Project Proposal 

#### *Shaun M Rebello*



# Data Labeling Approach

|  |  |
| --- | --- |
| **Project Overview and Goal**What is the industry problem you are trying to solve? Why use ML in solving this task? | This project is to create a data annotation job to support a machine learning solution that can detect signs of Pneumonia in human x-rays.  Having an ML solution to support x-ray analysis can help scale the treatment of Pneumonia cases by eliminating healthy subjects from review by humans, thereby improving the utilization of doctors and specialists.  e.g. Doctors will spend less time looking at healthy x-rays to eliminate them and more time treating patients whose images show signs of Pneumonia. |
| **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 had originally had only yes/unsure/no as options for “Are there any indications of Pneumonia in this image?” and then a sliding scale of certainty from “100% yes down to 100% no”  I felt that this was somewhat redundant and less useful than a simple yes/no and a 1-5 scale of how sure the annotator is. This way, when looking at annotation results (and possibly confusion matrices) we can cross reference certainty and see if our annotators were confidently wrong, or if there are certain types/severities of Pneumonia that were more likely to cause uncertainty. I also felt that putting “unsure” as a primary option could lead to annotators using it as an escape route from difficult questions. |

# 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 thought 5-10% was a good amount, so I created 10 with the goal of 5 each, Pneumonia and no Pneumonia. |
| **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? | I don’t quite understand this question. If only 2 annotators have tried to answer, and both got it wrong, I would consider rephrasing the question, viewing the image myself to make sure its not incorrect or too ambiguous, and also be mindful that only 2 responses is not a strong case (in my opinion) |
| **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.) | I would be curious as to how these scores compare to the median and average scores of other, similarly sized image annotation jobs.  I would then;  1 – Review my test questions for clarity and brevity  2 – Review the most often incorrectly answered questions to make sure they are clear, fair and that the images are correct.  3 – Determine the issue of ‘Ease’ by comparing to similar annotation jobs and determining if it is possible to adjust (with the understanding that viewing x-rays and looking for shadows is not as straightforward as identifying flowers v. no flowers or whether or not a parking sign is present) |

# 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 would think the data was selected with Pneumonia in mind; showing examples with and without Pneumonia. I suspect there are many x-rays that show shadows and other anomalies that are not Pneumonia-related which could confuse our project when encountered later.  We also need a lot more data. |
| **Designing for Longevity**How might you improve your data labeling job, test questions, or product in the long-term? | As per the above, I would introduce more x-rays that are “not healthy” but don’t necessarily indicate Pneumonia. We need edge cases and the ability to identify more conditions (or even simply a ‘Pneumonia, no Pneumonia, other’ option so that images can be labeled for later when we expand the project. |