



LREC 2026

Computational Affective Science (CAS) @ LREC 2026

Workshop Proceedings

Editors

**Christopher Bagdon, Krishnapriya Vishnubhotla,
Kristen A. Lindquist, Lyle Ungar, Roman Klinger,
Saif M. Mohammad**

16 May 2026

©ELRA Language Resources Association (ELRA), 2026

These proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-68-5

Preface

We are proud to present the first Computational Affective Science (CAS) workshop, a series, and the first dedicated venue, for exploring the intersection of Natural Language Processing (NLP) and Affective Science.

Affect refers to the fundamental neural processes that generate and regulate emotions, moods, and feeling states. Affect and emotions are central to how we organize meaning, to our behavior, to our health and well-being, and to our very survival. Despite this, and even though we are all intimately familiar with emotions in everyday life, there is much we do not know about how emotions work and how they impact our lives. Affective Science is a broad interdisciplinary field that explores these and other related questions about affect and emotions. Since language is a powerful mechanism of emotion expression, there is a growing use of language data and advanced natural language processing (NLP) algorithms to shed light on fundamental questions about emotions. Over the course of the series we hope to explore the many subareas and modalities in which affect is studied.

A common critique of NLP work is the lack of theory. CAS takes an interdisciplinary approach to affective science and aims at bringing together NLP researchers, scientists, and theorists from many research areas, including psychology, sociology, neuroscience, and philosophy. CAS is co-organized by researchers in NLP, Psychology, and Cognitive Science. We foresee this new workshop as an interdisciplinary hub that will foster innovative CAS research.

For this first interaction of the workshop, we invited long and short archival paper submissions, as well as non-archival extended abstracts on a broad range of topics at the intersection of affective science and natural language processing. We received 29 submissions, 19 long papers, 4 short papers, and 6 extended abstracts. Of these, we accepted 23 submissions (79%, plus two additional which withdrew after acceptance due to inability to attend). We are thrilled by the level of participation and quality from submissions, especially with the variety of backgrounds the papers come from. We received 16 submissions from NLP/Computer Science, 6 from Psychology, 4 from Digital Humanities, and 2 from Linguistics backgrounds.

We are excited to award our first Best Paper award to "Beyond Toxic Positivity: Interpersonal Affect Regulation in LLM-Based Dialogue Agents using Discourse Politeness Theory".

CAS is fortunate to have two fantastic keynote speakers: Dr. Katie Hoemann, an assistant professor of psychology at the University of Kansas, and Dr. Jonathan Gratch, a professor of computer science and psychology at the University of Southern California. Dr. Hoemann will give a talk on "The construction of emotional meaning in language"; Dr. Gratch will speak about "Beyond Emotion Recognition: Expressions as Social Action".

We would like to thank the LREC 2026 organizers for their support and providing a venue for our inaugural workshop. We especially thank our program committee, as without their hard work this workshop would not have been possible.

We look forward to continuing this workshop each year and hope to bring more researchers from various disciplines together to tackle the intricacies and challenges of affect.

CAS 2026 Program Chairs

Organizing Committee

Christopher Bagdon, University of Bamberg, Germany
Krishnapriya Vishnubhotla, National Research Council Canada
Kristen A. Lindquist, The Ohio State University, USA
Lyle Ungar, University of Pennsylvania, USA
Roman Klinger, University of Bamberg, Germany
Saif M. Mohammad, National Research Council Canada

Table of Contents

<i>Quality and Agreement in Multilabel Emotion Annotation: A Case Study and Evaluation Framework</i>	
Emily Sofi Ohman and Anna Koufakou	1
<i>Understanding Irony through Explanations and Background Knowledge</i>	
Aaron Maladry, Els Lefever, Cynthia Van Hee and Veronique Hoste	16
<i>Speech-Based Emotion Recognition and Classification Integrating a CNN and BiLSTM Network</i>	
Fatima Uroosa, Asim Abbas, Muhammad Tayyab Zamir and Grigori Sidorov	29
<i>A Corpus-Based Comparison of two Approaches for Emotion Annotation in French Texts</i>	
Valentina Dragos and Delphine Battistelli	38
<i>Clarifying the Role of Psychological Factors in Language Acquisition: A Psycholinguistic Lexical Ratings Dataset</i>	
Wanwan Zheng	47
<i>Unrequited Emotions: Investigating the Gaps in Motivation and Practice in Speech Emotion Recognition Research</i>	
Taryn Wong, Zeerak Talat, Hanan Aldarmaki and Anjalie Field	57
<i>Feeling First, Speaking Second: A Dual-Process Cognitive-Affective Architecture for LLM Agents</i>	
Nicolò Buscaroli and Fabio Tamburini	74
<i>Emotion Recogniton in Conversations - empirical study</i>	
Rufaida Kashif, Benjamin Piwowarski and Helena Gomez Adorno	93
<i>Exploring Cross-Modal Interactions in Unimodal and Multimodal Emotion Recognition: An Empirical Study</i>	
Quanqi Du, Loic De Langhe, Els Lefever and Veronique Hoste	105
<i>Annotation Matters: Resolving Cross-Corpus Performance Drops in Hebrew Offensive Language Detection</i>	
Gili Berger Hefetz, Yossef Haim Shrem, Natalia Vanetik and Chaya Liebeskind	116
<i>Multi-Source Emotion Annotation in Children's Language: When LLM Consensus Diverges from Human Judgment</i>	
Farida Said and Jeanne Villaneau	125
<i>EVOKE: Emotion Vocabulary Of Korean and English</i>	
Yoonwon Jung, Hagyeong Shin and Benjamin Bergen	136
<i>Linguistic Distancing on Social Media: Indicators of Emotion Regulation Across Age Groups</i>	
Daniela Teodorescu, Saif M. Mohammad and Alona Fyshe	151
<i>MOSAIC : a Corpus of Small-Group Interactions During a Collaborative Task</i>	
Amine Benamara, Celine Clavel, Brian Ravenet, Nicolas Sabouret, Mathilde Sassier-Roubin and Julien Saunier	164

