Report

Step 1: Randomized data

I created a function: Rater Data seeder, that links to my database (MongoDB), creates models and saves it. It can be found at lodestone/seed/rater\_data\_seeder.js. I set the number of data entries to be saved to 50,000, which took about 8 minutes to load. There are a total of ~45,000 documents after the seeding because the same raterId and taskId have a unique index constraint. I’ve also included a zip of the raw data in chart form.

Step 2: Visualization

I’ve created bar charts for all queries and attached images to each of them in the zip as well which can be found at lodestone/chart\_images

1. *What is the agreement rate between the engineer and all the raters for each day?*

* Daily agreement rate between engineers and raters for 3-label is about 34%
* Daily agreement rate between engineers and raters for 5-label is about 20%
* See Daily Agreement Rates Chart

1. *What is the agreement rate between the engineer and all the raters for each week?*

* Weekly agreement rate between engineers and raters for 3-label is about 34%
* Weekly agreement rate between engineers and raters for 5-label is about 20%
* See Weekly Agreement Rates Chart

1. *Identify raters that have the highest agreement rates with the engineer.*

* *Raters K, A, L, C, J have the highest agreement rate on average with engineers*

1. *Identify raters that have the lowest agreement rates with the engineer.*

* *Raters Y, M, I, R have the lowest agreement rate on average with engineers*

1. *Identify raters that have completed the most Task IDs.*

* *Raters T, J, W, H have the most tasks completed at ~1850, ~1820, ~1820, ~1790, respectively*

1. *Identify raters that have completed the least Task IDs.*

* *Raters Z, F, B, C have the least tasks completed at ~1719, ~1710, ~1705, ~1690, respectively*

1. *What is the precision for each of the 5 labels?*

* *5-Label precision is about 20%*

1. *What is the recall for each of the 5 labels?*

* *5-Label recall is about 80%*

1. *What is the precision for each of the 3 labels?*

* *3-Label precision is about 34%*

1. *What is the recall for each of the 3 labels?*

* *3-Label recall is about 66%*

1. *What is the overall agreement rate considering that the raters have to be in agreement with both the engineer's 3-label answer and the engineer's 5-label answer.*

* *Raters with overall best agreement rate are I, R, T, L with 7.9%, 7.8%, 7,8%, 7.8%, respectively*

*Questions worth considering:*

*What can you do to improve agreement rates over time?*

* I think it’s important to figure out the root cause of why the agreement rates don’t match. If the agreement rates between raters are consistent then every rater must think alike and every engineer thinks alike. Therefore, our data is biased because there is only one perspective on both sides.

How do you improve precision of a label over time?

* We can see that among raters that complete the same Task ID, their answers are consistently ~34%. This shows regardless of the task, raters can’t come to an agreement with each other, likewise for engineers as well. To figure out the cause and improve precision raters need to collaborate with each other and engineers with themselves also.

What changes are needed or required to improve your dataset to achieve over 90% agreement, precision, or recall?

* The main thing is communication between all parties involved. If raters know engineers want, agreement rates will improve. If raters think alike, precision will improve.

Why do some raters perform better than others?

* In general, the precision and agreement rates are low, but that’s because as a whole, the raters and engineers, are not on the same wavelength; however, there are some engineers that agree with some raters. Those outliers should be used as a seed for future generations as a means to improve agreement and precision rates.

*Step 5:*

*Identify 3 more potential questions to consider that can be used to identify issues among raters.*

How can tasks be distributed so that raters develop growth along the way thus affecting future tasks?

Are engineers correct 100% of the time and does that mean there’s no room for improvement in engineers or changes in their answers?

Is high agreement rate something that is always positive? Conflict fosters innovation and creativity. Maybe by having low agreement rates, it will allow all parties to evolve.

Step 6 (Optional):

SELECT \* FROM raterdatas

WHERE cast (datediff (day, 0, date) as datetime = ‘2005-10-06’