A Agreement Analysis

We report here on the agreement analysis performed on the product review (PR) dataset (Table A1), crisis tweets (CT) dataset (Table A2), and video human facial expressions (IRep) dataset (Table A3).

Table A1: Inter-rater reliability scores on each category the raters were able to choose in the PR annotation task. The IRR score is computed using the Krippendorff's α metric.

	R1	R2	R3	R4	R5
Fit & Aspect	0.24	0.17	0.20	0.11	0.22
Size & Aspect	0.44	0.37	0.35	0.24	0.41
No Issue with Size & Aspect	0.51	0.43	0.39	0.25	0.43

Table A2: Inter-rater reliability scores on each category the raters were able to choose in the CT annotation task. The IRR score is computed using the Krippendorff's α metric.

	R1	R2	R3	R4	R5
Infrastructure and utilities damage	0.44	0.56	0.53	0.57	0.52
Injured or dead people	0.82	0.93	0.90	0.95	0.88
Caution and advice	0.42	0.56	0.56	0.51	0.52
Donation needs or offers or volunteering	0.78	0.83	0.84	0.87	0.81
Sympathy and emotional support	0.70	0.80	0.81	0.86	0.73
Displaced people and evacuations	0.50	0.72	0.72	0.80	0.68
Missing, trapped, or found people	0.003	0.01	-0.003	-0.001	0.003
Not related or irrelevant	0.55	0.59	0.57	0.67	0.57
Other useful information	0.22	0.26	0.24	0.29	0.29

B Variability Analysis

Table B1 shows an overview of the raters variability in each repetition of the video concept relevance (VCR) tasks. For each video - concept pair, we computed the standard deviation of their score. The majority of the experiments have a mean standard deviation (MSTD) of around 0.3, with the task VCR_O having a higher value, of around 0.36. The standard deviation of deviations (STDD) is similar across tasks and repetitions, with the lowest value observed for the VCR_O task. These high values observed for MSTD and STDD show that video concepts relevance annotation is subjective, with raters often disagreeing. Concepts of type organization seem to generate the most disagreement among annotators.

C Power analysis

In this section, we show additional results on the power analysis performed on the following tasks and datasets: video concept relevance (VCR) in Figure C1, product reviews (PR) in Figure C2, crisis tweets (CT) in Figure C3, and word similarity (WS353) in Figure C4. We recall here that the power analysis is performed using bootstrap experiments on the number of raters. This allows us to observe the impact of the number of raters per unit (in each task and repetition) on the inter-rater reliability score, computed in our case using Krippendorff's α. Thus, in the aforementioned figures, we show the IRR distribution when we bootstrap for each number of raters 100 times, on every task and repetition. We observe that all repetitions of all tasks tend to display similar variability in terms of IRR scores.

Table A3: Inter-rater reliability scores on each category the raters were able to choose in the IRep annotation task. The IRR score is computed using the Krippendorff's α metric.

	R1	R2	R3	R4
Amusement	0.42	0.12	0.66	0.27
Anger	0.45	0.39	0.72	0.44
Awe	0.25	0.19	0.25	0.09
Boredom	0.49	0.38	0.58	0.30
Concentration	0.34	0.40	0.56	0.21
Confusion	0.23	0.22	0.41	0.13
Contemplation	0.13	0.09	0.64	0.10
Contempt	0.16	0.24	0.53	0.11
Contentment	0.23	0.50	0.71	0.10
Desire	0.54	0.69	0.83	0.14
Disappointment	0.19	0.18	0.62	0.16
Disgust	0.52	0.20	0.50	0.14
Distress	0.08	0.10	0.26	0.13
Doubt	0.19	0.18	0.22	-0.02
Ecstasy	0.13	0.24	0.50	0.17
Elation	0.12	0.21	0.66	0.15
Embarrassment	0.19	-0.007	0.28	0.06
Error.other	1.0	1.0	1.0	0.28
Fear	0.43	0.36	0.67	0.31
Interest	0.22	0.08	0.44	0.13
Love	0.66	0.61	0.72	0.32
Neutral	0.41	0.15	0.50	0.13
Pain	0.15	0.44	0.46	0.13
Pride	0.17	0.07	0.39	0.17
Realization	0.09	-0.003	0.33	0.03
Relief	0.17	0.28	0.53	0.05
Sadness	0.51	0.44	0.54	0.22
Shame	-0.006	-0.002	0.0	0.17
Surprise	0.35	0.44	0.67	0.20
Sympathy	0.19	0.07	0.51	-0.02
Triumph	0.42	0.22	0.28	0.08
Unsure	0.61	0.55	-0.002	1.0

