Project Proposal

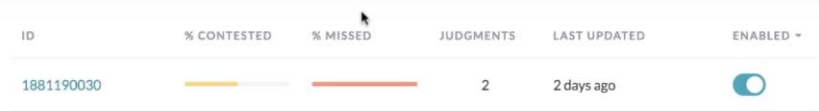
Eda AYDIN

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	Help doctors quickly determine if there are pneumonia in the chest x-rays images we provide. Using Machine Learning here helps doctors to quickly eliminate cases that do not have any pneumonia symptoms. Spend more time on the cases where there are symptoms.
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. (‘yes’, ‘no’, ‘unknown’). The first two labels were chosen as we need to decide if the chest x-ray image of patient have pneumonia. The third label ‘unknown’ is chosen to leave room for uncertainty.

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?
	25 test questions were developed.
	- The answer to 52% of the questions were no.
	- The answer to 48% of the questions were yes.
	- there is no bias towards any specific label.
	  ○ The answer to 88% of the questions were high confident.
	  ○ The answer to 12% of the questions were low confident.

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?	 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.)	 Provide more examples for each label. Provide more tips to determine whether there is pneumonia or not.

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 need some more data for ML model. 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. Dataset could also be improved to be diverse and have more variety like images with lighting conditions.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	Test questions can be improved according to new data with more corner/edge cases.