<i>Age and Affect in Language: How Emotion Expression on Social Media Varies Across Adulthood</i> Daniela Teodorescu, Jan Philip Wahle and Saif M. Mohammad	175
<i>Multimodal Affective Modeling in an LLM-based Intelligent Tutoring System for Foreign Language Learning</i> Dionysios Koulouris, Athasios Kallipolitis, Melina Tziomaka, Argyrios Zafeiriou, Stamatios Orfanos, Andreas Menychtas, Ilias Maglogiannis, George Tsoulouhas, Stamatia Michalopoulou, Athina Sioupi and Voula Giouli	190
<i>PoETIC: A Re-framing of Context Dependent Emotion Detection</i> Nirmal Surange and Manish Shrivastava	201
<i>From Sentiment to Valence in Metaphor: a Comparison of BERT-based Sentiment and Prompted Large Language Models</i> Rebecca Guolo, Ginevra Martinelli, Chiara Barattieri di San Pietro and Valentina Bambini	212
<i>Affect, Body, Cognition, Demographics, and Emotion: The ABCDE of Text Features for Computational Affective Science</i> Jan Philip Wahle, Krishnapriya Vishnubhotla, Bela Gipp and Saif M. Mohammad	217
<i>Beyond the Black Box: Ethical and Theoretical Grounding in Affective Computing</i> Wajdi Zaghouani	235
<i>Beyond Toxic Positivity: Interpersonal Affect Regulation in LLM-Based Dialogue Agents using Discourse Politeness Theory</i> Rina Sakagami, Emmanuel Ayedoun and Masataka Tokumaru	244

Workshop Program

Saturday, May 16, 2026

14:00–14:10 Opening Remarks

14:10–15:00 Keynote Talk: Dr. Katie Hoemann | The Construction of Emotional Meaning in Language

The experience of emotion is a form of meaning-making: it reveals one's relationship to the circumstances. Often, the emphasis is on the emotions explicitly named or subjective feelings conveyed. In this talk, I argue that we should use a broader set of tools to study emotional meaning in language. I put forward three sets of language features that capture: the contextual features or aspects of experience salient at each moment (attention); the conceptual vantage point which from events are viewed (construal); the evaluation of events along relevant dimensions (appraisal). Using empirical findings, I illustrate how these language features can be used to answer specific questions about emotional meaning-making and how it varies based on situation, person, and culture. This interdisciplinary approach—grounded in socio-, cognitive, and computational linguistics as well as discursive, cognitive, and emotion psychology—seeks to move the field to a higher dimensional, dynamical account of emotional meaning.

Dr. Katie Hoemann is an Assistant Professor and the Dale J. Weary Faculty Fellow of Social Psychology at the University of Kansas, where she directs the Emotions in Context (EMIC) Lab. Katie studies how people experience emotion, and how the ways they make meaning of these experiences can shape motivations, behaviors, and social interactions in each moment. She also looks at how these momentary processes map on to individual and cultural differences in emotion, and how they are predictive of health and well-being. The hallmark of her research is studying emotion in the real-world contexts people navigate in their daily lives. Katie holds a BA in Cultural Anthropology, Spanish, and Linguistics from Northwestern University, an MA in Cognitive Linguistics from Bangor University, and a PhD in Psychology from Northeastern University. She was previously a postdoctoral fellow at the Center for Social and Cultural Psychology at KU Leuven.

15:00–15:45 Oral Presentations

16:00–17:00 Poster Session
Room: Poster Area

Quality and Agreement in Multilabel Emotion Annotation: A Case Study and Evaluation Framework

Emily Sofi Ohman and Anna Koufakou

Saturday, May 16, 2026 (continued)

Understanding Irony through Explanations and Background Knowledge

Aaron Maladry, Els Lefever, Cynthia Van Hee and Veronique Hoste

Speech-Based Emotion Recognition and Classification Integrating a CNN and BiLSTM Network

Fatima Uroosa, Asim Abbas, Muhammad Tayyab Zamir and Grigori Sidorov

A Corpus-Based Comparison of two Approaches for Emotion Annotation in French Texts

Valentina Dragos and Delphine Battistelli

Clarifying the Role of Psychological Factors in Language Acquisition: A Psycholinguistic Lexical Ratings Dataset

Wanwan Zheng

Unrequited Emotions: Investigating the Gaps in Motivation and Practice in Speech Emotion Recognition Research

Taryn Wong, Zeerak Talat, Hanan Aldarmaki and Anjalie Field

Feeling First, Speaking Second: A Dual-Process Cognitive-Affective Architecture for LLM Agents

Nicolò Buscaroli and Fabio Tamburini

Emotion Recogniton in Conversations - empirical study

Rufaida Kashif, Benjamin Piwowarski and Helena Gomez Adorno

Exploring Cross-Modal Interactions in Unimodal and Multimodal Emotion Recognition: An Empirical Study

Quanqi Du, Loic De Langhe, Els Lefever and Veronique Hoste

Annotation Matters: Resolving Cross-Corpus Performance Drops in Hebrew Offensive Language Detection

Gili Berger Hefetz, Yossef Haim Shrem, Natalia Vanetik and Chaya Liebskind

Multi-Source Emotion Annotation in Children's Language: When LLM Consensus Diverges from Human Judgment

Farida Said and Jeanne Villaneau

EVOKE: Emotion Vocabulary Of Korean and English

Yoonwon Jung, Hagyeong Shin and Benjamin Bergen

Linguistic Distancing on Social Media: Indicators of Emotion Regulation Across Age Groups

Daniela Teodorescu, Saif M. Mohammad and Alona Fyshe

Saturday, May 16, 2026 (continued)

MOSAIC : a Corpus of Small-Group Interactions During a Collaborative Task

Amine Benamara, Celine Clavel, Brian Ravenet, Nicolas Sabouret, Mathilde Sassier–Roublin and Julien Saunier

Age and Affect in Language: How Emotion Expression on Social Media Varies Across Adulthood

Daniela Teodorescu, Jan Philip Wahle and Saif M. Mohammad

Multimodal Affective Modeling in an LLM-based Intelligent Tutoring System for Foreign Language Learning

Dionysios Koulouris, Athasios Kallipolitis, Melina Tziomaka, Argyrios Zafeiriou, Stamatios Orfanos, Andreas Menychtas, Ilias Maglogiannis, George Tsoulouhas, Stamatia Michalopoulou, Athina Sioupi and Voula Giouli