D Stability analysis

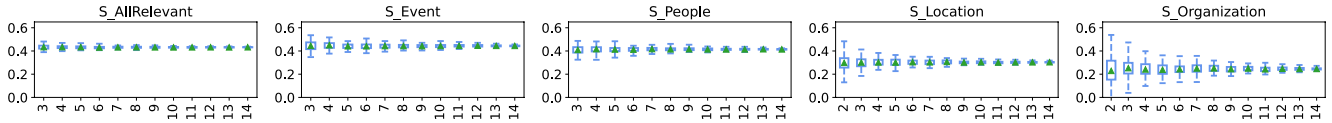
The stability analysis is performed by measuring the correlation between pairwise repetitions of a task. For the tasks at hand, we compute the correlations using the Spearman's rank correlation for the IRep annotation task and datasets (in Table D1) and using the Chi-square test of independence for the PR (in Table D2) and CT (in Table D3) annotation tasks and datasets.

E Replicability similarity analysis

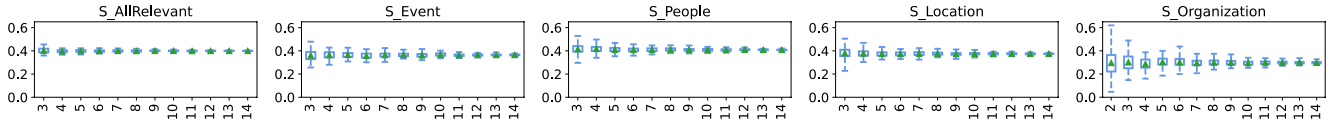
In this section, we report on the replicability similarity analysis, which indicated the degree of agreement between two rater pools, thus between two repetitions of the same task. This analysis is performed using the cross-replication reliability metric on the video concept relevance tasks

Table B1: Overview of rater variability metrics for all *VCR* annotation tasks and their repetitions (R_1, R_2 and R_3).

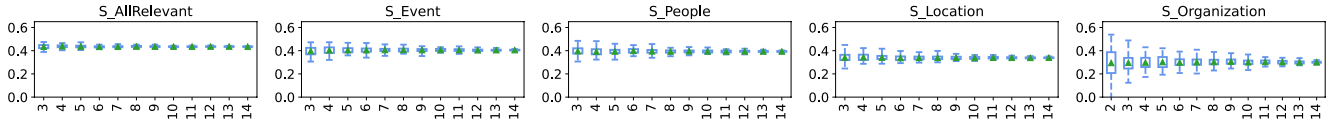
Measure	VCR_ALL			VCR_E			VCR_P			VCR_L			VCR_O		
	R_1	R_2	R_3	R_1	R_2	R_3	R_1	R_2	R_3	R_1	R_2	R_3	R_1	R_2	R_3
Mean Stdev (MSTD)	0.28	0.31	0.30	0.24	0.28	0.28	0.31	0.31	0.32	0.31	0.32	0.31	0.36	0.37	0.35
Stdev of Deviation (STDD)	0.20	0.20	0.19	0.22	0.20	0.21	0.19	0.20	0.19	0.20	0.19	0.20	0.17	0.16	0.15



(a) R_1



(b) R_2



(c) R_3

Figure C1: Bootstrap the number of raters in each repetition (R_1 to R_3) of the VCR annotation tasks. The plot shows the distribution of the IRR scores per each number of raters, where the number of raters ranges from 0 to 15. The green triangle represents the mean value.

