Project Proposal

Amisha Muni

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	Many patients need hospitalization, or even lose their lives every year due to Pneumonia. The best method today for detecting Pneumonia is by using a chest XRAY. These chest XRAYS need to be evaluated by an experienced radiologist, and there can be variations based on the experience of the person and the complexities of the particular case. As diagnostic aid that could complement the observations done by a trained radiologist could result in spotting Pneumonia cases earlier or even more uniformly in the case of edge cases. The project aims to use Machine Learning algorithms, which will be trained on healthy and Pneumonia diagnosed Chest XRAYS, so that they can then, when exposed to unclassified Chest XRAYS indicate which XRAYS show symptoms of Pneumonia. A good algorithm trained on well classified datasets procured from hospitals can greatly improve the detection efficiency of the average radiologist, and benefit more patients overall by being a diagnostic aid to the medical community.

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?	The labels chosen are: 1. Is this a chest XRAY? (Yes/No) 2. Is there a cloudy or opaque area in any of the lungs? (Yes/Maybe/No) 3. Is the diaphragm shadow visible? (Yes/Maybe/No) 4. Do you see symptoms of Pneumonia in this image? (Yes/Maybe/No) In the Rules I started by requesting that in case of doubt, the annotators flag it, and that most questions offer a Maybe option. The reason being annotation of these images by lay people is extremely challenging, and indicating this to an annotator right at the outset minimizes the risk they will indicate a result even when unsure. This ultimately improves the quality of the data that we would train the model on. In the tips I started by giving a background on the reason for looking out for opaque or cloudy areas in an XRAY, to aim to build understanding of why this analysis is important. The questions chosen also aim to make the annotator focus on the image, and the right items to look for. i.e. first they indicate really that the image in front of them is a chest XRAY. This avoids any spurious data or super unclear images, and also prompts them to check the orientation and alignment of the presented image. The subsequent question focuses on the lungs, and asks explicitly if cloudy or opaque areas are visible. After that I ask if the diaphragm shadow is visible. And finally, the concluding question is whether the symptoms of Pneumonia are visible in the image. I prefer this to asking for a conclusion right in the very first question, on whether they see Pneumonia or not, and then having follow up questions around the lung opacity or absence of diaphragm shadow.

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?	I created 8 test questions, so as to cover all the possible options presented in the answers. I also tried to upload a link of an image which had a Pelvic XRAY rather than a chest XRAY, so that a negative answer to the very first question was also covered.
Improving a Test Question Given the following test question which almost 100% of annotators missed, statistics, what steps might you take to improve or redesign this question?	(see below)

ID	% CONTESTED	% MISSED	JUDGMENTS	LAST UPDATED	ENABLED
1881190030			2	2 days ago	

A situation like this would mean the image presented in the test question is complex and ambiguous. I would ensure the answers are explained more clearly, and additionally include it in the Instructions itself so that a more detailed explanation with bounded boxes and colours can be given.

If there is a particular aspect that is complex or confusing, I would additionally consider improving my overall test design so that I collect as much information as possible in future runs, and maybe even have a freetext input so that annotators can indicate in detail some feedback on why they chose what they chose for the image.

<table>
<tr>
<td>
Contributor Satisfaction

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>


Given the scores, the Instructions as well as Test questions need improvement.
I may also need to consider if this type of image annotation needs more specialist annotators rather than lay annotators, given the ease of the job is rated low too. However this can be considered after the instructions and test questions are improved.

To improve the instructions, I would elaborate more examples, give a more detailed background and propose a step by step for their consideration.

To improve the test questions, I would aim to further break down the questions to capture as much detail as possible, potentially allow freetext input for questions the annotators consistently get wrong.
</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?	The data may be predominantly from a particular racial background, or skewed in terms of gender or age. There may also be noise, if the images used for training are not well annotated themselves. Or a heavy bias towards detecting pneumonia, if there are too many images with pneumonia vs healthy chest XRAYS. The reverse can also occur, if we have too few examples of images with Pneumonia, it can be harder to actually detect it.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	Long term improvements would imply making sure the dataset is continuously refreshed with newer examples, that there isnt a bias introduced by soliciting images only for certain ethnic groups or genders or ages. Test questions need to be revised too, to avoid annotators clicking through them and clearing them simply as they already know the answers – we need to test them against a changing set of questions. We should also keep in touch with industry advances, and ensure the labelling, instructions and questions keep up and change the content accordingly. Regular discussion with the medical community is critical to ensuring we stay on track.