

Project Proposal

Svajune Klimasauskaite



Data Labelling Approach

Project Overview and Goal

What is the industry problem you are trying to solve? Why use ML in solving this task?

Problem

Due to the lack of medical professionals and their high salaries, we want to optimise the pneumonia case identification process. The targeted business outcome is to save costs as well as improve decision-making process that should lead to increased customer satisfaction.

Goal

To build a product that helps doctors quickly identify cases of pneumonia in children. The classification system shall have the following key features:

- * Can help flag serious cases
- * Quickly identify healthy cases
- * And, generally, act as a diagnostic aid for doctors

Why ML?!

This is a challenging task because it is not always clear when pneumonia symptoms are present or not in an image. There are a few different visual symptoms that indicate pneumonia. The most important areas are the lungs and the diaphragm.

- * A **normal**, healthy image will depict clear lungs without any areas of abnormal cloudiness/opacity; there may be structured, web-like vasculature in the lungs but otherwise that area should be clear. In healthy images, you are also more likely to see a diaphragm shadow.
- * A **pneumonia** image may include a few things: areas of cloudiness/opacity in several concentrated areas or one large area. You may also see a general pattern of opacity that obscures the structure of the lungs, heart and diaphragm.

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?

“What”

A data annotation job is designed so that a non-expert can identify more noticeable cases of pneumonia.

The label with a scale for Healthy or Pneumonia case with including the confidence level.

- ✱ When the image has clear pneumonia or not case, then annotator should choose 1 or 5.
- ✱ When the image has a NOT such clear pneumonia (e.g. no example in the instruction), then annotator should choose 2 or 4.
- ✱ The selection of 3 would mean an “unknown” case.

“Why”

When we are dealing with a complicated and sensitive topic as medical images, we want to eliminate the scenarios when a Pneumonia case is identified as Healthy one (success criteria for our model is Recall parameter).

Therefore, for the first iteration, we want to train the model with the images which were annotated with the higher confidence level (in order to avoid the “shit in” - “shit out” scenario).

For the second iteration, I'll have to review the images which were annotated with the lower confidence interval in order to update annotation job instructions with more precise description and examples. The goal is to re-annotated those images with a higher confidence level.

Test Questions & Quality Assurance

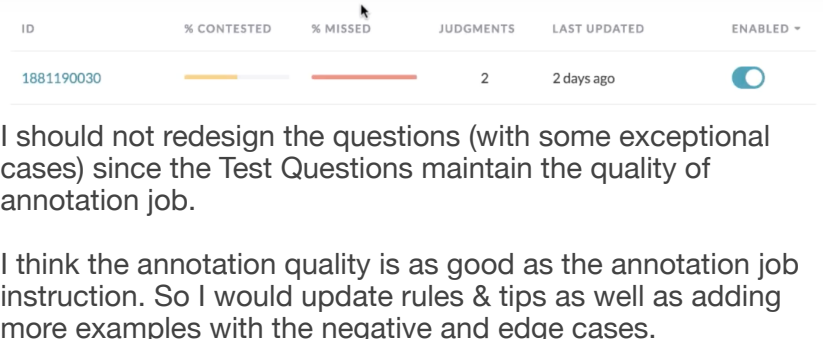
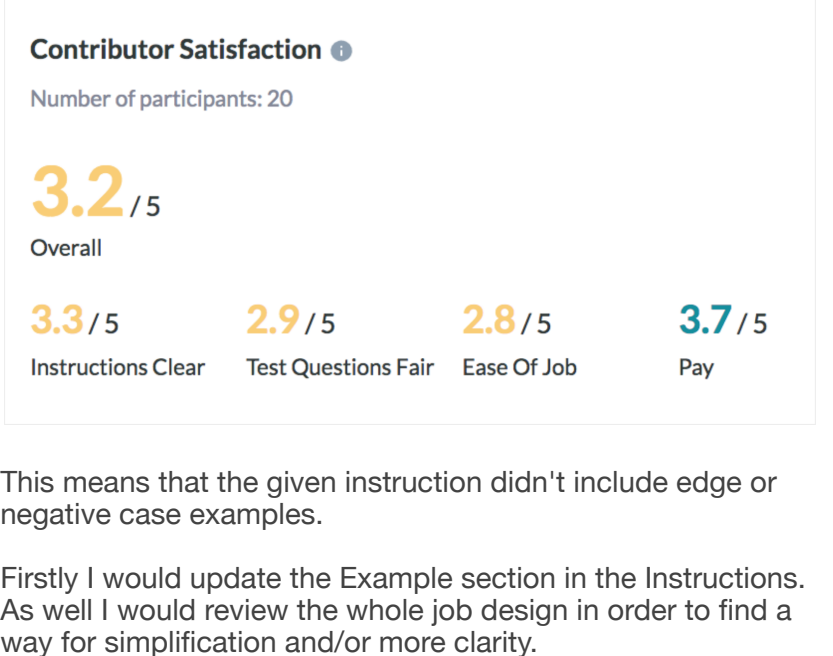
Number of Test Questions

Considering the size of this dataset, how many test questions did you develop to prepare for launching a data annotation job?

The tool recommended to create 8 Test Questions for the first job launch.

In general:

- ✱ It is recommended to have between 50-100 test questions in a job. Because contributors can only see a test question once, the more that are created, the faster the job will complete.
- ✱ The minimum total number of test question should be equal or more to the Number of rows per page +3.

<h3>Improving a Test Question</h3> <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>	
<h3>Contributor Satisfaction</h3> <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>	

Limitations & Improvements

<h3>Data Source</h3> <p>Consider the size and source of your data; what biases are built into the data and how might the data be improved?</p>	<p>Biases are built when we have not equally distributed data, e.g. when data leans towards more Healthy images or more Pneumonia images.</p> <p>We should maintain the 50/50 proportion to remove the bias.</p>
<h3>Designing for Longevity</h3> <p>How might you improve your data labeling job, test questions, or product in the long-term?</p>	<ul style="list-style-type: none"> ✱ The model should be dynamic so that it would continuously learn from the new input. ✱ Annotation job should be constantly updated with new image examples which were annotated with a low confidence level. ✱ In general, ideas for the improvements should come from monitoring job status and metrics.