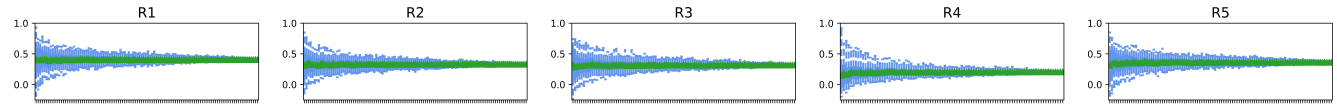


Figure C2: Bootstrap the number of raters in each repetition (R_1 to R_5) of the PR annotation task. The plot shows the distribution of the IRR scores per each number of raters, where the number of raters ranges from 0 to the maximum number of raters employed in the given repetition. The green triangle represents the mean value.

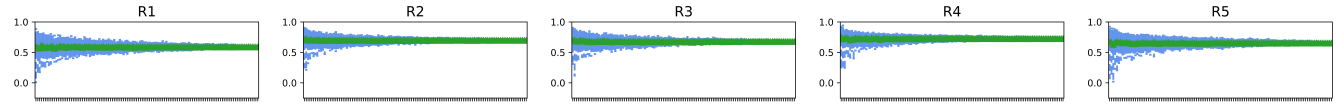


Figure C3: Bootstrap the number of raters in each repetition (R_1 to R_5) of the CT annotation task. The plot shows the distribution of the IRR scores per each number of raters, where the number of raters ranges from 0 to the maximum number of raters employed in the given repetition. The green triangle represents the mean value.

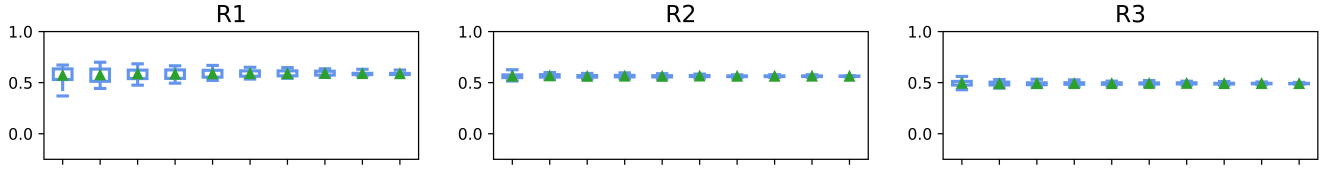


Figure C4: Bootstrap the number of raters in each repetition (R_1 to R_3) of the WS353 annotation task. The plot shows the distribution of the IRR scores per each number of raters, where the number of raters ranges from 0 to the maximum number of raters employed in the given repetition. The green triangle represents the mean value.

Table D1: Spearman's ρ rank correlation of the relevance of each emotion in the IRep annotation task and datasets.