PoETIC: A Re-framing of Context Dependent Emotion Detection

Nirmal Surange and Manish Shrivastava

From Sentiment to Valence in Metaphor: a Comparison of BERT-based Sentiment and Prompted Large Language Models

Rebecca Guolo, Ginevra Martinelli, Chiara Barattieri di San Pietro and Valentina Bambini

Affect, Body, Cognition, Demographics, and Emotion: The ABCDE of Text Features for Computational Affective Science

Jan Philip Wahle, Krishnapriya Vishnubhotla, Bela Gipp and Saif M. Mohammad

Beyond the Black Box: Ethical and Theoretical Grounding in Affective Computing

Wajdi Zaghouani

Beyond Toxic Positivity: Interpersonal Affect Regulation in LLM-Based Dialogue Agents using Discourse Politeness Theory

Rina Sakagami, Emmanuel Ayedoun and Masataka Tokumaru

Saturday, May 16, 2026 (continued)

17:00–17:50

Keynote Talk: Dr. Jonathan Gratch | Beyond Emotion Recognition: Expressions as Social Action

Emotion science increasingly agrees that expressions do not simply reveal internal states. Their meaning is contextual, directed at others, and co-constructed in interaction. Yet most computational approaches to "emotion recognition" still assume the opposite: that expressions encode latent states that can be inferred from morphology or timing alone. This assumption is embedded in common practices (such as training models on decontextualized annotations of presumed internal states) and even in emerging regulations like the EU AI Act, which treats emotion recognition as the recovery of hidden mental properties. My talk argues that the problem is not just methodological but conceptual. While contemporary theories emphasize context and co-construction, they offer little guidance on how to formalize these ideas computationally. What, precisely, counts as context? How do social goals, interactional dynamics, and cultural conventions shape what an expression means? Drawing on work in affective computing and social interaction, I propose a shift from recognizing emotions to modeling communicative behavior. I will illustrate how expressions function within interpersonal processes – negotiation, coordination – and how their meaning emerges through interaction rather than residing in the face or body alone. I conclude by outlining a research agenda for computational models that treat expressions as socially situated actions: specifying the structure of context, capturing co-construction over time, and enabling interactive systems that participate in, rather than merely observe, the creation of meaning.

Dr. Jonathan Gratch is a Research Full Professor of Computer Science and Psychology at the University of Southern California (USC) and Chief Science Officer at USC's Institute for Creative Technologies, but will be transitioning to a Professor of Computing at Vanderbilt University in the Fall. He completed his Ph.D. in Computer Science at the University of Illinois in Urbana-Champaign in 1995. Dr. Gratch's research focuses on computational models of human cognitive and social processes, especially emotion, and explores these models' potential to advance psychological theory and shape human-machine interaction. He is the founding Editor-in-Chief (retired) of IEEE's Transactions on Affective Computing, Associate Editor for Affective Science, Emotion Review, and former President of the Association for the Advancement of Affective Computing (AAAC). He is a Fellow of AAAI, AAAC, and the Cognitive Science Society.

17:50–18:00

Closing Remarks

Quality and Agreement in Multilabel Emotion Annotation: A Case Study and Evaluation Framework

Emily Ohman & Anna Koufakou

Waseda University, Florida Gulf Coast University

Tokyo, Japan; Fort Myers FL, USA

ohman@waseda.jp, akoufakou@fgcu.edu

Abstract

Emotion annotation is inherently subjective, yet most NLP pipelines still assume “gold” labels, typically produced by majority voting, and treat annotator variation as noise. In this paper, we present a multilabel emotion annotation case study and use it to examine how annotator behavior and aggregation choices affect both agreement estimates and downstream emotion classifiers. Rather than collapsing disagreement into a single label, we represent targets as soft vote-share labels (including an intensity-weighted variant) and evaluate models using both thresholded metrics (macro-/micro-F1) and probabilistic alignment (Bernoulli cross-entropy SoftBCE), alongside data-derived disagreement diagnostics. Across annotation regimes, we show that disagreement is structured and leaves measurable traces in model behavior: hard labels may maximize F1 metrics, while soft supervision yields predictions that better reflect empirical annotator variance and uncertainty. Our results provide practical guidance for designing, aggregating, and evaluating multilabel emotion datasets when multiple interpretations are plausible.

Keywords: emotion annotations, annotator disagreements, perspectives, evaluation metrics

1. Introduction

The difficulty of annotating for emotions is widely acknowledged in the field of emotion detection and sentiment analysis (see e.g. Andreevskaia and Bergler, 2007; Bermingham and Smeaton, 2009; Mohammad, 2016; Munezero et al., 2014; Strappavara and Mihalcea, 2010; Wiebe and Riloff, 2005; Öhman, 2021). Recent reviews also summarize the research landscape, available resources, challenges and gaps in this area (Plaza-del Arco et al., 2024; Koufakou and Nieves, 2025). In the following, we focus on specific aspects and challenges related to our work.

Overall, NLP research typically uses basic emotions (see e.g. Ekman, 1971; Plutchik, 1984) or simplifies fine-grained frameworks (see e.g. Cowen and Keltner, 2017; Öhman, 2020), relying on classification models with a small set of largely disjoint labels. This facilitates model training and improves performance: for example, the GoEmotions (Demszky et al., 2020) experiments showed higher performance when the authors merged and remapped fine-grained categories to basic emotions (Ekman). These techniques, however, fail to capture the full, fine-grained, and often overlapping nature of human emotional experience. In addition to discrete-category theories, affect is also commonly modeled in dimensional terms such as valence–arousal–dominance (VAD) and related circumplex accounts (Russell and Mehrabian, 1977).

Equally notable is that much of the related research in NLP relies on the notion of “gold” labels, typically derived through majority voting among annotators. Due to the highly subjective nature of in-

terpreting emotions, let alone emotions in a medium like text, humans usually do not agree on emotion labels. When annotators disagree, the corresponding records are often discarded. As Plank (2022) notes, “human variation in labeling is often considered noise.” Yet this raises an important question: *is such variation truly noise?* Aside from cases of genuinely careless or inconsistent annotators, whose contributions may indeed warrant removal, disagreement among humans often indicates that an item is open to interpretation. In such cases, rather than enforcing a single “correct” label, we might instead expect NLP models to reflect this nuance in their predictions.

