Sender Context in Contemporary Emotion Recognition Systems and Databases: a Systematic Review

Exploring the use of Personal Details for Recognizing Emotions

Christo Vasilev[1]
Supervisor: Sayak Mukherjee[1]
Responsible Professor: 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 21, 2026

Name of the student: Christo Vasilev
Final project course: CSE3000 Research Project
Thesis committee: Bernd Dudzik, Sayak Mukherjee, Stephanie Tan

An electronic version of this thesis is available at http://repository.tudelft.nl/.

Abstract

Recent studies on emotion recognition have extensively recorded the unreliability of recognizing emotions using only facial features of a target expressor of emotion. As a response, context-aware emotion recognition (CAER) systems have become more prevalent in contemporary emotion recognition research. This systematic review focuses on context related to the expressors of emotions, also known as senders. Such context includes age, culture and personality. More specifically this study examines how this context is represented in contemporary emotion-recognition datasets and how well it is integrated into CAER systems. Studies were collected from different literature databases and filtered using a hybrid AI-assisted and manual screening process. The results were limited to only papers from 2019 to prevent overlap with previous studies in this field. Only 8 such papers were found and included in the study. From these results a clear gap was found between the sender-context information available in datasets and its exploitation in recognition systems. Additionally, fairness was identified as a blind-spot in both databases and systems.

1 Introduction

Recognizing emotions from facial expressions alone is an unreliable task: people systematically disagree on what emotions faces express. Consequently facial emotion recognition systems trained on such inconsistent labels cannot achieve meaningful accuracy [5]. Context is necessary to make emotion perception reliable: visual scenes, voices, bodies, cultural background and many other non-facial features systematically shape how emotions are perceived [2]. This contextual information can be organized into three sources: knowledge about the person expressing the emotion also known as *sender context*, information about the surrounding physical and social environment known as *situation context*, and knowledge about the observer of the person expressing the emotion also known as *perceiver context* [16].

Several sender context types like age, gender and culture have been identified by literature in the area of human psychology as relevant for emotion recognition done by humans and for technology to succeed in the same task, it must mimic this human ability. Additionally, there exists research aiming to collect context like age or gender of people automatically from audiovisual data. Therefore it follows that sender context should be systematically recorded and made available in audiovisual CAER datasets. Despite this evidence, according to previous reviews, individual datasets don't specifically account for sender context through systematic variation [6].

This systematic review aims to observe what changes if any there are in the way sender context is captured in datasets and to map the way CAER systems use it for mimicking human emotional intelligence and what implications it has on generalization and fairness. While psychology has thoroughly investigated contextual influences on human emotion perception, this review adopts a computer science perspective, focusing on how machine learning systems can be designed and improved to leverage sender context for more robust and equitable emotion recognition.

Section 2 surveys existing systematic reviews and positions this work relative to them. Section 3 formalizes the research objectives. Section 4 describes the systematic review methodology, Section 5 presents the extracted findings organized around the two research questions, Section 6 interprets these findings and identifies open challenges, Section 7 addresses responsible research considerations, and Section 8 concludes with directions for future work.

2 Related Works

Dudzik et al. [6] previously examined how sender, situation, and perceiver context are made available in emotion recognition datasets, identifying which datasets expose sender demographic information. This systematic review builds on that work by extending the scope to literature published after 2019, acting as a continuation of the landscape survey. Additionally, while Dudzik focused on datasets, this review also analyzes how contemporary machine learning systems exploit sender context, examining the gap between the contextual richness of datasets and contextual awareness of systems.

Other reviews exist which explore context in affective datasets. Al-Azani and El-Alfy [**?**] reviewed the state of emotional AI datasets since the start of the century. One of their contributions was identifying the importance of participant diversity across gender, age, and ethnicity, however they didn't report what other sender context is recorded in the datasets and how sender context is distributed in the data in general.

Wang et al. [15] provide a broad systematic survey of affective computing covering emotion models, databases across five modality categories and over 380 papers on unimodal and multimodal recognition methods. While comprehensive in scope, the review does not treat sender demographic characteristics, such as age, gender, or cultural background, as a variable of interest in either dataset design or system behaviour, and therefore does not address whether systems exploit or are affected by such information.

3 Research Objective

The objective of this review is to systematically examine how sender context is made available in emotion recognition datasets and exploited by contemporary machine learning systems. This investigation is structured around two primary research questions, each with two sub-questions that together provide a comprehensive view of sender context in the emotion recognition landscape.

RQ1: What sender context is captured in existing datasets? This question examines the current state of emotion recognition datasets with respect to sender characteristics. It includes two sub-questions: *RQ1a: Are senders diverse across the datasets?* addresses whether there is variation in sender populations across demographic groups, avoiding narrow representations. *RQ1b: Are senders balanced across datasets?* investigates whether senders are represented in similar proportions, or whether certain demographic groups are overrepresented.

