Quality Match

Bicycle project crowd evaluation

By Nguyen Duc Tuan

Task 1 - a. Annotators

Total of 22 annotators:

- annotator_01
- annotator_02
- annotator_03
- annotator_04
- annotator_05
- annotator_06
- annotator_07
- annotator_08
- annotator_09
- annotator_10
- annotator_11
- annotator_12
- annotator_13
- annotator_14
- annotator_15
- annotator_16
- annotator_17
- annotator_18
- annotator_19
- annotator_20
- annotator_21
- annotator_22

Task 1 - b. Annotation time

Initially, the minimum and maximum annotation time were found as following:

- Min duration: -99999 ms
- Max duration: 42398 ms

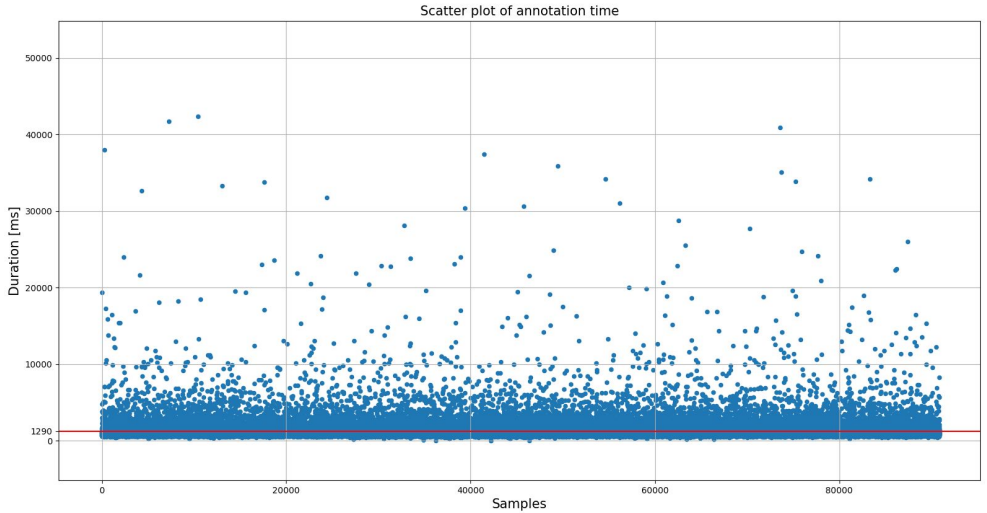
All the values “-99999” ms are considered as data issue and are removed from the dataset.

Additionally, results that are marked as corrupted were also removed.

Task 1 - b. Annotation time

With data issue and ‘corrupt_data’ removed

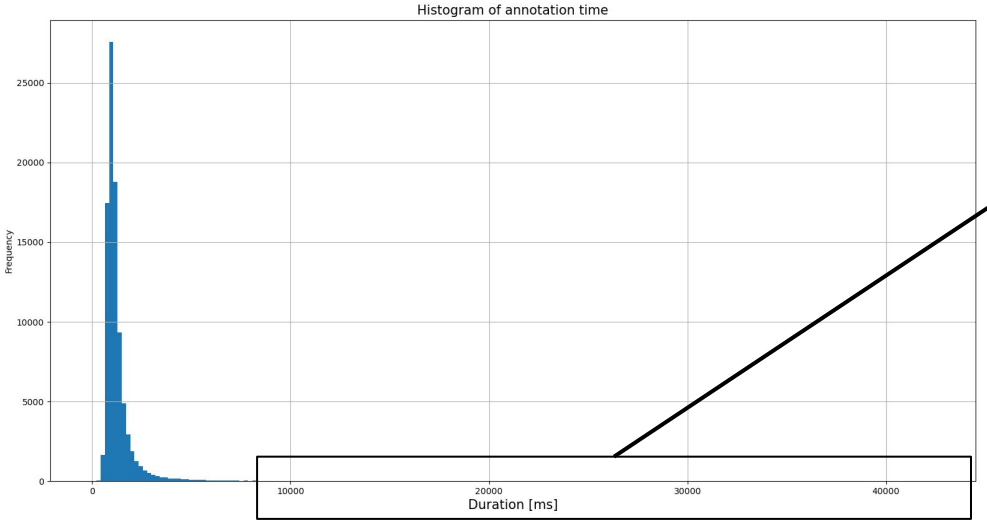
mean	1289.94
std	1124.01
min	10.00
25%	887.00
50%	1058.00
75%	1328.00
max	42398.00

