Annotation Practices in Affective Computing Research

What are these automated systems actually trained on?

1. Research Question

What are current data collection and reporting practices of human annotations in societally impactful applications of machine learning research in the area of affective computing?

2. Motivation

- A domain with potentially a huge societal impact
- The subjectivity of the domain. Individuals see emotions differently from each other. [1]

3. Methodology

- A literature review was performed on the 100 most cited papers on affective computing research from the last 5 years.
- To find these papers, a search string with relevant keywords was entered into Scopus.
- Keywords included for example 'emotion recognition', 'sentiment analysis' or 'affect classification'.
- For the papers included in the study, questions related to annotation practices were answered. A similar approach was taken by Geiger et al [2]

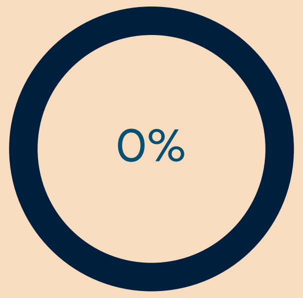
4. Results

215 datasets in 100 papers

Most popular 4 datasets appear about 20% in the total count of 215.

	Count	Proportion
Text	89	41.4%
Images	47	21.86%
EEG and other physiological measures	24	11.16%
Audio and video	20	9.30%
Audio, video and text	16	7.44%
Social media content	10	4.65%
Audio	7	3.25%
Other	2	0.93%

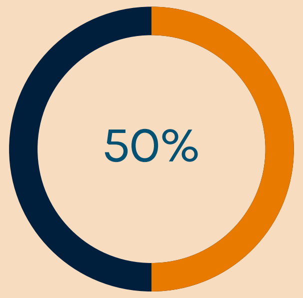
Table 1: Type of data



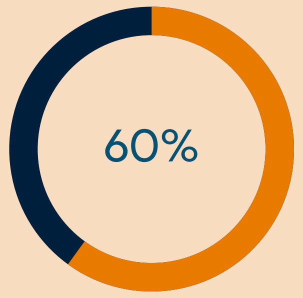
Estimate amount of annotators

	Used annot.	Original annot.
Positive / negative	48	29
Positive / negative / neutral	36	31
Positive / negative, 4-7 levels	13	28
Discrete emotions, less than 5	15	6
Discrete emotions, 5 - 10	78	81
Discrete emotions, 10 - 15	4	8
Valence / Arousal, high / low	7	3
Valence / Arousal, range	12	24
FACS	0	2
No information	3	9

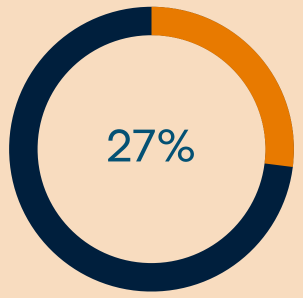
Table 2: Type of annotations



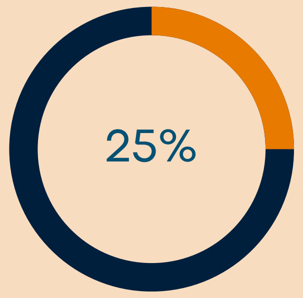
Multiple annotator overlap



Reporting of inter-annotator agreement



Expert annotation



Training provided

" Annotation practices in affective computing are of varying quality "

In general, practices could and should be improved.

5. Insights

- Overall, examples of very good and very bad annotation practices were found.
- The quality of the annotaton practices vary a lot from dataset to dataset.
- Multi annotator overlap is especially important for emotion datasets, because the ground truth can be hard to determine. Individual humans also perceive emotions differently from each other [1]. The encountered multiple annotator overlap is considered to be quite low.
- Often, no information was given and information on the dataset could also not be found elsewhere. Further emphasizing the lack of attention given to quality data collection and annotation.
- The annotations given to datasets differ a lot. The field of affective computing could use some unity in this, as this gives more clarity as to what a well-performing model should be able to predict.

6. Limitations

- Datasets are hard to compare because the type of data and the collection method of the data differs so much from each other.
- The datasets are only annotated by one person

7. References

[1] R.H. Swain, A.J. O'Hare, and K. Brandley. Individual differences in social intelligence and perception of emotion expression of masked and unmasked faces. Cogn Research, 7(54), 2022

[2] R. S. Geiger, K. Yu, Y. L. Yang, M. Dai, J. Qiu, R. Tang, and J. Huang. Garbage in, garbage out? Do machine learning application papers in social computing report where human-labeled training data comes from? Fat* '20: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pages 325–336, 2020. Bq8fj Times Cited:28 Cited References Count:68

Suzanne Backer https://www.linkedin.com/in/suzanne-backer

Supervisor: Andrew Demetriou

Responsible professor: Cynthia Liem

TU Delft Delft University of Technology