

Project Proposal

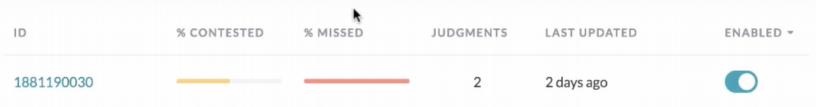



Ayşe DEMİREL

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	My goal in this project is to build a product that helps doctors quickly identify cases of pneumonia in children. I want to quickly identify healthy cases. There are a lot of x-ray samples and we don't have enough time to classify them before any children died. If we can classify them, every person can know the illness before detailed looking with doctor. So early diagnosis can be happened.
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?	"Pneumonia, healthy and unknown" labels are chosen. The goal is that find the pneumonia in this project, but there are also healthy images in the x-ray samples. Unknown label is to prevent misclassification. Every images is not very clear to classify so it can be affect the result if we have just two options.

Test Questions & Quality Assurance

<p>Number of Test Questions</p> <p>Considering the size of this dataset, how many test questions did you develop to prepare for launching a data annotation job?</p>	<p>The size of the dataset is 117. I created 10 test questions. It is more than %5 so I stopped to add more test question.</p>
<p>Improving a Test Question</p> <p>Given the following test question which almost 100% of annotators missed, statistics, what steps might you take to improve or redesign this question?</p>	 <p>I think we can redesign it with more explanation. If annotator can not understand the goal and the options, missed situation is more happened.</p>
<p>Contributor Satisfaction</p> <p>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.)</p>	 <p>I think examples are not enough or not clear. Test questions also should be more. If we have more data, adding them can be helpful. Also we should think about the lack of labeling. Which label are we use ? Is it really convenient ? If we have "other" label option and the option was selected more than the correct option, labeling problem is on our desk.</p>

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?	Healthy situation is on the background, but the project user will see both healthy and pneumonia in different time. And also another illness can be find with same diagnosis. We should think about how we can separate them.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	Firstly, annotators feedback is important. The feedback will guide the road of our goal. We have 3 label now, but it can be change. Test questions can be more in the long term. When we create the test question, considered %5 of all test data.