

Project Proposal





Jay Gohil

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	The industry problem to be rectified here is to build a product that helps doctors quickly identify cases of pneumonia in children. Moreover, using machine learning in the project helps in fast and accurate analysis (after proper model is developed) that will save huge amount of time, which is critical in medical field.
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?	I decided to keep 3 labels for the annotation job for this project, as the project required identification of pneumonia, which required a simple YES/NO as an answer along-with N/A for unidentifiable cases. Moreover, the decision to keep labels 3 was simple – yes for positive identification, no for negative identification and N/A for unidentifiable cases.

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>Considering the size of the dataset provided in the CSV file, I provided 10 test questions as I believed it to be a feasible number in the given project.</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>In case of such situation, the question must be rephrased to emphasize more on the details, the image used can be changed or updates (or checked for viability and relevance with the question) and its general outlook to an annotator should be rechecked.</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>In this case, the instructions should be corrected to be more precise and crisp (and explain each aspect of annotation job), test questions must be relevant to the job (with clear examples for each case and label) and examples provided must be made simple to understand (and should convey all instances clearly).</p>

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?	I think that the size (117) of the dataset is relatively smaller, and the source seems to be from April 2019 of children which makes it to be of a single timeframe. Thus, due to single source of data which is smaller and from single timeframe, I think that dataset cab be biased; for instance, there might not be all possible cases of pneumonia in the dataset (making detection process by ML model limited) and the cases that are of pneumonia can be of limited variation due to variation. Thus, this biases might be present in the dataset, and can be improved by enlarging the size of data with different timeframe and source of origin.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	In the long term, I would like to incorporate more varied examples and test questions that cover several different types of pneumonia cases, reduce the number of labels to 2 (in case initial results show negligible importance and occurrence for 'N/A' label) and make every other aspect of the annotation job short, intuitive and crisp.