the shoulders but does not fall significantly below the shoulders, select Medium. If the hair falls past the shoulders, select Long. If the person in the photo is completely bald, select Bald. If the person in the photo is partially bald, select this option. If not enough of the hair is visible to make a judgment about the length (if, for example, the hair is tied up or covered), select Not Visible.	Short, Medium, Long, Bald, Partially Bald, Not Visible
Eyebrows thickness	Does the person have thick/bushy eyebrows? If the person has eyebrows that are thicker than normal in terms of width, length, and density, select Yes. If the person has eyebrows that are thin or normal, select No.	Thick, Thin/Normal

Continuation of Table 3.1		
Category	Definition/description of the task	Category Sub-labels
Open mouth	The mouth is open if there is a gap between the lips such that the teeth, the tongue, or the inside of the mouth are visible. The mouth is closed if there is no space between the lips or minimal space between the lips. The teeth, tongue, gums, and inside of the mouth are not visible. The mouth is covered if not enough of it is visible to determine whether it is open or closed.	Open, Closed, Covered
Glasses	Is the person wearing eye glasses (spectacles) or sun glasses? Selection options are "yes" or "no".	Glasses, No Glasses
Makeup	If the person in the photo is wearing visible foundation, blush, bronzer, or other makeup on the skin, select yes. If the person does not appear to have any foundation, blush, bronzer, or other makeup on the skin, select no. Remember that the person may be wearing eye makeup but not makeup on the skin, so focus only on the skin.	Makeup, No Makeup

Continuation of Table 3.1		
Category	Definition/description of the task	Category Sub-labels
Eye makeup	How would you rate the level of eye makeup? If the person in the photo is wearing eye makeup, select Yes. If the person in the photo is not wearing eye makeup, select No.	Eye Makeup, No Eye Makeup
Lipstick	For this task only report/examine on the face/person that is in the red box. Notes: Focus on the lip region and determine if a person has applied lipstick or gloss. Lipstick colors varies from dark to nude. Look for variation in lip shade compared to the face.	Lipstick, No Lipstick
Mustache	If the person has a mustache (hair above the upper lip), select Yes. If the person does not have a mustache (hair above the upper lip) or the region has been clean shaven, select No.	Mustache, No Mustache
Beard	If the person in the photo has a beard, select this option. If the person in the photo has a goatee, select this option. If the person has stubble or five o'clock shadow on the jaw area, select this option. If the person in the photo has no facial hair on the jaw or chin area, select this option.	Beard, Goatee, Stubble, None/Clean Shaven
End of Table		

3.1.4 Creation of Hive Projects

When establishing these projects, they are set up as data labeling, "a specific set of instructions for a term [13]." Each project is titled as the label characteristic being applied to the images. Refer to the labels that are now included in the narrowed down process in chapter 6, Hive Labels. For the older projects, there was not a section to add another language in Hive, so each project was titled with the Spanish translation and then the English in brackets.

A task is a single row of data (photo, video, audio or clip) that workers will label. A worker is a person who labels data on the platform. Judgements are a one worker-provided answer to one task. Real tasks are the production data that you need labeled to create the dataset. Next is creating classes, "The classes list is the list of options from which workers will select categorizations for each task [13]." Again, all information is translated into Spanish.

A qualifier is a "set of instructions and qualification tasks that a worker must successfully pass to begin completing work on the job [13]." During the set up, each project must be set with a minimum qualifier accuracy which "is the percentage of qualifier tasks a worker is required to complete correctly to qualify for your job and receive honeypot and real tasks [13]." It is recommended to have a minimum qualifier accuracy between .75-.90 depending on the complexity of the qualifier and the overall job. For my work the minimum is set at .80.

Once a worker is qualified for a project they will have honeypot tasks to complete to qualify before their answers are accepted. When setting up the honeypots, the settings are there to encourage workers to stay focused and on task. They function as checks to ensure the workers are actually qualified for the task at hand, and not making random selections on the label. The Minimum Honeypot Count "is

the number of honeypots that will appear immediately after the qualifier test and before a worker is able to start real tasks. If a worker's honeypot accuracy falls below the Minimum Honeypot Accuracy, they will be temporarily banned from the job and must re-qualify [13]." Minimum Consensus Size is "the minimum number of worker judgements that must agree before a task's status finalizes and the task completes [13]." Maximum Moderation Count is "the maximum number of worker judgments a task can receive. If a task's judgements still do not consent with one another, the task will complete with an inconclusive status [13]." It is Recommended to have a setting at 2/3. These are valuable tools the Hive employs to assist in obtaining an accurate dataset for this project.

Chapter 4

Experiments

My approach to developing a procedure for data labeling has gone through three iterations. A data-scraping process collects the images to evaluate during these experiments. This process explained in depth in Chapter 3, Methodology, will collect images from Google following a user-defined query. To be added to the dataset, the image must follow the parameters. A system compares the images to validation images, and adds them to a dataset for evaluation. The first and second experiments evaluated images from West Asia as candidates for the dataset. The third experiment evaluated images from North Africa as candidates for the dataset. After each experiment, I re-evaluated the procedure and the Hive application experiment based on my results and the feedback from both the Hive company and my committee. The second experiment uses the failed (inconclusive) data from experiment one. The third experiment is an entirely new set of data from North Africa. With each experiment iteration and reflecting on lessons learned, the process for labeling was refined. The newly improved process is used for labeling this last set of data. North Africa image labeling is my final project, and the results help to support the assertion that the process developed dramatically improves the development of a crowdsource labeled

face dataset by reducing the occurrences of inconclusiveness using the tested protocol in this work.

4.1 West Asia Experiment

West Asia Experiment 1 consists of data collected via the process mapped out above and has 5,723 images from ten countries. Those countries are Azerbaijan, Egypt, Georgia, Kazakhstan, Kyrgyzstan, Lebanon, Maldives, Saudi Arabia, Tajikistan, and Turkmenistan. See figure 4.1 for a world map of the 10 countries.

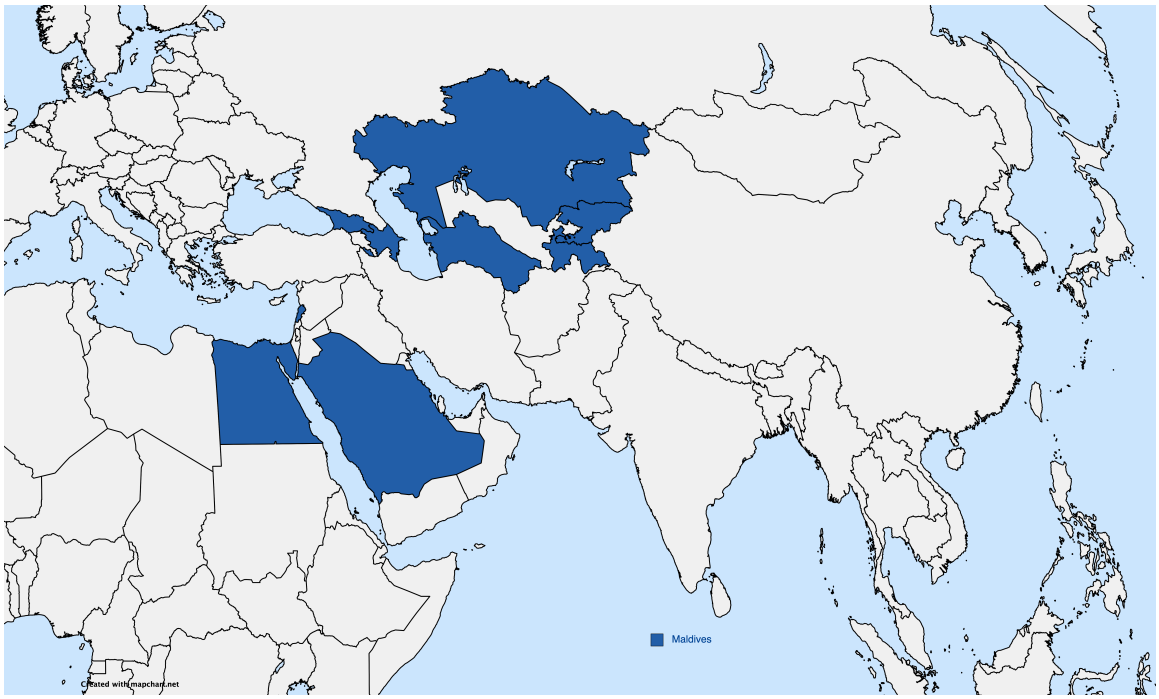

Figure 4.1: World map of West Asia Countries

In the setup for each project, the title is in Spanish and in brackets the English translation because the hive workers are located in spanish speaking regions. Where possible I used binary (two) classes for a label as it simplifies, reduces errors for labeling. Workers can become confused with many possible labels. (Bangs, lip-