	R_1 & R_2	R_1 & R_3	R_1 & R_4	R_2 & R_3	R_2 & R_4	R_3 & R_4
amusement	ρ=0.45, p=2.13e-55	ρ=0.60, p=5.27e-106	ρ=0.43, p=6.54e-49	ρ=0.38, p=3.94e-37	ρ=0.39, p=6.26e-40	ρ=0.42, p=1.71e-46
anger	ρ=0.49, p=2.40e-65	ρ=0.53, p=6.77e-80	ρ=0.44, p=7.31e-51	ρ=0.52, p=3.16e-74	ρ=0.50, p=5.88e-69	ρ=0.55, p=4.25e-84
awe	ρ=0.20, p=6.65e-11	ρ=0.10, p=0.001	ρ=0.07, p=0.02	ρ=0.15, p=5.72e-07	ρ=0.07, p=0.02	ρ=0.13, p=2.08e-05
boredom	ρ=0.34, p=9.15e-31	ρ=0.50, p=5.71e-70	ρ=0.34, p=1.54e-29	ρ=0.45, p=1.47e-54	ρ=0.31, p=7.30e-25	ρ=0.34, p=2.01e-30
concentration	ρ=0.37, p=1.23e-35	ρ=0.36, p=2.94e-33	ρ=0.30, p=6.93e-23	ρ=0.51, p=7.76e-71	ρ=0.41, p=1.77e-45	ρ=0.39, p=2.07e-40
confusion	ρ=0.30, p=2.55e-23	ρ=0.27, p=1.91e-19	ρ=0.25, p=2.17e-16	ρ=0.37, p=2.00e-36	ρ=0.20, p=1.10e-10	ρ=0.32, p=2.92e-26
contemplation	ρ=0.15, p=1.62e-06	ρ=0.08, p=0.006	ρ=0.14, p=6.79e-06	ρ=0.05, p=0.11	ρ=0.10, p=0.001	ρ=0.15, p=6.38e-07
contempt	ρ=0.14, p=3.75e-06	ρ=0.24, p=7.69e-15	ρ=0.23, p=7.75e-14	ρ=0.28, p=7.41e-21	ρ=0.13, p=2.48e-05	ρ=0.16, p=3.52e-07
contentment	ρ=0.43, p=6.06e-50	ρ=-0.02, p=0.48	ρ=0.13, p=2.91e-05	ρ=-0.09, p=0.002	ρ=0.13, p=3.32e-05	ρ=0.19, p=4.29e-10
desire	ρ=0.51, p=7.18e-72	ρ=0.51, p=1.37e-72	ρ=0.28, p=3.48e-20	ρ=0.63, p=1.76e-120	ρ=0.25, p=3.77e-17	ρ=0.31, p=8.54e-25
disappointment	ρ=0.16, p=7.40e-08	ρ=0.23, p=1.12e-14	ρ=0.14, p=8.11e-06	ρ=0.30, p=5.40e-24	ρ=0.12, p=0.0001	ρ=0.17, p=9.18e-09
disgust	ρ=0.26, p=1.03e-17	ρ=0.28, p=9.20e-21	ρ=0.22, p=3.81e-13	ρ=0.30, p=2.45e-23	ρ=0.23, p=7.11e-14	ρ=0.21, p=1.40e-12
distress	ρ=0.19, p=2.05e-10	ρ=0.22, p=3.27e-13	ρ=0.17, p=3.01e-08	ρ=0.17, p=1.16e-08	ρ=0.19, p=6.30e-10	ρ=0.20, p=1.24e-10
doubt	ρ=0.18, p=2.42e-09	ρ=0.02, p=0.46	ρ=0.02, p=0.51	ρ=0.09, p=0.004	ρ=0.04, p=0.23	ρ=0.01, p=0.68
ecstasy	ρ=0.22, p=2.67e-13	ρ=0.08, p=0.01	ρ=0.13, p=1.75e-05	ρ=-0.009, p=0.78	ρ=0.15, p=1.82e-06	ρ=-0.01, p=0.64