RQ2: How is sender context implemented in CAER systems? This question focuses on the way the datasets are exploited to create systems with accurate abilities for recognizing emotions and what role sender context plays in them. *RQ2a: Does sender context impact generalization?* asks whether systems incorporating sender information can recognize emotions of unseen senders more accurately if they are more informed about the sender's background. *RQ2b: Does sender context improve bias mitigation?* examines what bias-mitigation strategies the recognition systems employ, if any, and whether leveraging sender context reduces demographic biases or amplifies them.

Together, these research questions map the landscape of sender context from two complementary perspectives to identify gaps, opportunities, and implications for fairness and generalization.

4 Methodology

This research project employs a Systematic Literature Review (SLR) methodology adopting a ten-step process outlined by Boland et al. [3]. It comprises: (1) planning the review, (2) performing scoping searches and identifying the review question, (3) conducting literature searches, (4) screening titles and abstracts, (5) obtaining full-text papers, (6) applying selection criteria to select relevant studies, (7) extracting data from selected papers, (8) assessing study quality, (9) analyzing and synthesizing findings, and (10) writing and disseminating results. Additionally, this review adheres to the PRISMA 2020 reporting framework [13].

This section details the implementation of the systematic review process, documenting this review's approach to identifying, screening, and analyzing relevant literature on contextual cues in audio-visual emotion recognition. Section 4.1 explains the selection of digital libraries queried for this review. Section 4.2 then describes the search strategy, which defines the key concepts and selection criteria. The screening process is described in Section 4.3, comprising initial automated screening (Section 4.3.1) followed by manual validation (Section 4.3.2). Finally, Section 4.5 outlines the data extraction methodology, which captures information about the contextual cues employed in emotion recognition systems and datasets, enabling synthesis of patterns across the included studies.

4.1 Database Selection

Three digital libraries were selected for this review: Scopus, IEEE Xplore, and the ACM Digital Library. Scopus and IEEE Xplore are some of the databases recommended by the TU Delft Library for comprehensive coverage of computer science and engineering literature. Scopus was selected as the primary search engine for its broad cross-disciplinary indexing, while IEEE Xplore was included to ensure comprehensive coverage of computer vision and machine learning research. The ACM Digital Library was selected because it is recognized as a database including literature with a more specific focus on Affective Computing [6]. Web of Science was also considered as a data source but it wasn't used for this review because of the lack of access to it through the TU Delft Library.

4.2 Search Strategy

The search strategy comprised two components: defining selection criteria to determine which studies qualify for inclusion, and identifying key concepts with associated synonyms to construct the database queries.

4.2.1 Selection Criteria

A predefined set of exclusion criteria were created to select only meaningful studies (see Table 1). Exclusion criterion E1 requires a study to introduce a dataset or propose a recognition system, ensuring the review captures original artifacts rather than position papers or surveys. E2 requires the use of sender context such as age, gender, culture, or personality, as this is the core variable of interest for the review. E3 limits the literature pool to only systems and datasets which record or use during inference either visual or audio data. E4 and E5 restrict the publication window to 2019 or later and before 20 April 2026, focusing on recent advances while ensuring all included work was available for full-text retrieval before the start of the review. E6 restricts the language to English for accessibility. E7 also makes sure the literature is accessible by removing papers whose full text is inaccessible via

university access, as detailed analysis requires the full paper. E8 and E9 limit the corpus to peer-reviewed journal articles and conference papers, ensuring the included work meets academic quality standards and excludes books and book chapters due to the limited time available for the review.

Table 1: Exclusion Criteria for Literature Review

ID	Criteria	Motivation
E1	Doesn't introduce a new emotion dataset or an emotion recognition system	Ensures study introduces an original system or database.
E2	Doesn't use sender context (e.g., age, gender, culture, personality)	Core requirement for this study's focus
E3	Doesn't record or use for inference audio and video data	Grounds study in audio-visual modalities
E4	Published before 2019	Captures recent advances in the field
E5	Published after 20 April 2026	Ensures publication is before review start
E6	Not published in English	Ensures accessibility for review
E7	Full text not accessible via university access	Ensures feasibility of detailed analysis
E8	Not peer-reviewed	Maintains research quality standards
E9	Not a journal article or conference paper	Focuses on formally published research

4.2.2 Key Concepts

Emotion-Recognition was chosen as the primary concept because it captures the core activity of interest, namely systems and datasets that recognize, detect, infer, or predict emotional states, across the varied terminology used in different disciplines. **Sender** was included to ensure studies contain explicit information about who expresses the emotion. **Context** additionally focuses on studies that incorporate contextual cues, which is needed for analyzing how the expressor context influences recognition. Finally, **Recognizer/Dataset** grounds the review in concrete, reproducible artifacts, either implemented systems or annotated datasets, rather than purely theoretical work. For each concept, multiple synonyms were identified to maximize search recall across different terminologies used in the literature (see Table 2).

For each database, the search query was adjusted according to its syntax requirements and recorded. The date of consultation and amount of results were also recorded to ensure reproducibility of the search. See Table 5 for all search queries.

