A Survey of Interrater Agreement in Datasets for Audio Visual Automatic Affect Prediction:

A Systematic Literature Review

Alexandru Preda[1]
Supervisor(s): Bernd Dudzik[1]
[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Alexandru Preda
Final project course: CSE3000 Research Project
Thesis committee: Bernd Dudzik, Catharine Oertel

Abstract

With the rise in the number of human-computer interactions, the need for systems that can accurately infer and respond to users' emotions becomes increasingly important. One can achieve this by examining audio-visual signals, aiming to identify the underlying emotions from an individual's gestures, auditory cues, and surroundings. Such automatic affect prediction systems depend heavily on labeled datasets. However, the subjective nature of emotion interpretation often introduces uncertainty, making it challenging to create reliable and high-quality datasets. To combat this issue, researchers have tried employing multiple raters to judge the affective state of a person, while employing interrater agreement measures to monitor uncertainty. To this moment, it is still unclear to what extend does reaching a good level of interrater agreement impact the performance of audio-visual Automatic Affect Prediction models. As a first step towards understanding the potential influences on performance, this paper conducts a systematic literature review to investigate the objective annotation procedure in audio-visual databases. The survey extracted relevant literature from 4 scientific databases: Scopus, IEEE Xplore, Web of Science and ACM Digital Library. The results are aggregated from 55 papers and presented by following the PRISMA guidelines. They indicate that most databases use multiple annotators, a little more than half measure interrater agreement, and most train the raters to increase the uniformity of the labels.

1 Introduction

Affect represents a wide range of emotional experiences, from intense, short-lived emotions to diffuse, long-lasting moods and stable affective dispositions [1]. Automatic Affect Recognition represents the process of using machine learning to infer the emotional state of an individual by analysing their expressions, tone of voice and body language extracted from audio-video content [2].

With the increasing prevalence of technology and social media in our daily lives, along with the implicit increase in the availability of video content [3, 4], audio-video content has evolved beyond a form of entertainment or information retrieval to also being a tool for communication and marketing. Therefore, it is essential to tag videos by their affective content to benefit both users, because they can more easily find the content that matches their emotional needs(e.g. videos to relieve boredom) within large collections, but also the businesses that create videos to express their intent more accurately [4]. However, finding a general solution to predict the affective state of an individual while watching a video is extremely hard [5].

In order to tag audio-visual content, Automatic Affect Prediction (AAP) models could be employed [2]. According to [6], most AAP models were build using supervised learning methods. A fundamental part in developing these models are datasets. Without a good dataset the model cannot generalize to other contexts accurately and, thus, impact the performance [7, 8]. Unfortunately, this is no easy task as emotions are intricate and multifaceted, often open to diverse interpretations [1, 5]. A good example of this phenomena is highlighted in figure 6 of [9] which shows that the same facial expression can be interpreted in many plausible ways. This can introduce uncertainty in the labeling of a database as there is no way for an individual to "correctly" label the entries [9]. Therefore, the accuracy of the model is influenced by the way in which the datasets are created. To address this challenge, researchers are exploring the use of multiple raters and measuring their agreement during dataset labeling [10, 11]. By gathering ratings from different individuals and assessing the level of consensus among them, researchers, such as [10, 11], seek to quantify and mitigate the uncertainty associated with emotion labeling. The level of consensus, also known as *interrater agreement* (IRA) or *interrater reliability* (IRR), reflects the extent to which raters agree on the same label for an entry [12]. More about IRA is included in Results.

In a previous study [9] it was concluded that a low interrater agreement score in facial databases translate to poor accuracy of facial Automatic Affect Prediction system. However, the review was not conducted in a systematic manner and it did not focus on the larger context provided by audio-visual content. Therefore, there is a clear knowledge gap in the sense that it is still unknown how fostering interrater agreement in audio-visual databases impacts the performance of Automatic Affect Prediction systems.

To answer this question, a two-step system is required. The first step consists of gathering all the available audio-visual datasets for Automatic Affect Prediction to understand the targeted affective states, the array of representation schemes and details of the annotation process focusing on the extent to which multiple raters are employed and if interrater agreement is facilitated and how. The second step consists of tracing all the papers that built an Automatic Affect Prediction system based on a dataset introduced in any paper found in the first part and studying their performance. This process is outlined in Table 1 below, with the first 5 sub-questions referring to the first step and the last sub-question referring to the second step. As this research is limited to only 9 weeks, the work is limited to the first step, namely the first 5 sub-questions. Therefore, this paper will answer **To what extent is interrater agreement used in datasets for audio visual Automatic Affect Prediction and how is it implemented?** through a systematic literature review. The results will be reported following the 2020 PRISMA guidelines [13] and will lay the groundwork to answer the last sub-question.

Table 1: List of sub-questions

Num	**Research Sub-Question**
SQ1	What types of affective states have been targeted by datasets (e.g., only emotions or mood?)
SQ2	What different affect representation schemes have been used in these datasets (and what is the specific motivation for using specific schemes)?
SQ3.A	Do datasets collect multiple ratings for a record (and how many)?
SQ3.B	If so, do datasets measure interrater agreement?
SQ3.C	What measures do they use for this (and what is the level of agreement)?
SQ3.D	Do dataset creators use any strategies to facilitate/facilitate interrater agreement (and what are these)?
SQ4	Is there a change in how datasets measure interrater agreement over time?
SQ5	Is there a relationship between the affect representation scheme used by datasets and their interrater agreement?
SQ6	Is there a relationship between the interrater agreement in datasets and the empirical performance of affect prediction systems that use them for training and evaluation?

The structure of this paper is as follows: Section 2 describes the methodology and steps took through out the systematic literature review. Section 3 presents our results. Section 4 outlines the ethical considerations related to the study and results. Section 5 discusses and evaluates the results. Lastly, section 6 presents the conclusions of this research and introduces recommendation for future work.

2 Methodology

This research performs a Systematic Literature Review to answer the research question outlined in the Introduction. Systematic reviews are suitable for this kind of research because, unlike other reviews that take a more general approach, they describe the collection process of the papers along with the filtering strategies that were employed, ensuring the reproducibility of the results [14]. Additionally, a literature review is appropriate in this context because it allows for a comprehensive examination of existing research on audio visual databases for Automatic Affect Prediction (AAP) models [15]. By systematically reviewing the literature, patterns, gaps, and inconsistencies in previous studies can be identified, providing a robust foundation for this research. Moreover, the research followed the PRISMA 2020 Guidelines [13] to ensure that it transparently reports why the review was done, how it was done and what it found.

To conduct a proper Systematic Literature Review, the following process was followed:

1. **Searching:** Formulated a query to locate relevant literature within scientific databases.
2. **Filtering:** Manually screened the retrieved papers using the criteria outlined in Eligibility Criteria.
3. **Extraction:** Reviewed the selected papers to extract information relevant to the research question.
4. **Synthesise and interpret data:** Presented the synthesized data in accordance with PRISMA Guidelines [13] and discussed the conclusions.

This section describes the methodology employed through out the research period of 9 weeks. The structure is as follows: Subsection 2.1 introduces the criteria for including and excluding papers during manual screening. Subsection 2.2 presents the search strategy, including the selection of literature databases and the motivation behind their choice. Subsection 2.3 discusses design decisions made to ensure the research could be completed within the limited time frame. Subsection 2.4 presents the steps of retrieving the papers and assessing the inclusion of each one. Finally, subsection 2.5 presents the search results obtained after completing the process, while subsection 2.6 presents the information extracted from the included papers.

2.1 Eligibility Criteria

In order to ensure consistency in selecting the papers to review, and implicitly allow for reproducibility, eligibility criteria must be established. For this survey, the eligibility criteria used are presented in Table 1.

Table 2: Criteria for including or excluding literature from the review.

Inclusion Criteria	**Motivation**
I1: Paper introduces a new Audio-Visual affect prediction dataset	Main focus of the research, helps to ensure the uniqueness of the databases in the set of the literature to review.
I2: Data has been labeled by at least one human rater on both modalities and paper discusses the procedure used for labelling	To discuss interrater agreement, it is important to understand how the data was labeled.
Exclusion Criteria	**Motivation**
E1: Paper, with the exception of the dataset, is not in English	For the purpose of reproducibility of the review, the papers should be accessible for the vast majority of people.
E2: Papers released after 20.05.2024	The date at which the queries for this research were conducted. Having this exclusion criteria helps ensure the reproducibility of this study.
E3: Dataset is labeled using self-reports instead of third-party annotations	The focus of this research is to study interrater agreement and that can only be present if the labeling focuses on how third party individuals perceive that feeling or emotion and not how the subject actually feels.
E4: The subject that generates the affective state is not human	Anything can provoke an affective state, but the study is focused on affect prediction for humans.

2.2 Search strategy

To perform the searches, we used the following search engines: **Scopus**[1], **Web of Science**[2], **IEEE Xplore**[3], **and ACM Digital Library**[4]. These databases were selected for their popularity, extensive coverage of technical literature, and recommended by TU Delft for computer science research (source). IEEE Xplore was included because it is the publisher of ACII and the Transactions on Affective Computing, both of which are premier venues for the field of affective computing. Additionally, the ACM Digital Library was specifically chosen due to its substantial collection of papers in the field of affective computing.

