A SURVEY OF INTERRATER AGREEMENT IN DATASETS FOR AUDIO-VISUAL AUTOMATIC AFFECT PREDICTION: A SYSTEMATIC LITERATURE REVIEW

Author: Alexandru Preda **Responsible professor:** Bernd Dudzik

1. Background

- With the rise in the number of human-computer interactions, the need for systems that can accurately infer and respond to users' affective state becomes increasingly important.
- **Affect** represents a wide range of mental responses. (e.g. emotions, moods, attitudes, preferences, feelings etc.) [1]
- **Automatic Affect Prediction (AAP)** represent the process of using machine learning to infer the affective state of an individual [2].
- Effective AAP models **are highly dependent** on **labeled datasets** [7].
- **Emotions** are intricate and multifaced, open to **multiple interpretations** Fig. 1) [1]. Because of that, the **datasets** are usually **labeled manually**, which can introduce uncertainty in the data.
- **Interrater agreement (IRA)** represents the extent to which raters agree on the same label for an entry.

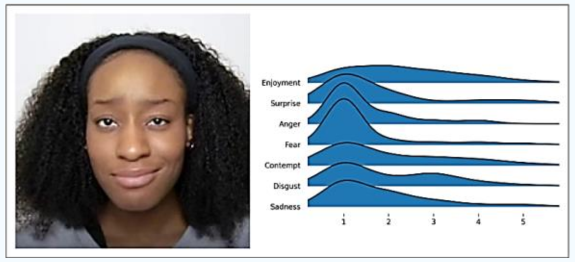


Fig. 1:Multiple interpretations of the evoked emotion [6]

2. Research Question

To what extent is interrater agreement used in datasets for audio visual AAP and how is it implemented?

- Targeted affective states
- Affect Representation Schemes (ARS)
- Number of Raters Used
- The measures used to compute the level of interrater agreement
- Strategies to facilitate IRA
- Relationship between the ARS and the level of IRA

3. Methodology

- To answer the research question, a Systematic Literature Review [4] was conducted.
- Steps of a Systematic Literature Review:

Define Protocol → Screen the results → Data Extraction → Data Synthesis → Report findings

- The 2020 PRISMA guidelines were followed to transparently report the procedure and the results of this review [5].

3.1. Search Strategy

Literature Databases: **Scopus, IEEE Xplore, Web of Science and ACM Digital Library**

Query development:

- To develop the query, the main topic was split into 4 concepts (Fig. 2)
- For each concept, a set of descriptive words was created and included in the query.
- Set of 7 predetermined studies were used to assess and optimize the performance of the query

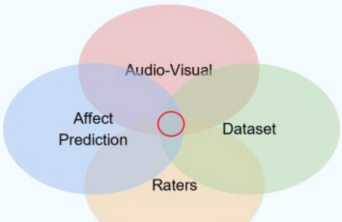


Fig. 2: Key concepts of the topic

3.2.Eligibility Criteria

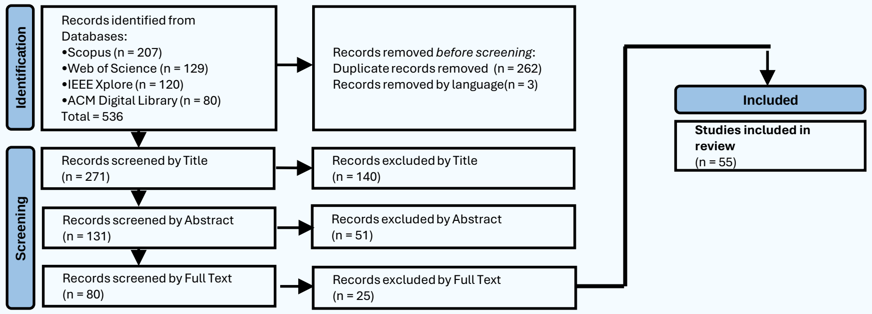
To consistently filter the results of the query, the following eligibility criteria were developed:

Inclusion criteria:
- Introduces an audio-visual dataset
- Data was labeled by humans

Exclusion Criteria
- Paper is not in English
- Dataset is labeled through self-reports
- The affect generator is not human
- Released after 20.05.2024

3.3 Reviewed papers

After running the query, then scanning the results by title, abstract and full text, 55 studies were included in the review. A detailed overview of this process can be observed bellow:

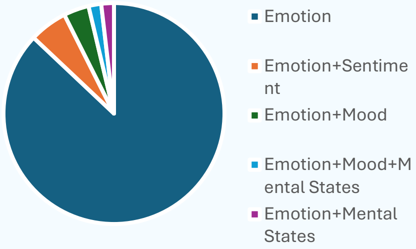


3. Results

3.1 Targeted affective states

Out of the 55 papers reviewed:

- **All** of them focused on labelling **Emotion**
- **47** only focused on emotion
- **3** labeled Emotions and Sentiments
- **2** labeled Emotion and Mood
- **1** labeled Emotion, Mood and Metal States
- **1** labeled Emotion and Mental States



3.2 Affect Representation Schemes

43 distinct ARS were identified from 55 studies.