stick, mustache, and glasses all have binary options, 'Yes' or 'No'.) However, age has the following options available for selection: 'Child', 'Teen', 'Young adult', 'Middle age', 'Older'; expression has a choice of 'Smiling', 'Angry or Frowning', 'Sad', or 'Blank expression'; Makeup and eye makeup have 3 each, 'Heavy', 'Light', or 'None'. Hair length is 'Short', 'Medium', 'Long', 'Bald', 'Partially bald', or 'Cannot be determined'; Hair color options are 'Black', 'Brown', 'Blonde', 'Grey or White', 'Red', 'Bald', 'Other', or 'Cannot be determined'; The beard options are: 'Beard', 'Goatee', 'Stubble', or 'None/clean shaven'; eyebrow thickness is 'Yes', 'No', or 'Not visible'; the open mouth category is described as 'Open', 'Close', or 'Covered'.

In the qualifier tab the accuracy is set to 80% and in the honeypot tab the count is set to 10 with a minimum accuracy of 90%. In real tasks, the minimum consensus size[1] is 4 and the maximum worker count[2] is 5. All images are loaded into the projects with the exception for the categories of makeup, eye makeup, and lipstick which only has females and the categories of beard, and mustache which only has males. Included in the set up is a set of worker instructions. Each instruction tab is in both English and Spanish, an example is found in Figure 4.2.

[1]the minimum number of worker judgements that must agree before a task's status finalizes and the task completes

[2]the maximum number of workers the task will be sent to before it is return as inconclusive

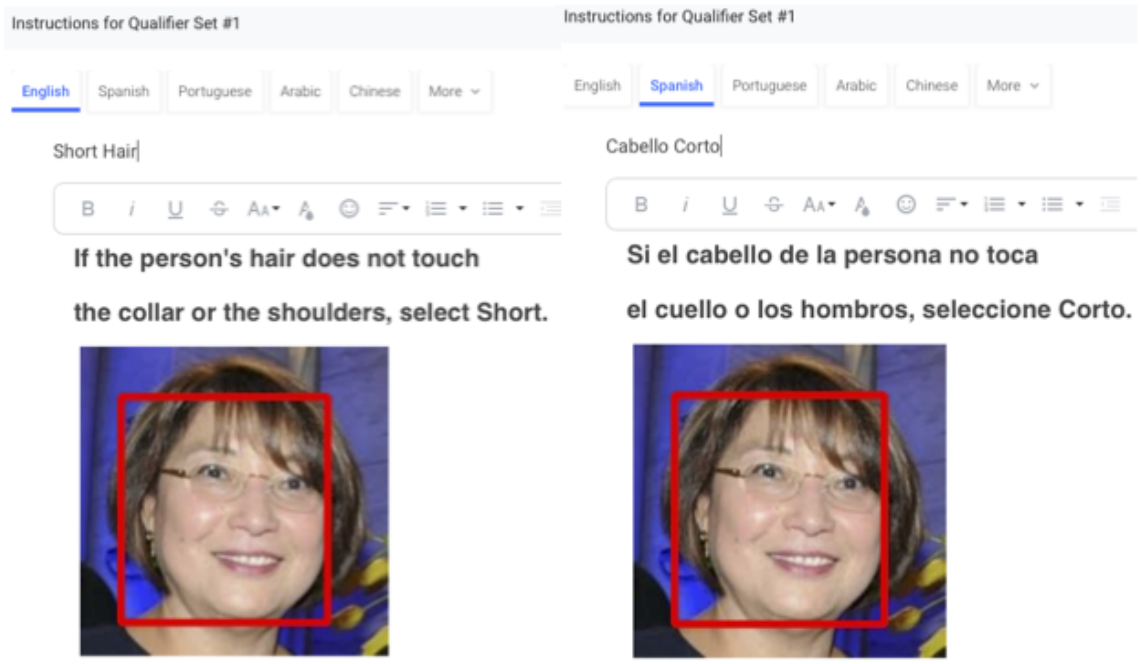

Figure 4.2: Hive instruction example illustrating hair length

Their purpose of creating these standards is so they can be re-used in each project to minimize the time spent setting up and for consistency among the various projects. The honeypots are made the same way; every list has the same honeypots already labeled and were used when duplicating the project.

To begin labeling a new region, all that is required is to copy the project, change the name of the region (WA for West Asia), then add the data.

Percentage of Inconclusive by Attribute	
Attribute Class	**Percentage of Inconclusive**
Mouth Open	6%
Hair Color	28%
Hair Length	21%
Eyebrow Thickness	26%
Makeup	37%
Eye Makeup	34%

Figure 4.3: Inconclusive results from experiment 1

After this experiment was completed, there was a suggestion from the Hive staff that we change the accuracy, the real task and the honeypots.

4.2 West Asia Experiment 2.0

West Asia 2.0 is the second experiment completed. A copy is constructed of the inconclusive data from the previous run and is now a new project for re-evaluation. Based on the feedback from the staff at The Hive AI, there are a few modifications made to improve the data labeling process. The modifications to all attributes include a change to the qualifier slide in the instructions, changing the honeypot tasks and

altering the consensus. The suggestion to make it as simple as possible for the workers led to the creation of a new qualifier for each class or a set of commonly confused classes. Tasks that previously received an inconclusive result were good candidates to include in the qualifier. Another adjustment was to expand the qualifier instruction sets and replace images with samples from the dataset.