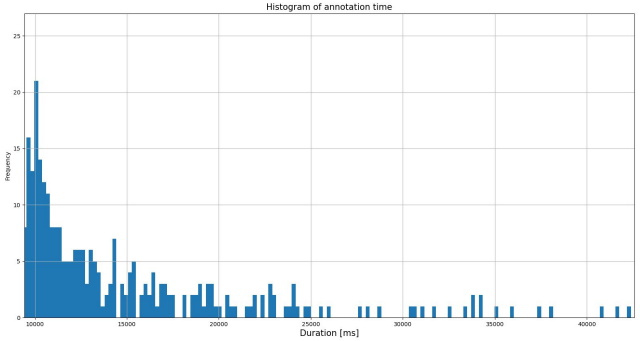


(measured in ms)

Task 1 - b. Annotation time

Histogram plot





Task 1 - b. Annotation time

Annotation time sorted by annotators

Max annotation time	42398.00 ms	annotator_06
Min annotation time	10.00 ms	annotator_04
Max average annotation time	1687.78 ms	annotator_19
Min average annotation time	879.43 ms	annotator_22

Task 1 - b. Annotation time

Annotation time sorted by images

Max annotation time	42398.00 ms	img_5245
Min annotation time	10.00 ms	img_5100
Max average annotation time	5577.50 ms	img_1340
Min average annotation time	738.90 ms	img_8427

Task 1 - b. Annotation time

Images

img_5245

Max annotation time by an annotator

img_5100

Min annotation time by an annotator

img_1340

Max average annotation time

img_8427

Min average annotation time

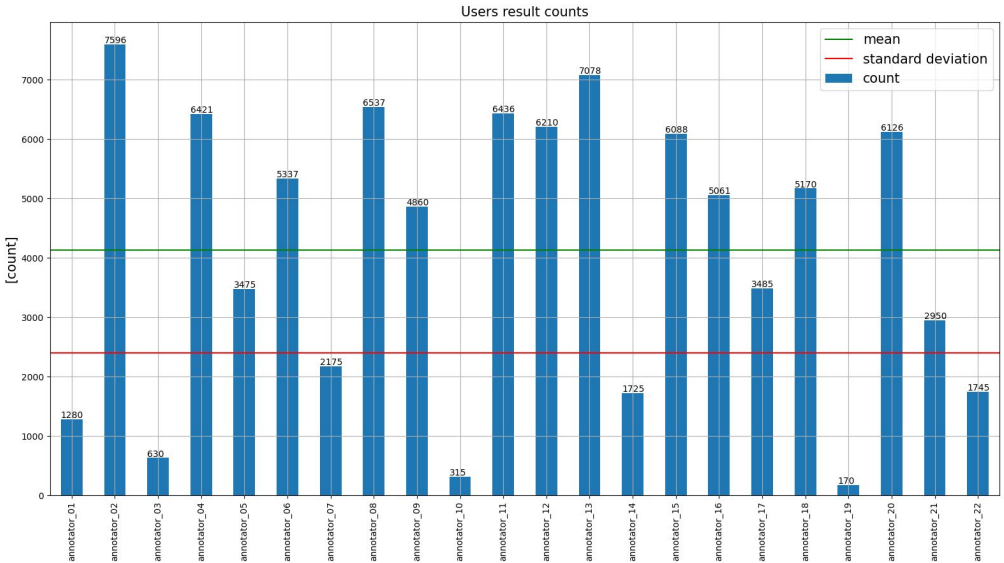
Task 1 - c. Annotators' results

There are huge differences between the sample sizes of each annotator. Some annotators have up to approximately 6000 samples, while some others have only less than 500 samples.

