Toward More Responsible Labeling of ML Training Datasets

For Software Engineers

**Note: there are two options for this lesson plan, catering to slightly different audiences.*

Who this is for: Product managers or team leads, to use with software engineers labeling their own data or writing labeling tasks / annotation guides

Pre-reading: Both the instructor and participants should read the **Responsible Language Guide for Artificial Intelligence & Machine Learning** first.

Background: Software engineers working on labeling language data, or providing instructions to teams labeling data need to be aware of how the language used in labels can exacerbate or mitigate harmful racial biases. This lesson plan is designed for facilitators to demonstrate ways in which labels applied to data classifying humans can result in inequities, as well as outline responsible practices and guidelines for using language that advances racial equity and inclusion.

Target participants: Software engineers at US-based offices of tech companies who (i) choose to build and label their own datasets for their projects, (ii) are labeling "golden data", or (iii) are responsible for writing detailed annotation instructions for other employees or external vendors to label their data.

***Persona**: This individual has expert-level understanding of how algorithms / AI systems use the data they label to make decisions. They have undergone general diversity, equity, and inclusion training within their organization, and may or may not have considered the impact of their project on different social groups. Regardless, they are not yet fully equipped to ensure that the language used in the labels for their data advances equity and inclusion as well. This individual will make high-level decisions about what features of their data need to be labeled, what social group categories should be used, etc., and will communicate this information through annotation guidelines.*

Goals: (1) Equip participants with an understanding of how human interventions and decisions can result in biased labels for data, particularly as they pertain to language data; (2) enable participants to identify potential pitfalls and areas of bias in their own data labeling and/or task creation work; and (3) provide participants with good practices to follow to counteract bias in data labels and/or write annotation guidelines that propagate language to advance racial equity and inclusion.

Preparation:

- Familiarize yourself with this lesson plan.
- Prior to the session, go over the "suggested activities" throughout this plan and include / customize those that are relevant to your audience. When making adjustments, make sure to also adjust the session time to account for how long each activity is slated to take.
- Set up the physical room (or virtual space) so participants are able to break out into smaller groups of 3-4 for the activities.
- Prepare the facilitation materials (see list below) and ensure participants have their participant packets before the training starts (see materials below).

Materials:

- For facilitators: Printed copy of this lesson plan, screen, projector, **presentation deck** to accompany lesson plan (also in Appendix i)
- For participants: A notepad, pen, and the participant packet which includes a printed/soft copy of the **Responsible Practices handout** (also in Appendix ii), a printed/soft copy of the **Sample annotation guidelines** (also in Appendix iii) and a printed/soft copy of the **Personal action plan template** (also in Appendix iv).

Feedback survey templates

- Appendix v: **Immediate follow-up survey**

Note:

- *Bullets vs. Numbers / Letters:* In this lesson plan, anything that is bulleted is a talking point for the facilitator. While anything that is numbered or denoted by letters is an instruction for the facilitator.

Time: 180 minutes

Introduction [15 MINUTES]

1. Welcome participants to the session. Present slide 2 and use the following talking points:
    - Welcome! We are excited to welcome you to this workshop. The workshop will delve into how language used to label data about humans -- particularly data containing text / speech -- can be biased, and the harms such labels can cause. We will also be exploring how the annotation guidelines you develop and provide can mitigate bias and promote language that advances racial equity and inclusion.
    - We will focus on racial bias in language and labels, but we will also explore and keep in mind that gender and other forms of bias come into play.
    - Please make sure you have a sheet of paper and pen from your participant packets ready!
2. Do a quick round of introductions. Have everyone say their name, pronouns, one thing they hope to learn about today, and one thing that's bringing them joy at the moment. Model this by going first.
3. Tell participants that you will now work together to co-create a community agreement. Explain:

• A community agreement includes rules of conduct for the training. Community agreements engage participants in developing these ground rules so they have ownership of the rules and are responsible for them—individually and as a group.

