

Project Proposal



Taif ALharbi

Data Labeling Approach

Project Overview and Goal

What is the industry problem you are trying to solve? Why use ML in solving this task?

Create a product that aids doctors in efficiently identifying pneumonia cases in children. This solution aids in promptly identifying critical cases, distinguishing healthy cases, and serving as a diagnostic tool for doctors.

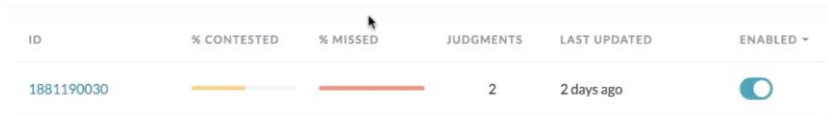

Using ML enables doctors to identify pneumonia cases more efficiently, can save valuable time for doctors, allowing them to focus on critical cases that require immediate attention, can identify common indicators of pneumonia more effectively than traditional manual methods, help doctors identify cases that require urgent intervention, ensuring timely and appropriate medical attention. Additionally, the implementation of machine learning can provide doctors with alternative perspectives, encouraging them to reconsider their initial assessments if the model suggests a different diagnosis.

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?

To classify the presence of pneumonia symptoms based on an image, three labels were designated: "yes," "no," and "unknown." The selection of the first two labels facilitates the decision-making process regarding the existence of pneumonia symptoms. The inclusion of the "unknown" label allows for the consideration of uncertainty, providing flexibility when a definitive determination cannot be made and then choosing the percentage to be pneumonia.

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>10 test questions were developed. The answer to 45% of the questions was "yes" and the other 45% was "no", There is no bias towards any specific label.</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>Provide a comprehensive explanation to the annotator, outlining the specific reasons behind the labeling decision. Review the specified rules to ensure their clarity and lack of ambiguity. Additionally, Revise the question to eliminate any potential ambiguities.</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>In this case, it should improve the instructions by making them more concise and specific, clarifying every aspect of the annotation task. The test questions should directly relate to the job and include explicit examples for each case and label. Moreover, the provided examples should be simplified for better comprehension, while effectively conveying all possible instances. Also, provide additional instances that represent each label and assess the clarity and absence of ambiguity in the stated rules and tips. Furthermore, expand upon the provided tips to ensure they are comprehensive and easily understandable.</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?	The dataset appears to be relatively small, consisting of 117 observations. The size of the dataset that we are currently 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. This limited size and single timeframe of the dataset may introduce bias. For example, it is possible that not all possible cases of pneumonia are included in the dataset, which could restrict the capabilities of a machine-learning model in detecting pneumonia accurately. Additionally, the cases of pneumonia present in the dataset may exhibit limited variation due to the constrained timeframe. Consequently, biases might be present in the dataset, and these biases could be mitigated by expanding the dataset to include data from different diverse sources of origin.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	In the long term, my objective is to introduce a greater variety of examples and test questions that encompass multiple types of pneumonia cases. Additionally, I aim to streamline and simplify all other aspects of the annotation task, ensuring they are concise, intuitive, and clear. As I encounter new data and encounter more complex cases, I will continuously enhance the test questions. It may also be necessary to update the rules and tips to ensure they remain relevant and aligned.