The sample size will be used as one of the attributes to evaluate the performance of the annotator.

mean	4130.45
std	2403.23
min	170.00
25%	1852.50
50%	4960.50
75%	6189.00
max	7596.00

(measured in samples)

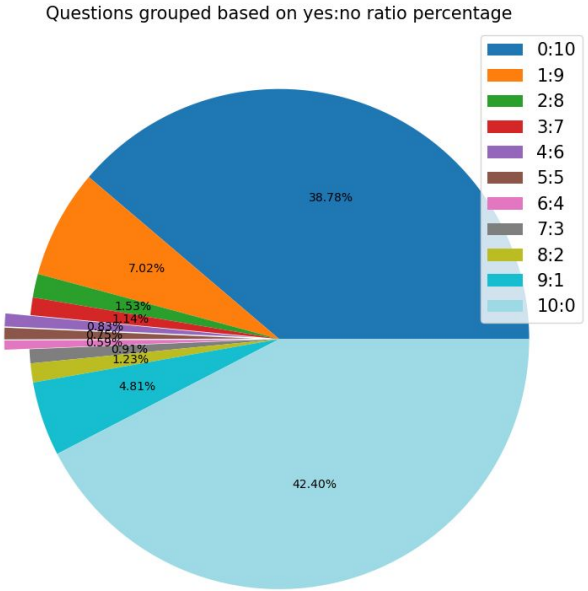


Task 1 - d. Highly disagree questions

The highly disagree rate is decided based on the ratio of yes/no answers of each image

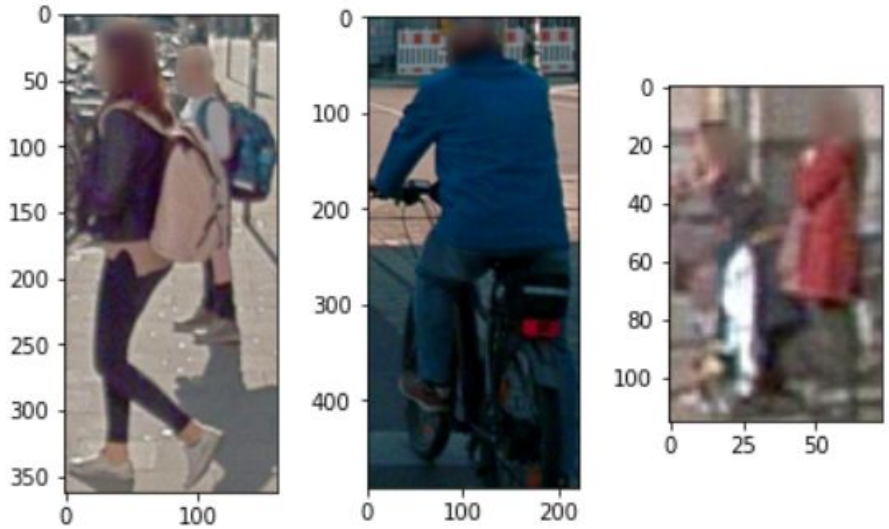
The sum of group 4:6, 5:5 and 6:4 indicates approximately 2.17% of the total questions are highly disagreed.

yes:no	count	percent
0:10	3524	38.78%
1:9	638	7.02%
2:8	139	1.53%
3:7	104	1.14%
4:6	75	0.83%
5:5	68	0.75%
6:4	54	0.59%
7:3	83	0.91%
8:2	112	1.23%
9:1	437	4.81%
10:0	3853	42.40%