4. Ask participants: What are some good rules or principles we should set as a group for this space? Have participants share verbally or via chat (if virtual) and add their suggestions to Slide 4.

   a. Be sure to share your screen so participants can view Slide 4. For in person settings, use a flipchart and write "Community Agreement" at the top.

   b. Add some suggestions to the community agreement yourself. Such as:

      - Engage in brave conversations. (Be thoughtful about what you say and be open to the fact that even comments that have good intentions can have negative impacts. Try to understand others' perspectives and be open to feedback.)

      - Take space, make space. (Allow everyone to have time to share and react. Be brave and speak up, but also step back and allow others time to contribute.)

      - Share your own experiences and honor others' experiences.

      - Practice active listening

      - Honor confidentiality.

      - (For virtual settings) Keep video on to the extent that is comfortable or possible for you, and remain focused (refrain from checking email / scrolling online).

5. Ask participants: Is there anything you would like to take off this list, or anything else to add to it?

6. To conclude, tell participants:

   • Please raise your hand (or give a thumbs up in a virtual setting) if you agree to this community agreement. If in a physical space, tape the community agreement to the wall in a place where people can see it.

Activity A [20 MINUTES]

7. Introduce activity A and present slide 5:

   • We will kick off this lesson with an activity!

   • I will read out a scenario and then have you break out into smaller groups of 3-4 to answer questions pertaining to the scenario.

8. Present the scenario on slide 5. Read it aloud to participants, or invite participants to read it to themselves.

Jarvis, a Black food blogger, plans to write a twitter thread on the connection between Black history, identity, and food culture. Before he publishes the tweets, he decides it would be good to run his content through an AI powered writing tool that can catch spelling and grammatical errors, and detect the writer's tone of voice. Jarvis writes the following into the AI powered grammar tool, "*When throwing down in the kitchen, it ain't enough to just know the recipe. Black food is called "soul food" for a reason. You gotta feel the connection between the ingredients, our ancestors, and our history. Mac and cheese, greens, yams, jerk chicken, and pound cake - these ain't just foods. They are central to us.*"

Jarvis is shocked when the platform says that the tone sounds angry, and flags some of his language as incorrect.

9. Have participants break out into smaller groups of 3 people and discuss for 5 minutes: Why might the algorithm have classified the tone of Jarvis' essay as angry? What might the algorithm be flagging as incorrect and why?)

a. If in a virtual environment, make sure to provide a link to the presentation or specifically to slide 5 in the chat prior to creating the breakout rooms.

10. Bring the entire group back together and have 2-3 breakout groups share.

11. Use the following points to explain what happened in this case:

• There are two main issues here. First, the word "jerk" generally has a negative connotation or is considered offensive.[1] So, it may have been labeled or flagged as offensive. The algorithm picked up on that word without picking up on the context. It did not know that jerk chicken is a food dish and classified the whole essay as sounding negative.[2]

• It's possible that "pound" may have been flagged as violent, since it is a word sometimes used to describe physical fighting.

• Second, the algorithm flagged the term "ain't" as incorrect. However, Jarvis is using African American English in his post. African American English uses words like "ain't", as well as double negatives that we don't see in "Standard" American English.[3] African American English is a language variety (as is "Standard" American English").

• Importantly, all language varieties are linguistically equal. No language variety is linguistically better or more correct. Rather, each language variety follows its own sets of rules. So terms like "ain't" and double negatives are not incorrect linguistically.

• Despite this, "Standard" American English is often seen as the 'right' or 'proper' way of speaking. This is why we put "Standard" within quotations. It is not inherently standard. Rather, it has been granted this position.

• Language datasets tend to favor "Standard" language varieties (like "Standard" American English). Other language varieties present in these datasets (such as African American English) tend to be wrongly labeled as incorrect, or "unintelligible". This happens especially if annotators don't speak these varieties. As a result, African American English is often misclassified as being incorrect by AI systems.[4]

• This creates situations like the one with Jarvis, where the system is unable to interpret the context to understand that Jarvis is purposefully using African American English to speak in a voice that resonates with his audience.