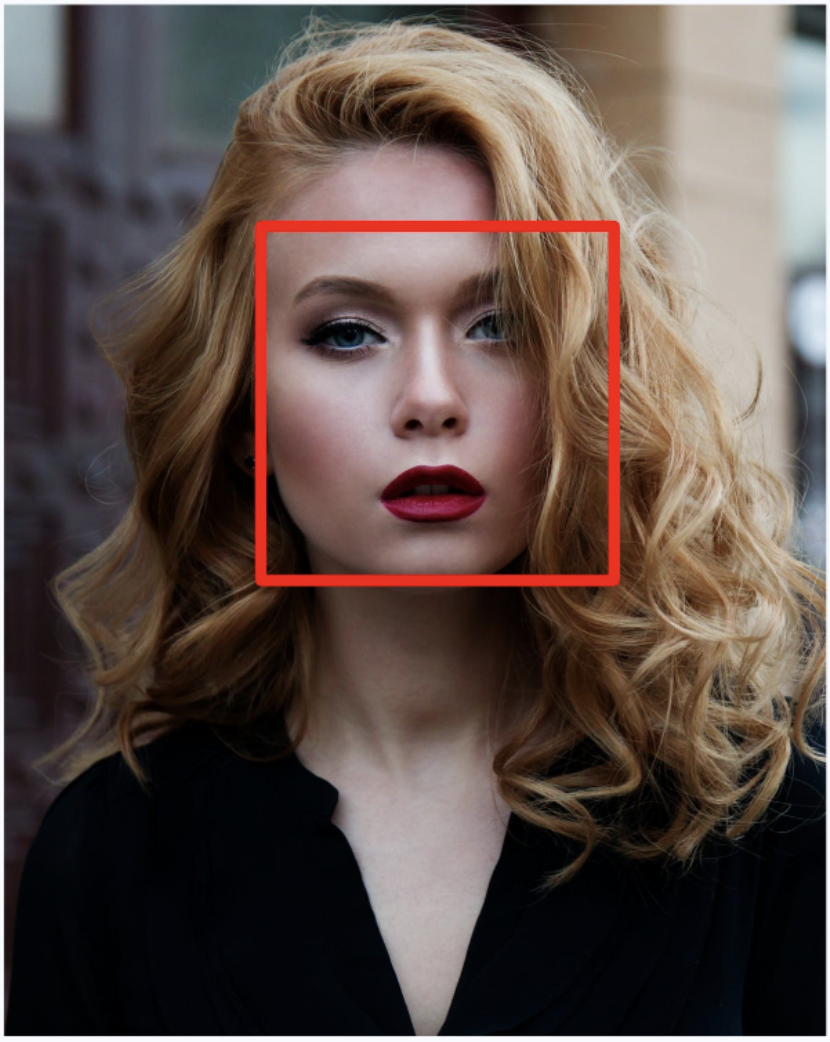

Figure 4.4: An example of a qualifier task for eye makeup before changing to regional examples

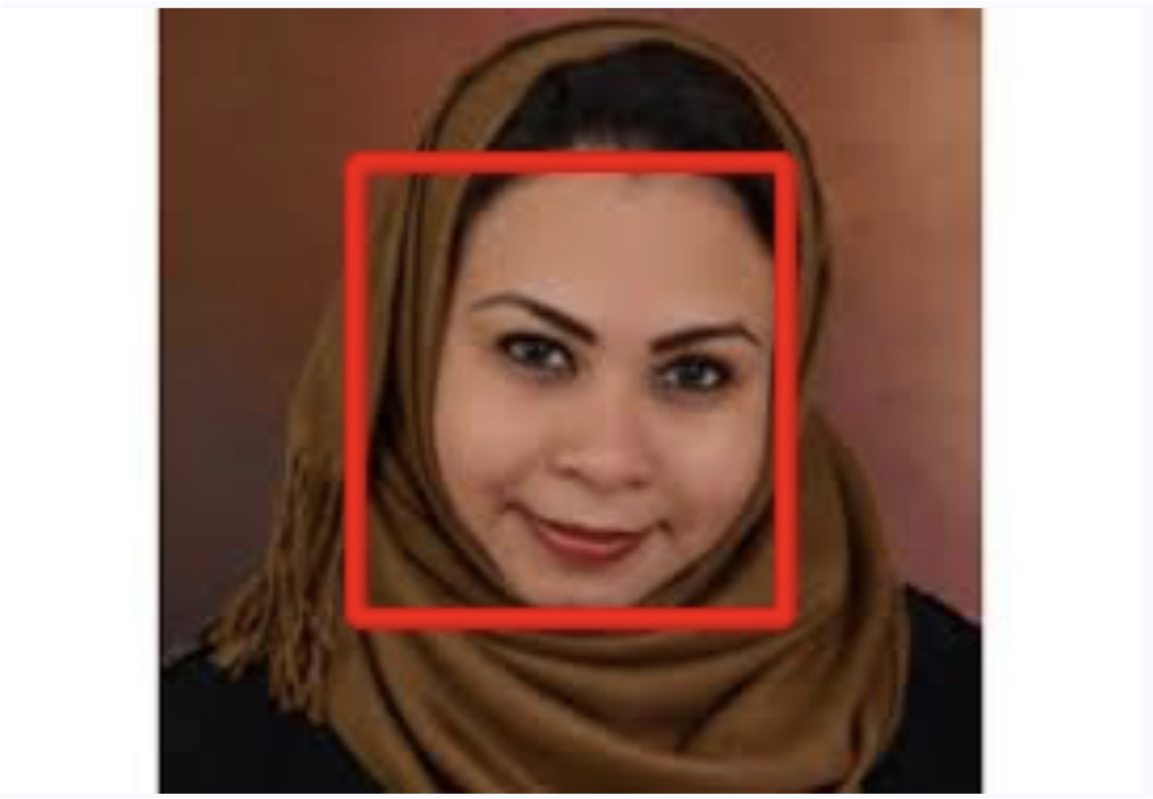

Figure 4.5: An example of a qualifier task for eye makeup after changing to regional example

There were also changes made to the honeypots that were from regions other than the one presently being evaluated by the workers. These new honeypot images were less likely to be detected as such as they appeared more similar to the images in being labeled from the same region. This change forces the workers to pay attention and check the right box. There was a suggestion to replace honeypot images with images from the dataset originally returned with inconclusive labels. Most of the inconclusive tasks would have consented with a minimum consensus size of 3 and maximum worker count of 5 (instead of 4/5). Another way to gauge that the honeypot tasks are the perfect amount of difficulty is that the speed at which the workers complete a task (real task) and the speed at which they complete a honeypot should be equal. If the honeypot task is taking longer, there is a chance that it is a more difficult

task, or the workers might know that it is a honeypot and will spend more time labeling it correctly, so they avoid getting banned. Figure 4.6 provides a comparison of real-task versus honeypot speed. The figure indicates that the honeypot speeds are much higher than the real task speed, meaning workers take longer to complete the task. The difference shows that the honeypots are either too complex compared to the real tasks or the workers are taking their time on these tasks to avoid making mistakes and being banned as they are easy to identify.

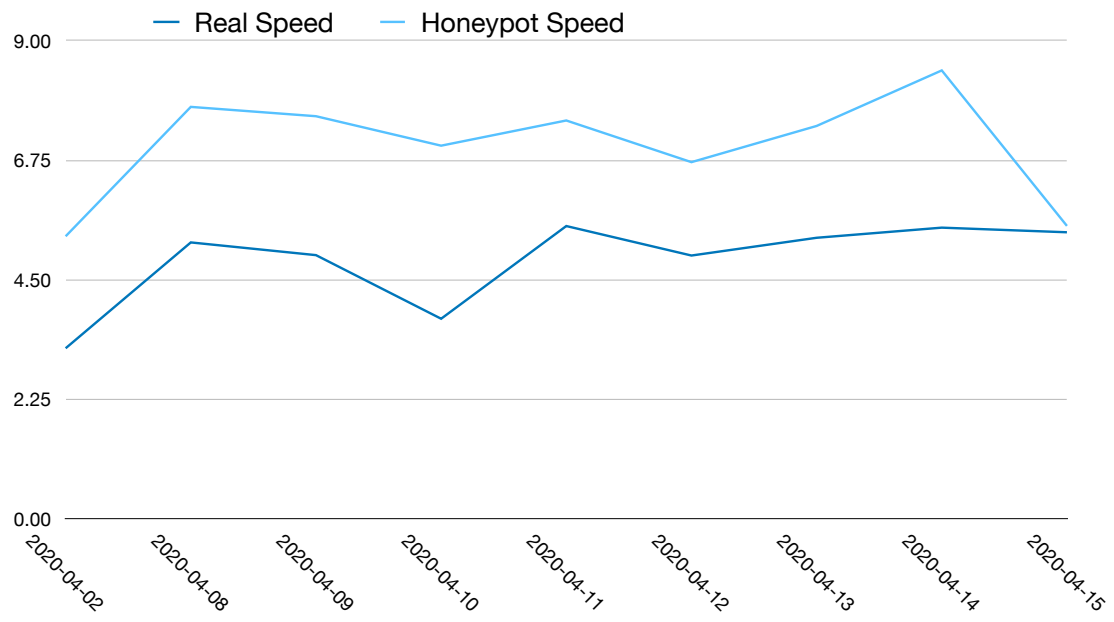

Figure 4.6: An example of the difference in the honeypot and real task speed. It is assumed that it was easy to identify the honeypots