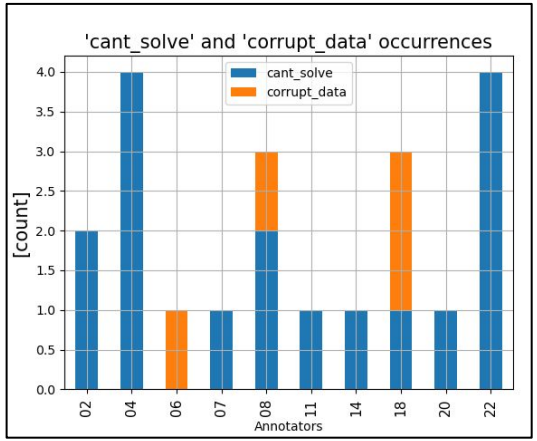
Task 1 - d. Highly disagree questions

Examples of highly disagree questions



Task 2 - ‘cant_solve’ and ‘corrupt_data’

The percentages of 'cant_solve' and 'corrupt_data' to happen are 0.5035% and 0.0727% respectively.



vendor_user_id	cant_solve _percent	corrupt_data _percent
annotator_02	0.0263%	0.0000%
annotator_04	0.0623%	0.0000%
annotator_06	0.0000%	0.0187%
annotator_07	0.0460%	0.0000%
annotator_08	0.0306%	0.0153%
annotator_11	0.0155%	0.0000%
annotator_14	0.0580%	0.0000%
annotator_18	0.0193%	0.0387%
annotator_20	0.0163%	0.0000%
annotator_22	0.2292%	0.0000%

- 'annotator_04' and 'annotator_22', each marked 4 samples as 'cant_solve'.
- 'annotator_18' is the one made use of the option 'corrupt_data' most (2 samples).

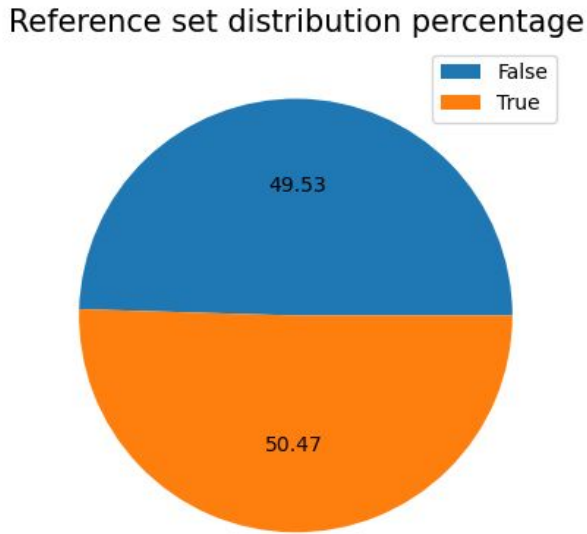
There is no repetition of any sample among those marked with 'cant_solve' and 'corrupt_data'. Each sample in this region is unique.

These numbers can be considered as immaterial as their percentages are insignificant.