To find the relevant literature for our survey, a search query needed to be developed. In order to do that, the topic was split in 4 broad concepts: Audio-Visual, Affect Prediction, Dataset and Raters. Their intersection represents the relevant set of papers, see Figure 1. For each category a set of keywords was defined using the following procedure, with the specific terms available in Table 3. Initially, a set of 7 relevant papers for this research were manually selected. The set consist of: [16, 17, 18, 19, 20, 21, 22]. Then for each concept, relevant words were determined and assembled into a query. The query was ran in each literature database, with small adaptations for each. The results were analysed to check if they matched the expectations, whether the query was wide/narrow enough and if they returned the set of papers that was manually selected, if they were present in that specific database. The process was repeated until a adequate query was found. The queries for each database along with the words used for each concept can be found in Appendix A.

Figure 1: Venn diagram presenting the main concepts of the query. The red circle highlights their intersection, the set of relevant papers.

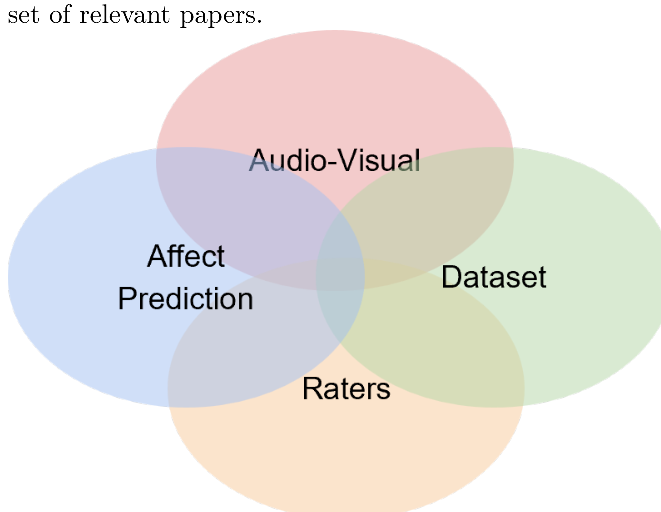


Table 3: Concepts and their associated keywords

Concept	Keywords
Audio-Visual	video, audiovisual, audio visual, audio-visual, audio video, speech video, speech visual
Affect Prediction	**Affect:** affect, affective, emotion, emotional, mood, mental state **Prediction:** analysis, recognition, predict*, computing, detect*, classif*
Dataset	dataset, database, corpus
Rater	manual label*, manual annotat*, manual rat*, manual evaluators, human label*, human annotat*, human rat*, human evaluators, multiple label*, multiple annotat*, multiple rat*, multiple evaluators, rat*, interrat*, inter-rat*, inter-evaluator

[1] Scopus: https://www.scopus.com/home.uri
[2] Web of Science: https://webofscience.com/
[3] IEEE Xplore: https://ieeexplore.ieee.org/Xplore/home.jsp
[4] ACM Digital Library: https://dl.acm.org/

2.3 Feasibility Constraints

Due to the limited time frame available for this research, namely 9 weeks, a few design choices were made in the query building process to ensure that the project can be achieved successfully. Normally in a Systematic Literature Review, it is preferable to have a wider query and then filter the papers manually in order to get the relevant results. Given the exceptionally high number of available papers in the Affective Computing field, especially for audio visual automatic affect prediction, it was decided that the query terms for each concept will be searched only in specific areas of the paper. The search was done as follows: In the title, abstract, and keywords, searches will focus on terms related to audio-visual and prediction. Terms related to datasets and affect will be searched for only in the title. Since terms related to the rater concept are typically not found in the title or abstract, they will be searched for throughout the entire body of the text.

2.4 Selection Process

1. **Run literature queries:** Perform a query on each of the literature databases mentioned in 2.2, using the queries from Appendix A.
2. **Screen by title:** Decide whether to include or exclude the papers obtained from step 1 using the criteria defined in 2.1. In case the title is vague, the paper will be analysed further in the next steps.
3. **Screen by abstract:** Decide whether to include or exclude the papers obtained from step 2 using the criteria defined in 2.1.
4. **Screen by full-text:** Screen the full text of the paper during the final review of the papers and exclude records that do not fit the Eligibility Criteria.

2.5 Search Results

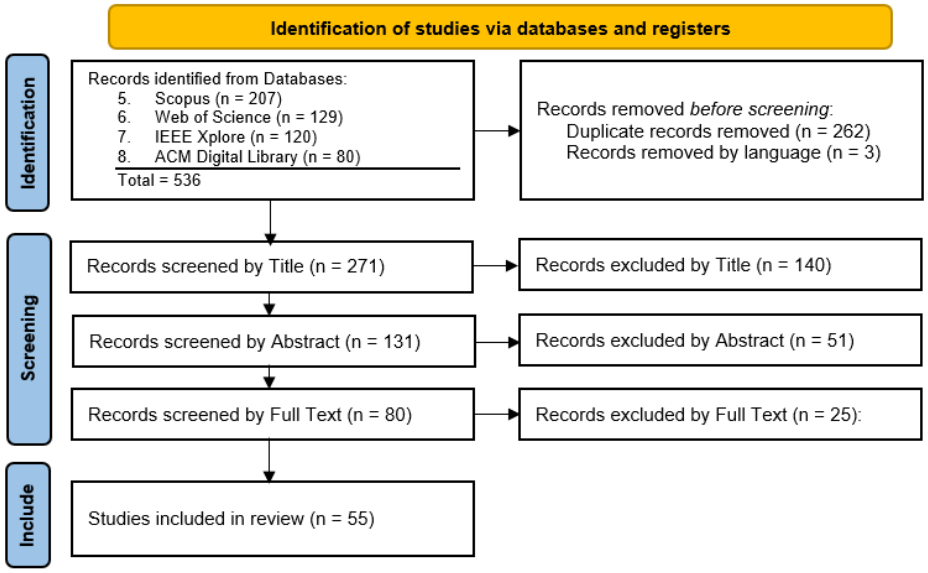
After executing the query, as described in Search strategy, a total of 536 papers were retrieved on May 20, 2024. The sources were as follows: 207 from Scopus[1], 129 from Web of Science[2], 120 from IEEE Xplore[3], and 80 from ACM Digital Library[4].

Out of them, 262 duplicates and 3 non-English papers were removed. Next, following the Eligibility Criteria, 140 papers were excluded during title screening, 51 during abstract screening, and 25 during full-text screening. Ultimately, 55 papers were included in the review. These steps are visually summarized in Figure 2.

Of the 25 papers excluded during full-text screening, the reasons were as follows:

- 9 lacked data annotated on both modalities
- 8 lacked the description of the annotation process or used the intended labels (i.e. the affective state that the video meant to provoke)
- 5 had self-reported labels
- 2 did not use humans in the video stimulus
- 1 did not introduce a new dataset

Figure 2: PRISMA Flow Diagram[13, p.10] for summarizing the searching and filtering process outlined in subsection 2.4



2.6 Data Extraction

The data extraction process began right after executing the queries. The 536 results were exported to Zotero[5], a reference manager. Using the built-in features of Zotero[5], the duplicates where removed, followed by a manual pass of the entries sorted by title to eliminate the duplicates that were missed by the automatic tool. After finalising the Selection Process, relevant data to the research question was extracted from each of the included paper with the help of a Microsoft Excel[6] spreadsheet. From each entry, the following information was extracted:

- Publication Year
- Targeted Affective States
- Affect Representation Scheme type
- Affect Representation Scheme
- The motivation behind using the Affect Representation Scheme, if present
- The number of raters used for the annotation process
- The Interrater Agreement measures used, if any
- Strategies to facilitate Interrater Agreement, if present

3 Results

This chapter highlights the results of the study. It aims to provide an objective, data based, answer to the research question. The data was extracted from a total of 55 datasets, covering the 16-year span from 2008 to 2024. All the relevant data can be found in Appendix B. It is important to note that that some of the literature used additional methods to label data such as machine learning tools for automatic labelling or self reports of the subjects. In these

[5]Zotero: https://www.zotero.org/
[6]Excel: https://www.microsoft.com/en-us/microsoft-365/excel

cases, only the part that is labeled by objective human raters us considered for this study.

The structure of this section is the following. Section 3.1 presents the types of affective states targeted by the datasets, aiming to answer SQ1. Section 3.2 provides an overview of the affect representation schemes, aiming to solve SQ2. Section 3.3 goes into the details of the annotation procedure such as the number of annotators used, whether Interrater Agreement is measured and what, if any, strategies for facilitating Interrater Agreement are used, thus, addressing SQ3. Section 3.4 explores the changes in how interrater agreement is measured in datasets over the years, answering SQ4. Lastly, section 3.5 provides an answer to SQ5 by analyzing a possible relationship between the Affect Representation Scheme used by datasets and their Interrater Agreement.

3.1 Targeted Affective States

This subsection discusses the types of affective states that were aimed to be identified by the included studies. A detailed view of the targeted affective state types, along with the relevant data for each paper, can be found in Appendix B.