There were also attribute specific modifications implemented. Both Makeup and Eye Makeup previously tasked workers with selecting from the following: 'None',

'Light' and 'Heavy'. They are now offered the binary choice between ''Makeup and 'No Makeup'. This will hopefully narrow down the choice and minimize confusion. Hair Color previously tasked workers with selecting from the following: 'Blonde', 'Brown', 'Black', 'Gray/White', 'Bald', 'Red', and 'Cannot Be Determined/Not Visible'. They are now offered the following reduced options: 'Light Brown/Blonde', 'Dark Brown/Black', 'Gray/White', 'Bald', 'Cannot Be Determined/Not Visible'. It is easier to narrow down the task to shades versus an absolute color. In the qualifier tab, the accuracy is set to 80% and in the honeypot tab, the count is set to 10 with a minimum accuracy of 80%. For the real task settings, the minimum consensus size is 3 and the maximum worker count is 5. With these changes implemented, I used that same set of inconclusive data and then started the project. The final number of inconclusive came back as 87, meaning 5% of the images reviewed were inconclusive. This number is a more manageable amount for a team to then go back and label manually.

Percentage of Inconclusive by Attribute			
Attribute Class	**WA Before Modification**	**WA After Modification**	▲
Mouth Open	6%	2%	-4%
Hair Color	28%	5%	-23%
Hair Length	21%	2%	-24%
Eyebrow Thickness	26%	2%	-24%
Makeup	37%	0%	-37%
Eye Makeup	34%	0%	-34%

Figure 4.7: Comparison of West Asia Experiments

After this experiment, I have a plan to implement the procedure above with the new dataset, North Africa. Using this refined procedure, I will observe if there is a reduction in the number of inconclusive results due to the implementation of the binary labels, as well as the language improvements.

4.3 North Africa Experiment

After evaluating the process and results from the West Asia Experiment, the next experiment to conduct is North Africa. It is a group composed of 634 images from four countries. The countries are Libya, Morocco, Sudan, and Tunisia.

Figure 4.8: North African Countries

All changes in the procedure that were implemented above in West Asia Experiment 2.0 were the starting point of this experiment. However, there are some additional changes. The commonly confused classes get a qualifier slide and a consensus requirement. The Spanish translation was also reviewed again by a native speaker from Latin America, the region where the workers reside.

New translations were provided so that the wording was less confusing to a

native speaker. The translation improvement to revisit the language barrier and updated the projects to the version that allows a separate section for the translation. As stated above, Hive's platform did not previously have a separate section for translations. After updating their platform, they now have sections for translations (rather than having everything in Spanish and in parenthesis English). Each instruction slide needed to be updated to only have one language per slide and have the translation in the Spanish tab. This is the same for the labels, each class was in English and the corresponding Spanish tab had the appropriate translation. In doing this, if a Spanish speaking worker is using the platform, they will select Spanish as the language and only see the instructions and labels in Spanish.

Rather than copying the previous honeypots, Hive allows the labeling of uploaded data for the project. So, after uploading the data from North Africa, I selected a few images per category and labeled them to use as honeypots. Having the honeypot images derived from the region being evaluated for the dataset should provide a more accurate test, while still saving time on the set up of each project. The hope is that this will continue to ensure there is no way to differentiate honeypots from the real task by the end users.

There was a suggestion from the committee to include the gender in the labels. Previously, they were excluded for cost savings. An example of how adding gender to the image label is beneficial, can be seen when evaluating the presence of makeup. Initially, the thought was that there were more women who wore makeup in certain countries. So, to avoid having extra costs, it was assumed that men would not have makeup on. This change will lead to more comprehensive and inclusive data. Now all 634 images will be labeled in each category. Hopefully this will not change the inconclusive count as there is a greater volume of data being added to the categories.

In addition to adding gender, the hair color description was also altered. The

hair color is now on a scale much like skin color. The options are: 'Dark brown/black', 'Light brown/blonde', 'Gray/white', 'Bald', and 'Not visible'. This allows a more appropriate labeling option for countries where hair coverings, such as hijabs, are common and for pictures that do not include the hair. The category also helps diminish inconclusive results for images where the individual is bald.

The results from the data shows that it took 28 hours for all images to be examined. The average honeypot speed came back as 4.386 and compared to the average real speed at 4.94 the honeypots are a success in terms of disguise. Again, the honeypots should take the same time as a real task or less than. It means that the workers are performing the labeling task appropriatly and are not gaming the job/task.

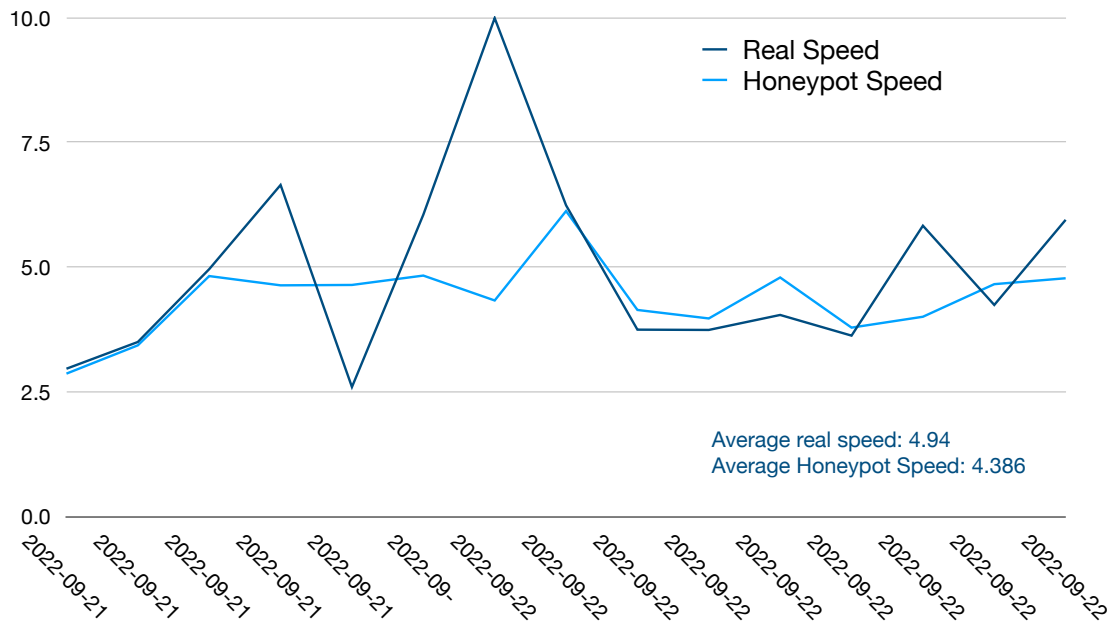

Figure 4.9: North Africa Honeypot VS Real Task Speed

The data demonstrated that there are still some sections of data not being labeled. However, there is an overall reduction in the number of inconclusive data. The inconclusive count is 'Eye makeup': 0, 'Hair color': 10, 'Age': 2, 'Expression': 36, 'Lipstick': 0, 'Hair length': 8, 'Beard': 14, 'Mustache': 0, 'Mouth open': 0, 'Eyebrow thickness': 11, 'Makeup': 0, 'Glasses': 0.