Agreement and alignment among annotators is measured by Inter-Annotator Agreement (IAA). There are several metrics used to assess IAA; most common are Cohen’s κ , Fleiss’ κ , and Krippendorff’s α . The choice of IAA metric is influenced by various factors, for example the complexity of the annotation scheme (e.g. binary vs multi-label), the number of annotators, and how the work is assigned to the annotators. Regardless of the IAA metric, highly subjective tasks such as perceived emotions might lead to moderate or even low IAA. On the other hand, high agreement does not always mean high-quality annotations, for example, annotators could consistently agree on the wrong label. Besides the agreement between annotators, there might also be questions as to removing or filtering certain labels based on annotator effort or quality, which are not as straightforward to answer as we show in our case study (see Section 2).

Instead of eliminating disagreement, embracing

it can reveal the range of emotional interpretations a text may evoke. *Perspectivism* in NLP (Basile et al., 2021; Cabitza et al., 2023; Frenda et al., 2025) challenges the assumption of a single, objective ground truth and recognizes that multiple valid interpretations arise from annotators’ diverse backgrounds, experiences, and perspectives. By preserving disagreement, *perspectivism* enables more inclusive and representative emotion recognition systems and reframes annotation not merely as labeling but as a means of exploring the full spectrum of emotions relevant to a task. In the past, only a few datasets released full individual annotator labels, e.g. (Demszky et al., 2020), but this is starting to change¹(Barz et al., 2025; Muhammad et al., 2025). Additionally, there have been efforts to gain insights into the annotator disagreements, for example in environmental communications (Barz et al., 2025), while Weber-Genzel et al. (2024) showed the challenges of distinguishing between human label variation that may be important to preserve and annotation errors.

To preserve human disagreement, rather than focusing on traditional evaluation metrics defined over gold or *hard* labels, prior work has proposed the use of *soft* labels, for example, see earlier SemEval Tasks *Learning with Disagreement* (Leonardelli et al., 2023, 2025) based mostly on data with binary or Likert scale labels. To the best of our knowledge, soft-label approaches have not yet been explored specifically for emotion annotation.

In this paper, we present a multilabel emotion annotation case study and examine how soft labels can be used to preserve annotator disagreement, inform annotation quality assessment, and support appropriate NLP model evaluation. The case-study anchors the following contributions we make:

- A critical review of annotator agreement metrics and common practice of assigning hard labels for NLP
- A multilabel emotion annotation case study illustrating real disagreement patterns and questions of annotator quality
- A soft-label framework that preserves annotator disagreement, together with an evaluation framework and metrics specifically aligned with soft labels in a multilabel setting, covering both emotion categories and intensity annotations.

Section 2 introduces the case study, followed by our methodology for aggregating multi-annotator

labels in Section 3.1, and then the evaluation framework for agreement analysis and downstream NLP experiments in Section 4. Finally, we provide our concluding remarks in Section 5.

2. Case Study

We present a case study of three annotators annotating Hemingway’s *The Old Man and the Sea* (in English) using an expanded version of Plutchik’s wheel of emotions.

Field	Value
Resource name	The Old Man and the Sea emotion annotations (TOMATS)
Resource type	Sentence-level multilabel emotion annotations with intensity
Language	English
Unit of annotation	Roughly sentence-level text segments
Size	1,677 data points (sentences)
Annotators	3 expert annotators
Label set (core)	{anger, anticipation, disgust, fear, joy, sadness, surprise, trust}
Additional labels	Optional free-text “other” emotions/notes (not modeled)
Intensity	Pre-selected-emotion rating on 0–10; rescaled to [0,1] for modeling
Annotation format	Spreadsheet (one row per item; up to 3 emotions + intensities + other)

Table 1: Resource card for the TOMATS emotion annotation case study.

2.1. Dataset and Annotation Protocol

All three annotators were research assistants of the first author and native speakers of English, majoring in NLP or adjacent fields. Annotators 2 and 3 received most of their education in the US, while Annotator 1 attended international schools with English as the primary language of instruction. Although demographically similar, women in their mid-20s, they differ in cultural upbringing and educational exposure, which we expected to influence annotation style.

Ideally, a broader demographic profile (for example, including older and male annotators) would have improved diversity, but the present study uses the available annotators to highlight how even superficially “similar” annotators can diverge substantially in emotion annotation.

The annotators were asked to first read the text in full before considering the emotions and then start from the beginning to label each sentence² in context. They could assign multiple emotions or select

¹We release all processing code and derived annotation tables keyed by sentence index (without redistributing copyrighted text). The experiments can be reproduced by applying the scripts to a locally obtained copy of the novel. The code is available at https://github.com/esohman/SoftBCE_Emotions.

²roughly sentence level with minor splitting/merging where needed for coherence

“no emotion.” For the core annotation, annotators selected from the eight Plutchik emotions: *anger*, *anticipation*, *disgust*, *fear*, *joy*, *sadness*, *surprise*, *trust*. Annotators could additionally enter free-text “other” emotions/notes when none of the core labels captured their interpretation. These “other” emotions are listed in Appendix B.

We also introduced Plutchik’s dyad system to encourage reflection on secondary and tertiary emotions and allowed annotators to add additional categories or notes when they felt the taxonomy was insufficient. Annotators also rated the intensity of each selected emotion on a 0–10 scale. We used a 0–10 scale to provide annotators a familiar, high-resolution ordinal range that supports within-item comparisons (e.g., weak vs. strong evidence for a selected emotion). For modeling, we linearly rescaled intensities to [0,1] so they can be used directly as weights in vote-share aggregation while keeping all soft targets within the Bernoulli objective’s natural range.

2.2. Annotator Behavior and Disagreement

Sum	primary	secondary	tertiary
A1	1593	1258	443
A2	184	13	3
A3	1369	1200	566

Table 2: Annotation counts for each annotator

Table 2 shows the number of data points annotated by each annotator. We can see that Annotator 1 and Annotator 3 annotated emotions at similar rates across the 1677 sentences and assigned comparable numbers of secondary and tertiary labels. Their primary-emotion agreement is 42%, but agreement rises to 62% when considering any overlapping emotion label.