12. Share the following:

• This example illustrates how bias in labeling language data comes into play. Algorithms rely on patterns in how we use and label language. But what we say and how we say it varies based on language variety, speaker, and context. If a certain language variety is labeled as incorrect or unintelligible, it can carry over to the algorithm's outputs and perpetuate bias against this language variety and people who speak it. If a word is labeled as offensive but the machine doesn't recognize it's not offensive in certain contexts, that can be problematic too.

13. Present slide 6 and ask participants: What are some other issues that might come up when labeling text and speech? What are the implications for AI systems?

14. Build off their answers with the following:

• *Language identification.*

- Issue: Because language datasets are often centered around "Standard" language varieties, African American English is (1) underrepresented in training datasets, and/or (2) erroneously labeled as incorrect or unintelligible.

- Implication: Texts written in African American English are routinely misidentified as being written in a language other than English, at a rate far higher than texts written in "Standard" American English.[5] This has various implications such as texts written in AAE not showing up as much in search algorithms.

• *Hate speech detection.*

- Issue: Labeling hate speech is subjective. What is considered hate speech evolves and depends on contexts.[6] For example, slurs that have been reclaimed by groups against whom they were originally weaponized can erroneously be classified as hate speech.

- Implication: Hate speech algorithms more often incorrectly identify African American English as hate speech (hateful or offensive) than "Standard" American English.[7] Black people aren't the only folks affected. For example, the LGBTQ+ community has also reclaimed terms that were previously used as slurs against members of their community. These nuances and contexts are harder for algorithms to pick up on.

Topic & transition [10 MINUTES]

15. Transition out of the activity and tell participants:

• We've seen how labels matter. We've just explored the impacts that racial bias in labeling language data can have on AI systems and subsequently, people. To understand how to address this, we are going to explore what bias is and how it comes into play.

16. Present slide 7 and ask participants: Thinking about the outcomes of Activity A, what comes to mind when you hear the term bias? Have some participants share their answers.

17. Present slide 8 and build off their answers to explain:

• Unconscious biases are cognitive shortcuts: split-second judgments. Humans experience biases all the time, often unintentionally and unconsciously. Our brains are wired to be biased, and these biases show up in the initial reactions we have.

• Humans have 2 key modes of thinking: System 1, which is our automatic, quick instincts; and System 2, which involves more deliberate effort, agency, and choice. We need our System 1 thinking to organize and manage all the stimuli we constantly face, but this is also where our biases come into play.

• We make our quick judgments based on personal experiences, education, upbringing and communities -- and the stereotypes and norms that accompany them.

• Of course, we can have conscious biases as well. But when labeling data, particularly in time-sensitive situations, we tend to fall back on the quick System 1 thinking. We tend to label data based on what is familiar to us, and any stereotypes or biases we hold can come into play unchecked.

• This can be the case with labeling African American English. Those who do not speak this language variety may fall back on incorrect beliefs that it is not the "right" way of speaking.

• Now that we've seen how bias can affect the way language data is labeled, let's try another activity to understand more clearly the role that task creators play in the process!

Activity B [30 MINUTES]

18. Introduce the following activity:

- Unless working with crowdsourced datasets, task creators typically provide detailed guidance / instructions to the raters they're working with. Raters work with the broad options for labels they are given, so their agency over choices they can make is limited by what is defined in their guidelines.
- A big part of ensuring labels are ultimately equitable and inclusive goes back to this guidance. In particular, managers get to determine the array of labels for raters to choose from, define instructions for determining them, and choose what examples / context is provided for raters.

19. Present slide 9 and read the scenario aloud to participants.
    - Let's look at the following example around demographic and health data. While not specific to annotation guidelines, this example will still illustrate the challenges with creating labeling tasks and categories for raters to work with.