Task 3 - Reference set balance

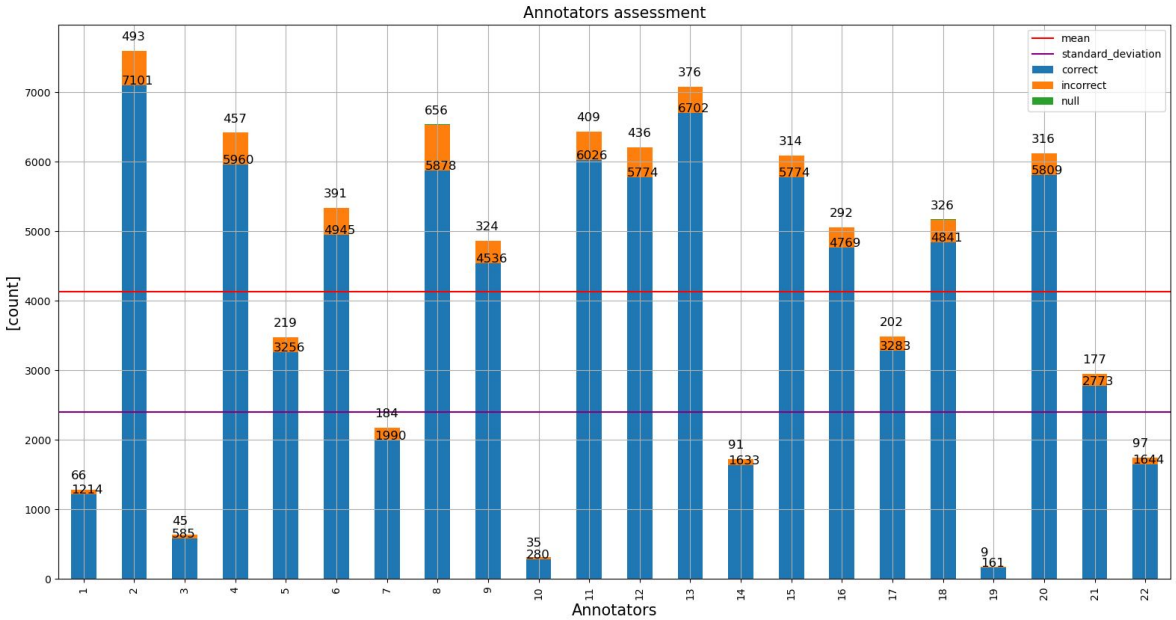
Conclusion: the reference set is balanced.

