

Project Proposal

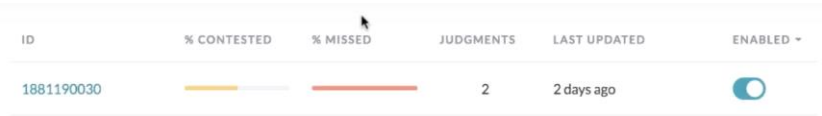



<YAN QIN>

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	<Instead of relying on doctors' expertise to justify cases of pneumonia, we want to build a system as a diagnostic aid for doctors to flag serious cases, quickly identify healthy cases. ML can be used to classify images and label them.>
Choice of Data Labels What labels did you decide to add to your data? And why did you decide on these labels vs any other option?	<p>Three labels. Yes, No and Not Sure. With Not Sure, certain confidence scale 1-4 is offered.</p> <p>The purpose is to build a dataset to distinguishes x-ray images with and without pneumonia symptom. Images without clear indication will be processed based on confidence level.</p> <p>Upside of this labeling method is that options are fairly straight forward. Contributors can easily follow the examples then cross check x-ray images and see if certain area is opaque or any shadow.</p> <p>Downside of this method is that contributors could be biased. They could think most of images are pneumonia relevant. Or most of cases they would just choose Not Sure to complete the questions quickly.</p>

Test Questions & Quality Assurance

<p>Number of Test Questions</p> <p>Considering the size of this dataset, how many test questions did you develop to prepare for launching a data annotation job?</p>	<p>Total 10 test questions. It's 125 entities with 16 pre-labeled. So initially total 26 reference data points are provided for contributors.</p>
<p>Improving a Test Question</p> <p>Given the following test question which almost 100% of annotators missed, statistics, what steps might you take to improve or redesign this question?</p>	 <p><1. Clear instructions to not to miss any question. 2. Review the question and see if it's too difficult or any flaw. 3. Add every question as required field to complete before proceed to next.></p>
<p>Contributor Satisfaction</p> <p>Say you've run a test launch and gotten back results from your annotators; the instructions and test questions are rated below 3.5, what areas of your Instruction document would you try to improve (Examples, Test Questions, etc.)</p>	 <p><1. Instructions need to be clear and concise. All labeling and selection criteria should be clearly explained. 2. Paralleled examples should be provided with enough similarity. Contributors should read through all examples before proceed. 3. The difficulty of test questions should be in similar level as examples. Contributors can cross check test questions and do labeling with right match.></p>

Limitations & Improvements

<p>Data Source</p> <p>Consider the size and source of your data; what biases are built into the data and how might the data be improved?</p>	<ol style="list-style-type: none"> 1. 125 data points may not be representative for pneumonia research. The source of data may not be diverse and random enough to be more inclusive. 2. Positive and negative symptoms may not be balanced in the sample. 3. Need to consider demographic of all children by gender, age, location and history.
<p>Designing for Longevity</p> <p>How might you improve your data labeling job, test questions, or product in the long-term?</p>	<ol style="list-style-type: none"> 1. Acquire high quality data source with needed variables before data processing and training. 2. Build an adaptive process to reflect data source and process data accordingly during labeling job and test questions. 3. Build a feedback loop, involve all stakeholders (engineers, medical staff and project managers) to review and justify if data collecting and labeling process are on the track.