elation	ρ=0.29, p=1.66e-21	ρ=0.36, p=1.27e-33	ρ=0.18, p=2.04e-09	ρ=0.40, p=9.74e-42	ρ=0.36, p=3.78e-34	ρ=0.26, p=1.99e-17
embarrassment	ρ=0.14, p=3.61e-06	ρ=0.21, p=6.35e-12	ρ=0.12, p=5.34e-05	ρ=0.21, p=9.86e-12	ρ=0.12, p=0.0001	ρ=0.08, p=0.01
fear	ρ=0.50, p=8.38e-70	ρ=0.53, p=1.37e-78	ρ=0.38, p=3.87e-38	ρ=0.70, p=4.58e-156	ρ=0.46, p=1.16e-55	ρ=0.48, p=6.67e-64
interest	ρ=0.26, p=3.87e-18	ρ=0.24, p=2.92e-15	ρ=0.17, p=9.02e-09	ρ=0.13, p=1.34e-05	ρ=0.17, p=2.64e-08	ρ=0.15, p=4.44e-07
love	ρ=0.61, p=2.62e-110	ρ=0.57, p=7.00e-93	ρ=0.42, p=1.27e-45	ρ=0.72, p=6.70e-172	ρ=0.57, p=4.42e-94	ρ=0.57, p=4.97e-93
neutral	ρ=0.27, p=2.47e-19	ρ=0.16, p=0.0001	ρ=0.23, p=6.46e-14	ρ=0.09, p=0.002	ρ=0.22, p=1.00e-12	ρ=0.09, p=0.004
pain	ρ=0.43, p=6.23e-50	ρ=0.27, p=7.23e-19	ρ=0.29, p=6.68e-22	ρ=0.47, p=6.94e-61	ρ=0.29, p=5.90e-22	ρ=0.24, p=2.34e-15
pride	ρ=0.14, p=2.09e-06	ρ=0.29, p=7.77e-22	ρ=0.17, p=1.47e-08	ρ=0.45, p=1.73e-54	ρ=0.16, p=1.56e-07	ρ=0.16, p=1.05e-07
realization	ρ=-0.02, p=0.46	ρ=0.03, p=0.26	ρ=0.00004, p=0.99	ρ=-0.006, p=0.86	ρ=-0.03, p=0.37	ρ=0.07, p=0.03
relief	ρ=0.09, p=0.003	ρ=0.11, p=0.0003	ρ=0.12, p=0.0001	ρ=0.25, p=1.28e-16	ρ=0.10, p=0.001	ρ=0.17, p=3.96e-08
sadness	ρ=0.51, p=1.42e-72	ρ=0.57, p=2.14e-93	ρ=0.40, p=7.26e-43	ρ=0.52, p=3.07e-75	ρ=0.43, p=3.25e-50	ρ=0.47, p=3.01e-58
shame	ρ=0.33, p=3.11e-29	ρ=0.28, p=4.83e-20	ρ=0.16, p=1.004e-07	ρ=-0.002, p=0.94	ρ=-0.01, p=0.64	ρ=0.16, p=1.91e-07
surprise	ρ=0.49, p=6.36e-66	ρ=0.55, p=7.92e-87	ρ=0.35, p=1.91e-31	ρ=0.55, p=2.84e-86	ρ=0.30, p=3.35e-24	ρ=0.38, p=6.81e-39
sympathy	ρ=0.08, p=0.01	ρ=0.14, p=2.72e-06	ρ=0.03, p=0.35	ρ=0.26, p=3.31e-18	ρ=0.12, p=5.03e-05	ρ=0.15, p=1.37e-06
triumph	ρ=0.44, p=6.66e-53	ρ=0.53, p=2.21e-74	ρ=0.26, p=2.05e-18	ρ=0.65, p=5.37e-127	ρ=0.22, p=4.55e-13	ρ=0.30, p=1.01e-23
unsure	ρ=0.66, p=5.72e-135	ρ=0.16, p=2.12e-07		ρ=0.11, p=0.0002		