Annotator 2 presents a different case. They only considered roughly 10% of the data to have any emotion-association and almost no secondary or tertiary emotion associations (13 and 3 respectively). A closer look at their annotations further reveal that not only did they not annotate many data points at all, the data points they did annotate were not, contrary to expectations, data points with clear high-intensity emotion-linked content. Taken together with similarly sparse completion patterns observed in other research tasks, this points to limited protocol engagement rather than an alternative emotional reading.

Preliminary agreement scores calculated using Krippendorff’s α in Table 3 indicate low agreement. We include these metrics to highlight how unsuitable they are for multilabel emotions despite being commonly requested by reviewers.

Emotion	Krippendorff’s α
Anger	0.1674
Anticipation	0.2160
Disgust	0.1632
Fear	0.1599
Joy	0.1984
Sadness	0.2793
Surprise	0.2075
Trust	0.0920

Table 3: Krippendorff’s α for binary presence/absence of each Plutchik emotion.

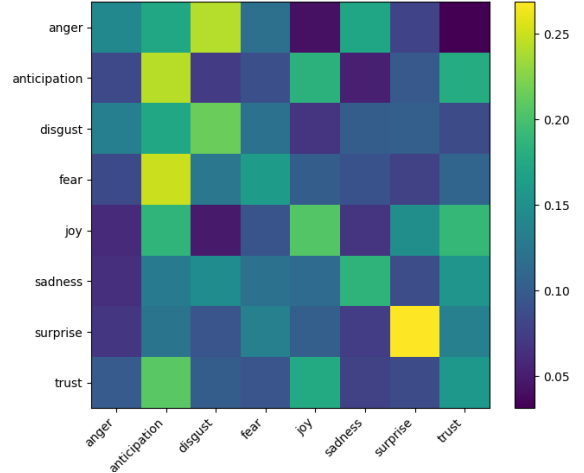


Figure 1: Annotation overlap (A1 vs A3)

To contextualize model confusion patterns, we additionally compute an annotation overlap matrix that captures systematic disagreement between annotators. We construct an emotion–emotion co-occurrence matrix (Fig. 1) by counting how often an emotion selected by one annotator overlaps with an emotion selected by another annotator for the same item, allowing up to three labels per annotator. The resulting matrix is row-normalized and visualized as a heatmap. We can see that *fear* is often confused with *anticipation*, and *anger* with *disgust* something that was also noted by e.g. Öhman et al. (2024). This representation does not encode correctness, but rather reflects structured ambiguity in the annotation process itself. Appendix A shows intra-annotator emotion overlap matrices.

2.3. Agreement Analysis and Implications

To assess annotation reliability, we compute pairwise agreement scores between the three annotators across all eight Plutchik emotions. Because the task is multilabel, each instance may exhibit zero or more emotions; classical inter-annotator agreement coefficients (e.g., Cohen’s κ or Fleiss’ κ) are therefore not directly applicable, as they assume

mutually exclusive categorical judgments. Instead, we adopt a per-emotion binary evaluation framework commonly used in multilabel affect annotation, where agreement is computed independently for each emotion (e.g., [Mohammad and Turney, 2013](#); [Mohammad and Bravo-Marquez, 2017](#)) (see also [Artstein and Poesio, 2008](#) for a general discussion of agreement in non-exclusive annotation tasks).

For each annotator pair (a, b) , for emotion e , and item i , we consider binary labels

$$y_{i,e}^{(a)}, y_{i,e}^{(b)} \in \{0, 1\}, \quad (1)$$

indicating the presence or absence of emotion e . We compute the per-emotion F1 score as this metric directly captures how consistently annotators agree on the presence of an emotion and is less sensitive to extreme class imbalance than accuracy. From the set of eight per-emotion F1 scores, we report:

- **macro-F1**: the unweighted mean of the per-emotion F1 scores;
- **micro-F1**: the F1 score computed after flattening all $8 \times N$ binary decisions into a single contingency table;
- **Per-emotion F1**: the individual agreement scores for each Plutchik emotion.

Annotator 2 left the majority of items unannotated ("none"). Here, "none"/blank entries denote no provided label (nonresponse/placeholder) rather than an explicit 'no emotion' judgment, which was available as a selectable option. Treating such blanks as negative judgments would artificially deflate agreement. To distinguish disagreement from missing information, we therefore measure agreement under two conditions: (i) on the full dataset, where blank entries count as negatives, and (ii) on a filtered subset consisting only of items where Annotator 2 produced at least one genuine emotion label. In the latter setting, we treat all other entries for Annotator 2 as missing.

In Table 4, we observe substantial differences in pairwise annotator agreement depending on whether the analysis is performed on the full dataset or restricted to the subset of items that Annotator 2 actually annotated with a genuine emotion label. On the full dataset, agreement between Annotator 2 and the other annotators is extremely low (macro-F1 ≈ 0.08), reflecting the fact that Annotator 2 left most items unannotated and thus systematically disagreed with the others. However, when we restrict the evaluation to the 184 items that Annotator 2 labelled, agreement rises sharply (macro-F1 ≈ 0.38 - 0.40), approaching the level of agreement observed between the two reliable annotators (Annotator 1 and Annotator 3: macro-F1 ≈ 0.49). This indicates that Annotator 2 is not inherently inconsistent; rather, the majority of their missing labels

should be treated as missing data rather than as true negatives. Consequently, for all downstream modeling we treat Annotator 2's missing annotations as missing information rather than evidence of emotion absence, and only incorporate their annotations on items they explicitly rated.

As a sidenote, it should not be ignored that the low agreement scores are in part because the annotators, Annotator 3 in particular, utilized the "other emotion" column. For example, in one case Annotator 1 annotated the primary emotion as *anticipation* and *fear* as the secondary emotion. For that same data point Annotator 3 had left the primary slot empty and added *pessimism* as the other emotion. It could be argued that *pessimism* is indeed an overlap of *anticipation* and *fear*.

In another example, the sentence "*Chew it well, he thought, and get all the juices.*" occurs during Santiago's (the protagonist) prolonged struggle at sea, he eats a small fish raw to maintain his strength. The internal monologue reflects pragmatic self-discipline and physical endurance rather than pleasure, emphasizing survival under extreme deprivation. Annotator 1 had selected *disgust*, *anger*, *anticipation* (in that order) and Annotator 3 selected *anger*, *anticipation*, and added *determination*. Again, the annotations seem aligned with a human interpretation, but harder to accurately show quantitatively.

