Project Proposal 

#### *Noureddine Ettayyeby*



# Data Labeling Approach

|  |  |
| --- | --- |
| **Project Overview and Goal**What is the industry problem you are trying to solve? Why use ML in solving this task? | In this project we are trying to build an Ai product to assist doctor on determining if a patient has pneumonia or not, the aim of this project is to offer a solution for eliminating healthy patient and making the process of detecting pneumonia faster by giving doctors a good guess whether a patient has pneumonia. This a good project for using ML because the data is there and the process can be automated and would save a lot of time for doctors and help get to treatment to patient who needs it. |
| **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 choose to go with three labels YES if the patient has pneumonia, NO if the patient is healthy and NOT SURE if you can’t decide, And I think this choice is better for this situation since or trying to build a classifier it made sense to lunch a classifying job so we can get the labels for the clear one and then see how we can the job be modified to get labels for the undecided images. |

# 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 have around 118 images in the dataset so I developed 8 test questions for the job, which around 6% of the dataset. |
| **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 would take a closer look to this question and determine the cause why annotators missed it and either improves the instructions to point cases like these especially if there are lot of them, or include in the examples so people can identify cases like easily. |
| **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 definitely try to improve the instructions portion of the document, because it rated below 3.5 and clear instructions  Would surely help in making the job successful and also try to improve the example portions to include cases that are hard and, in the meantime, improve the test questions by maybe including cases like the test question in the examples, to make easy for the annotators to pick those cases in the job and respond to the test question at the same time. |

# 

# 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? | Since we have in this job data collected from X-ray images of patient, we can expect that data is imbalanced, naturally it we would find more cases of the YES class, which cloud tip the annotators to classifies all images with a YES label, one why to address this is collect more data from healthy patient so we can have a more imbalanced dataset and lunch a more successful job and train a better model overall, another way to address this problem is oversample the No class so we can have more instances of this class. |
| **Designing for Longevity**How might you improve your data labeling job, test questions, or product in the long-term? | We can improve the job by designing it to incorporate the feedback from annotators to improve test questions and instructions, also we should watch for the changes in the data and retrained as needed to improve the products so we should account for the possibility that we may need to lunch the jobs multiple times as we collect more data or as data changes. |