As mentioned in the Introduction, according to [1], the term affect encompasses a wide range of emotional experiences, from intense, short-lived emotions to diffuse, long-lasting moods and stable affective dispositions. Scherer emphasizes the complexity of precisely defining emotions, as they are often confused with moods or feelings. This difficulty arises because emotions, moods, and feelings, although related, are distinct affect phenomena that operate on different temporal scales and levels of intensity. Emotions can be defined as intense and brief episodes that are triggered by specific events, while moods are more diffuse and long lasting, often emerging without an apparent cause [1]. Lastly, attitudes are long term, consistent predispositions towards specific objects or persons [1], while sentiments are simpler, often reflecting positive or negative feelings towards specific stimuli, such as words or phrases, in a given context [23].

Following the review of the 55 included papers, it was discovered that 47 focused solely on emotions, while the other 8 combined the study of emotions with attitude, sentiments, moods or mental states. Even though mental states are not considered an affective state, they were used by [24, 25] in conjunction with categorical emotions to annotate videos. Table 4 presents an overview of this information.

Table 4: Types of affective states and the research papers that target them

Targeted Affective States	Number of Papers	Papers
Emotion	47	[26, 27, 28, 29, 30, 31, 32, 10, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 21, 48, 49, 17, 50, 16, 51, 19, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 20, 62, 63, 64, 65, 18]
Emotion + Sentiment	3	[66, 67, 68]
Emotion + Attitude	1	[69]
Emotion + Mood	2	[22, 11]
Emotion + Mood + Mental States	1	[25]
Emotion + Mental States	1	[24]

3.2 Affect Representation Schemes

This section summarizes the affect representation schemes utilized in the reviewed papers, aiming to provide an answer to SQ2. An overview of all individual affect representation schemes, along with their encoding and their associated datasets can be found in Appendix B. Table 3.2 provides a summary of the identified Affect Representation Schemes and which papers utilize them.

Table 5: Summary of the Affect Representation Schemes used by each individual paper

Categorical	Ekman's basic emotions [70]	[64, 65, 24, 59, 60, 22, 25, 19, 16, 21, 47, 46, 45, 68, 67, 37, 36, 39, 33, 30, 29, 58]
	Plutchik's Wheel of Emotions [71]	[40, 69]
	"emotion zones for regulation" framework [72]	[10]
	Other	[44]
Dimensional	Valence-Arousal	[20, 62, 57, 53, 55, 56, 51, 49, 11, 35, 26]
	Valence-Arousal + Others	[28, 66, 42, 50, 61]
	Valence	[38]
	SAM's Pleasure-Arousal-Dominance [73]	[32]
Mix	V+labels	[54, 17, 43]
	VA+labels	[27, 34, 41, 48, 52]
	VAD+labels	[18, 63, 31]

The included paper used a total of 43 different affect representation schemes. From this total, 24 are categorical ARS, 8 are dimensional and 12 are a mix of the two.
The categorical schemes primarily rely on discrete labels to represent different emotions. In the reviewed papers, it was observed that one of the most widely adopted categorical ARS was introduced by Paul Ekman, who defined six basic emotions: anger, disgust, fear, joy, surprise, and sadness [70]. Out of the 24 categorical schemes, one is directly based on Ekman's model [45] while other 19 are derivatives of Ekman's original model, often using a subset or superset of the emotion labels and a neutral category. Moreover, it was observed that the labels "joy" and "happiness" were used interchangeably. On the second position in terms of popularity in the studies is Plutchik's Wheel of Emotions, which categorizes emotions into eight groups: anger, disgust, fear, joy, surprise, sadness, trust, and anticipation. Two papers utilized this model, one being a direct application of Plutchik's scheme [40], and the other being a variation of this model [69]. Out of the 2 ARS that are left, one interesting approach found in the literature is implemented by CALMED [10]. It uses the "emotion zones for regulation" framework [72], which categorizes emotions into four distinct zones: green, yellow, red, and blue. Each zone represents a different set of emotions, making it easier to identify one's emotional state. The green zone encompasses calm and positive emotions, such as happiness. The yellow zone includes less comfortable emotions, like excitement and worried. The red zone covers intense and often disruptive emotions, such as anger and fear, while the blue zone represents low-energy emotions, like sadness and fatigue. The last categorical ARS was implemented by [44] and it aimed to classify emotions using the labels neutral, positive, frustrated, and anxious, as, according to the author, these were the most common emotional classes used in related studies.

Dimensional schemes, on the other hand, represent emotions along continuous dimensions. From the papers included in this review, it was determined that 8 distinct dimensional ARS were used. The most popular dimensions were Valence (positive or negative affect) and Arousal (intensity), both being present in all, but 2 affect representation schemes. Other dimensions used in conjuncture with valence and arousal are: dominance, liking/disliking, impact, engagement and aggression. From the remaining 2 dimensional ARS that do not use both Valence and Arousal, one decided to only used valence to classify emotions, while the other one decided to use SAM's pleasure, arousal and dominance model [73] in which pleasure reflects how positively or negatively one feels in response to a stimulus or situation, while dominance indicates the degree of control or influence a person feels in a given situation.

The mixed schemes combine elements of both categorical and dimensional approaches, providing a more comprehensive representation of affect. Out of the 11 different mixed ARS, 3 combine Valence with a variation of categorical labels such as Ekman's basic emotions [70], 5 combine both Valence and Arousal with a variation of categorical labels, while the last 3 combine Valence, Arousal and Dominance with the categorical emotions.

Most of the papers did not provide the motivation behind the ARS. There are, however some exceptions. CHEAVD [19] decided to use a categorical approach due to the short length of the audio visual stimuli they used, as it would create difficulties in the labelling process. According to them, using the categorical approach would also benefit the raters as it would be easier to understand the process. Another study, [45], argues for the use of Ekman's basic emotions [70] as the facial stimuli are similar across many cultures and this set of labels represent more of a universal framework. Lastly, [28, 49] argue that using the Valence Arousal system would be easily adaptable to either discrete or dimensional emotion definitions.

3.3 Interrater Agreement

This sections presents the results related to the interrater agreement. It begins by examining the number of raters involved in the labeling process, followed by an overview of the methods used to calculate interrater agreement and lastly, whether the researcher took any measures to improve the agreement between the raters. This section aims to answer SQ3.

Number of Annotators This research aimed to study interrater agreement in audio visual datasets in order to lay the ground work for studying its effect on performance. To that extent, and due to the limited the time frame of the study, the queries were focused on finding datasets were the data was labelled by humans. Therefore, in almost all cases, the researchers acknowledged the subjectivity in interpreting affective states, and decided to use multiple raters. The only case in which this did not happened is [56], where the authors decided to only use on rater.

When searching for annotators, researchers opted for 2 solution. One of them was to collect as many as raters as possible from their local areas, such as students at their universities or local experts. The second option was to use a crowd sourcing platform such as Amazon MTurk[7]. One such case is [57] in which the research recruited a total of 1517 raters to annotate the dataset on the the valence scale and 2442 raters for the arousal scale. Table 6 presents a summary of the amount of raters used by the databases. In the cases where authors explicitly mentioned that the audio-visual stimuli were rated by only a certain amount of people from the total, the minimum number of raters per video was taken. Appendix B presents the amount of raters for each individual dataset.

Interrater Agreement Measures From the reviewed papers, 34 out of the 55 measure interrater agreement (IRA). Most of them calculate IRA using one of the following metrics: Cohen's kappa [74], Fleiss' kappa [75], Krippendorff's alpha [76], or Cronbach's alpha [77]. However some of the papers also consider using the percentage of the same categorical choice as the metric, Intraclass Correlation Coefficient (ICC), FINN coefficient, percent consensus, Randolph's kappa, Root Mean Square Error (RMSE), Spearman's rho, Concordance Correlation Coefficient (CCC), or the Pearson correlation. Besides that, there is also a study, namely [65] that does not mention the method used to calculate the IRA. The measures used by each paper can be found in Appendix B

[7]Amazon MTurk: https://www.mturk.com/

Table 6: Summary of the amount of raters used by the dataset

Num. of Raters	Num. of Datasets	Datasets
1	1	[56]
2	2	[64, 10]
3	13	[60, 46, 44, 67, 69, 54, 43, 27, 61, 50, 49, 11, 28]
4	3	[19, 68, 32]
5	10	[65, 25, 16, 21, 18, 58, 48, 41, 53, 66]
6	3	[40, 36, 20]
7	2	[33, 51]
8	1	[59]
10-19	5	[22, 45, 39, 29, 55]
20-29	5	[30, 63, 62, 38, 26]
30-49	3	[37, 17, 34]
>50	7	[24, 47, 52, 31, 57, 42, 35]

Figure 3: Popularity of the IRA measures (a) lifetime (b) over the years.

