Project Proposal 

#### *<Gauri Ruhal>*



# Data Labeling Approach

|  |  |
| --- | --- |
| **Project Overview and Goal**What is the industry problem you are trying to solve? Why use ML in solving this task? | It helps doctors to quickly determine if there are pneumonia symptoms in the images we provide. Using ML here helps the doctors to quickly eliminate cases that do not have any pneumonia symptoms and spend more time on the cases where there are symptoms. Also using ML might also help the doctor to rethink his decision if the model says otherwise than what doctor thinks. |
| **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? | There are three labels that is “Yes”, “No” & “Unknown”. The first two labels are chosen when we need to decide if there are pneumonia symptoms given the image. The third label that is “Unknown” is chosen to leave room for uncertainty. We could have also chosen the labels as “pneumonia”, “healthy” or “unknown”. To keep it simple, I decided to choose yes/no. |

# 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? | 12 test questions were developed. The answer to 4 of the questions were yes and to the other 4 were no and the rest of the 4 were unknown so that there is no bias towards any specific label. |
| **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 might do one thing by rephrasing the question to remove any ambiguities that might be present. Then take a step back and see if the rules specified are clear and unambiguous. And also provide a more detailed description so that the annotator knows why it was labeled the way it is. |
| **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 provide more examples for each label and also try to see if the rules and tips stated are clear and non-ambiguous. Also I would try to provide more tips. |

# 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 size of the dataset currently that we are dealing with is not large enough for a machine learning model to learn patterns. We might need some more data for the ML model to be robust enough to deal with all possible scenarios. Only 24 labels are given, 8 are one label and the other 8 & 8 are of other labels (no and unknown). So, from high level there does not seem to be any bias as of now. But if there are biases in the dataset you need to account for it either by augmenting the class that does not have more labels or throwing away some data from the class that has more data. |
| **Designing for Longevity**How might you improve your data labeling job, test questions, or product in the long-term? | Test questions can be improved as you come across new data and with more corner cases.  Rules and tips also might need to be updated to reflect that changes. |