4.3 Screening Process

Given the time constraints of the review, a two-stage screening process was adopted: an initial AI-assisted stage to efficiently reduce the candidate pool, followed by a manual screening

Table 2: Key Concepts and Synonyms Used in Database Queries

Emotion-Recognition	**Sender**	**Context**	**Recognizer/ Dataset**
emotion	sender	personal context	recogniser
emotional	expresser	individual context	recognizer
affect	expressor	sender context	detector
affective	poser	contextual cue	machine learning
recognition	experiencer	contextual information	predictor
detection	participant	social context	database
inference	speaker	cultural context	dataset
prediction			

stage to validate the results. The full flow of records is summarised in Figure 1.

4.3.1 AI Screening

Each record was evaluated by a custom Python script using the Anthropic API with the Claude Haiku 4.5 model [1]. For each paper, the script submitted the title, abstract, publication year, and keywords to the model, together with a structured system prompt listing the nine exclusion criteria (see Appendix B.1). The model was instructed to return a JSON object assigning `y` (passes), `n` (fails), or `?` (uncertain) for every selection criteria. Based on the criteria results it returned an overall decision of `INCLUDE`, `EXCLUDE`, or `MAYBE` and a one-sentence explanation. The decision rules were: `EXCLUDE` if the model was certain the paper met at least one exclusion criterion; `INCLUDE` if none of the criteria were met; and `MAYBE` in any case of uncertainty. All records marked as `INCLUDE` or `MAYBE` were carried forward to the manual screening stage.

4.3.2 Manual Screening

To guard against hallucinations and automated misclassifications, all records marked as `INCLUDE` or `MAYBE` by the AI screening stage were carried forward to manual review. In addition, a random 5% sample of `EXCLUDE` records (with fixed seed for reproducibility) was included to estimate the false-exclusion rate. For each record in this queue, a single reviewer assessed the title and abstract against the nine selection criteria. Where the title and abstract alone were insufficient to reach a confident decision, the full text was retrieved and consulted before assigning a final verdict of `INCLUDE`, `EXCLUDE`, or `MAYBE`.

4.4 Search Results

The database searches yielded 75 records from Scopus, 78 from IEEE Xplore, and 63 from the ACM Digital Library, giving 216 results in total and 171 unique records after deduplication. They were passed to the AI screening stage. The model classified 8 records as `INCLUDE`, 8 as `MAYBE`, and 155 as `EXCLUDE`. All 16 `INCLUDE` and `MAYBE` records, together with a 5% random sample of `EXCLUDE` records, were carried forward to manual screening, yielding a manual review stack of 24. Following manual assessment, 8 studies were confirmed as eligible and their full texts were retrieved for data extraction. Of those 8, the AI marked 5 as `INCLUDE`, 3 as `MAYBE` and none as `EXCLUDE`

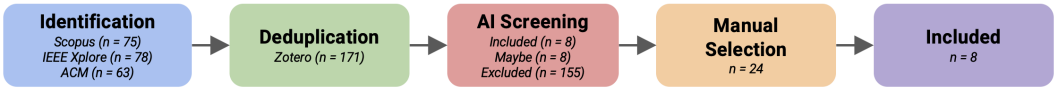

Figure 1: Flow of records through the selection process.

4.5 Extraction Method

For each included dataset the goal was to categorize the sender-context information into discrete labels which appear most commonly in the recovered datasets. Additionally, the diversity and distribution of these characteristics within the dataset will be analyzed where available, allowing an assessment of representation balance and potential fairness concerns.

For each included system the type of sender-context used during inference was extracted. For each system also the performance of the system when using sender context and the performance of the system without sender context was extracted, if available. Additionally, any bias-mitigation strategies of senders were extracted where available. Finally, systems were classified according to a predefined set of usage strategies, such as: at inference feature inclusion, conditioning mechanisms, multimodal fusion, post-hoc adjustment or hybrid.

5 Results

The systematic search yielded 8 included records spanning 2019 to 2026. Of these, 1 contributes a dataset with sender context labels, 4 propose a recognition system using sender data and 3 do both. This results in 4 dataset contributions and 7 system contributions in total. From the following subsections Section 5.1 presents the included datasets and Section 5.2 structures assessments of diversity and balance in the datasets, providing information for answering *RQ1a* and *RQ1b*. Section 5.3 presents the included systems and Section 5.4 structures generalizability and bias mitigation in systems to help answer *RQ2a* and *RQ2b*.

5.1 Included Datasets

Table 3 summarises the 4 papers that contribute a labelled dataset. Each was screened for whether sender characteristics are recorded alongside the audio-visual data. All context types present in the dataset are indicated in the table columns.

Paper / Dataset	Personality	Age	Gender	Culture	Mood	Health	Experience
MultiEMP [11]		✔	✔	✔			
Indian-AV [8]		✔	✔	✔			
BIRAFFE2 [9]	✔	✔	✔				✔
AMIGOS [7]	✔	✔	✔		✔	✔	

Table 3: Sender characteristics recorded in included datasets. ✔ = characteristic recorded; blank = not present.

5.2 Dataset Diversity and Balance

Demographic diversity is somewhat varied across the four included datasets. Figure 2 contrasts the diversity in 2 sender context reported by almost all dataset studies.

