FACULTY OF ARTS AND PHILOSOPHY

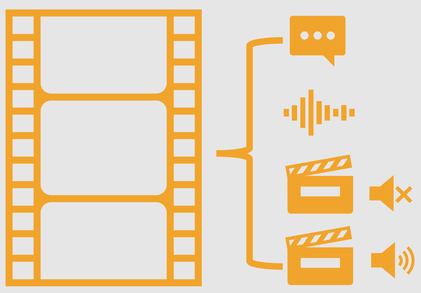
LT3, LANGUAGE AND TRANSLATION TECHNOLOGY TEAM

Quanqi Du, Thomas Demeester & Veronique Hoste

BEYOND KAPPA AND ALPHA: A SIMPLE BUT EFFECTIVE METHOD FOR ANNOTATION AGREEMENT MEASUREMENT AND PREDICTION

Background: Building a multimodal emotion dataset, with annotations on both unimodal and multimodal setups.
Problem: We are confident on our annotation design and implementation, while the Kappa and Alpha scores indicate a modest agreement among annotators. The interpretation of Kappa/Alpha scores might not reflect the real agreement in our case.
Question: As the Kappa/Alpha are not easy to interpret, maybe we can try some simple ad easy-reading solution, e.g., absolute annotation difference (AAD)?

Experiment 1: inter- and intro-annotator (dis)agreement measurement with AAD



Valence-5, Arousal-3, Joy

Valence-4, Arousal-4, Surprise

Valence-5, Arousal-1, Confusion

Valence-3, Arousal-2, Contentment

Dataset: 94 video clips centred around the topic *Review* from YouTube, reviewed and selected to be rich in authentic emotional expressions.

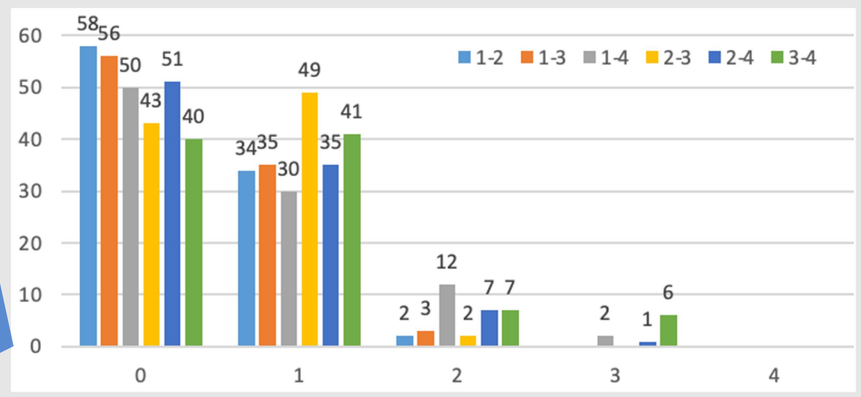
Up to fair agreement indicated by Kappa, and not realiable suggested by Alpha

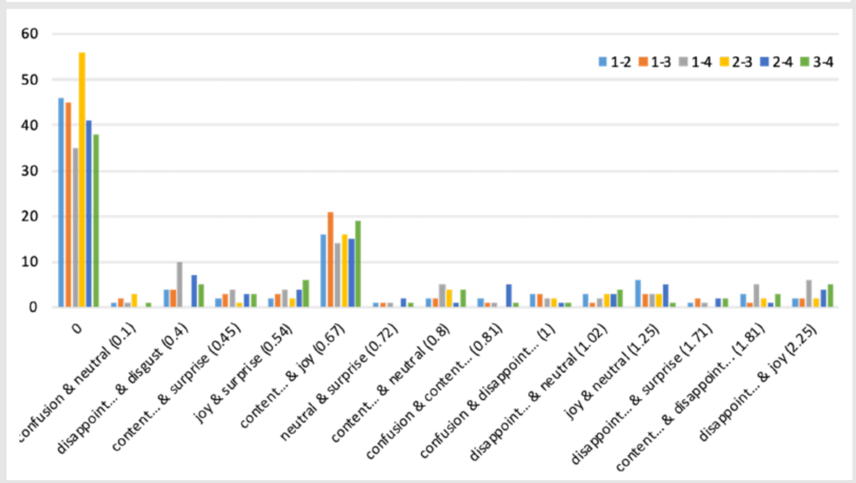
		text	audio	vdieo	all
Kappa	Valence	.33	.23	.21	.27
	Emotion	.32	.27	.19	.29
Alpha	Valence - nominal	.33	.23	.22	.27
	Valence - ordinal	.64	.48	.46	.52
	Valence - interval	.64	.48	.46	.52
	Valence - ratio	.59	.42	.38	.46

Higher agreement revealed by AAD:

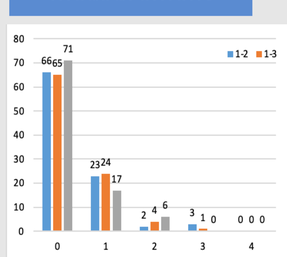
1. Most of the valence differences between the six pairs of annotators are indeed limited to 0 or 1.

