**PROJECT PROPOSAL**

#### *Gouthamsimha Koppula*



# Data Labeling Approach

|  |  |
| --- | --- |
| **Project Overview and Goal**What is the industry problem you are trying to solve? Why use ML in solving this task? | **This model helps doctors quickly identify cases of pneumonia in children. Using ML here helps doctors in quickly identifying the pneumonia cases and it reduces the time doctor spend in the cases which doesn’t show any 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? | I decided to use 3 labels i.e., YES, NO and Confidence level on the scale of 1-10.  Yes- if there are Pneumonia symptoms present.  No- If there are no Pneumonia symptoms.  The third one is to know how confidence of annotator about the symptoms on the scale of 1-10. |

# 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? | I developed a total of 8 questions for this Data Annotation job. Answers to 4 questions were YES and 4 questions were No ,so that there is no bias towards any specific label |
| **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? | This is an opportunity to identify what’s going wrong and where it’s going wrong. We can redesign by augmenting the instructions and adding new examples. It helps if we also the remove the uncertainties that might be present. |
| **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.) | This shows where the contributor misunderstood the Job. We can increase the satisfaction by adding more relevant examples and more test questions. Make sure the rules are clear and add more tips. |

# 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 data currently have is very less for a ML Model to learn patterns.  Data can we improved by cropping the image perfectly, Lightning conditions also effect the data. |
| **Designing for Longevity**How might you improve your data labeling job, test questions, or product in the long-term? | For evolving data, We will be using a Dynamic model. It continuously trains on new-data. Test questions can improved as you face new set of data, Examples, tips and rules. |