Inconclusive Frequency Per Label		
Label Name	Number of Categories	Inconclusive count
Age	5	2
Expression	4	36
Hair Color	5	10
Hair Length	6	8
Eyebrow Thickness	2	11
Open Mouth	3	0
Glasses	2	0
Makeup	2	0
Eye Makeup	2	0
Lipstick	2	0
Mustache	2	0
Beard	4	14

Table 4.1: inconclusive Frequency Per Label

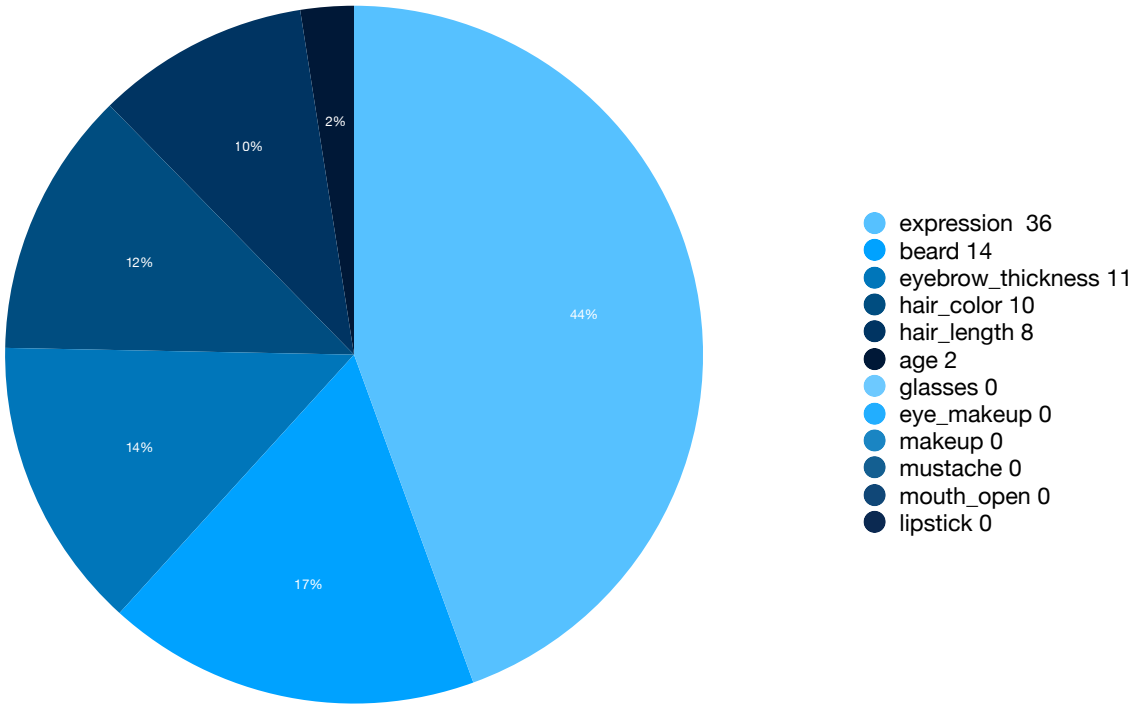

Figure 4.10: North Africa Inconclusive Count

Compared to the other experiments, North Africa does have the lowest count of inconclusive data. Figure 4 shows that there is a decrease from West Asia 2.0 in

the categories Mouth Open, Hair Color, and Hair Length. Makeup and Eye Makeup stayed the same at 0% inconclusive proving that the change in not dividing gender categories seems to not have a change on worker input. The category Eyebrow Thickness had no change in the percentage of inconclusive results.

Percentage of Inconclusive by Attribute				
Attribute Class	**WA Before Modification**	**WA After Modification**	▲	**North Africa**
Mouth Open	6%	2%	-4%	0%
Hair Color	28%	5%	-23%	1.5%
Hair Length	21%	2%	-24%	1%
Eyebrow Thickness	26%	2%	-24%	2%
Makeup	37%	0%	-37%	0%
Eye Makeup	34%	0%	-34%	0%

Figure 4.11: Experiment Comparison

Chapter 5

Conclusion and Future Work

During the development of the labeling process, there were a few alterations that resulted in significant impacts to the outcomes when reviewing the data from each iteration. At the end of the project, I developed a process for labeling images that is effective and efficient based on the response times of honeypots and the frequency of inconclusives. This is important because the objective of using a crowdsourced tool is to decrease the cost and effort associated with labeling while improving the quality of the labels. I have examined and presented my findings of inconclusive data. The results indicate this final version of the process is a procedure that will produce accurately labeled data with less manual intervention, when looking at the volume of inconclusive data. For example, the results from West Asia 1.0 required thousands of images to be hand labeled. In contrast, North Africa only had 81 inconclusive results, but required honeypots to be hand labeled. The setup for each project became increasingly more time-consuming and difficult. However, the results show that the change was impactful. Taking the time to change the honeypots takes less time than manually labeling hundreds of inconclusive images. This process is an efficient way to label large volumes of data. The results demonstrate that the

changes made during the development of the process are beneficial in producing a usable product.

5.1 Future work

5.1.1 The platform

If there had been additional funding, I would have compared my work in Hive with other platforms such as Mechanical Turk. Each platform is different with different implementations of tools to improve the labeling process. Future work can look at the tools used by different platforms to determine the platform that is best suited for this labeling process. Though Hive did have a simple cost structure, I do wonder if sacrificing that part of simplicity could allow for a change in other areas. One area I think about is language. Hive requires you to translate to another language outside of English and they do indicate they have a large number of Spanish speakers. I would like to see if using a platform like MTurk would lower the issues with language barriers and lessen the time spent in translation. This would be dependent on their built in translation capability and the number of workers they have that are non-native English speakers. An additional way time could be saved is if the platform could translate for the user and would allow the worker pool to be more diverse.

5.1.2 Face Labels

Though most categories with inconclusive results were reduced, the category 'Expression' did not see as much variability. 'Expression' and 'Eyebrow Thickness' had the greatest percentages of inconclusive data. (One factor might be they are cultural and gender dependent.)

"Eyebrow thickness" is an area that needs significant adjustment, as eyebrow thickness ideas can change based on gender. For example, what would be considered thick eyebrows on a woman might be seen as normal for a man.

The results show that the 'Expression' category is a bit more subjective and could be further refined, as it is still the category with the most confusion. There should be future research on how people perceive emotions to see if there is any way of aiding workers tasked with labeling images for emotions. To continue to increase the volume of labeled data, I think modifying the classes might also be beneficial. For future initiatives, investigating how humans gauge facial expressions and reviewing available research on this subject may aid in creating changes to this label. I predict an additional contributor to inconclusive data within the North Africa test is due to issues with the photos such as graininess, images not properly situated in the bounding boxes, and the use of black and white images. Creating some adjustments in photo quality and in the labeling process for the more subjective classes would be beneficial.

5.1.3 Data Scraping

For future work, I would recommend revisiting the data scraping tool and process to focus on selecting higher quality images. Images that have larger faces with sharp focus and good contrast over the face. Having clearer images would help workers tasked with labeling. Implementing a method that is incorporated into the tool's functionality to exclude pictures with poor resolution might assist in reducing the volume of images that are categorized as having inconclusive results. One potential issue from this, which has been historically noted, is the fact that in many underdeveloped countries, there is a limited amount of resource photos with informa-

tion available. This could result in a significantly lower pool of images of individuals that fits the requirements listed in the methodology. One additional thought is to investigate if there is a way to change the photo requirements without compromising the integrity of the data.

Appendices

Appendix A Hive Labels

Age: determine the age category of the person. If the person in the photo appears to be between 0 and 12 years old, select Child. If the person in the photo appears to be between 13 and 19 years old, select Teen. If the person in the photo appears to be between 20 and 29 years old, select Young Adult. If the person in the photo appears to be between 30 and 50 years old, select Middle Aged. If the person in the photo appears to be older than 50 years old, select Older.