is_bicycle	count
False	4501
True	4586

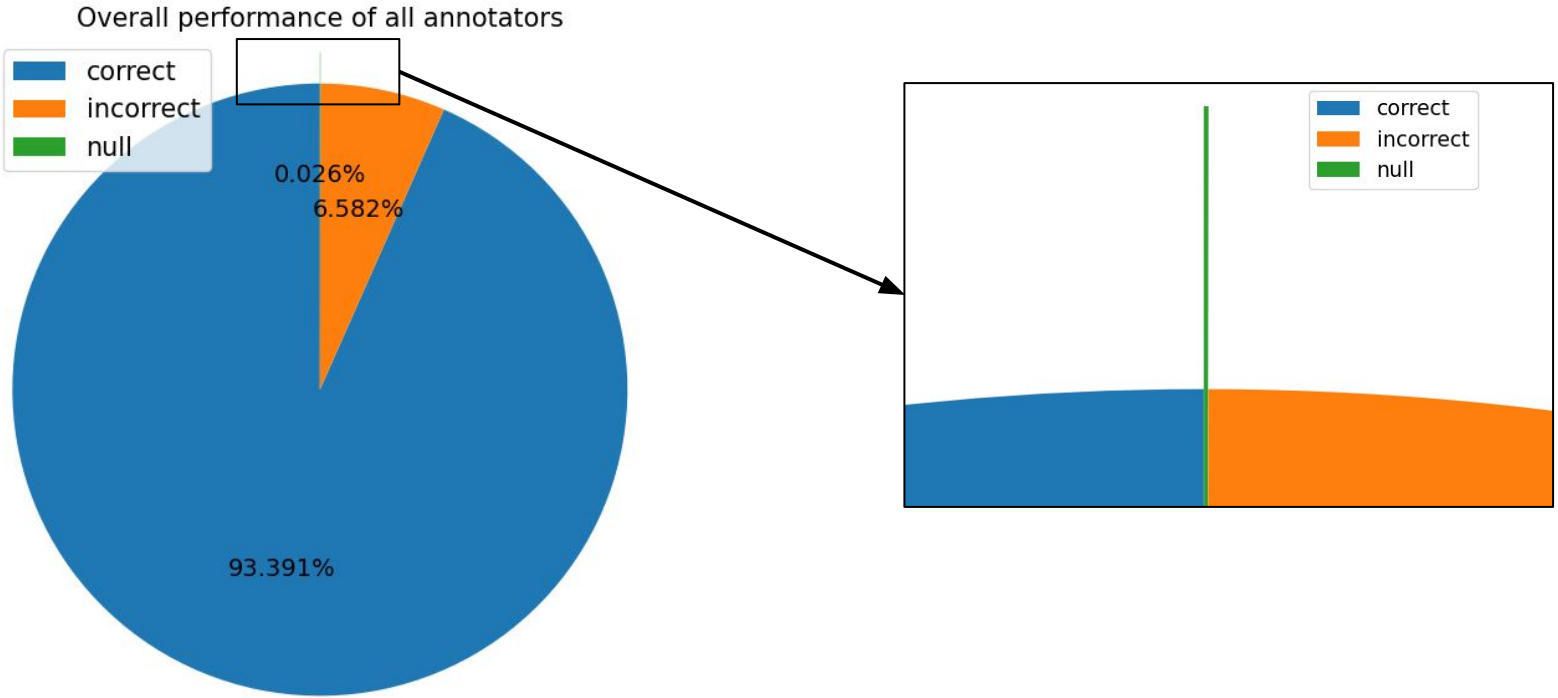


Task 4 - Annotators' results assessment

Note: 'null' is used to indicate samples with 'cant_solve' or 'corrupt_data' are labeled as True



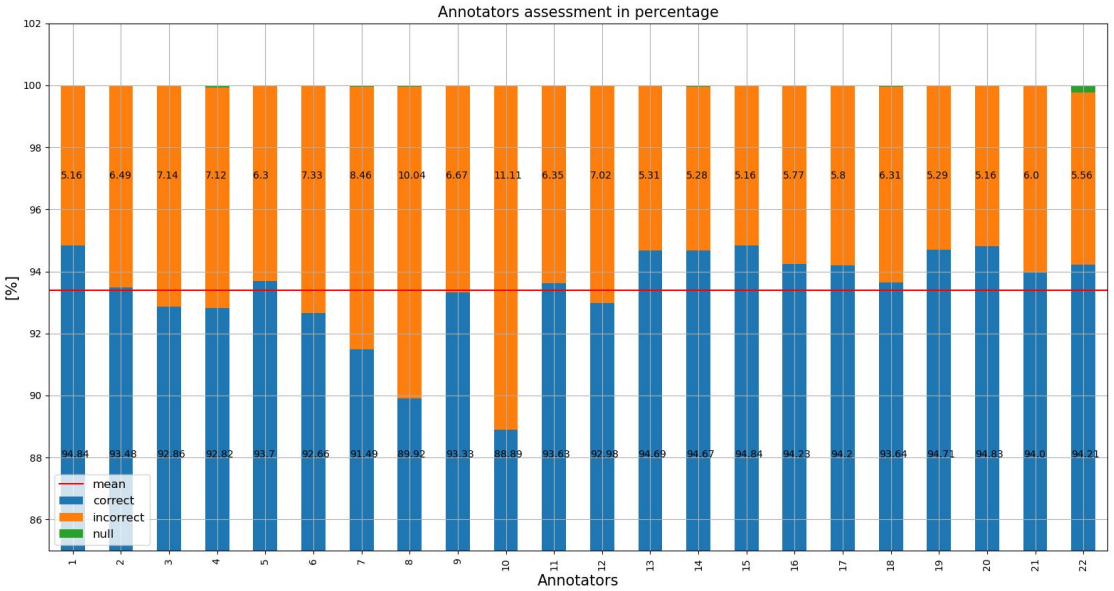
Task 4 - Annotators’ results assessment



Task 4 - Annotators' results assessment

Initially, all annotators are taken in to evaluation. From the percentage graph, the annotators are sorted into 2 groups, good and bad annotators, based on their individual correct percentages comparing with the average correct percentage (93.39%).