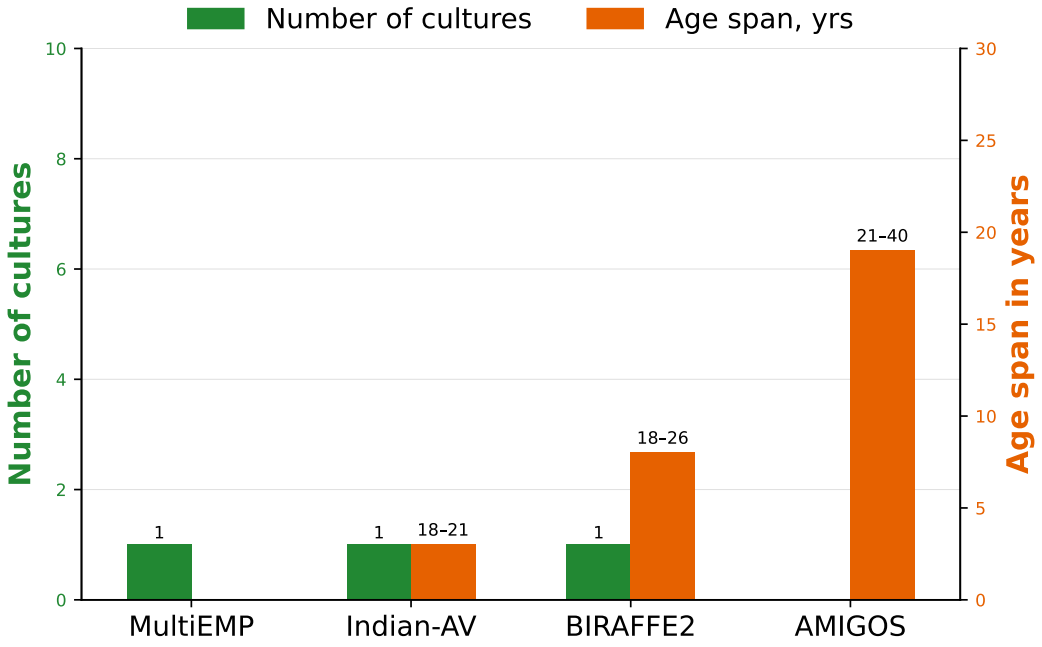

Figure 2: Cultural and age diversity of the included datasets for which this information is reported: *MultiEMP* [11], *Indian-AV* [8], *BIRAFFE2* [9] and *AMIGOS* [7]. Green bars (left axis) give the number of cultures represented, counted at the national level; orange bars (right axis) give the age span (max−min years), with the actual range annotated on top.

From the results no dataset recording cultural data from senders represents more than one culture. Contrastingly, ages are better represented in some datasets with *AMIGOS* [7] recording the biggest variety of sender ages. It also records personality using the Big-Five format but doesn't divide the senders personalities in discrete groups. Its gender balance is skewed in favour of more male participants: 27 compared to only 13 female. *Multi-EMP* [11] achieves gender balance through matched dyads, yet is restricted to Korean young adults, and *Indian-AV* [8] is similarly constrained to an 18–21 age range with a male-skewed distribution. *BIRAFFE2* [9] is limited to a small sample of young Polish university students with similar background, limiting cultural generalizability.

5.3 Included Recognition Systems

Table 4 summarises the 7 papers that propose or evaluate an emotion-recognition system using sender context at inference time. Each was screened for whether sender characteristics are used at inference time. All context types actively incorporated by the system are indicated in the table columns.

Paper	Personality	Gender	Culture	Voice	Mood
PGIF [17]	✔				
Persona-CTG [14]	✔				
Intercultural [10]			✔		
Meta-SER [18]		✔		✔	
AMIGOS [7]	✔				✔
BIRAFFE2 [9]	✔				
MultiEMP [11]				✔	

Table 4: Sender-context usage in included recognition systems. ✔ = context used at inference.

5.4 System Generalizability and Bias Mitigation

Generalizability strategies differ substantially across the seven included systems. Figure 3 shows the performance achieved by incorporating sender context for systems that report a direct numerical comparison against a baseline without sender context. *Meta-SER* [18] validates across three encoders under a speaker-independent setup, establishing state-of-the-art results, yet the metadata auxiliary tasks (gender, speaker identity, speech style) are sourced exclusively from IEMOCAP, so it is unknown whether the performance gains hold on corpora that lack such metadata annotations. *PGIF* [17] performs better than all 15 baselines selected in the paper for comparing its performance on the IEMOCAP and MELD benchmarks. However, evaluation remains within English-language acted corpora. *Persona-CTG* [14] is one of the few that explicitly target cross-cultural generalizability: the latter applies temporal causal feature selection to identify culture-invariant audio-visual cues.

None of the included systems report explicit bias-mitigation procedures such as fairness-aware training objectives or demographic parity constraints.

6 Discussion

This section interprets the findings from Section 5 in relation to the research questions. Section 6.1 discusses the sender context captured across the included datasets and addresses RQ1, with RQ1a (diversity) and RQ1b (balance) treated in separate subsections. Section 6.2 examines how the included systems make use of sender context, with RQ2a split into context usage and generalization subsections, and evaluates bias mitigation practices in relation to RQ2b. Section 6.3 acknowledges the limitations of this review.