In March 2020, the World Health Organization declared COVID-19 a pandemic. Recognizing the importance of collecting data on the spread of the virus, a United States organization -- the Health Data Institute -- began tracking COVID-19 data across the country. Early on, a team within Health Data Institute recognized disparities in COVID-19 health outcomes faced by different demographic groups. To better track and understand these disparities, the team decided to disaggregate by race the incidence and death rate data being collected. However, problems ensued shortly after. The team had included an array of labels for various racial groups that individuals were able to choose from in self-reported data collection forms, but the labels offered on the form resulted in confusion and questions. For instance: individuals who identified as Black Dominicans had to choose between "Black" and "Hispanic" on the form, but couldn't select both.[8]

The team also worked with healthcare organizations in some states to consolidate data that was already being collected -- but this meant relying on pre-existing forms and label options which varied, resulting in inconsistencies. For example, the Native Hawaiians and Pacific Islanders (NHPI) racial group was included on some forms as part of the "Asian" category, on some forms within an umbrella category of "other", and on some forms it wasn't listed at all.[9]

Potential build for facilitator: If looking to make this more interactive, we suggest building out a dataset of COVID-19 incidence in the US. You can draw from the **CDC's website** if looking for real world data, or build out a sample dataset of your own. Use labels such as "White", "Black", "Asian", "Hispanic", "Native Hawaiian and Pacific Islander (NHPI)", and "Other". Allow participants to play around with the labels. For instance, they may group some labels together -- such as "Asian" and "NHPI" under the umbrella category "Other". Encourage them to see how the demographic data trends change with more vague labels, and think about what information this may obscure.

20. Have participants break out into groups of 3 and discuss the following questions for 5 minutes: What are the potential healthcare impacts (including physical and mental wellbeing) of categorizing individuals using the labels listed in this scenario? Can you think of other examples in which inconsistent or ambiguous labels have been used to capture people's identities? What are the impacts?

a. If in a virtual environment, make sure to provide a link to the presentation or specifically to slides 9 and 10 in the chat prior to creating the breakout rooms.

21. Have 2-3 breakout groups share. Then use the following talking points to supplement the discussion:

- The labels used here are inconsistent. "Hispanic", for instance, is listed as an ethnicity in other government forms -- when asked by the Health Data Institute to self-identify as either Black or Hispanic, Black Dominican individuals were not given an opportunity to capture the full spectrum of their identities. Further, non-specific or umbrella labels like "other" are imprecise, conflating the experience of multiple racial groups that actually face significantly different health outcomes.
- These inconsistent and imprecise labels were problematic from a healthcare standpoint, because Black Hispanic people (including Black Dominicans) have been known to have different health outcomes from White Hispanic people. So restricting Black Hispanic people's ability to self-identify as only "Hispanic" or "Black" made it difficult to track COVID-19 disparities and adequately assign resources.
- More broadly: Incorrect categories can mask important social disparities between racial groups. For example: different government databases use at least seven different terms to identify/label Native Americans, including terms like Native Americans, American Indian/Alaska Native, among others. Federal and state statistics also tend to misclassify Native Americans under other race / ethnicity categories. This has resulted in an undercounting of the Native American population and inaccurate reporting of key indicators that are used to allocate federal resources.[10]
- Using vague, ambiguous terms can diminish the experiences of specific racial groups. For example: in the 2020 election exit polls, CNN used the labels "White", "Black", "Asian", and "something else" to record voted information.[11] Something else" conflates the experience of multiple racial groups, and fails to credit the powerful impact that specific voter demographics (Indigenous groups in particular) had during the election.
- Incomplete or insufficient categories can also cause erasure of already marginalized communities. Gender is a prominent example – most datasets only try to capture male and female gender labels, failing to incorporate the rest of the gender spectrum. This compounds the erasure of non-binary and transgender individuals, impacting their physical and mental health and wellbeing.
- Note that there is no prescribed way to classify race and ethnicity. One way of thinking about these two concepts is to look at them as distinct yet overlapping terms, but another way to conceptualize them is to view race as a subsection of ethnicity. Ultimately, it is important to be as precise as possible, and consistent. To the extent possible, people's self-identification should be respected.
- Note also that this lesson plan is focused on the US context, and that terms for different races / ethnicities may vary across geographies as well. For instance: in Spain, "Indigenous" is considered derogatory, whereas in Mexico it is the default for "Indigenous groups".