min	88.89 %
max	94.84 %
mean	93.39 %



Good annotators:

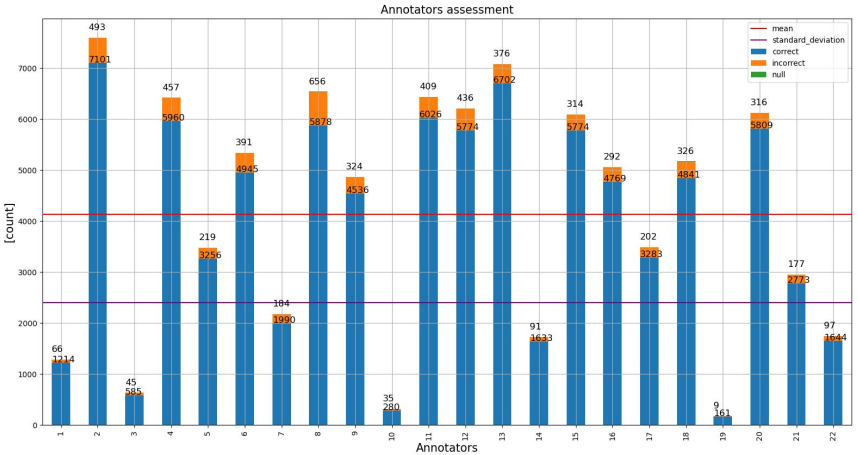
01, 13, 14, 15, 16, 17, 19, 20

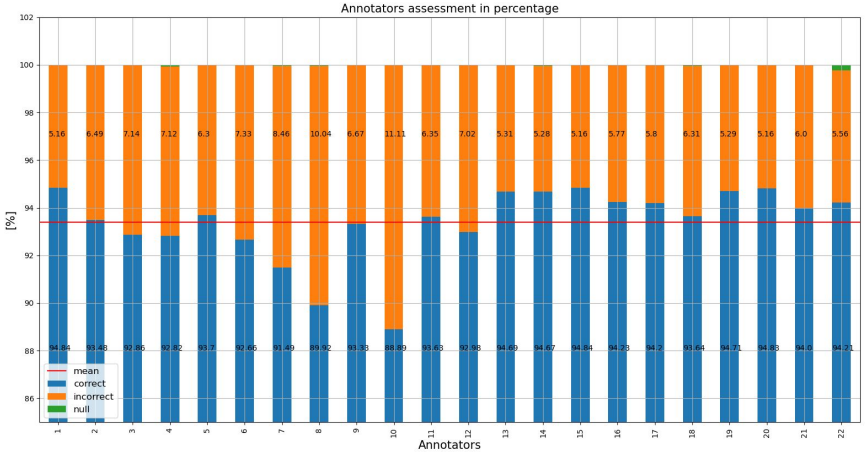
Bad annotators:

07, 08, 10

Task 4 - Annotators' results assessment

The assessment is unfair as some annotators' total samples are far below the average.





E.g.: annotator_01 is assessed as good while his/her total sample is only ~1200 comparing to annotator_02 with more than 7000 samples.