6.1 Sender Context in Datasets (RQ1)

Across the four included datasets, the most consistently recorded sender characteristics are age, gender, personality, and culture. Gender and age are captured by all four, reflecting their straightforward operationalization as demographic metadata. Personality, primarily in the form of the Big Five traits, appears in two datasets [7, 9], also culture is recorded in two corpora [8, 11]. This distribution aligns with prior literature establishing that age, culture,

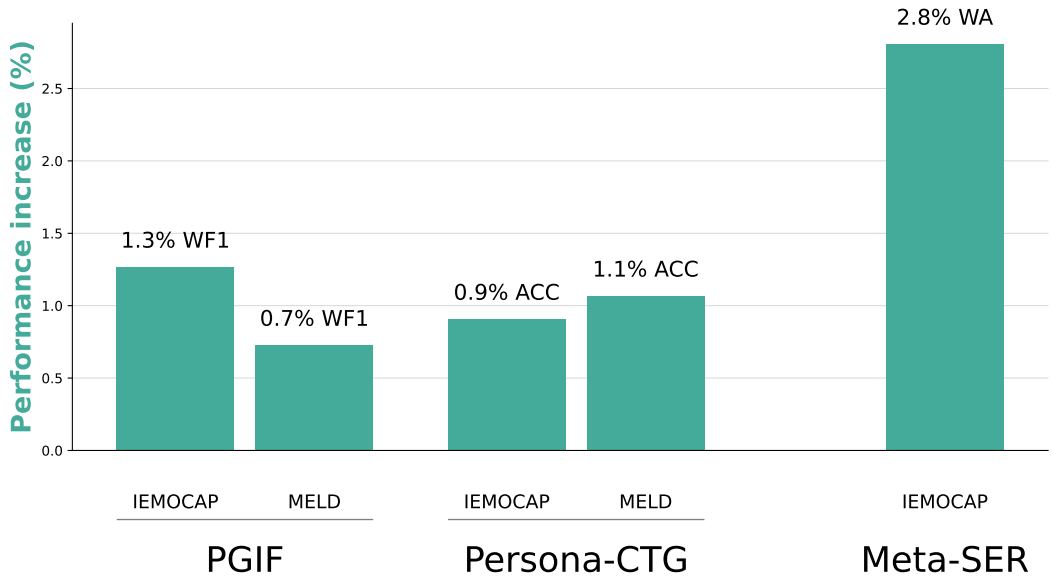

Figure 3: Performance improvement from incorporating sender context in percent, measured against the best reported baseline without sender context reported in each study. Systems that report a direct numerical comparison: *PGIF* [17], *Persona-CTG* [14] and *Meta-SER* [18]. Systems with multiple evaluations show one bar per benchmark and benchmark names are shown at the bottom of bars. System names are bellow them, grouping bars that belong to the same paper. *WF1*: weighted F1-score; *ACC*: accuracy; *WA*: weighted accuracy; *UA*: unweighted accuracy.

and gender each exert measurable influence on emotional expression and recognition, giving theoretical motivation for their inclusion as sender context. Similar high frequency results in some contexts were observed in previous dataset reviews [6].

6.1.1 RQ1a: Diversity

Across the included datasets, some sender domains are systematically lacking in diversity. No study spans multiple national or cultural groups with a significant amount of participants, while two [11, 8] are each restricted to a single nationality and a similar young-adult age range. Another dataset [9] draws exclusively from Western university populations, limiting their cultural and demographic breadth despite capturing multiple sender contexts. One additional dataset [7] covers a wide age range but records no cultural embedding and therefore there is no way of measuring its cultural diversity. Overall, no dataset comprehensively captures the full spectrum of sender characteristics identified as relevant. No corpus has senders with diverse cultures and few studies prioritize a single context to be systematically varied.

6.1.2 RQ1b: Balance

The balance of participants across context domains is problematic in nearly all included datasets. Two of the datasets [9, 7] are systematically imbalanced towards including male

participants, both having more than twice as many male participants, as female ones. Another dataset [8] has a comparably low skew towards male participants but is limited to participants from only 3 different ages, so it cannot achieve meaningful variety with balance in that domain. Only one dataset [11] reports almost equal gender representation with only 2 more male participants than female, but is lacking in age representation, including mostly people younger than 24.

Taken together, the evidence suggests that existing datasets tend to optimise one aspect of representativeness at the expense of the other, and that genuinely diverse and balanced datasets, covering wide ranges of sender context remain few.

6.2 Sender Context in Emotion-Recognition Systems (RQ2)

The seven included systems collectively demonstrate that a variety of sender contexts can be integrated into emotion recognition: four systems [17, 14, 9, 7] use personality, one [10] uses culture and two [18, 11] use speech-level metadata. However, no single system simultaneously leverages a broad combination of these context types. Each system focuses on at most two categories. This fragmentation suggests that the field has not yet focused on systems capable of jointly exploiting the full range of available sender information.