3. Aggregating multi-annotator labels

Here, we describe our methodology for constructing hard and soft targets from human multilabel annotations.

Each instance is annotated by up to three annotators, who may select multiple emotions and optionally provide intensity ratings. Because Annotator 2 labels a smaller subset of instances, we treat missing annotations as missing rather than as explicit negative votes. For each instance i and emotion k , we compute a soft target as the empirical vote share among contributing annotators:

$$\tilde{y}_{i,k} = \frac{1}{|\mathcal{A}_i|} \sum_{a \in \mathcal{A}_i} \mathbf{1}[k \in S_{a,i}] \quad (2)$$

where $S_{a,i}$ is annotator a 's selected label set and $\mathcal{A}_i$ denotes annotators who provided at least one label for instance i . This produces multilabel targets $\tilde{y}_{i,k} \in [0, 1]$ that can be interpreted as per-label Bernoulli probabilities.

To incorporate intensity information, we additionally define an intensity-weighted soft target. For each annotator, selected emotions are weighted by their reported intensity (scaled to $[0, 1]$), and aggregated across contributing annotators. This preserves graded signal about strength of affect

Pair	Setting	macro-F1	micro-F1	Avg. Per-emotion F1	Min/Max Per-Emotion F1
A1–A2	Full	0.0788	0.0882	0.0795	0.0093 / 0.1760
A1–A2	Filtered	0.3949	0.4857	0.3932	0.1053 / 0.7015
A1–A3	Full	0.3794	0.3952	0.3804	0.2857 / 0.4980
A1–A3	Filtered	0.4899	0.5187	0.4870	0.3051 / 0.6410
A2–A3	Full	0.0829	0.0899	0.0820	0.0058 / 0.1469
A2–A3	Filtered	0.3797	0.4664	0.3830	0.0455 / 0.7706

Table 4: Summary comparison of pairwise annotator agreement in the full vs. filtered subsets.

while still producing multilabel targets in $[0, 1]$ suitable for Bernoulli objectives.

3.1. Targets and disagreement diagnostics

Let $y_{i,k}^{(a)} \in \{0, 1\}$ denote whether annotator a assigned emotion k to instance i , and let $\mathcal{A}_i$ denote the set of annotators who provided at least one label for instance i (i.e., missing annotations are treated as missing). We define the per-label vote share

$$p_{i,k} = \frac{1}{|\mathcal{A}_i|} \sum_{a \in \mathcal{A}_i} y_{i,k}^{(a)} \quad (3)$$

which yields multilabel soft targets $p_{i,k} \in [0, 1]$ that can be interpreted as Bernoulli probabilities. Hard targets are obtained by union-of-labels over contributing annotators:

$$y_{i,k}^{\text{hard}} = \mathbf{1}[p_{i,k} > 0] \quad (4)$$

To characterize ambiguity in the annotations independently of any model, we compute two data-derived disagreement measures from annotator label sets. First, we use mean per-label Bernoulli variance,

$$D_{\text{var}}(i) = \frac{1}{K} \sum_{k=1}^K p_{i,k} (1 - p_{i,k}) \quad (5)$$

which is maximized when annotators split evenly on a label and minimized when they agree ($p_{i,k} \in \{0, 1\}$). Second, we compute mean pairwise Jaccard disagreement between annotators’ label sets,

$$D_{\text{Jac}}(i) = 1 - \frac{1}{\binom{|\mathcal{A}_i|}{2}} \sum_{a < b} \frac{|S_{a,i} \cap S_{b,i}|}{|S_{a,i} \cup S_{b,i}|} \quad (6)$$

where the sum is taken over all annotator pairs (a, b) contributing to instance i . These measures are computed from annotations alone and are therefore constant across model variants trained under the same annotator-inclusion regime.

Because our downstream models operate over the 8 Plutchik emotions, we filter free-text “other” labels and compute soft-label distributions over this fixed set. This effectively reallocates probability mass from custom labels to the remaining categories via renormalization, which can reduce apparent disagreement (lower Dvar) in cases where

annotators expressed qualitatively different states outside the taxonomy. Future work could mitigate this by introducing an explicit “other” class, mapping free-text labels into a hierarchical taxonomy, or analyzing Dvar both with and without the “Other” bucket.

3.2. Bernoulli cross-entropy for multilabel soft targets

Because our task is multilabel, we model each emotion as an independent Bernoulli variable. Let $\hat{p}_{i,k} = \sigma(z_{i,k})$ denote the model’s predicted probability for label k (with logit $z_{i,k}$). We quantify probabilistic alignment between model predictions and the aggregated soft targets $p_{i,k}$ using Bernoulli cross-entropy (binary cross-entropy; BCE):

$$\text{SoftBCE}(i, k) = -\left(p_{i,k} \log \hat{p}_{i,k} + (1 - p_{i,k}) \log (1 - \hat{p}_{i,k})\right) \quad (7)$$

We report SoftBCE averaged across labels and instances as an auxiliary model–annotation alignment metric; lower values indicate closer alignment. In addition to SoftBCE, downstream models are evaluated using macro-/micro-F1 with decision thresholds tuned on a validation split. Similar cross-entropy-style diagnostics have been used in prior disagreement-oriented work (e.g., [Leonardelli et al., 2023, 2025](#)), but our setting differs in that targets are multilabel Bernoulli vote shares (with missingness treated explicitly), rather than single-label distributions.

Concrete stratified examples of the resulting hard and soft targets (low/medium/high disagreement) are provided in Table 5.

4. Evaluation Pipeline

Our evaluation pipeline is designed to examine how annotation behavior and disagreement structure affect both agreement estimates and downstream emotion classification. Rather than optimizing predictive performance, the goal is to assess how different assumptions about annotation completeness and disagreement propagate through the modeling process.

The pipeline consists of four stages.

(1) Label extraction and harmonization. Annotations are provided in separate spreadsheet sheets, one per annotator, with up to three Plutchik emotion labels per item and optional free-text notes. For each annotator a and item i , we extract the set of explicitly assigned Plutchik emotions, discarding “none”, blanks, and non-Plutchik entries. This yields a mapping $(i) \rightarrow \{e_1, e_2, \dots\}$, that reflects only positive emotion assignments.