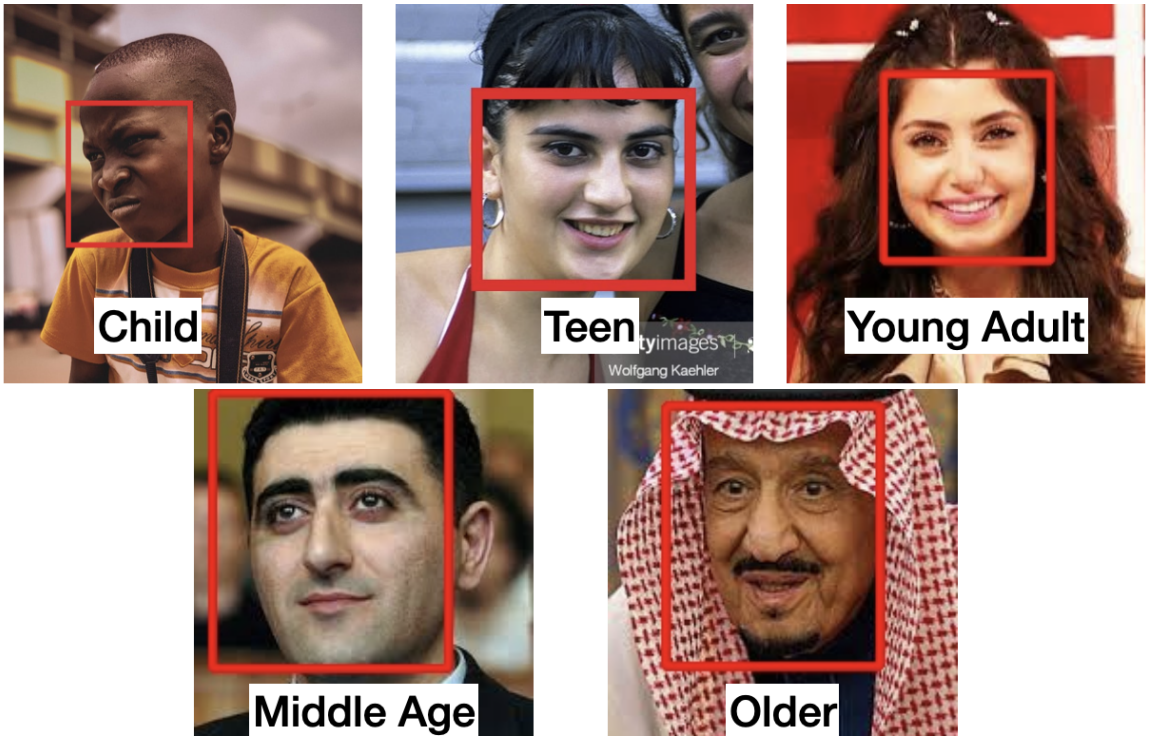

Figure 1: Example photos labeled as a worker would see in instructions

Expression: If the person in the photo is smiling, select Smiling. If the person in the photo appears to be frowning or angry, select Frowning/Angry. If the person in the photo has a blank expression (the facial muscles are relaxed), select Blank Expression. If the person in the photo is making a sad expression, select Sad.

Figure 2: Example photos labeled as a worker would see in instructions

Hair Color: Look only at the hair on the head. If the person in the photo has dark brown or black hair, select Dark Brown/Black. Look only at the hair on the head. If the person in the photo has light brown or blonde hair, select Light Brown/Blonde.Look only at the hair on the head. If the person in the photo has gray or white hair, select Gray/White. Look only at the hair on the head. If the person's hair in the photo is not visible, select Not Visible. If the person in the photo is bald, select Bald.

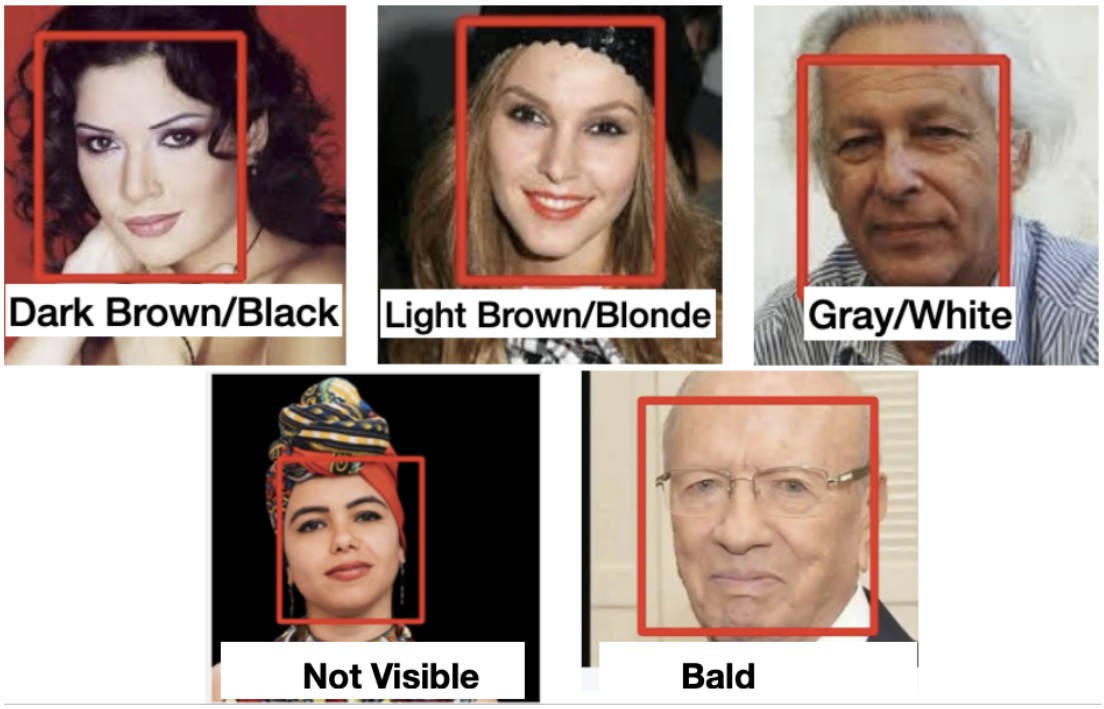

Figure 3: Example photos labeled as a worker would see in instructions

Hair length: If the person's hair does not touch the collar or the shoulders, select Short. If the hair is long enough to touch the collar or the shoulders but does not fall significantly below the shoulders, select Medium. If the hair falls past the shoulders, select Long. If the person in the photo is completely bald, select Bald. If the person in the photo is partially bald, select this option. If not enough of the hair is visible to make a judgment about the length (if, for example, the hair is tied up or covered), select Not Visible.