6.2.1 RQ2a: Generalization

Several systems that incorporate sender context achieve competitive performance on unseen senders or datasets. Four systems [18, 17, 14, 10] achieve competitive or state-of-the-art performance when evaluated on senders, unseen during training, or on cultural groups not represented in the training data. This indicates that sender context can be a positive factor for generalization. Despite the quality of the generalisation of the systems, the evidence is limited to a small number of studies, tested on a narrow set of datasets, emotion categories, and elicitation conditions. Therefore it is unclear whether the findings hold more broadly or found ways to exploit the benchmarks.

6.2.2 RQ2b: Bias Mitigation

None of the included systems report explicit bias-mitigation procedures like: training with fairness-aware penalties or enforcing equal prediction rates across subgroups. While some systems incorporate sender attributes as auxiliary signals to improve accuracy, none evaluate whether doing so amplifies or reduces performance disparities across demographic subgroups. This represents a clear gap in the implementations: the integration of sender context into emotion recognition systems has so far been driven by performance gains rather than fairness considerations.

6.3 Review Limitations

Several limitations of this review should be acknowledged: (1) The search was restricted to three digital libraries (Scopus, IEEE Xplore, ACM Digital Library) and relevant work published exclusively in other repositories may have been missed. (2) Although all records classified as `INCLUDE` or `MAYBE` by the AI screener were manually verified, including 5% of records classified as `EXCLUDE`, the remaining excluded papers received only a single model pass. Papers close to being included may therefore have been incorrectly placed in the `EXCLUDE` pool. This risk is only partially mitigated by the random audit. (3) The review

covers only eight included records, a consequence of the specificity of the exclusion criteria. This small corpus limits the statistical generalizability of any cross-study patterns observed. (4) All characterisations of the included datasets and systems are based solely on what is reported in the corresponding papers; the underlying artifacts (dataset files, source code, model weights) were not directly inspected. Claims about demographic composition, balance, or system behaviour therefore reflect the authors' own descriptions and may not fully capture the properties of the datasets and models themselves. (5) Data extraction was performed by a single reviewer within the time limitation of 10 weeks, without independent verification, introducing risks of subjective interpretation or human errors during the extraction of data from some literature.

7 Responsible Research

This review operates at the intersection of two responsibilities: conducting a fair, transparent synthesis of a high-risk AI domain and ensuring that the synthesis itself is reproducible and honestly reported.

7.1 Ethical Considerations

The subject of this review is itself ethically sensitive. Emotion recognition is classified as a high-risk application under the EU AI Act, particularly in domains such as healthcare, education, and the workplace [12]. The central motivation of this work is fairness: systems trained on demographically skewed data have been shown to fail for anyone who deviates from the majority group [4]. By surveying how sender attributes such as age, culture, and personality are represented and used, this review aims to make existing imbalances visible rather than to enable finer-grained profiling of individuals. The resulting recommendations from this paper are framed toward auditing coverage and improving generalisation, not toward maximising predictive power over protected attributes. As a secondary-research study working only with published, peer-reviewed literature, this review did not collect data from human participants and required no ethics approval. The ethical responsibility it carries lies in how the primary studies it synthesises are represented and how their limitations are reported.

7.2 Reproducibility

The review is designed to be reproducible in line with the PRISMA 2020 reporting standard [13]. The search strategy is reported in full: the concept clusters, the Boolean query template, the per-database adaptations, the date of each consultation and the number of results. Following it, the selection process is documented end-to-end in the flow of Figure 1, from identification through screening to inclusion.

One stage was approached with particular caution — the initial AI-assisted screening. Large Language Models can misclassify records and reproduce training biases. To mitigate this the AI was strictly instructed to only use titles, abstracts and keywords to give ratings on each selection criteria. Additionally, the AI selections were tested by deliberately screening some excluded studies. None of the results screened this way passed the selection criteria during manual screening, proving partial accuracy of the system in applying the selection criteria. To support reproducibility, the prompts used during the AI-reviewing stage are made available (see B.1).

8 Conclusions and Future Work

This systematic review examined how sender context is captured in datasets and exploited by systems for audio-visual emotion recognition. Across the included work, a wide array of sender contexts including personality, culture, age, gender, mood and speech-level metadata, have been incorporated into recognition systems. It was identified that systems that leverage such context generally demonstrate competitive generalisation to unseen speakers and cultural groups. This suggests that sender context is a viable and beneficial source for building more robust emotion recognition models.

At the same time, two critical gaps remain: (1) The included datasets rarely achieve diversity and balance simultaneously across more than one sender-context dimension: datasets in general lack cultural variety; those which are diverse with respect to age tend to exhibit an imbalanced gender representation, while those which balance genders, lack diversity in ages. Future work should prioritise the collection of datasets that are representative across several sender characteristics at once, rather than optimising a single dimension in isolation. (2) None of the included systems evaluate or address bias. Although sender attributes are used to improve accuracy, no system examines whether performance differences arise across demographic subgroups and no systems implement fairness-aware training.