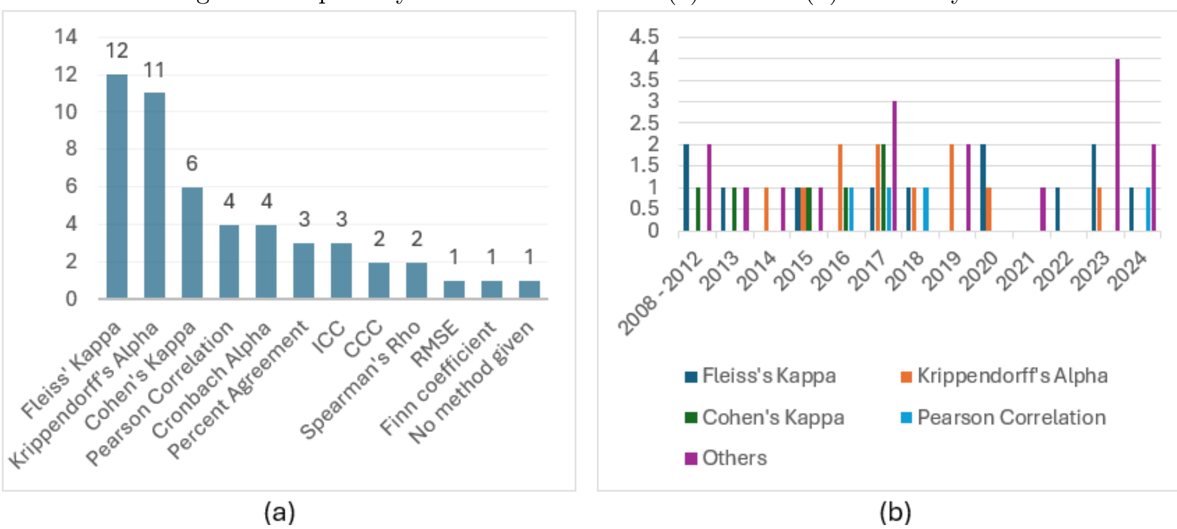


Figure 3a highlights which methods were used the most for assessing the interrater agreement in the studied papers. It is clear that the preferred methods are Fleiss' Kappa [75] and Krippendorff's Alpha[76], with both being used (almost) twice as much as the Cohen's Kappa [74]. It is important to note that some papers used multiple methods to calculate agreement.

Given the wide array of interrater agreement measures and the varying interpretations provided by their respective authors, comparing the agreement scores can be challenging and unreliable. To that extent, the collected agreement score for each paper are presented in Appendix B.

Strategies for Facilitating Agreement In order to facilitate IRA, many researcher have implemented certain strategies. However, no author decided to measure IRA before and after implementing the strategies. Therefore, it is not possible to measure their effectiveness, but they should bring consistency into the labelling process. The data suggests that the most commonly used strategy to facilitate agreement among annotators is providing instructions, with 9 datasets implementing it. The instructions usually include tips on how to use the annotating platform and definitions of each emotional label or dimension. This is followed by giving the annotators a list of examples (7 instances) and then by eliminating annotations where the rater seemed to be distracted or unreliable (6 instances). Other 5 papers decide to eliminate the videos where raters couldn't agree on a label. Expert raters were used in 5 cases, while testing the annotators for consistency and accurate understanding of the emotion in order to eliminate those that perform poorly was employed 5 times. Conducting training workshops, and hiring raters that are from the same region or culture as the subject were also employed 4 times each. A less common strategy was to include using raters that are familiar with the subject. In [10], the authors use children with autism as the audio visual stimulus, while there parents label the emotion that they believe their own child experiences. Although IRA is not measured in this study, involving parents likely enhances IRA because they can consistently label emotions based on their experience with their child.

3.4 Interrater Agreement measures over time

This section will discuss the trends in the preferred method to calculate IRA over the course of the years. Figure 3b presents the measures that were used in every year from 2008 to 2024. Aiming to answer SQ4

Due to the small number of datasets prior to 2012, all datasets from 2008 to 2012 were merged in a single column. The landscape of interrater Agreement methods saw a dynamic evolution over the years. In the early times, researchers experimented with Fleiss's Kappa and Cohen's Kappa, while Krippendorff's Alpha emerged around 2014 in the competition for reliable IRA measures. Fleiss's Kappa appears to have maintained a costant popularity over the years, while in the past 2 years, researchers have began to experiment with other options.

3.5 The Relationship between the Affect Representation Scheme used by datasets and their Interrater Agreement

This section aims to search for links between the affect representation schemes and their respective interrater agreement level, fulfilling SQ5.

After reviewing the literature, 43 different ARS were found in a set of 55 papers. This could pose a challenge in finding a connection between ARS and IRA due to the limited data available for each specific. However, there is the Valence-Arousal (VA) ARS of which 11 entries are present and the many derivations of the Ekman's basic emotions [70]. After analysing each of this possible combinations, it was concluded that no relationship can be determined between ARS and Interrater agreement. This is because, if we look at the VA, both low [51, 55], and high [11, 20] levels of agreement can be found. This is similar to the variations of Ekman's basic emotions [70] where we can also find both low [30] and high [68, 69] levels of IRA. Therefore, this section concludes that there is no relevant connection between the ARS and the the level of IRA.

4 Responsible Research

In order to conduct research in a responsible manner, it is necessary to consider the ethical implication. This section outlines two parts: subsection 4.1 provides a reflection about the used methodology, while subsection 4.2 discusses the ethical consideration of affect prediction.

4.1 Reflection upon the Methodology

This research involves conducting a systematic literature review. This approach aims to ensure a comprehensive and unbiased synthesis using the available literature on the subject. It is important to note that systematic literature reviews happen across a longer periods of time, usually months, sometimes including multiple researchers, while this research only spanned 10 weeks and was done by only one researcher. This means that the interpretation of results or the decisions made during the screening process might have been affected by bias or missed some of the necessary rigour. Despite efforts to maintain bias as low as possible during the selection process, the possibility of bias cannot be completely ruled out. Moreover, the researcher is a 3rd year TU Delft student with little to no experience in systematic literature reviews.

Moreover, because the queries implement a selective approach when it comes to in which parts does it search for specific terms can also have an impact in the search results, as some papers could have been missed. Unfortunately, given the limited time frame of this research, no measure could have been taken to prevent this.

4.2 Ethical Viewpoint on Affect Prediction

This research aims to help research improve the datasets used for Automatic Affect Prediction systems and, implicitly, their performance. Therefore, the ethical implications involved with the usage of this system must be considered.

These systems are based on the datasets they are trained on and because of this, many unwanted effects can occur. One of them is the possibility of classification which can have a varying impacts depending on the industry they are used in. Another potential issue is that interpretations can vary across different languages and cultures which raises concerns of whether the system is appropriate for the location in which it will be used. This issues can arise during the use of such systems, but they can be mitigated by the developers, if they are taken into account during the development or implementation process.

5 Discussion

While 34 out of the 55 papers included in the study measure interrater agreement (IRA), none reported a second labeling run aimed at improving the agreement score. This is surprising since the initial calculation of the agreement might suggest potential inconsistencies among raters. Ideally, datasets with low IRA scores could benefit significantly from a re-evaluation and refinement of the labels. Such a process would not only improve the reliability of the dataset but could possibly also enhance the quality of models trained on these datasets. 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. This lack of standardization in ARS contributes to difficulties in assessing how well it aligns with IRA, ultimately impacting the reliability of the datasets and the performance of the affect prediction models trained on them.

Moreover, the lack of standardized ARS highlight the need for dataset annotation guidelines. This could align the focuses of the community and accelerate the development of affective databases.

To summarise, the practice of measuring IRA without subsequent refinements to the labeling process appears to be a missed opportunity for improving dataset quality. Combined with the deviations from well established ARS, this highlights the need for a standardized approach in affective dataset creation.

6 Conclusions and Future Work

The aim of this study was to investigate whether audio-visual datasets designed for automatic affect prediction attempt to facilitate interrater agreement. This paper conducted a systematic literature review which fully reviewed a total of 55 papers that introduced an audio visual dataset.

The review revealed that 54 out of the 55 papers used multiple raters for dataset labeling, and 34 of these measured interrater agreement. It was observed that the level of interrater agreement was found to be independent of the affective representation scheme employed. However, none of the reviewed papers measured interrater agreement before and after implementing strategies to improve it, preventing an evaluation of the effectiveness of these strategies. Despite this, section 3.3 of the paper provides an overview of the most commonly used tactics to facilitate interrater agreement.

This study contributes to the understanding of the role of interrater agreement in the development of reliable audio-visual datasets for automatic affect prediction. It lays the foundation for assessing the impact of interrater agreement on model performance. To investigate this further, future work should involve tracking all the studies that reference the dataset papers used in this study and that develop an automatic affect prediction model based on them. Then, the performance differences between these models should be compared in order to draw meaningful conclusions about the influence of interrater agreement on affective prediction accuracy. One might wonder why the analysis shouldn't be confined to the datasets used in this research. The reason is straightforward: not all papers implement a baseline model, and even those that do may not be generally representative. Hence, a broader set of models is necessary for a comprehensive study.

References

[1] K. R. Scherer, "What are emotions? and how can they be measured?," *Social Science Information*, vol. 44, pp. 695–729, 12 2005.

[2] E. Avots, T. Sapiński, M. Bachmann, and D. Kamińska, "Audiovisual emotion recognition in wild," *Machine Vision and Applications*, vol. 30, pp. 975–985, July 2019.

[3] E. Cambria, H. Wang, and B. White, "Guest editorial: Big social data analysis," *Knowledge-Based Systems*, vol. 69, pp. 1–2, 10 2014.

[4] S. Wang and Q. Ji, "Video affective content analysis: A survey of state-of-the-art methods," *IEEE Transactions on Affective Computing*, vol. 6, pp. 410–430, 10 2015.