(2) Multilabel representation. We convert the extracted labels into a binary multilabel tensor

$$Y \in \{0, 1\}^{A \times N \times K}$$

where $Y_{a,i,k} = 1$ indicates that annotator a assigned emotion k to item i . This representation treats each annotator as an independent multilabel classifier and supports both pairwise agreement analysis and model training. Missing annotations (i.e., annotator–item pairs where no labels were provided) are encoded as empty label sets and tracked via a contribution mask, which is used in the next stage when aggregating vote shares over the set of contributing annotators $\mathcal{A}_i$.

(3) Agreement analysis with missing-label filtering. Because one annotator exhibits systematic sparsity, interpreting blank entries as negative judgments would conflate missingness with disagreement. We therefore compute agreement under two conditions:

1. **Full dataset:** all missing labels are treated as “no emotion”.
2. **Filtered dataset:** items for which Annotator 2 assigned at least one genuine Plutchik emotion are retained; all other items are removed for all annotators.

This comparison allows us to disentangle structural disagreement caused by missing labels from genuine divergence in emotional interpretation.

(4) Downstream modeling and evaluation. To assess whether differences in annotation behavior propagate into downstream models, we evaluate a small set of representative classifiers under different annotation regimes. The goal is not to optimize predictive performance, but to use modeling behavior as a diagnostic lens on annotation consistency. We further introduce an evaluation framework for soft labels in multilabel emotion annotation and demonstrate its practical application.

We consider four classifier families:

- (a) a transformer-based multilabel classifier trained with Bernoulli cross-entropy (binary cross-entropy) on aggregated hard targets;

- (b) the same architecture trained with Bernoulli cross-entropy using vote-share soft targets $\tilde{y}_{i,k} \in [0, 1]$, thereby preserving annotator disagreement in the supervision signal;
- (c) the same architecture trained with intensity-weighted soft targets, where vote shares are modulated by annotator-provided intensity scores (scaled to $[0, 1]$);
- (d) a one-vs-rest SVM baseline using frozen transformer embeddings.

Each model is trained under two annotation conditions: using all available annotators, and using only the two annotators identified as reliable by the filtered agreement analysis. We report both thresholded classification metrics (macro-/micro-F1) and probabilistic alignment metrics. Data-derived disagreement measures (e.g., D_{var} and D_{Jac}) are computed from annotations only, while SoftBCE denotes the Bernoulli cross-entropy between model probabilities and the soft targets and is used as an auxiliary model–annotation alignment metric (see subsections 3.1 and 3.2 respectively).

4.1. Training objectives

We train all transformer models with a multilabel Bernoulli objective (`BCEWithLogitsLoss`). In the hard-label condition, targets are binary vectors aggregated from annotator label sets. In the soft-label conditions, we replace binary targets with either (i) vote-share soft targets $\tilde{y}$ or (ii) intensity-weighted soft targets, yielding a soft-target variant of BCE that exposes annotator uncertainty and graded affect strength during training. Examples are stratified by D_{Jac} tertiles (computed from annotator label sets).

We report both thresholded classification metrics (macro-/micro-F1) and probabilistic alignment metrics. While F1 evaluates discrete decisions after thresholding, SoftBCE evaluates the full predicted probability vector against the aggregated soft targets; improvements in SoftBCE therefore indicate better-calibrated probabilities that more faithfully reflect inter-annotator variation.

Table 6 lists implementation details for reproducibility.

All results are from a single run with fixed seed; we did not perform multi-seed averaging.

4.2. Evaluation

Table 7 compares BERT models trained with hard targets versus two soft-target variants, under settings that either include or exclude Annotator 2. Hard-label training yields the highest F1 overall when Annotator 2 is included (macro-F1 = 0.616; micro-F1 = 0.625), but the intensity-weighted

Example text	Annotations	Hard labels	Soft labels
<i>Low disagreement</i>			
“That’s very kind of you,” the old man said.	A1: Sa, Su, Tr A2: Jo, Tr A3: Jo, Tr	[0, 0, 0, 0, 1, 1, 1, 1]	$[0, 0, 0, 0, \frac{2}{3}, \frac{1}{3}, \frac{1}{3}, 1]$
“The bird is a great help,” the old man said.	A1: Jo, Tr A2: Jo A3: Jo, Su	[0, 0, 0, 0, 1, 0, 1, 1]	$[0, 0, 0, 0, 1, 0, \frac{1}{3}, \frac{1}{3}]$
<i>Medium disagreement</i>			
He was gone and the old man felt nothing.	A1: Sa A2: Sa A3: Su	[0, 0, 0, 0, 0, 1, 1, 0]	$[0, 0, 0, 0, 0, \frac{2}{3}, \frac{1}{3}, 0]$
Come up easy and let me put the harpoon into you.	A1: Ant, Jo, Tr A2: Ant A3: Ant, Di, Jo	[0, 1, 1, 0, 1, 0, 0, 1]	$[0, 1, \frac{1}{3}, 0, \frac{2}{3}, 0, 0, \frac{1}{3}]$
<i>High disagreement</i>			
“I think perhaps I can too. But I try not to borrow. First you borrow. Then you beg.”	A1: Fe, Sa, Tr A2: Di A3: Ant, Jo, Tr	[0, 1, 1, 1, 1, 1, 0, 1]	$[0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0, \frac{2}{3}]$
“They say his father was a fisherman. Maybe he was as poor as we are and would understand.”	A1: Ant, Jo, Sa A2: Tr A3: Su, Tr	[0, 1, 0, 0, 1, 1, 1, 1]	$[0, \frac{1}{3}, 0, 0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \frac{2}{3}]$

Table 5: Stratified examples illustrating how hard and soft targets are derived from human annotations. Hard labels are the union-of-labels over annotators (1 if any annotator selected the emotion). Soft labels are vote-share targets (fraction of contributing annotators selecting each emotion). Vectors follow the fixed emotion order: [Ang:anger, Ant:anticipation, Di:disgust, Fe:fear, Jo:joy, Sa:sadness, Su:surprise, Tr:trust].