Closing these gaps requires annotating cross-cultural audiovisual data, aiming for cross-generational age ranges and balanced gender representations. Additionally, future work should focus on bias mitigation for emotion recognition, integrating fairness constraints and demographic-stratified evaluation as standard practice alongside accuracy reporting. These directions can produce fairer and safer CAER systems.

A Search Queries

DB	Query	Date	N	Description
Scopus	TITLE-ABS-KEY (((emotion OR emotional OR affect OR affective) W/1 (recognition OR detection OR inference OR prediction)) AND (context OR contextual OR metadata OR background OR personality) AND (model OR detector OR "machine learning" OR predictor OR database OR dataset OR infoset) AND (sender OR expresser OR expressor OR subject OR actor OR target OR poser OR experiencer)) AND PUBYEAR > 2018	7.6.2026	671	Wide net; broad sender terms (subject, actor, target) and context terms (background, personality) introduce noise.
Scopus	TITLE-ABS-KEY (((emotion OR emotional OR affect OR affective) W/1 (recognition OR detection OR inference OR prediction)) AND (sender OR expresser OR expressor OR poser OR experiencer) AND (context OR contextual) AND (recogniser OR detector OR "machine learning" OR predictor OR database OR dataset OR infoset)) AND PUBYEAR > 2018	7.6.2026	6	Over-restricted; only rare domain jargon retained for senders—almost nothing matches.

DB	Query	Date	N	Description
Scopus	TITLE-ABS-KEY (((emotion OR emotional OR affect OR affective) W/1 (recognition OR detection OR inference OR prediction)) AND (sender OR expresser OR expressor OR poser OR experiencer OR participant OR individual OR speaker) AND (context OR contextual OR metadata) AND (recogniser OR recognizer OR detector OR "machine learning" OR predictor OR database OR dataset)) AND PUBYEAR > 2018	7.6.2026	749	Recall recovered by adding participant, speaker; but individual and metadata brought noise back.
Scopus	TITLE-ABS-KEY (((emotion OR emotional OR affect OR affective) W/1 (recognition OR detection OR inference OR prediction)) AND (sender OR expresser OR expressor OR poser OR experiencer OR participant OR speaker) AND (context OR contextual) AND (recogniser OR recognizer OR detector OR "machine learning" OR predictor OR database OR dataset)) AND PUBYEAR > 2018	7.6.2026	449	Pruned individual and metadata; still, bare "context" matches any domain.
Scopus	TITLE-ABS-KEY (((emotion OR emotional OR affect OR affective) W/1 (recognition OR detection OR inference OR prediction)) AND (sender OR expresser OR expressor OR poser OR experiencer OR participant OR speaker) AND ("personal context" OR "individual context" OR "sender context" OR "contextual cue" OR "contextual information" OR "social context" OR "cultural context") AND (recogniser OR recognizer OR detector OR "machine learning" OR predictor OR database OR dataset)) AND PUBYEAR > 2018	7.6.2026	75	Context anchored to sender-specific phrases; best precision.
IEEE Xplore	((emotion OR emotional OR affect OR affective) NEAR/1 (recognition OR detection OR inference OR prediction)) AND (sender OR expresser OR expressor OR poser OR experiencer OR participant OR speaker) AND ("personal context" OR "individual context" OR "sender context" OR "contextual cue" OR "contextual information" OR "social context" OR "cultural context") AND (recogniser OR recognizer OR detector OR "machine learning" OR predictor OR database OR dataset)	7.6.2026	78	Scopus query translated for IEEE Xplore: W/1 → NEAR/1, no field tag (All Metadata in UI), date as filter.

DB	Query	Date	N	Description
ACM DL	(Title:(("emotion recognition" OR "emotion detection" OR "emotion inference" OR "emotion prediction" OR "affect recognition" OR "affect detection" OR "affective recognition" OR "affective detection") AND (sender OR expresser OR expressor OR poser OR experiencer OR participant OR speaker) AND ("personal context" OR "individual context" OR "sender context" OR "contextual cue" OR "contextual information" OR "social context" OR "cultural context") AND (recogniser OR recognizer OR detector OR "machine learning" OR predictor OR database OR dataset)) OR Abstract:(("emotion recognition" OR "emotion detection" OR "emotion inference" OR "emotion prediction" OR "affect recognition" OR "affect detection" OR "affective recognition" OR "affective detection") AND (sender OR expresser OR expressor OR poser OR experiencer OR participant OR speaker) AND ("personal context" OR "individual context" OR "sender context" OR "contextual cue" OR "contextual information" OR "social context" OR "cultural context") AND (recogniser OR recognizer OR detector OR "machine learning" OR predictor OR database OR dataset)))	7.6.2026	63	No proximity operator in ACM DL; W/1 replaced by explicit compound phrases. Title and Abstract fields targeted explicitly.

Table 5: Search queries used for each database, with the date of consultation and number of results.