Table D2: Stability analysis on the PR annotation task and dataset. The table indicates the correlation of the aggregated raters' annotations (*i.e.*, using majority vote) between pairwise repetitions of the PR task.

	R2	R3	R4	R5
R1	χ = 24.87, p = 5.35e-05	χ = 35.0, p = 4.65e-07	χ = 34.55, p = 5.76e-07	χ = 29.80, p = 5.38e-06
R2		χ = 29.88, p = 5.19e-06	χ = 29.70, p = 5.64e-06	χ = 24.87, p = 5.35e-05
R3			χ = 30.91, p = 3.19e-06	χ = 35.0, p = 4.65e-07
R4				χ = 25.68, p = 3.67e-05

Table D3: Stability analysis on the CT annotation task and dataset. The table indicates the correlation of the aggregated raters' annotations (*i.e.*, using majority vote) between pairwise repetitions of the CT task.

	R2	R3	R4	R5
R1	χ = 23.63, p = 9.48e-05	χ = 19.56, p = 0.0006	χ = 23.63, p = 9.48e-05	χ = 24.04, p = 7.81e-05
R2		χ = 35.58, p = 3.52e-07	χ = 40.70, p = 4.33e-08	χ = 27.94, p = 1.28e-05
R3			χ = 35.58, p = 3.52e-07	χ = 32.22, p = 1.72e-06
R4				χ = 27.94, p = 1.28e-05

(VCR_ALL in Figure E1, VCR_E in Figure E2, VCR_P in Figure E3, VCR_L in Figure E4, VCR_O in Figure E5), product reviews task (in Figure E6), crisis tweets task (in Figure E7), and video human facial expressions task (in Figure E8).

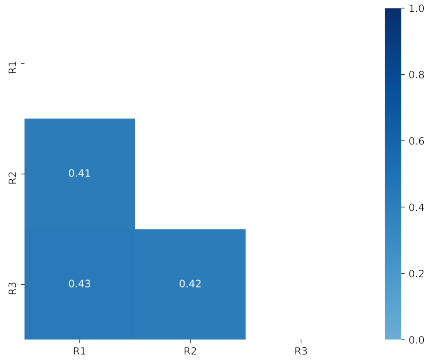


Figure E1: Cross-rater reliability analysis on the VCR_ALL annotation task and datasets. Each cell in the heatmap represents the xRR score between pairwise sets of repetitions.

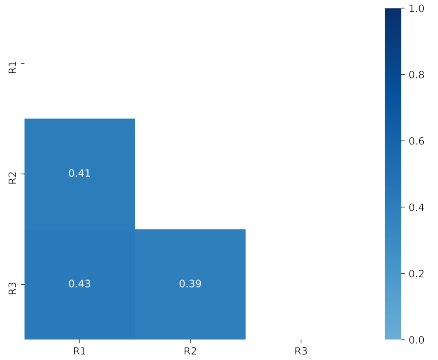


Figure E2: Cross-rater reliability analysis on the VCR_E annotation task and datasets. Each cell in the heatmap represents the xRR score between pairwise sets of repetitions.

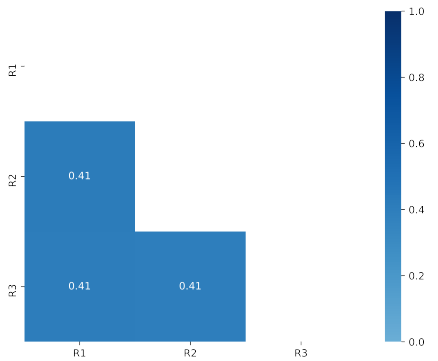


Figure E3: Cross-rater reliability analysis on the VCR_P annotation task and datasets. Each cell in the heatmap represents the xRR score between pairwise sets of repetitions.

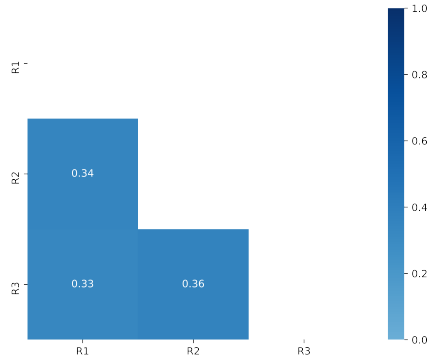


Figure E4: Cross-rater reliability analysis on the VCR_L annotation task and datasets. Each cell in the heatmap represents the xRR score between pairwise sets of repetitions.

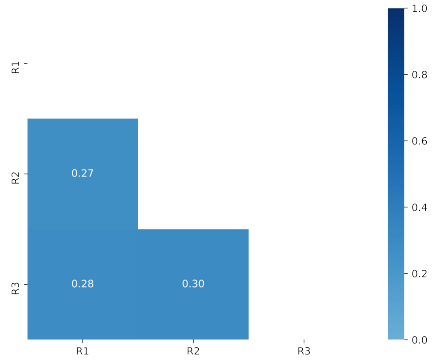


Figure E5: Cross-rater reliability analysis on the VCR_O annotation task and datasets. Each cell in the heatmap represents the xRR score between pairwise sets of repetitions.