[5] Y. Baveye, C. Chamaret, E. Dellandrea, and L. Chen, "Affective Video Content Analysis: A Multidisciplinary Insight," *IEEE Transactions on Affective Computing*, vol. 9, pp. 396–409, Oct. 2018.

[6] R. A. Calvo and S. D'Mello, "Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications," *IEEE Transactions on Affective Computing*, vol. 1, pp. 18–37, Jan. 2010.

[7] A. Jain, H. Patel, L. Nagalapatti, N. Gupta, S. Mehta, S. Guttula, S. Mujumdar, S. Afzal, R. Sharma Mittal, and V. Munigala, "Overview and Importance of Data Quality for Machine Learning Tasks," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, (Virtual Event CA USA), pp. 3561–3562, ACM, Aug. 2020.

[8] V. N. Gudivada, A. Apon, and J. Ding, "Data Quality Considerations for Big Data and Machine Learning: Going Beyond Data Cleaning and Transformations," 2017.

[9] F. Cabitza, A. Campagner, and M. Mattioli, "The unbearable (technical) unreliability of automated facial emotion recognition," *Big Data and Society*, vol. 9, July 2022.

[10] A. Sousa, K. Young, M. dâAquin, M. Zarrouk, and J. Holloway, "Introducing CALMED: Multimodal Annotated Dataset forÂ Emotion Detection inÂ Children withÂ Autism," in *Lect. Notes Comput. Sci.*, vol. 14020 LNCS, pp. 657–677, Springer Science and Business Media Deutschland GmbH, 2023.

[11] J. Miranda-Correa, M. Abadi, N. Sebe, and I. Patras, "AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups," *IEEE Transactions on Affective Computing*, vol. 12, no. 2, pp. 479–493, 2021.

[12] N. Gisev, J. S. Bell, and T. F. Chen, "Interrater agreement and interrater reliability: Key concepts, approaches, and applications," *Research in Social and Administrative Pharmacy*, vol. 9, pp. 330–338, May 2013.

[13] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, *et al.*, "The prisma 2020 statement: an updated guideline for reporting systematic reviews," *Bmj*, vol. 372, 2021.

[14] A. Boland, G. Cherry, and Dickson Rumona, *Doing a systematic review: a student's guide.* SAGE, 2014.

[15] J. W. Knopf, "Doing a Literature Review," *PS: Political Science & Politics*, vol. 39, pp. 127–132, Jan. 2006.

[16] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception," *IEEE Transactions on Affective Computing*, vol. 8, pp. 67–80, Jan. 2017.

[17] H.-C. Chou, W.-C. Lin, L.-C. Chang, C.-C. Li, H.-P. Ma, and C.-C. Lee, "NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 292–298, Oct. 2017.

[18] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335–359, Dec. 2008.

[19] Y. Li, J. Tao, L. Chao, W. Bao, and Y. Liu, "CHEAVD: a Chinese natural emotional audioâvisual database," *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, pp. 913–924, Nov. 2017.

[20] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–8, Apr. 2013.

[21] O. Perepelkina, E. Kazimirova, and M. Konstantinova, "RAMAS: Russian Multimodal Corpus of Dyadic Interaction for Affective Computing," in *Speech and Computer* (A. Karpov, O. Jokisch, and R. Potapova, eds.), (Cham), pp. 501–510, Springer International Publishing, 2018.

[22] K. Wang, Z. Zhu, S. Wang, X. Sun, and L. Li, "A database for emotional interactions of the elderly," in *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, pp. 1–6, June 2016.

[23] E. H. Hovy, "What are Sentiment, Affect, and Emotion? Applying the Methodology of Michael Zock to Sentiment Analysis," in *Language Production, Cognition, and the Lexicon* (N. Gala, R. Rapp, and G. Bel-Enguix, eds.), pp. 13–24, Cham: Springer International Publishing, 2015.

[24] M. Mahmoud, T. BaltruÅ¡aitis, P. Robinson, and L. Riek, "3D Corpus of spontaneous complex mental states," in *Lect. Notes Comput. Sci.*, vol. 6974 LNCS, pp. 205–214, 2011.

[25] S. Zhalehpour, O. Onder, Z. Akhtar, and C. Erdem, "BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 300–313, 2017.

[26] Z. Ren, J. Ortega, Y. Wang, Z. Chen, Y. Guo, S. Yu, and D. Whitney, "VEATIC: Video-based Emotion and Affect Tracking in Context Dataset," in *Proc. - IEEE Winter Conf. Appl. Comput. Vis., WACV*, pp. 4455–4465, Institute of Electrical and Electronics Engineers Inc., 2024.

[27] S. Wang, X. Li, F. Zheng, J. Pan, X. Li, Y. Chang, Z. Zhu, Q. Li, J. Wang, and Y. Xiao, "VAD: A Video Affective Dataset with Danmu," *IEEE Transactions on Affective Computing*, pp. 1–17, 2024.

[28] H. Chen, S. Alghowinem, S. Jang, C. Breazeal, and H. Park, "Dyadic Affect in Parent-Child Multimodal Interaction: Introducing the DAMI-P2C Dataset and its Preliminary Analysis," *IEEE Transactions on Affective Computing*, vol. 14, no. 4, pp. 3345–3361, 2023.

[29] J. Ong, F. Leung, and F. Liu, "The Reading Everyday Emotion Database (REED): a set of audio-visual recordings of emotions in music and language," *Language Resources and Evaluation*, 2023.

[30] C. von Eiff, J. Kauk, and S. Schweinberger, "The Jena Audiovisual Stimuli of Morphed Emotional Pseudospeech (JAVMEPS): A database for emotional auditory-only, visual-only, and congruent and incongruent audiovisual voice and dynamic face stimuli with varying voice intensities," *Behavior Research Methods*, 2023.

[31] S. Mishra, N. Srinivasan, M. Asif, and U. Tiwary, "Affective film dataset from India (AFDI): creation and validation with an Indian sample," *Journal of Cultural Cognitive Science*, vol. 7, no. 3, pp. 255–267, 2023.

[32] K. Zhang, X. Wu, X. Xie, X. Zhang, H. Zhang, X. Chen, and L. Sun, "Werewolf-XL: A Database for Identifying Spontaneous Affect in Large Competitive Group Interactions," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1201–1214, 2023.

[33] A. Ray, S. Mishra, A. Nunna, and P. Bhattacharyya, "A Multimodal Corpus for Emotion Recognition in Sarcasm," in *Lang. Resour. Eval. Conf., LREC*, pp. 6992–7003, European Language Resources Association (ELRA), 2022.

[34] F. Shalchizadeh, S. Shamekhi, R. Naghdi Sadeh, and A. Darvish, "Persian emotion elicitation film set and signal database," *Biomedical Signal Processing and Control*, vol. 72, 2022.

[35] M. Gnacek, I. Mavridou, J. Broulidakis, C. Nduka, E. Balaguer-Ballester, T. Kostoulas, and E. Seiss, "AVDOS-Affective Video Database Online Study Video database for affective research emotionally validated through an online survey," in *Int. Conf. Affect. Comput. Intell. Interact., ACII*, Institute of Electrical and Electronics Engineers Inc., 2022.

[36] T. Khanh, S.-H. Kim, G. Lee, H.-J. Yang, and E.-T. Baek, "Korean video dataset for emotion recognition in the wild," *Multimedia Tools and Applications*, vol. 80, no. 6, pp. 9479–9492, 2021.

[37] H. Chen, K. Chin, and C. Tan, "Selection and validation of emotional videos: Dataset of professional and amateur videos that elicit basic emotions," *Data in Brief*, vol. 34, 2021.

[38] D. Ong, Z. Wu, T. Zhi-Xuan, M. Reddan, I. Kahhale, A. Mattek, and J. Zaki, "Modeling Emotion in Complex Stories: The Stanford Emotional Narratives Dataset," *IEEE Transactions on Affective Computing*, vol. 12, no. 3, pp. 579–594, 2021.

[39] J. Chen, C. Wang, K. Wang, C. Yin, C. Zhao, T. Xu, X. Zhang, Z. Huang, M. Liu, and T. Yang, "HEU Emotion: a large-scale database for multimodal emotion recognition in the wild," *Neural Computing and Applications*, vol. 33, no. 14, pp. 8669–8685, 2021.

[40] G. Shen, X. Wang, X. Duan, H. Li, and W. Zhu, "MEmoR: A Dataset for Multimodal Emotion Reasoning in Videos," in *MM - Proc. ACM Int. Conf. Multimed.*, p. 493502, Association for Computing Machinery, Inc, 2020.

[41] C. Park, N. Cha, S. Kang, A. Kim, A. Khandoker, L. Hadjileontiadis, A. Oh, Y. Jeong, and U. Lee, "K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic conversations," *Scientific Data*, vol. 7, no. 1, 2020.

[42] K. Baraly, L. Muyingo, C. Beaudoin, S. Karami, M. Langevin, and P. Davidson, "Database of Emotional Videos from Ottawa (DEVO)," *COLLABRA-PSYCHOLOGY*, vol. 6, Feb. 2020.

[43] K. Ma, X. Wang, X. Yang, M. Zhang, J. Girard, and L.-P. Morency, "ElderReact: A multimodal dataset for recognizing emotional response in aging adults," in *ICMI - Proc. Int. Conf. Multimodal Interact.*, pp. 349–357, Association for Computing Machinery, Inc, 2019.