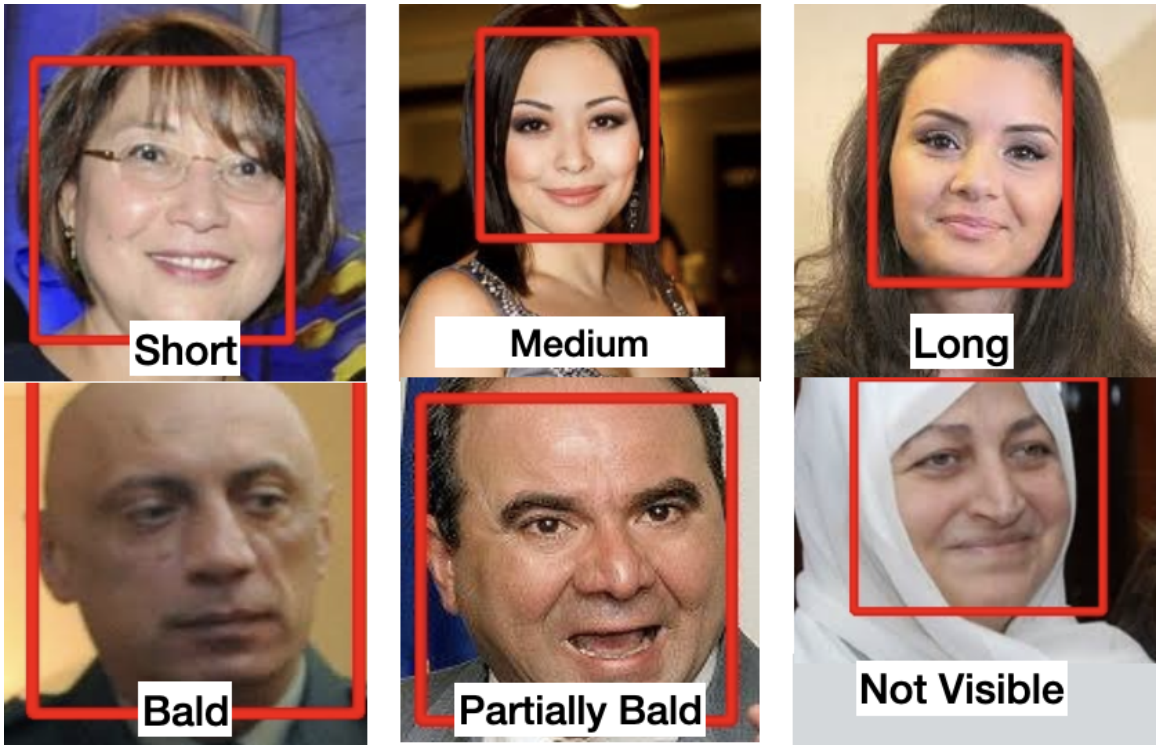

Figure 4: Example photos labeled as a worker would see in instructions.

Eyebrows thickness: Does the person have thick/bushy eyebrows? If the person has eyebrows that are thicker than normal in terms of width, length, and density, select Yes. If the person has eyebrows that are thin or normal, select No.

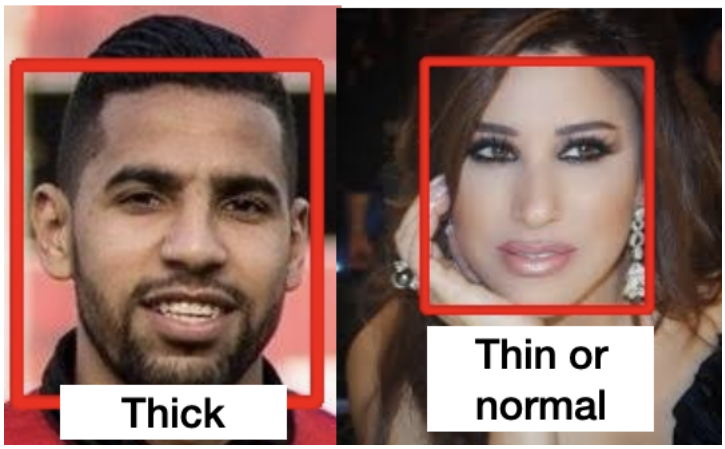

Figure 5: Example photos labeled as a worker would see in instructions

Open mouth: The mouth is open if there is a gap between the lips such that the teeth, the tongue, or the inside of the mouth are visible. The mouth is closed if there is no space between the lips or minimal space between the lips. The teeth, tongue, gums, and inside of the mouth are not visible. The mouth is covered if not enough of it is visible to determine whether it is open or closed.

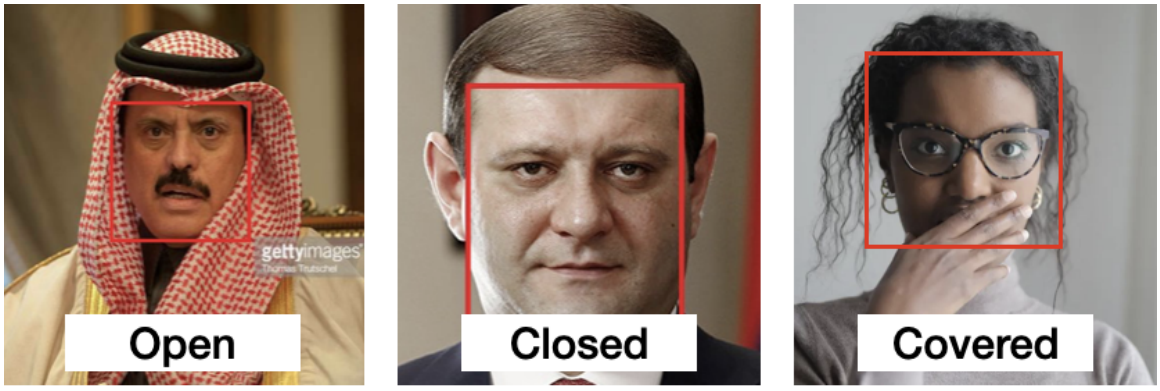

Figure 6: Example photos labeled as a worker would see in instructions.

Glasses: Is the person wearing eye glasses (spectacles) or sun glasses? Selection options are "yes" or "no".

Figure 7: Example photos labeled as a worker would see in instructions

Makeup : If the person in the photo is wearing visible foundation, blush, bronzer, or other makeup on the skin, select yes. If the person does not appear to have any foundation, blush, bronzer, or other makeup on the skin, select no. Remember that the person may be wearing eye makeup but not makeup on the skin, so focus only on the skin.

Figure 8: Example photos labeled as a worker would see in instructions.

Eye makeup: How would you rate the level of eye makeup? If the person in the photo is wearing eye makeup, select Yes. If the person in the photo is not wearing eye makeup, select No.

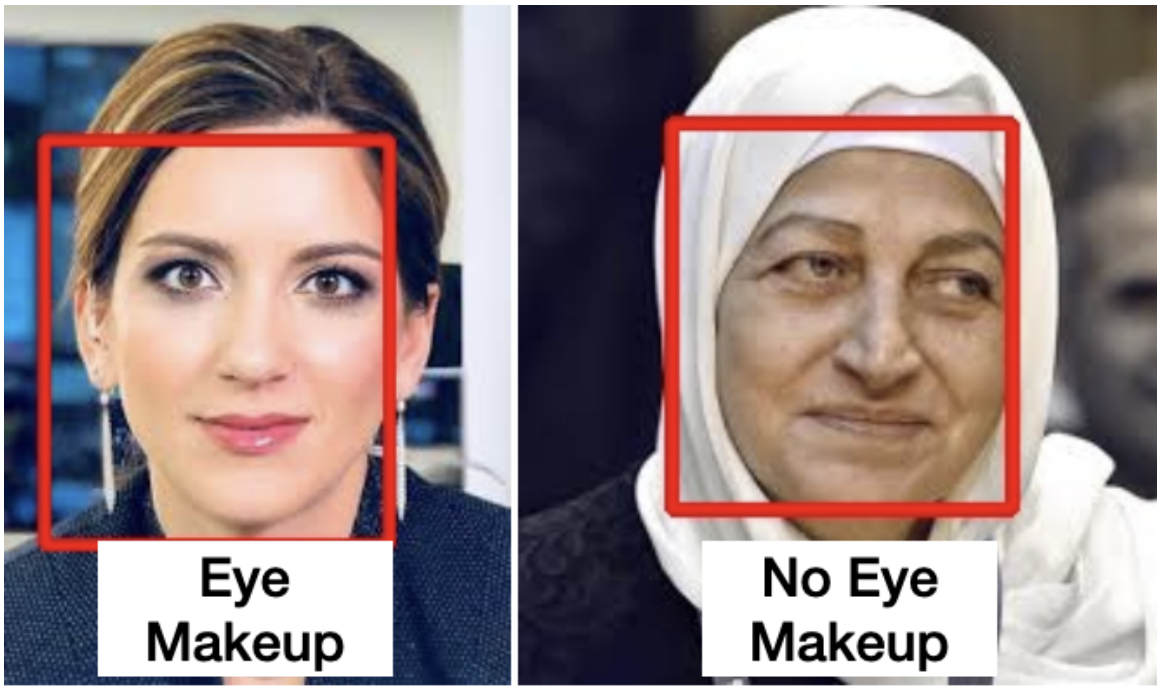

Figure 9: Example photos labeled as a worker would see in instructions.