Categorical ARS

- 24 out of 43
- 20 of them are derivates of Ekman's basic emotions [8], one of which is the actual one
- 2 used Plutchik's Wheel of Emotions [9]
- 1 used the "emotion zones for regulation" framework[10]
- 1 used what they defined as the most used labels in other studies

Dimensional ARS

- 8 out of 43
- Most popular was Valence-Arousal with 11 papers using it
- 5 used VA with other dimensions such as dominance, liking, impact, engagement and aggression
- 1 used only Valence
- 1 used SAM's Pleasure-Arousal-Dominance [11]

Mixed ARS

- 11 out of 43

3.3 Interrater Agreement

- Most studies used between 3 and 5 annotators
- 34 out of the 55 papers measured IRA
- The preferred methods to computer IRA are Fleiss' Kappa and Krippendorff's Alpha, Fig. 3 highlighting the overall distribution

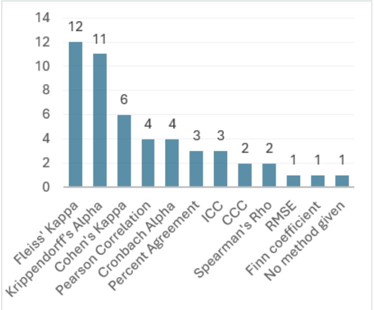


Fig. 3 Popularity of IRA measures

3.4 Interrater Agreement over time

- Early methods: Fleiss's Kappa and Cohen's Kappa
- Krippendorff's Alpha emerged around 2014 and become one of the favorite methods
- Fleiss's Kappa maintained constant popularity
- Past 2 years: increased experimentation with other IRA methods

3.5 Relationships between ARS of a dataset and their IRA

- Due to the extremely high number of ARS, no individual relationship could have been determined.
- Comparing either Valence Arousal derivates or Ekman's basic emotion derivatives did not reveal any relationships of a specific group of ARS with a level of IRA

4. Conclusions & Discussions

- The majority of papers compute Interrater Agreement
- Most popular methods of computing IRA are Fleiss' Kappa, Krippendorff's Alpha, and Cohen's Kappa
- IRA appears to be independent from the ARS. However, due to the high number of representation schemes no definitive argument can be made
- Interestingly, despite the number of papers that calculate agreement, many of which try to facilitate IRA, no paper does a second run of annotation with the aim of improving the score.
- The absence of a second labeling run raises questions about the purpose behind measuring IRA. If the ultimate goal is to ensure high-quality and reliable annotations, the natural progression would be to use IRA scores as feedback to refine the labeling process.
- Additionally, the study uncovered that many affect representation schemes (ARS) deviated from well-established models without providing a clear motivation. These deviations make the process of correlating ARS with IRA very difficult, as they introduce uncertainty that is not related to the emotional content being measured but rather to the subjective choices of the researchers.

5. Future work

- This study laid the groundwork for a new study on how would interrater agreement affect the performance of audio-visual automatic affect prediction system.
- The high number of different ARS that were used in the researched studies without a proper motivation highlight the need of developing a standardized ARS that could possibly enhance the quality of datasets and align the focus of the community to accelerate the development of affective databases

Resources

[1] K. R. Scherer, "What are emotions? And how can they be measured?," Social Science Information, vol. 44, no. 4, pp. 695–729, Dec. 2005
[2] R. Picard, Affective Computing. MIT Press. MIT Press, 2000. Accessed: May 08, 2024.
[3] I. Lefter, G. J. Burghouts, and L. J. M. Rothkrantz, "An audio-visual dataset of human–human interactions in stressful situations," J Multimodal User Interfaces, vol. 8, no. 1, pp. 29–41, Mar. 2014, doi: 10.1007/s12193-014-0150-7.
[4] A. Boland, G. Cherry, and Dickson Rumona, Doing a systematic review: a student's guide. SAGE, 2014.
[5] M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews".
[6] F. Cabitza, A. Campagner, and M. Mattioli, "The unbearable (technical) unreliability of automated facial emotion recognition," Big Data and Society, vol. 9, no. 2, Jul. 2022.
[7] V. N. Gudivada, A. Apon, and J. Ding, "Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations," 2017.
[8] P. Ekman, "An argument for basic emotions," Cognition and Emotion, vol. 6, pp. 169–200, May 1992.
[9] R. Plutchik, "The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain
their complexity and provide tools for clinical practice," American Scientist, vol. 89, no. 4, pp. 344–350, 2001.
[10] L. Kuypers, "The zones of regulation: A framework to foster self-regulation," Sensory Integration Special
Interest Section Quarterly, vol. 36, no. 4, pp. 1–4, 2013.
[11] A. Mehrabian and J. A. Russell, An approach to environmental psychology. the MIT Press, 1974.