[44] A. Lotz, K. Ihme, A. Charnoz, P. Maroudis, I. Dmitriev, and A. Wendemuth, "Recognizing behavioral factors while driving: A real-world multimodal corpus to monitor the driver's affective state," in *LREC - Int. Conf. Lang. Resour. Evaluation*, pp. 1589–1596, European Language Resources Association (ELRA), 2019.

[45] T. Sapiński, D. Kamińska, A. Pelikant, C. Ozcinar, E. Avots, and G. Anbarjafari, "Multimodal Database of Emotional Speech, Video and Gestures," in *Lect. Notes Comput. Sci.*, vol. 11188 LNCS, pp. 153–163, Springer Verlag, 2019.

[46] F. Abu Shaqra, R. Duwairi, and M. Al-Ayyoub, "The audio-visual Arabic dataset for natural emotions," in *Proc. - Int. Conf. Future Internet Things Cloud, FiCloud*, pp. 324–329, Institute of Electrical and Electronics Engineers Inc., 2019.

[47] S. Livingstone and F. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, May 2018.

[48] P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, and S. Wermter, "The OMG-Emotion Behavior Dataset," in *Proc Int Jt Conf Neural Networks*, vol. 2018-July, Institute of Electrical and Electronics Engineers Inc., 2018.

[49] N. Lubis, D. Lestari, S. Sakti, A. Purwarianti, and S. Nakamura, "Construction of spontaneous emotion corpus from Indonesian TV talk shows and its application on multimodal emotion recognition," *IEICE Transactions on Information and Systems*, vol. E101D, no. 8, pp. 2092–2100, 2018.

[50] I. Lefter, C. Jonker, S. Tuente, W. Veling, and S. Bogaerts, "NAA: A multimodal database of negative affect and aggression," in *Int. Conf. Affect. Comput. Intell. Interaction, ACII*, vol. 2018-January, pp. 21–27, Institute of Electrical and Electronics Engineers Inc., 2017.

[51] A. Zlatintsi, P. Koutras, G. Evangelopoulos, N. Malandrakis, N. Efthymiou, K. Pastra, A. Potamianos, and P. Maragos, "COGNIMUSE: a multimodal video database annotated with saliency, events, semantics and emotion with application to summarization," *Eurasip Journal on Image and Video Processing*, vol. 2017, no. 1, 2017.

[52] L. Zhang, S. Walter, X. Ma, P. Werner, A. Al-Hamadi, H. Traue, and S. Gruss, "'BioVid Emo DB': A multimodal database for emotion analyses validated by subjective ratings," in *IEEE Symp. Ser. Comput. Intell., SSCI*, Institute of Electrical and Electronics Engineers Inc., 2017.

[53] J. Shukla, M. Barreda-Ángeles, J. Oliver, and D. Puig, "MuDERI: Multimodal database for emotion recognition among intellectually disabled individuals," in *Lect. Notes Comput. Sci.*, vol. 9979 LNAI, pp. 264–273, Springer Verlag, 2016.

[54] B. Nojavanasghari, T. Baltrušaitis, C. Hughes, and L.-P. Morency, "Emo react: A multimodal approach and dataset for recognizing emotional responses in children," in *ICMI - Proc. ACM Int. Conf. Multimodal Interact.*, pp. 137–144, Association for Computing Machinery, Inc, 2016.

[55] C. Katsimerou, J. Albeda, A. Huldtgren, I. Heynderickx, and J. Redi, "Crowdsourcing empathetic intelligence: The case of the annotation of EMMA database for emotion and mood recognition," *ACM Transactions on Intelligent Systems and Technology*, vol. 7, no. 4, 2016.

[56] N. Lubis, R. Gomez, S. Sakti, K. Nakamura, K. Yoshino, S. Nakamura, and K. Nakadai, "Construction of Japanese audio-visual emotion database and its application in emotion recognition," in *Int. Conf. Lang. Resourc. and Eval. - LREC*, pp. 2180–2184, European Language Resources Association (ELRA), 2016.

[57] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: A video database for affective content analysis," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 43–55, 2015.

[58] C. Eroglu Erdem, C. Turan, and Z. Aydin, “BAUM-2: a multilingual audio-visual affective face database,” *Multimedia Tools and Applications*, vol. 74, no. 18, pp. 7429–7459, 2015.

[59] H. Cao, D. Cooper, M. Keutmann, R. Gur, A. Nenkova, and R. Verma, “CREMA-D: Crowd-sourced emotional multimodal actors dataset,” *IEEE Transactions on Affective Computing*, vol. 5, no. 4, pp. 377–390, 2014.

[60] W. Bao, Y. Li, M. Gu, M. Yang, H. Li, L. Chao, and J. Tao, “Building a Chinese Natural Emotional Audio-Visual Database,” in *2014 12th International Conference on Signal Processing (ICSP)*, pp. 583–587, Oct. 2014.

[61] H. Wei, D. Monaghan, N. O’Connor, and P. Scanlon, “A new multi-modal dataset for human affect analysis,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8749, pp. 42–51, 2014.

[62] J. Cu, K. Solomon, M. Suarez, and M. Sta. Maria, “A multimodal emotion corpus for Filipino and its uses,” *Journal on Multimodal User Interfaces*, vol. 7, no. 1-2, pp. 135–142, 2013.

[63] M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic, “A multimodal database for affect recognition and implicit tagging,” *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 42–55, 2012.

[64] A. Esposito and M. T. Riviello, “The new italian audio and video emotional database,” in *Proceedings of the Second International Conference on Development of Multimodal Interfaces: Active Listening and Synchrony*, COST’09, (Berlin, Heidelberg), pp. 406–422, Springer-Verlag, 2009.

[65] A. Esposito, M. T. Riviello, and G. Di Maio, “The COST 2102 Italian Audio and Video Emotional Database,” in *Proceedings of the 2009 Conference on Neural Nets WIRN09: Proceedings of the 19th Italian Workshop on Neural Nets, Vietri Sul Mare, Salerno, Italy, May 28â30 2009*, (NLD), pp. 51–61, IOS Press, 2009.

[66] J. Kossaifi, R. Walecki, Y. Panagakis, J. Shen, M. Schmitt, F. Ringeval, J. Han, V. Pandit, A. Toisoul, B. Schuller, K. Star, E. Hajiyev, and M. Pantic, “SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 3, pp. 1022–1040, 2021.

[67] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, “MELD: A multimodal multi-party dataset for emotion recognition in conversations,” in *ACL - Annu. Meet. Assoc. Comput. Linguist., Proc. Conf.*, pp. 527–536, Association for Computational Linguistics (ACL), 2020.

[68] M. Firdaus, H. Chauhan, A. Ekbal, and P. Bhattacharyya, “MEISD: A Multimodal Multi-Label Emotion, Intensity and Sentiment Dialogue Dataset for Emotion Recognition and Sentiment Analysis in Conversations,” in *COLING - Int. Conf. Comput. Linguist., Proc. Conf.*, pp. 4441–4453, Association for Computational Linguistics (ACL), 2020.

[69] Y.-P. Ruan, S.-K. Zheng, J. Huang, X. Zhang, Y. Liu, and T. Li, “CH-MEAD: A Chinese Multimodal Conversational Emotion Analysis Dataset with Fine-Grained Emotion Taxonomy,” in *Asia Pac. Signal Inf. Process. Assoc. Annu. Summit Conf., APSIPA ASC*, pp. 498–505, Institute of Electrical and Electronics Engineers Inc., 2023.

[70] P. Ekman, “An argument for basic emotions,” *Cognition and Emotion*, vol. 6, pp. 169–200, May 1992.

[71] 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.

[72] 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.

[73] A. Mehrabian and J. A. Russell, *An approach to environmental psychology.* the MIT Press, 1974.

[74] J. Cohen, “A coefficient of agreement for nominal scales,” *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[75] J. L. Fleiss, "Measuring nominal scale agreement among many raters.," *Psychological bulletin*, vol. 76, no. 5, p. 378, 1971.

[76] K. Krippendorff, *Content analysis: An introduction to its methodology.* Sage publications, 2018.

[77] L. J. Cronbach, "Coefficient alpha and the internal structure of tests," *psychometrika*, vol. 16, no. 3, pp. 297–334, 1951.

