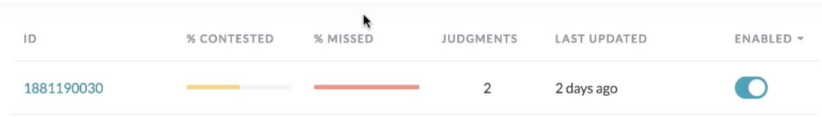
Project Proposal

Jude Anlasun

Data Labeling Approach

Project Overview and Goal What is the industry problem you are trying to solve? Why use ML in solving this task?	According to the United Nations Children's Fund (UNICEF), a child dies of pneumonia every 39 seconds globally. Pneumonia is the number one killer infectious disease among children under 5 claming the lives of about 2,200 children every day. The most common diagnosis of pneumonia by healthcare providers is the full examination of patient health history and physical exams and if pneumonia is suspected, it is complemented by Chest X-Ray, which may show an infiltrate, which is a collection of pus, blood, or protein in the lung tissue. During the examination of Chest X-Rays healthcare providers can make mistakes or take alot of time in their diagnosis process. To reduce the mistakes and the time taken by health care providers in diagnosing pneumonia in children, a Machine Learning product that can help doctors quickly identify cases of pneumonia in children would be built. Using Machine Learning (ML) would make it easier and quicker for doctors to identify serious cases of pneumonia using images of Chest X-Ray. Provide a quicker way to identify healthy cases and also act as a diagnostic aid.
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?	The choice of data labels choosen for this data annotation job can be classified as: • Yes – evidence of pneumonia given that there is cloudiness/opacity in several concentrated areas or one large area. • No – no evidence of pneumonia or healthy image that dipicts clear lungs without any areas of abnormal cloudiness or opacity. • Unknown – Difficult to identify if there is evidence of pneumonia or not.

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?	Twenty (20) Test questions were developed for this data annotation job.
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?	 • Questions should be rephrased to for improved understanding. • Make sure the steps and rules are clear and well described for annotators to know exactly what to do.
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.)	 The "Ease Of Job" and "Test Question Fair" score is very low which may bother on user experience and clarity, I would therefore focus on improving the explanation of the steps for for clarity and also improve on how the question is asked.

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 do not seems to be large enough for Machine Learning model as machine learning requires alot more data to make accurate predictions and to reduce biases. The dataset lacks additional attributes/variables that can describe the ages, gender or location of the population. Having these attributes/variables available would help the machine learning algorithm to have more patterns to learn from.
Designing for Longevity How might you improve your data labeling job, test questions, or product in the long-term?	1. Continuous improvement of the test questions especially the ones that performed poorly with incorrect answers. 2. Improving on the steps and provid further describtive infographics for clear explanation would help. 3. Monitoring contributors review scores to focus on improving under performing areas of the product.

References:

[1] https://data.unicef.org/topic/child-health/pneumonia/

[2] https://www.verywellhealth.com/diagnosis-of-pneumonia-4160855