Lipstick: For this task only report/examine on the face/person that is in the red box. Notes: Focus on the lip region and determine if a person has applied lipstick or gloss. Lipstick colors varies from dark to nude. Look for variation in lip shade compared to the face.

Figure 10: Example photos labeled as a worker would see in instructions.

Mustache: If the person has a mustache (hair above the upper lip), select Yes. If the person does not have a mustache (hair above the upper lip) or the region has been clean shaven, select No.

Figure 11: Example photos labeled as a worker would see in instructions.

Beard: If the person in the photo has a beard, select this option. If the person in the photo has a goatee, select this option. If the person has stubble or five o'clock shadow on the jaw area, select this option. If the person in the photo has no facial hair on the jaw or chin area, select this option.

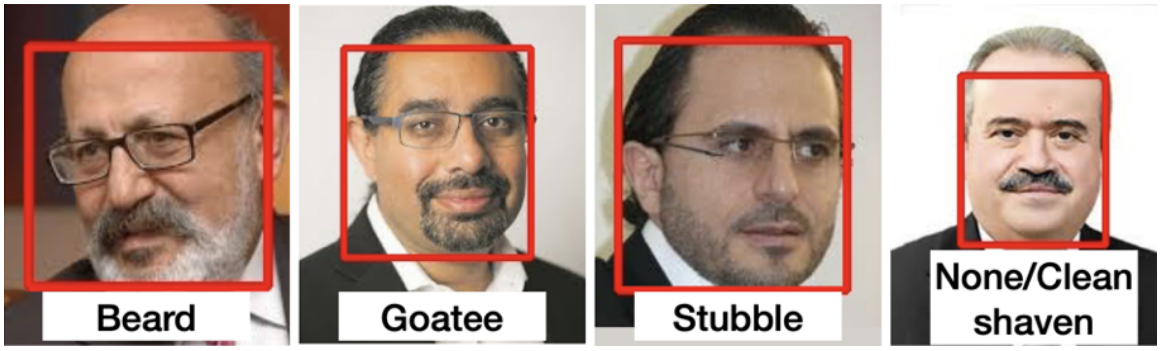

Figure 12: Example photos labeled as a worker would see in instructions.

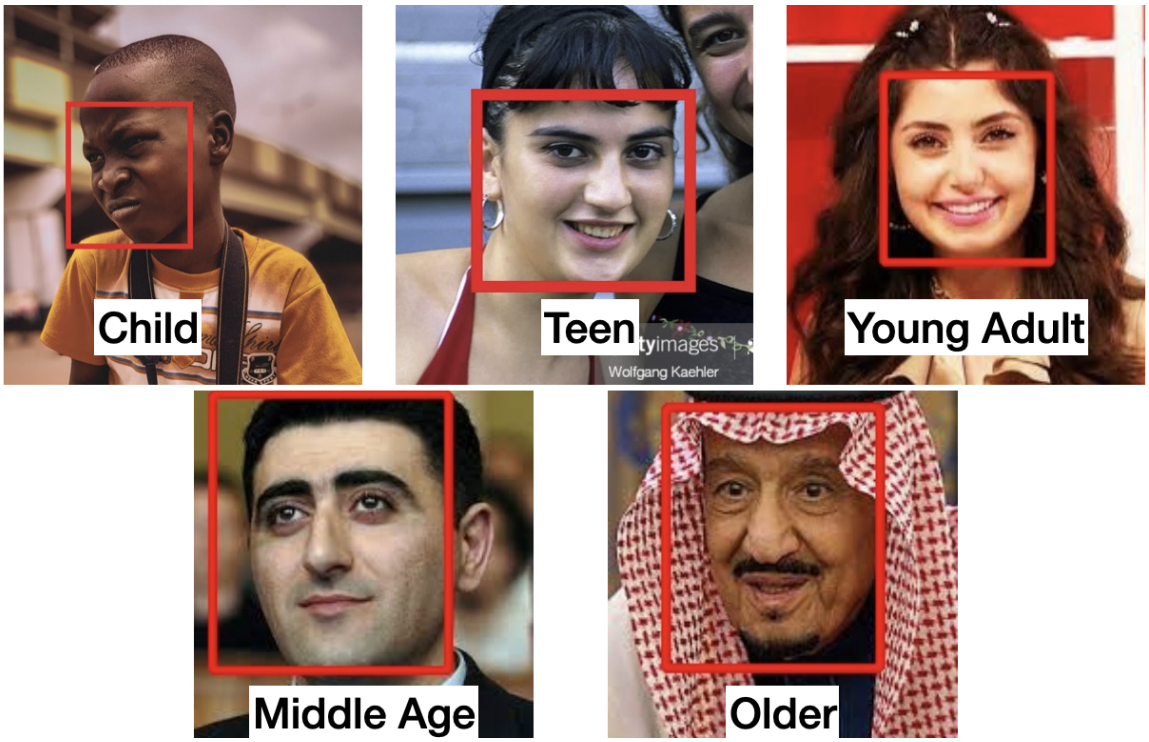

Figure 13: Example of an employees view of their progress towards a given set of achievements.

Bibliography

1. J. Boulamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," Proceedings of Machine Learning Research, vol. 81, pp. 1–15, 2018. [Online]. Available: https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf[Accessed:16-Jul-2022]

2. J. Buolamwini, "Gender shades," 2018. [Online]. Available: http://gendershades.org/overview.html

3. "What is data labeling?" [Online]. Available: https://aws.amazon.com/sagemaker/data-labeling/what-is-data-labeling/.

4. M. Sharma, "How does crowdsourcing data labelling work?" Analytics India Magazine, 2021. [Online]. Available: https://analyticsindiamag.com/how-does-crowdsourcing-data-labelling-work/.[Accessed:16-Jul-2022]

5. "Inconclusive," English Definition and Meaning — Lexico.com., 2022. [Online]. Available: ⟨https://www.lexico.com/en/definition/inconclusive⟩ [Accessed16July2022].

6. "Crowdsourcing," 2022. [Online]. Available: https://www.merriam-webster.com/dictionary/crowdsourcing.[Accessed:16-Jul-2022].

7. "Crowdfunding." [Online]. Available: https://www.lexico.com/en/definition/crowdfunding.[Accessed:16-Jul-2022].

8. S. Shapiro, "Types of crowdsourcing - stephen shapiro." [Online]. Available: https://stephenshapiro.com/types-of-crowdsourcing/.[[Accessed:16-Jul-2022].

9. "Hive — enterprise ai solutions." [Online]. Available: https://thehive.ai/hive-data.[Accessed:16-Jul-2022].

10. "Hive — enterprise ai solutions." [Online]. Available: https://thehive.ai/about-us.[Accessed:16-Jul-2022].

11. "Requester — amazon mechanical turk." [Online]. Available: https://requester.mturk.com/pricing.[Accessed:16-Jul-2022].

12. “Our complete solutions for image annotation.” [Online]. Available: https://www.isahit.com/image-annotation.[Accessed:25-Feb-2023].

13. “Welcome to hive: Information guide,” 2020.