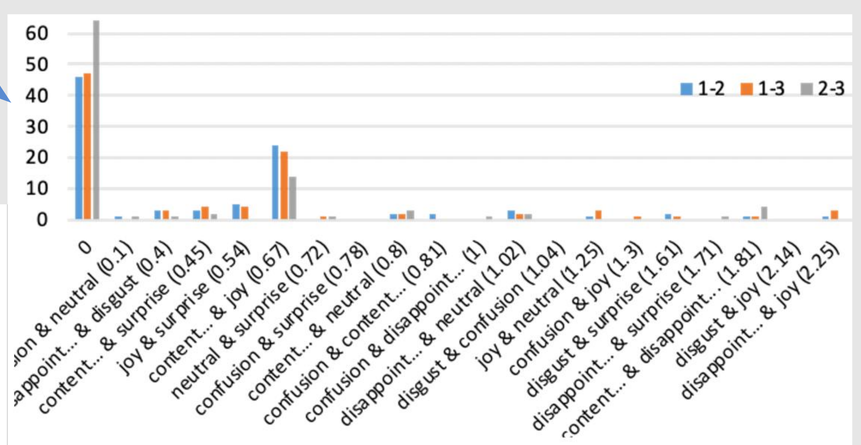
2. Most instances are annotated with identical emotions.





Similar for intra-annotator agreement measurement





Experiment 2: (dis)agreement prediction with AAD

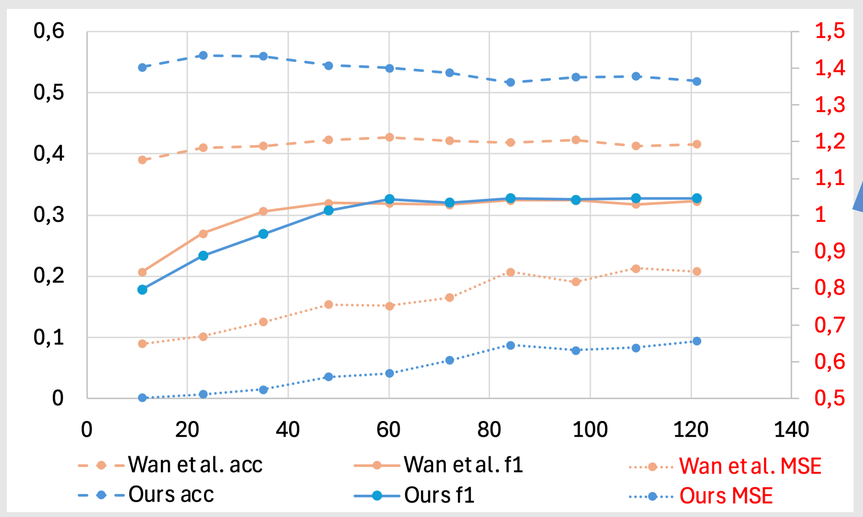
Dataset: DynaSent, more than 100,000 textual instances with 5 annotators.
Disagreement rating:

$$D = \frac{\frac{n_{minority}}{N_{total}}}{\frac{3}{N_{total}}} = \frac{n_{minority}}{3} \quad > \quad < \quad D^i = \sqrt{\frac{1}{\binom{n}{2}} \sum_{(x,y)\in\mathcal{N}} (x_i - y_i)^2}$$

Annotation distribution	Binary label	Wan's	Ours
🙂🙂🙂😐😐	disagree	0.67	0.77
🙂🙂🙂😐☹️	disagree	0.67	1.26
🙂🙂🙂😕☹️	disagree	0.67	N/A

☹️-negative 😐-neutral 🙂-positve 😕-mixed

Experiment: fine-tuning a RoBERTa-base model with a fixed learning rate 1e-5, and batch size 8 for 10 epochs, using NVIDIA Tesla V100-SXM2-16GB GPUs.



During the training process, our AAD-based rating strategy seems to perform better in terms of accuracy and MSE,

	Acc ↑	f1 ↑	MSE ↓
Wan's	41.71	32.1	0.097
Ours	51.64	32.3	0.072

In the results, our AAD-based rating still outperforms the other.

	Instances	acc	f1	precision	recall
Reg_2	94	60.64	58.57	64.26	61.14
Reg_4	94	45.74	30.97	34.07	32.89
label-1	31	N/A	50.57	39.29	70.97
label-2	13	N/A	24.00	25.00	23.08
label-3	2	N/A	0	0	0

Predictions on the 94 instances.

Contact
Quanqi.du@@ugent.be
https://lt3.ugent.be/people/du-quanqi/

- Reference:
- Du, Quanqi, Sofie Labat, Thomas Demeester, and Veronique Hoste. "UniC: a Dataset for Emotion Analysis of Videos with Multimodal and Unimodal Labels." (2024). Under review
- Du, Quanqi, Sofie Labat, Thomas Demeester, and Veronique Hoste. "Unimodalities Count as Perspectives in Multimodal Emotion Annotation." In Proceedings of the 2nd Workshop on Perspectivist Approaches to NLP @ ECAI 2023. (2023).
- Du, Quanqi, Sofie Labat, Thomas Demeester, and Veronique Hoste. " Beyond Kappa and Alpha: A Simple but Effective Method For Annotation Agreement Measurement and Prediction." Submitted to the 31st International Conference on Computational Linguistics (COLING 2025). Under review. (2024)

GHENT UNIVERSITY