Project Proposal

Hussain Al-Balhareth

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	- Build a product that helps doctors quickly identify cases of pneumonia in children - Build a labeled dataset that distinguishes between healthy and pneumonia x-ray images that can be used by ML engineers later on down the line to build a classification product. - Create a data labeling job using Appen's platform.
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?	- Label 0 for healthy and label 1 for pneumonia case. - Such labels are numeric and helpful in binary classification using ML

Test Questions & Quality Assurance

Test questions:

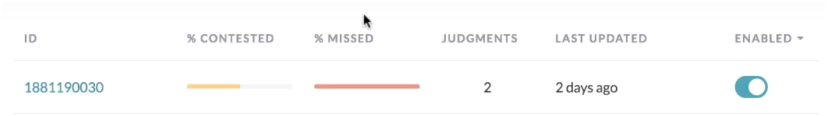
- Does this xray image indicate pneumonia case? (required)

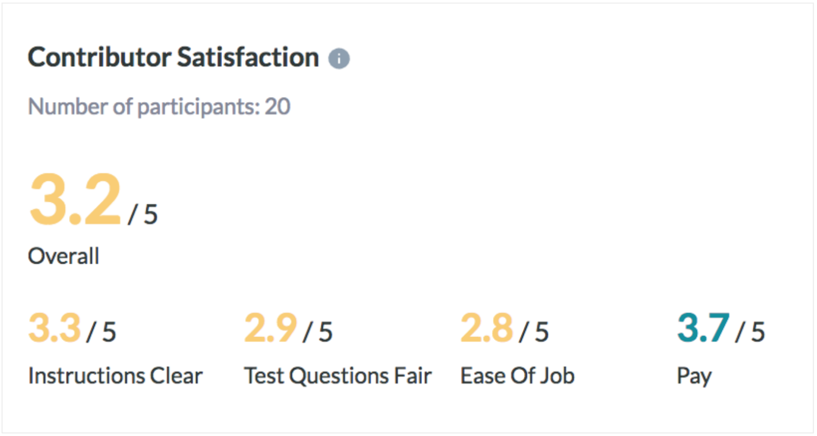
Quality assurance:

- How confident are you with your assessment? (required)

<table>
<tr>
<td><b>Number of Test Questions</b>

Considering the size of this dataset, how many test questions did you develop to prepare for launching a data annotation job?</td>
<td>9 test questions out of 117 cases which is more than 5%.</td>
</tr>
<tr>
<td><b>Improving a Test Question</b>

Given the following test question which almost 100% of annotators missed, statistics, what steps might you take to improve or redesign this question?</td>
<td>
<table>
<tr><th>ID</th><th>% CONTESTED</th><th>% MISSED</th><th>JUDGMENTS</th><th>LAST UPDATED</th><th>ENABLED ▾</th></tr>
<tr><td>1881190030</td><td></td><td></td><td>2</td><td>2 days ago</td><td></td></tr>
</table>
We may augment the instructions or include more examples or such tricky cases.</td>
</tr>
<tr>
<td><b>Contributor Satisfaction</b>

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.)</td>
<td>
<b>Contributor Satisfaction</b>
Number of participants: 20
3.2 / 5
Overall
3.3 / 5 Instructions Clear
2.9 / 5 Test Questions Fair
2.8 / 5 Ease Of Job
3.7 / 5 Pay

I will focus on all of them but on a priority basis starting with more clarifying examples andTest Questions, then Overview/Steps/Rules Tips.</td>
</tr>
</table>

Limitations & Improvements

Data Source Consider the size and source of your data; what biases are built into the data and how might the data be improved?	- Images are for chest x-ray with different sizes and exposure times. - Classification job is a bit tricky as annotators are not specialist in such field like real doctors. So, the challenge is to make this task doable for non-experts as much as possible. - Also, it's best that the data and the images be evenly distributed between Yes and No, High and Low confidence with variety and diversity.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	By accounting for changes in data. I assume in this case the data does not change so a static model is adequate.