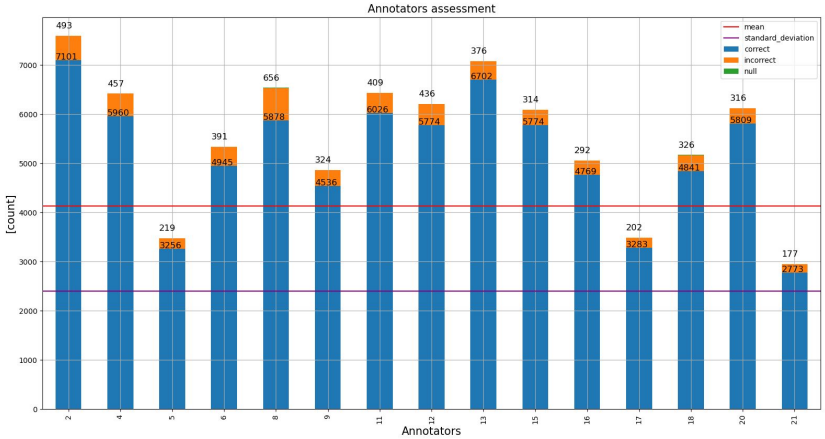
In order to evaluate correctly, only the annotators whose sample sizes are greater than the standard deviation value σ = 2403.23 are selected.

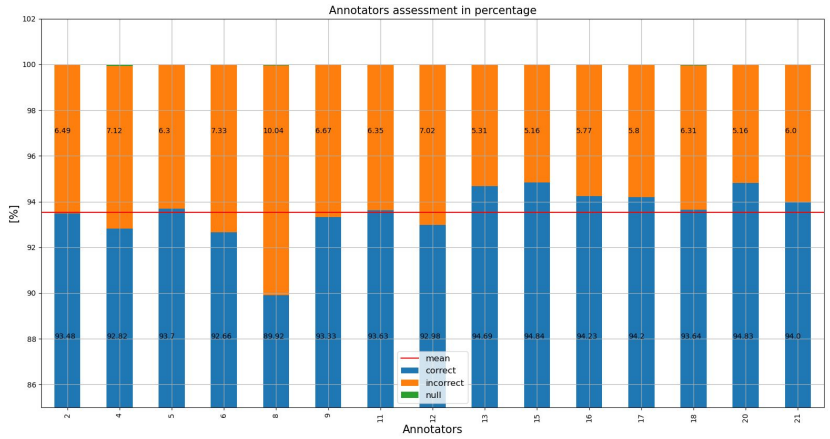
Task 4 - Annotators' results assessment

With this change, the good and bad annotators are reconsidered. In particular, despite the high correct percentages, *annotator_01* and *annotator_14* are removed from the good annotators group as their sample sizes are too small. *Annotator_21* is taken into consideration.

The bad annotator group removed *annotator_07* and *annotator_10* with the same reason and add 3 new annotators, *annotator_04*, *annotator_06* and *annotator_12*.

New good annotators: 13, 15, 16, 17, 20, 21 - New bad annotators: 04, 06, 08, 12





Task 4 - Annotators' results assessment

From inspection, ***annotator_13*** and ***annotator_20*** are considered as **good annotators** as their sample sizes are far above the average level of 4130 samples. Their correct percentages are more than 94% (nearly 95%), higher than the average of 93.52%. Their annotation times are also below the average, which is 1289.9 ms.

Despite the fast average annotation time and high accuracy of 94%, the low sample size make it difficult to make a conclusion of the performance of annotator_17.

vendor_user_id	mean_duration	result_count	correct_p
annotator_13	1155.01	7078	94.69
annotator_15	1365.29	6088	94.84
annotator_16	1269.79	5061	94.23
annotator_17	991.89	3485	94.20
annotator_20	1173.15	6126	94.83
annotator_21	1238.92	2950	94.00

Task 4 - Annotators' results assessment

All four annotators have high sample size with the accuracy below the average of 93.52%. However, beside *annotator_08*, the others three have the accuracy of 92%, which is close to the average line.

Only ***annotator_08*** has the lowest accuracy percentage of 89.9% and high average annotation time of 1434 ms. This causes ***annotator_08*** to be considered as low performance annotator.

vendor_user_id	mean_duration	result_count	correct_p
annotator_04	1113.93	6421	92.82
annotator_06	1496.94	5337	92.66
annotator_08	1434.70	6537	89.92
annotator_12	1306.31	6210	92.98

End.

Thank you for your time.