22. Have participants get back into their smaller groups and discuss the following question for 5 minutes: What recommendations and labeling guidance would you give to the Health Data Institute? If in a virtual environment, make sure to provide a link to the presentation or specifically to slide 11 in the chat prior to creating the breakout rooms.

23. Have 2-3 groups share ideas.

24. Thank participants for sharing their ideas. Tell participants that while this activity did not directly involve traditional annotation guidelines, it showcases the harms that inconsistent and

unclear labeling guidance can have on people. Ultimately, raters work with the instructions they are given, making it critical for you as task creators to ensure that these instructions use and promote the use of responsible language.

25. Tell participants that before moving onto good practices and solutions for mitigating bias in data labeling and collection processes, you will summarize some issues discussed so far.

Summary [15 MINUTES]

26. Present slide 12 and summarize the ways in which bias can come into play in labeling data.

- Through our activities, we've seen multiple ways in which bias can enter the data labeling process and how people use language more broadly.
- When labeling language / language varieties: this can be because raters are not familiar with the language variety being used, or because they're unaware of the context in which certain terms have been used.
    - Additional call-out! When labeling images of humans: although we don't delve into labeling images in this lesson plan, it's important to recognize how our biases can come into play here as well. Labels for images can often be subjective too -- such as when describing the range of human emotions (like whether someone is or looks angry). There are false stereotypes of the "angry Black woman" so Black women can be more often incorrectly labeled as angry. Labeling or describing individuals' gender or race can also be subjective.
- When dataset developers / product managers use imprecise or incomplete language to provide annotation instructions. The previous activity demonstrates how imprecise or incomplete instructions around language for labeling demographic data can result in bias and harm. Thinking back to the first activity, we can see how important annotation guidelines are for language data as well-- what might raters be instructed to look out for when identifying hate speech? Is there any context provided around African American English terms that have the potential to be misidentified?

Break [10 MINUTES]

Responsible practices [40 MINUTES]

27. Tell participants we will now explore ways to use language that advances racial equity and inclusion in data labeling processes. Have participants break up into groups of 3. Have half of the groups brainstorm responsible practices for writing guidelines for data labelers. Have the other groups brainstorm responsible practices for the data labeling process in general that would result in more inclusive language. Give the groups 10 minutes.

28. Bring participants back together and have 1-2 groups share who brainstormed responsible practices for writing guidelines for labelers. Write the recommendations on a white board, flipchart, or slide 13 under the header "Responsible practices for writing guidelines for labelers."

29. Build off of their answers with the following responsible practices for writing labeling guidelines using slide 14:

- 1. Recognize that labeling images of humans or language data can come with subjectivity. Be clear about what subjective labels look like and potential pitfalls to using them. Clearly define the labels necessary for the datasets and provide detailed examples to demonstrate. For instance: if the task involves labeling language data as unintelligible

or intelligible, provide examples in and information about different language varieties (such as "standard" American English, African American English, etc.) and remind raters not to label a variety as unintelligible just because they are unfamiliar with the rules and grammar associated with it.

• 2. Think critically about what data your project needs and why. Check if perceived racial or gender-based categorizations are necessary. Do you need race and ethnicity data? Or gender data? What are potential implications of collecting this data? How might you collect data that has multiple options without using "other"? Get clear around these questions for your project and provide adequate contexts to your raters.

- If race or gender-based categorizations are not necessary, raters should not use them. Labeling guidance should make this clear and where appropriate, suggest using more general labels like "person".

• 3. If race and /or gender labels are required - such as for fairness testing or dataset balancing purposes…

- First, reference self-identification of these social categories if possible. When dealing with people's identities, it is best not to make assumptions, so label them as they identify themselves.