B AI Prompts

B.1 AI Screening System Prompt

The following system prompt was used in the automated screening script (Section 4.3.1):

```
You are a systematic literature review screener.  Your task is to evaluate
academic papers against a set of exclusion criteria based solely on their title,
abstract, and keywords.

For each criterion, you will answer y (yes/passes), n (no/fails), or ?  (uncertain).

Exclusion criteria -- failing any of these EXCLUDES the paper:
  E1:  Doesn't introduce a new emotion dataset or an emotion recognition system
  E2:  Doesn't use sender context (e.g., age, gender, culture, personality)
  E3:  Doesn't record or use for inference audio and video data
  E4:  Published before 2019
  E5:  Published after 20 April 2026
  E6:  Not published in English
  E7:  Full text is inaccessible via university access
  E8:  The paper is not peer-reviewed
  E9:  The paper is not a journal article or conference paper

Decision rules:
- EXCLUDE: you are CERTAIN the paper meets an exclusion criterion
- INCLUDE: the paper clearly meets none of the exclusion criteria
- MAYBE: you are uncertain about any criterion

IMPORTANT NOTE: For E7/E8/E9 answer n (not excluded) if you have no reason
to believe the criterion applies.  Most papers in the dataset are peer-reviewed
```

```
journal or conference papers.

Respond ONLY with a JSON object in this exact format (no markdown, no extra
text):
{
  "E1":  "y|n|?",
  "E2":  "y|n|?",
  "E3":  "y|n|?",
  "E4":  "y|n|?",
  "E5":  "y|n|?",
  "E6":  "y|n|?",
  "E7":  "y|n|?",
  "E8":  "y|n|?",
  "E9":  "y|n|?",
  "decision":  "INCLUDE|EXCLUDE|MAYBE",
  "explanation":  "one concise sentence"
}
```

References

[1] Anthropic. Claude Haiku 4.5, 2025.

[2] Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. Context in emotion perception. 20(5):286–290.

[3] A. Boland, M. Cherry, R. Dickson, and Julia Carden. Doing a systematic review: A student's guide. *International Coaching Psychology Review*, 15:119–120, 09 2020.

[4] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, pages 77–91. PMLR.

[5] Federico Cabitza, Andrea Campagner, and Martina Mattioli. The unbearable (technical) unreliability of automated facial emotion recognition. 9(2):20539517221129549.

[6] Bernd Dudzik, Michel-Pierre Jansen, Franziska Burger, Frank Kaptein, Joost Broekens, Dirk K.J. Heylen, Hayley Hung, Mark A. Neerincx, and Khiet P. Truong. Context in human emotion perception for automatic affect detection: A survey of audiovisual databases. In *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 206–212. IEEE.

[7] J. A. Miranda-Correa, M. K. Abadi, N. Sebe, and I. Patras. AMIGOS: A dataset for affect, personality and mood research on individuals and groups. 12(2):479–493.

[8] R. Karani, V. Harkare, K. Kamath, K. Gupta, O. Shukla, and S. Desai. A multimodal deep learning approach for emotion recognition in a diverse indian cultural context. volume 1264, pages 293–306.

[9] K. Kutt, D. Drążyk, L. Żuchowska, M. Szeląże k, S. Bobek, and G.J. Nalepa. BIRAFFE2, a multimodal dataset for emotion-based personalization in rich affective game environments. 9(1).

[10] L. Mathur, R. Adolphs, and M. J. Matarić. Towards intercultural affect recognition: Audio-visual affect recognition in the wild across six cultures. pages 1–6.

[11] E. Lim, H. Lee, J.-E. Shin, H.-J. Yang, S.-H. Kim, S. Kim, and A. Kim. Multi-modal adaptive empathy assessment in online dyadic interaction using bi-directional multi-layer perceptron-mixer and dynamic weights fusion. 167.

[12] Sayak Mukherjee. Towards context-sensitive emotion recognition. In *Proceedings of the 27th International Conference on Multimodal Interaction*, pages 730–734. Association for Computing Machinery.

[13] Catrin Sohrabi, Thomas Franchi, Ginimol Mathew, Ahmed Kerwan, Maria Nicola, Michelle Griffin, Maliha Agha, and Riaz Agha. PRISMA 2020 statement: What's new and the importance of reporting guidelines. 88:105918.

[14] G. Tu, F. Xiong, B. Liang, and R. Xu. A persona-infused cross-task graph network for multimodal emotion recognition with emotion shift detection in conversations. pages 2266–2270.

[15] Yan Wang, Wei Song, Wei Tao, Antonio Liotta, Dawei Yang, Xinlei Li, Shuyong Gao, Yixuan Sun, Weifeng Ge, Wei Zhang, and Wenqiang Zhang. A systematic review on affective computing: emotion models, databases, and recent advances. 83-84:19–52.

[16] Matthias J. Wieser and Tobias Brosch. Faces in context: A review and systematization of contextual influences on affective face processing. 3.

[17] Y. Xie and R. Mao. PGIF: A personality-guided iterative feedback graph network for multimodal conversational emotion recognition. 12(5):3583–3595.

[18] Z. Wan, Z. Qiu, Y. Liu, and W. -Q. Zhang. Metadata-enhanced speech emotion recognition: Augmented residual integration and co-attention in two-stage fine-tuning. pages 1–5.