Project Proposal 

#### *Mohammed Al-Dhabaab*



# Data Labeling Approach

|  |  |
| --- | --- |
| **Project Overview and Goal**What is the industry problem you are trying to solve? Why use ML in solving this task? | Developing digital product via ML technology, That act as an assistant for doctors to recognize cases of pneumonia in children with saving much time! As well as could assistance with dangerous situations of pneumonia. |
| **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? | **Binary classification:**  Infected, For if there are some sign of pneumonia as well as the percentage of it. **(Handling Uncertainty)**  Normal, For if the child not infected via pneumonia.  Not certain, For if the human annotators not sure.  I decided on these labels, because our goal to identify if the child infected via pneumonia or not. Also, it is easy for annotators. Other data labels might not be easy for annotators and help us in our case. |

|  |  |
| --- | --- |
| **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? | Developed about 12 test questions approximal 10% of total number of data (Images). Balancing them between all my labels (Infected, normal and not certain).  It is recommended to have between 50-100 test questions in a job. as well as it recommend every answer to be represented at least once in your first 8 Test Questions. (According to [Appen](https://success.appen.com/hc/en-us/articles/213078963-Test-Question-Best-Practices)) |
| **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? | * Examine my instructions and test questions with trying to figure out in which part\parts I fell short **(Solving right problem)**, then fix it\them. If it is at instructions I going to make more obvious instructions as well as adding more different example to help out the annotators. * Even, I would try to show the importance of doing the job honestly and add some impactful things (words, image …etc.). * As well, add notes or tips if the annotators make mistakes, or add space to encourage annotators to write their opinion about the job. * And the list goes on…. |
| **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.) | It depends in which part of the job I has fall short, But first I would like to see annotators comments then I will examine my instructions and test questions. After that I will decide where is the problem to improve or fix it **(Solving right problem).** |

# Test Questions & Quality Assurance

# 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? | * Some of the x-ray images are of poor quality, we can improve them by taking or compensating with clearer and higher quality x-ray images. **(Sample bias)** * Some annotators maybe will not be objective when they doing their job that result in subjective thoughts control their labeling habits **(Observer bias)**. Here we can improve our data via having clear guidelines and methodology set up for the experiment as well as ensuring that annotators have objective thoughts . * Some annotators not trained for this kind of annotation job they might be make many mistakes, so we can avoid this bias **(Measurement bias)** to improve our data via hiring humans who are trained. * …. |
| **Designing for Longevity**How might you improve your data labeling job, test questions, or product in the long-term? | It depends on my data:   * If my data is not changeable I can use static model, that means there are no need for update. Which is trained once and used indefinitely until an update is required. * If my data is ever evolving I would use dynamic model, that means continuously update my data. which is continuously trained on new data, so it can keep learning from new input. * If my data is change frequently, I could need to continue making changes to my annotation job to reflect more relevant definitions or examples. |