- When self-identification is unavailable, ask raters to label features when possible (versus perceived race or gender). If your project requires perceiving gender or gender presentation, instruct raters label features such as facial hair, presence of makeup, clothing, skin tone, etc. For perceived race, have raters label features such as skin tone.[12] Use standard scales (typically provided in the labeling guidelines) such as the Fitzpatrick Skin Type.

- If labeling race or gender is still required, state that you/the raters are using perceived rather than self-reported identity. Be transparent about criteria used to assign perceived gender to individuals. This applies to skin tone as well.

- Ensure information on how race and gender categories are labeled is captured and made transparent; such as through incorporating this information in data statements[13] or Data Cards[14].

• 4. Be aware of how groups may be referenced as the default. For example: stating that someone has a "darker skin tone" while labeling a lighter skinned person as having "light skin", reinforces light skin as the primary reference category. To avoid these types of issues, provide raters with a standard scale such as the Fitzpatrick Skin Type. However, be aware that the skin tones outlined in this and other standard scales do not map onto racial or ethnic categories -- make sure not to conflate the two.[15] Also, note that this scale includes only a limited range of skin tones, and document any differences.[16]

• 5. Use precise and accurate language. When defining categories and labels for raters to select from, ensure that they are correct, specific, and consistent where possible. For instance: rather than reducing a diversity of groups into a homogenizing category like "non-White," be precise in naming the group or groups you are talking about.

• 6. Stay updated around the evolving language around racial and other social categories. Reference this **Glossary** for a set of definitions and terms to be aware of when labeling race / gender categories, and the tab on **"Frequently asked questions" in this Terminology Guide** to understand some key choices to be made.

- Note that this resource is focused on the US context and that terms for racial/ethnic categories may vary across geographies. For instance, in Spain, "Indigenous" is considered derogatory, whereas in Mexico it is the default for "Indigenous groups". In Spain, the accepted term is "autochthonous".

30. Have 1-2 groups share who brainstormed responsible practices for the rest of the data labeling process. Write the recommendations on a white board, flipchart, or on slide 15 with the header "Responsible labeling practices for inclusive language and labels".

31. Build off their answers with the following practices for labeling data using slide 14:

- 1. When labeling language data, utilize raters that speak the language variety (as much as possible). Raters should at least be familiar with the language variety.
- 2. Know / document as much about your rater demographics as possible.
- 3. Use domain and subject matter experts and social scientists to label data, as needed. For instance, consult with sociologists and / or community members for precise, up-to-date demographic terminology (particularly around race, ethnicity, and gender) when providing guidelines for labeling human data.
- 4. When using automated labeling for human subjects, make sure there is a human in the loop. These individuals can check the generated labels for accuracy and quality. Having a human in the loop is critical to ensure that biased labels are identified as early as possible. These humans in the loop should also be trained on bias that can come up in labeling. As mentioned earlier, automated systems are not neutral, and are likely to replicate existing biases -- use them to supplement and not replace your tasks.
- 5. When working with confidential data -- particularly demographic data -- ensure that the labeling tasks do not break privacy regulations.

32. Highlight how the responsible practices align with the different activities the participants did in the workshop.

33. Share the following points and ask:

- It is important to keep in mind the responsible practices for annotation guidelines, as well as responsible practices for the data labeling process more broadly in order to work towards inclusive language in labels.
- Beyond these responsible practices, a real challenge is that data labelers are incentivized to work very quickly to label as much as possible in the smallest amount of time. However, having to think quickly is exactly when our biases can come into play (our systems 1 thinking). Raters may not be incentivized to be mindful when doing labeling tasks when being mindful takes more time.
- What might be some ways to address this challenge?

