Project Proposal

Lyca Constantino

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 product is to help doctors quickly identify if there are signs of pneumonia based on the images we provide. With the help of ML, the process of classifying which ones need immediate help and eliminating healthy cases will be quicker. In addition, using ML might be helpful for the doctors to double-check their initial diagnostics.
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?	Positive, negative, and unknown are the data labels. Positive and negative are chosen as we need to identify if the image has pneumonia signs. However, the third label, "unknown", is chosen to pave way for the images that are quite unclear and need more investigation to determine whether it's negative or positive.

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?	9 test questions were developed. Three questions per data label have been made so there is no bias towards any specific label.
Improving a Test Question	Run through all the questions to check if there are some unclear ones, if so, these must be rephrased. Adding a detailed

Given the following test question which almost 100% of annotators missed, statistics, what steps might you take to improve or redesign this question?	description will be a big help to the annotator to understand more 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.)	Providing more examples for each label may possibly add more credibility to each label. Making sure that the instructions or steps are clearly stated. I also suggest using very basic English for everyone to understand it more. Adding tips and more examples will also be helpful.

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 believe that the size of the dataset is not large enough to fully teach patterns to the ML model. It is a need to add more data for it to be more robust on all possible scenarios. There may be no biases currently but definitely need to have almost equal parts of each data label. In addition, the data source could also be improved by having different lighting conditions, body sizes, image orientations, and image sizes.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	Test questions can continuously be improved especially when the trends are consistently changing. That means new data will always be generated so the whole data labeling job might also need to be updated continuously to keep up with the trend and be up to date.