A Final Queries and Concepts

A.1 Final Queries

- **Scopus:**

```
( TITLE-ABS-KEY ( ( "video" OR "audiovisual" OR ( ( "audio" OR "speech" )
    AND ( "video" OR "visual" ) ) ) )

AND TITLE ( ( "affect" OR "affective" OR "emotion" OR "emotional" OR "mood"
    OR "mental state" ) )

AND TITLE-ABS-KEY ( ( "analysis" OR "recognition" OR "predict*" OR "computing"
    OR "detect*" OR "classif*" ) )

AND TITLE ( ( "dataset" OR "database" OR "corpus" ) )

AND ALL ( ( ( ( "manual" OR "human" OR "multiple" ) AND ( "label*"
    OR "annotat*" OR "rat*" OR "evaluators" ) ) OR "rat*" OR "interrat*"
    OR "inter-rat*" OR "inter-evaluator" ) ) )
```

- **Web of Science:**

```
TS=(video OR audiovisual OR ((audio OR speech) AND (video OR visual)))

AND TI=(affect OR affective OR emotion OR emotional OR mood OR mental state )

AND TS=( analysis OR recognition OR predict* OR computing OR detect* OR classif*)

AND TI=(dataset OR database OR corpus)

AND ALL=(label* OR annotat* OR rat* OR evaluator OR rat* OR interrat*
    OR inter-rat* OR inter-evaluator)
```

- **IEEE Xplore:**

```
((("All Metadata":video OR "All Metadata":audiovisual
    OR ( ("All Metadata":audio OR "All Metadata":speech)
    AND ("All Metadata":video OR "All Metadata":visual) ))

AND ("Document Title":dataset OR "Document Title":database
    OR "Document Title":corpus)
```

```
AND ("Document Title":affect OR "Document Title":affective
    OR "Document Title":emotion OR "Document Title":emotional
    OR "Document Title":mood OR "Document Title":mental state )

AND ("All Metadata":analysis OR "All Metadata":recognition
    OR "All Metadata":predict* OR "All Metadata":computing
    OR "All Metadata":detect* OR "All Metadata":classif*)

AND ("Full Text & Metadata":rat* OR "Full Text & Metadata":interrat*
    OR "Full Text & Metadata":inter-rat*
    OR "Full Text & Metadata":inter-evaluator
    OR (( "Full Text & Metadata":manual
    OR "Full Text & Metadata":human OR "Full Text & Metadata":multiple)
    AND ("Full Text & Metadata":label* OR "Full Text & Metadata":annotat*
    OR "Full Text & Metadata":rat* OR "Full Text & Metadata":evaluator)))))
```

- **ACM Digital Library:**

```
Abstract:((video OR audiovisual OR ((audio OR speech)
    AND (video OR visual)))

AND ( analysis OR recognition OR predict* OR computing
    OR detect* OR classif*))

AND Title:((dataset OR database OR corpus)

AND (affect OR affective OR emotion OR emotional
    OR mood OR mental state ))

AND AllField:((( manual OR human OR multiple)
    AND (label* OR annotat* OR rat* OR evaluator))
    OR rat* OR interrat* OR inter-rat* OR inter-evaluator)
```

A.2 Concepts

- **Audio-Visual:** video, audiovisual, audio visual, audio-visual, audio video, speech video, speech visual
- **Affect Prediction:** Will be split in two parts and any combination will do.
  - **Affect:** affect, affective, emotion, emotional, mood, mental state
  - **Prediction:** analysis, recognition, predict*, computing, detect*, classif*
- **Dataset:** dataset, database, corpus
- **Rater:** manual label*, manual annotat*, manual rat*, manual evaluators, human label*, human annotat*, human rat*, human evaluators, multiple label*, multiple annotat*, multiple rat*, multiple evaluators, rat*, interrat*, inter-rat*, inter-evaluator

B Results Overview

Table 7 highlights the encodings for each affect representation scheme used in Table of Results, along with the number of times they have been used by the included studies, while Table of Results presents all the relevant data that was extracted from each paper.

Table 7: Affective Representation Scheme encodings for Table of Results, along with the amount of times they have been used by the included studies.

Labels	Amount of times used	
1EBS-4HSS-FI	V + anger, happiness, surprise, sadness + frustration + neutral + intensity	1
1EBS-6H	V + anger, disgust, fear, happiness, surprise, sadness	1
1EBS-6HN-9	V + anger, disgust, fear, happiness, surprise, sadness + neutral + curiosity, uncertainty, excitement, attentiveness, exploration, confusion, anxiety, embarrassment, frustration	1
2EBS-4A	VA + anger, disgust, fear, sadness, amusement	1
2EBS-5EMOCON	VA + anger, disgust, happiness, surprise, sadness, cheerful, nervous, boredom, confusion, delight, engaged concentration, frustration, none, confusion, contempt, dejection, eureka, pride, sorrow	1
2EBS-6H	VA + anger, disgust, fear, happiness, surprise, sadness	1
2EBS-6HN	VA + anger, disgust, fear, happiness, surprise, sadness + neutral	1
2WoE-8H	VA, anger, disgust, fear, happiness, surprise, trust, anticipation, sadness	1
3-3SNF	VAD + anger, happiness, sadness, frustration + neutral	1
3AF	VAD + aggression levels, fear	1
3EBS-6H-LF-C	VAD + liking, familiarity + anger, disgust, fear, happiness, surprise, sadness, contentment	1
3EBS-6NAA	VA + dominance + EBS-6 + neutral + anxiety, amusement	1
3SN	anger, happiness, sadness + neutral	1
3SN-SE	neutral, positive, frustrated, anxious	1
4C	(Emotion zones for regulation framework [72]) green, yellow, red and blue	1
EBS-4HN-2	anger, disgust, happiness, sadness + neutral + anxiety, boredom	1
EBS-5H	anger, disgust, fear, happiness, sadness	1
EBS-5HNI	EBS-5H + neutral + intensity	1
EBS-5SEN	anger, disgust, happiness, surprise, sadness + neutral	1
EBS-6	(Ekman's Basic Emotions [70]) anger, disgust, fear, joy, surprise, sadness	1
EBS-6HN	anger, disgust, fear, happiness, surprise, sadness + neutral	2
EBS-6HN-15-2	EBS-6HN + worried, anxious, blamed, sarcastic, aggrieved, curious, embarrassing, confused, proud, helpless, hesitant, contemptuous, frustrated, anticipated, shy, suspicious, fearful	1
EBS-6HN-15-4	EBS-6HN + worried, anxious, blamed, sarcastic, aggrieved, curious, embarrassing, confused, proud, helpless, hesitant, contemptuous, frustrated, anticipated, shy, guilty, exclamation, nervous, serious	1
EBS-6HN-2I	EBS-6HN + genuineness, calm, intensity	1
EBS-6HN-3	EBS-6HN + frustration, ridicule, excitement	1
EBS-6HN-4	EBS-6HN + bored, confused, disappointed, mixed	1
EBS-6HN-5	EBS-6HN + embarrassment, hopefulness, jealousy, pride, sarcasm, stress	1
EBS-6HN-CI	EBS-6HN + contempt + intensity	1

Continuation of Table 7		
Encoding	Labels	Amount of times used
EBS-6HN-MMS	EBS-6HN + boredom, contempt + unsure, thinking, concentrating, bothered	1
EBS-6HN-O	EBS-6HN + other (only when there are not at least 4 raters with the same label)	1
EBS-6HSI	happiness, sarcasm/irony, fear, anger, sadness, surprise	2
EBS-6-S	EBS-6 + sentiment: positive, negative and neutral	1
EBS-6-S-N2	EBS-6-S + neutral, acceptance + intensity	1
EMS	Thinking, concentrating, unsure, confused, triumphant, frustrated, angry, bored, neutral, surprised, happy, interested	1
PAD	SAM's pleasure, arousal and dominance [73]	1
V	valence	1
VA	valence, arousal	11
VA-ACI	VA + agreement, content, interest	1
VA-ECE	VA + engagement, coordinated engagement	1
VA-I	VA + impact	1
VA-LD	VA + liking, disliking	1
WoE-8	(Plutchik Wheel of Emotions [71]) anger, disgust, fear, joy, surprise, sadness, trust, anticipation	1
WoE-8N-17	WoE-8 + neutral + love, alarm, remorse, contempt, optimism, curiosity, pride, guilt, envy, disappointment, pessimism, anxiety, shy, puzzled, hesitant, sympathetic, schadenfreude	1
End of Table		