34. Build off of their answers to explain:

- Helping labelers practice quick reactions and exposing bias that can come into play helps folks recognize their own biases so they can work against it in the future. We have a **lesson plan for data labelers** you can check out and use.
- Beyond training, encourage processes for labelers to slow down to use systems 2 (critical thinking) when labeling image data or language data where biases can come into play. Ultimately, having data that is labeled with bias can come back to hurt your project and your organization. Consider building in incentives - financial or otherwise - for accurate and ethical data labeling.

Putting practices into action [30 MINUTES]

35. Let participants know it is time to reflect on their learnings from this session. Direct them to the sample annotation guidelines provided (see Appendix iii) and have them read the assignment alone for 3 minutes.

36. Have participants break into groups of 3. In their small groups, have them brainstorm for

10 minutes about what can be improved upon, and which good practices can be applied to do so.

37. Bring the entire group back together. Have each breakout group share 1-2 ways the guidelines could be improved and which good practices can be applied to do so.

38. When each group has shared. Build off their answers to explain:

- Language varieties other than "Standard" American English ("S"AE) are likely to be labeled as "junk"
  - This reduces the breadth of data that can be analyzed! Research on Child Directed Speech (CDS) finds the use of unique diminutive with reduplicative morphology ("ea ea--eat, eat", "no no--nose nose") in African American English (AAE) CDS. This is not characteristic of European-American CDS, although it is observed in CDS in other cultures. On the basis of these guidelines, this is likely to be labeled as junk.
  - The very label of "junk" propagates the harmful notion that "S"AE is the default, and other variants are somehow less than or deviant from it.
- Categorizing speech as CDS or ADS is itself a subjective process.
  - These guidelines don't define parameters for "child" or "adult". Research finds that Black children often get treated as adults from a younger age than White children.
- These guidelines only focus on binary male / female gender labels
  - Labeling non-binary and gender non conforming folks as "junk" is derogatory and offensive.
  - Only using a simplistic binary idea of gender is restrictive.
- Some good practices can include:
  - Highlight a few of the examples we've just discussed, as well as accompanying research in order to let labelers know what to look out for, especially if they don't speak language varieties other than "S"AE.
  - Note that different cultures have different ways of communicating with children and include diverse audio examples.

Personal action plan & wrap up [20 MINUTES]

39. Direct participants to fill in their Personal Action Plan (see Appendix iv for template). Give them 10 minutes to reflect on the questions provided and individually fill in their Personal Action Plans.

40. Have participants pair up for 5 minutes and share their Personal Action Plans with each other.

41. Wrap up the session. Go around the room and have every participant share:

a. What's one good practice / learning you'll take with you to your upcoming projects?

b. What's one word that describes how you're feeling right now?

Facilitator Notes

1. Conduct an immediate follow up survey (see Appendix v) to understand the effectiveness of the training, as well as the support participants to implement the good practices and lessons.

2. Develop a structured check-in process to follow up with participants after this lesson. This could look like the following:

- After 1 month: Use a survey to collect the following information from participants:
  - What good data labeling practices have you been able to implement?
  - How did that go and what did you learn?
  - Are there any questions you have remaining around labeling datasets and potential bias?
  - What other support would you like?
- After 6 months: Use a survey to collect the following information from participants:
  - How have you actively made use of language that advances racial equity and inclusion in your labeling tasks?
  - Are there any challenges you have faced?
  - What have you learned?

Appendix

For facilitators

- Appendix i. **Deck accompanying lesson plan**

For participant packet

- Appendix ii: **Responsible Practices handout**
- Appendix iii: **Sample annotation guidelines**
- Appendix iv: **Personal action plan template**

Feedback survey templates

- Appendix v: **Immediate follow-up survey**

This case study is an accompanying resource to the guide, **Responsible Language in AI & ML**. It was authored by Ishita Rustagi, Genevieve Smith, and Julia Nee with the Center for Equity, Gender & Leadership (EGAL) at the UC Berkeley's Haas School of Business. It benefited from invaluable feedback and contributions from practitioners at leading tech companies. We appreciate the prototyping and valuable feedback from: Alessio Frenda, Christen Madsen, Dominique Wimmer, Erin Rusaw, Lucie Levavasseur, Paul Nicholas, Tommy Denby, and Valerie O'Brien.