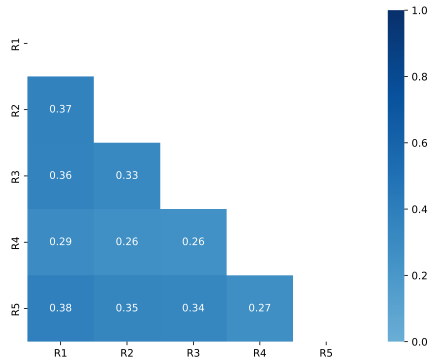


Figure E6: Cross-rater reliability analysis on the PR annotation task and datasets. Each cell in the heatmap represents the xRR score between pairwise sets of repetitions.

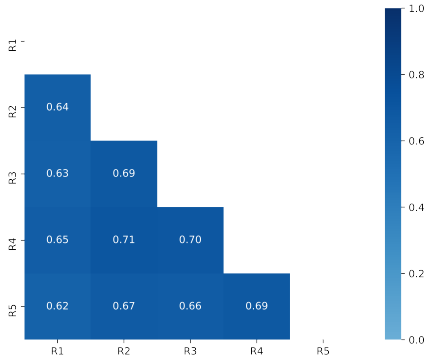


Figure E7: Cross-rater reliability analysis on the CT annotation task and datasets. Each cell in the heatmap represents the xRR score between pairwise sets of repetitions.

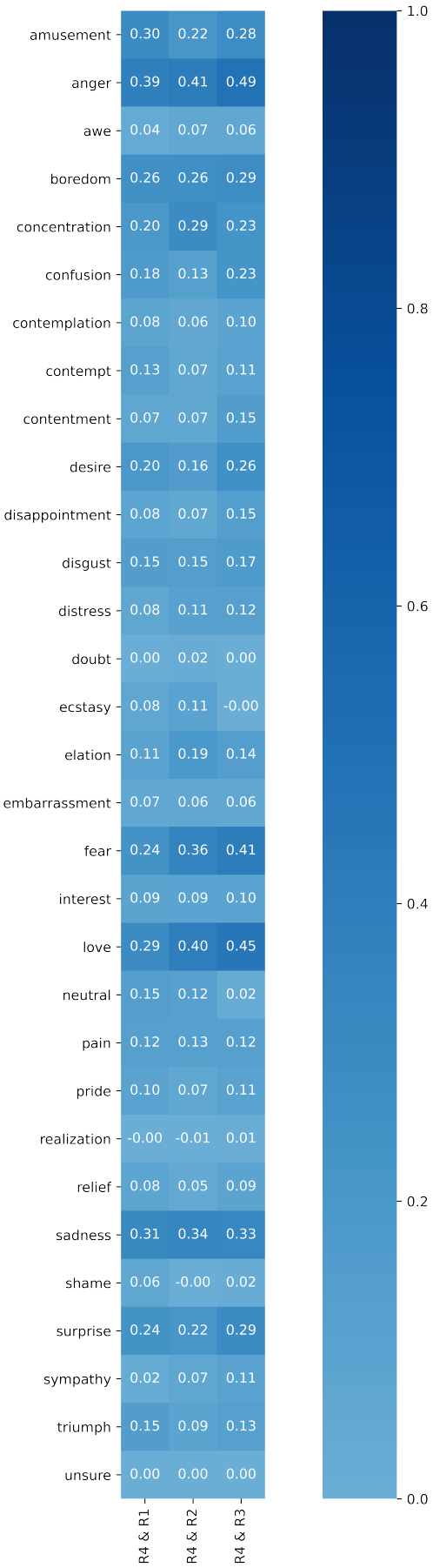


Figure E8: Cross-rater reliability analysis on $R4$ of the IRep annotation task and dataset. Each cell in the heatmap represents the xRR score between pairwise sets of repetitions.