Title	Targeted Affective	ARS Type	ARS	Nr of Ratings	What measure do they use to measure interrater agreement?	Level of agreement	Year
VAD: A Video Affective Dataset with Danmu	Emotion	Categorical + Dimensional	2WoE-8H	total: 21, between 3-5/video	Fleiss Kappa, ICC, percent cosensus	Moderately High: Fleiss - between 0.561 and 0.804, mean = 0.63 , median = 0.59|	2024
VEATIC: Video-based Emotion and Affect Tracking in Context Dataset	Emotion	Dimensional	VA	total: 192, between 25-73/video	Pearson Correlation	Only one annotator had corellation lower than 0.2 Standard deviation: valance:	2024
Affective film dataset from India (AFDI): creation and validation with an Indian sample	Emotion	Categorical + Dimensional	3EBS-6H-LF-C	407, 271, two stage process	-	-	2023
CH-MEAD: A Chinese Multimodal Conversational Emotion Analysis Dataset with Fine-Grained	Emotion + Attitude	Categorical	WoE-8N-17	3/video	Fleiss' Kappa	0.79	2023
Dyadic Affect in Parent-Child Multimodal Interaction: Introducing the DAMI-P2C Dataset	Emotion	Dimensional	VA-ECE	total: 5, 3/video	ICC	Good or Excellent	2023
Introducing CALMED: Multimodal Annotated Dataset for Emotion Detection in Children	Emotion	Categorical	4C	total: 8, 2/video	-	-	2023
The Jena Audiovisual Stimuli of Morphed Emotional Pseudospeech (JAVMEPS): A	Emotion	Categorical	EBS-6HN	26	Fleiss' Kappa	0.274-fair	2023
The Reading Everyday Emotion Database (REED): a set of audio-visual recordings of emotions in	Emotion	Categorical	EBS-6HN-5	total: 168, min 11/video	-	-	2023
Werewolf-XL: A Database for Identifying Spontaneous Affect in Large Competitive Group	Emotion	Dimensional	PAD	4	RMSE, Spearman's Rho, Krippendorff's Alpha, CCC	fair	2023
A Multimodal Corpus for Emotion Recognition in Sarcasm	Emotion	Categorical	EBS-6HN-3	7	Fleiss' Kappa	0.595	2022
AVDOS-Affective Video Database Online Study Video database for affective research	Emotion	Dimensional	VA	86	-	-	2022
Persian emotion elicitation film set and signal database	Emotion	Categorical + Dimensional	2EBS-6H	total: 88, min 35/video	-	-	2022
AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups	Emotion + Mood	Dimensional	VA	3	Cronbach's α	Mean 0.98 for valence and 0.96 for arousal | very strong IRA	2021
HEU Emotion: a large-scale database for multimodal emotion recognition in the wild	Emotion	Categorical	EBS-6HN-4	15	-	-	2021
Korean video dataset for emotion recognition in the wild	Emotion	Categorical	EBS-6HN-O	6	-	-	2021
Modeling Emotion in Complex Stories: The Stanford Emotional Narratives Dataset (SEND)	Emotion	Dimensional	V	total: 700, min 20/video	-	-	2021
Selection and validation of emotional videos: Dataset of professional and amateur videos that	Emotion	Categorical	EBS-5H	30	-	-	2021
SEWA DB: A Rich Database for Audio-Visual Emotion and Sentiment Research in the Wild	Emotion + Sentiment	Dimensional	VA-LD	total: 30, 5/culture	-	-	2021
Database of Emotional Videos from Ottawa (DEVO)	Emotion	Dimensional	VA-I	278	-	-	2020
K-EmoCon, a multimodal sensor dataset for continuous emotion recognition in naturalistic	Emotion	Categorical + Dimensional	2EBS-5EMOCON	5	Krippendorff 's Alpha	low	2020
MEISD: A Multimodal Multi-Label Emotion, Intensity and Sentiment Dialogue Dataset for	Emotion + Sentiment	Categorical	EBS-6-S-N2	4	Fleiss' Kappa	0.67 emotions, 0.75 sentiment, 0.72 intensity	2020
MELD: A multimodal multi-party dataset for emotion recognition in conversations	Emotion + Sentiment	Categorical	EBS-6-S	3	Fleiss' Kappa	0.43, 0.91 for sentiment	2020
MEmoR: A Dataset for Multimodal Emotion Reasoning in Videos	Emotion	Categorical	WoE-8	6	- `	-	2020
ElderReact: A multimodal dataset for recognizing emotional response in aging adults	Emotion	Categorical + Dimensional	1EBS-6H	3	ICC, FINN, Krippendorff's Alpha	low	2019
Multimodal Database of Emotional Speech, Video and Gestures	Emotion	Categorical	EBS-6	12	-	-	2019
Recognizing behavioral factors while driving: A real-world multimodal corpus to monitor the	Emotion	Categorical	3SN-SE	min 3/video	Krippendorff 's Alpha	0.27	2019
The audio-visual Arabic dataset for natural emotions	Emotion	Categorical	EBS-5SEN	3	-	-	2019
Construction of spontaneous emotion corpus from Indonesian TV talk shows and its	Emotion	Dimensional	VA	3	Pearson Correlation	Valence: 0.33, Arousal: 0.37	2018
RAMAS: Russian Multimodal Corpus of Dyadic Interaction for Affective Computing	Emotion	Categorical	EBS-6HN	total: 21, min 5/video	Krippendorff 's Alpha	0.44	2018
The OMG-Emotion Behavior Dataset	Emotion	Categorical + Dimensional	2EBS-6HN	5/video	-	-	2018
The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic,	Emotion	Categorical	EBS-6HN-2I	319	Fleiss' Kappa	substantial, 0.665	2018
BAUM-1: A Spontaneous Audio-Visual Face Database of Affective and Mental States	Emotion + Mood	Categorical	EBS-6HN-MMS	5	Cohen's Kappa	substantial, 0.67	2017
BioVid Emo DB': A multimodal database for emotion analyses validated by subjective ratings	Emotion	Categorical + Dimensional	2EBS-4A	94	-	-	2017
CHEAVD: a Chinese natural emotional audio–visual database	Emotion	Categorical	EBS-6HN-15-4	4	Cohen's kappa	0.485	2017
COGNIMUSE: a multimodal video database annotated with saliency, events, semantics and	Emotion	Dimensional	VA	7	Pearson correlation, Krippendorff alpha, Cohen's k	low	2017
MSP-IMPROV: An Acted Corpus of Dyadic Interactions to Study Emotion Perception	Emotion	Categorical	3SN	min 5/video	Fleiss' Kappa	0.487	2017
NAA: A multimodal database of negative affect and aggression	Emotion	Dimensional	3AF	total: 15, 3/dimension	Krippendorff 's alpha	Mean for: Actors: 0.72, Students: 0.308	2017
NNIME: The NTHU-NTUA Chinese interactive multimodal emotion corpus	Emotion	Categorical + Dimensional	1EBS-4HSS-FI	48	percentage of the same categorical choice as the metric.(For categories) | Spearman and CCC for continuous	good	2017
A database for emotional interactions of the elderly	Emotion + Mood	Categorical	EBS-4HN-2	18	-	-	2016
Construction of Japanese audio-visual emotion database and its application in emotion	Emotion	Dimensional	VA	1	-	-	2016
Crowdsourcing empathetic intelligence: The case of the annotation of EMMA database for	Emotion	Dimensional	VA	Min 10/video	Krippendorff 's alpha, Avg. Pearson Corr	Valence: 0,45/0.21+-0.20 Arousal: 0.44/ 0.14+-0.18	2016
Emo react: A multimodal approach and dataset for recognizing emotional responses in children	Emotion	Categorical + Dimensional	1EBS-6HN-9	total: 6, 3/video	Krippendorff 's alpha	0.511	2016
MuDERI: Multimodal database for emotion recognition among intellectually disabled	Emotion	Dimensional	VA	5	-	-	2016
BAUM-2: a multilingual audio-visual affective face database	Emotion	Categorical	EBS-6HN-CI	between 5-7/video	Cohen's kappa	Moderate - 0.55	2015

LIRIS-ACCEDE: A video database for affective content analysis	Emotion	Dimensional	VA	1517 for Valence, 2442 for arousal	Percent agreement, fleiss kappa. Krippendorff alpha,	The percent agreement indicates that annotators agreed on 83.5% and 86.2% of	2015
A new multi-modal dataset for human affect analysis	Emotion	Dimensional	VA-ACI	3	Cronbach's α	Arousal: 0.89, Valence: 0.66, Agreement: 0.63, Content: 0.60, Interest: 0.83	2014
Building a Chinese Natural Emotional Audio-Visual Database (CASIA)	Emotion	Categorical	EBS-6HN-15-2	3	-	-	2014
CREMA-D: Crowd-sourced emotional multimodal actors dataset	Emotion	Categorical	EBS-5HNI	total: 2443, min 8/video	Krippendorff's alpha because it can handle categorical responses (selected emotion label) as well as ratio responses	0.42	2014
A multimodal emotion corpus for Filipino and its uses	Emotion	Dimensional	VA	20	Fleiss' Kappa	0.6	2013
Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions	Emotion	Dimensional	VA	6	Cronbach's Alpha, Cohen's Kappa	Cronbach: Arousal: 0.8 Valence:0.74 Cohen: mean 0.244	2013
A multimodal database for affect recognition and implicit tagging	Emotion	Categorical + Dimensional	3EBS-6NAA	20	Cohen's Kappa	fair - 0.32	2012
3D Corpus of spontaneous complex mental states	Emotion + Mental	Categorical	EMS	77	Fleiss' Kappa	moderate - 0.45	2011
The COST 2102 Italian Audio and Video Emotional Database	Emotion	Categorical	EBS-6HSI	5	Not mentioned	0.5 approx	2009
The new italian audio and video emotional database	Emotion	Categorical	EBS-6HSI	2	-		2009
IEMOCAP: Interactive emotional dyadic motion capture database	Emotion	Categorical + Dimensional	3-3SNF	total: 6, 3/video + 2 for VAD	Fleiss' Kappa for categorical, cronbach alpha for vad between 2 raters	fleiss: 0.4, cronbach: Valence:0.809 Arousal:0.607 Dominance:0.608	2008

C Use of Large Language Models

ChatGPT was used through out this research paper to aid in the writing process and the development of the Latex document. In terms of writing, ChatGPT was asked mostly to provide rephrasing options or synonyms. For Latex, it was used to transform Excel data to the Latex format or to help in finding specific commands.

Example prompt: “How to color text in latex?”

Answer:
In LaTeX, you can color the font using the xcolor package. Here are the steps to color the font:

Include the xcolor package in the preamble:
Add xcolor to the preamble of your LaTeX document.

Use the ommand:
Use the text command to color specific text.

Here’s an example of a simple LaTeX document that demonstrates how to color text: ...