Endnotes

1 Crabb, J. (2019, May 28). Classifying Hate Speech: An overview. Retrieved from https://towardsdatascience.com/classifying-hate-speech-an-overview-d307356b9eba

2 This scenario is based on a real world situation in which an African American-Jewish culinary historian penned an essay on Black food history, resistance, and Black joy. When he ran in through Grammarly, the software's tone detector erroneously suggested that the tone of his piece was angry. He Tweeted about the incident, and representatives from Grammarly responded. They looked into the contents of his work and found that the software had misinterpreted the word "jerk" (as in "jerk chicken"). They issued an apology and claimed to have fixed the issue. Read more at https://twitter.com/Grammarly/status/1314348739822342146

3 We call the variety of English that is commonly promoted in places like business, media, and education "Standard" American English. "Standard" American English is based on the language used by those in power and is not objectively better than any other language variety. For this reason we put "Standard" in quotes. This variety is also referred to as White dominant language. African American English (AAVE) refers to the language varieties used by Black descendants of enslaved people in the US. Black Americans speak a diversity of language varieties, but these varieties do share some traits, including how they are often unjustly treated by those in power.

4 Jurgens et al. (2017). Incorporating dialectal variability for socially equitable identification.

5 Blodgett, S. L. & O'Connor, B. (2017). Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English

6 Matsakis, L. (2018). To break a hate-speech detection algorithm, try 'love'. Wired. Retrieved from https://www.wired.com/story/break-hate-speech-algorithm-try-love/.

7 Davidson, T., Bhattacharya, D. (2020). Examining Racial Bias in an Online Abuse Corpus with Structural Topic Modeling. Retrieved from https://arxiv.org/abs/2005.13041.

8 Meraji, S. M. (2020, April 22). The News Beyond The COVID Numbers. Retrieved from https://www.npr.org/2020/04/21/840609912/the-news-beyond-the-covid-numbers.

9 Ramirez, R. (2020, December 14). How Pacific Islanders have been left to fend for themselves in the pandemic. Retrieved from https://www.vox.com/2020/12/14/22168249/pacific-islanders-native-hawaiians-covid-19-pandemic.

10 Smith, G. & Rustagi, I. (2020). Mitigating bias in artificial intelligence. Center for Equity, Gender & Leadership. Retrieved from https://haas.berkeley.edu/wp-content/uploads/UCB_Playbook_R10_V2_spreads2.pdf#page=33.

11 Arauz Peña, P. 2020, November 9. "Alaskans react to CNN poll labeling Native voters 'something else'". Alaska Public Media. https://www.alaskapublic.org/2020/11/09/alaskans-react-to-cnn-poll-labeling-native-voters-something-else/.

12 While "skin tone" is often labeled instead of race, it is important to also be aware of the limitations / subjectivity of labeling skin tone. Read more here: Dixon, Angela R., and Edward E. Telles. "Skin Color and Colorism: Global Research, Concepts, and Measurement." Annual Review of Sociology, vol. 43, no. 1, 2017, pp. 405–424., doi:10.1146/annurev-soc-060116-053315.

13 More information on data statements for NLP here: https://www.aclweb.org/anthology/Q18-1041.pdf.

14 More information on PAIR's Data Cards here: https://pair-code.github.io/datacardsplaybook/.

15 Ware OR, Dawson JE, Shinohara MM, Taylor SC. Racial limitations of fitzpatrick skin type. Cutis. 2020 Feb;105(2):77-80. PMID: 32186531.

16 Pichon, L. C., Landrine, H., Corral, I., Hao, Y., Mayer, J. A., & Hoerster, K. D. (2010). Measuring skin cancer risk in African Americans: is the Fitzpatrick Skin Type Classification Scale culturally sensitive?. Ethnicity & disease, 20(2), 174–179.