Component	Setting
Backbone	<code>bert-base-uncased</code>
Tokenizer	<code>AutoTokenizer</code> (uncased WordPiece)
Input length	Max length 256; truncation; padding to <code>max_length</code>
Optimizer	AdamW, learning rate 2×10^{-5}
Batch size	16
Epochs	Up to 10 with early stopping
Early stopping	Patience 2 on validation macro-F1; keep best checkpoint (min $\Delta = 10^{-4}$)
Threshold selection	Tune global threshold t on validation macro-F1; grid $t \in \{0.10, 0.15, \dots, 0.90\}$
Data split	Random 80/20 train/validation split; <code>shuffle=True</code>
Random seed	42 (<code>numpy + torch</code>)
SVM baseline	One-vs-rest LinearSVC on frozen BERT [CLS] embeddings (embedding batch size 32; other params default)

Table 6: Model and training configuration used in all transformer experiments.

soft-label model remains competitive (macro-F1 = 0.607; micro-F1 = 0.617) while achieving substantially lower SoftBCE (0.548 vs. 0.681), indicating markedly better alignment between predicted probabilities and the aggregated soft targets. Uniformly

weighted soft labels reduce F1 more noticeably, though they still improve SoftBCE relative to hard training. Excluding Annotator 2 slightly decreases F1 across objectives and also increases SoftBCE for the hard-label model, suggesting that Annotator 2 contributes useful signal despite partial coverage. Finally, the disagreement measures D_{var} and D_{Jac} are stable within each annotation regime, as expected because they are computed from the annotations themselves rather than model outputs; they confirm that the underlying level of annotator divergence is comparable across the compared training objectives.

Because macro-/micro-F1 are computed after binarizing predictions, they depend on a tuned decision threshold and primarily reflect ranking performance around that cutoff. SoftBCE, in contrast, evaluates the full vector of predicted probabilities against the aggregated soft targets without thresholding. The two metrics therefore capture complementary aspects of model quality: thresholded classification performance versus probabilistic alignment with annotator uncertainty.

Incorporating intensity information improves the soft-label objective. Intensity-weighted soft labels yield higher macro- and micro-F1 than uniformly weighted soft labels in both the *with A2* and *without A2* settings, and they also produce markedly lower SoftBCE. This suggests that intensity ratings provide a useful inductive bias that sharpens the

Setting	Model	Thr.	macro-F1	micro-F1	SoftBCE	D_{var}	D_{Jac}	$\bar{A}$
With A2	Hard labels	0.35	0.616	0.625	0.681	0.066	0.708	1.98
With A2	Soft labels (uniform)	0.10	0.589	0.599	0.577	0.066	0.708	1.98
With A2	Soft labels (intensity)	0.10	0.607	0.617	0.548	0.066	0.708	1.98
With A2	SVM baseline	-	0.503	0.519	-	-	-	-
Without A2	Hard labels	0.15	0.606	0.616	0.717	0.065	0.707	1.85
Without A2	Soft labels (uniform)	0.10	0.588	0.597	0.564	0.065	0.707	1.85
Without A2	Soft labels (intensity)	0.10	0.591	0.600	0.578	0.065	0.707	1.85
Without A2	SVM baseline	-	0.501	0.517	-	-	-	-

Table 7: Downstream results for BERT under different supervision targets and annotator-inclusion regimes. Thr. is the tuned decision threshold (validation set). SoftBCE is the Bernoulli cross-entropy between predicted probabilities and aggregated soft targets (lower is better). D_{var} denotes mean per-label Bernoulli variance $\frac{1}{K} \sum_k \tilde{y}_k(1 - \tilde{y}_k)$, D_{Jac} denotes mean pairwise Jaccard disagreement (both data-derived; higher means more disagreement), and $\bar{A}$ is the average number of contributing annotators per item.

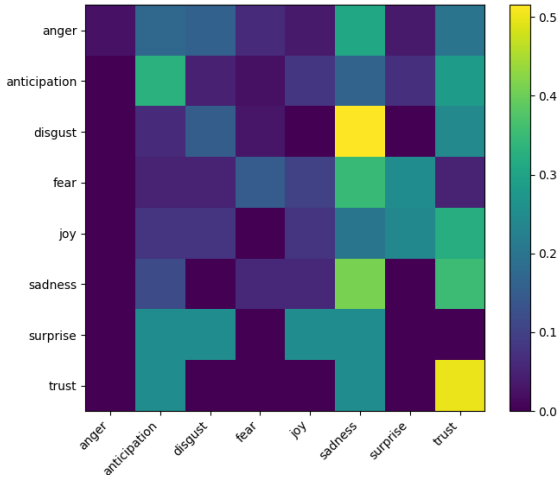


Figure 2: Confusion matrix for BERT with hard labels including A2 labels

soft targets by distinguishing weak affective cues from strong ones, improving both downstream decisions and probabilistic alignment with annotator uncertainty. Overall, these results support the use of soft labels as a diagnostic tool for understanding how annotation regimes propagate into model behavior, while highlighting that additional annotation detail does not automatically translate into improved downstream performance.

4.3. Confusion

Figure 2 presents a row-normalized single-label confusion matrix for the results from training BERT on hard labels with all annotators. Several emotions exhibit strong diagonal dominance, most notably *sadness* and *trust*, indicating that when these emotions are the primary label, the model predicts them consistently. The most prominent off-diagonal confusion occurs between *disgust* and *sadness*, with a substantial proportion of *disgust*-labeled instances predicted as *sadness*, suggesting overlap

in negative-valence affective cues. This has also been shown in previous work (Rossi and Öhman, 2025). Other emotions, such as *anger* and *anticipation*, display more diffuse prediction patterns without a single dominant confusion partner. Overall, the matrix reveals clusters of related emotions rather than random error, supporting its use as a diagnostic tool for identifying structured affective overlap rather than discrete misclassification.

4.4. Error Analysis

To complement aggregate F1 scores, we inspect structured error patterns using a row-normalized single-label confusion matrix, obtained by projecting multilabel outputs to a dominant label per instance. Confusions are not random: the largest off-diagonal mass occurs between semantically and valence-adjacent emotions, including *sadness* vs. *fear*, *trust* vs. *anticipation*, and *anger* vs. *disgust*. These patterns align with the dataset’s disagreement diagnostics (D_{var} , D_{Jac}), suggesting that model errors concentrate in regions of the label space that are also ambiguous for annotators. Predictive performance is weakest for rare emotions (e.g., *surprise*), consistent with challenges in multilabel learning under extreme imbalance.

5. Conclusions

This paper examined how annotation design choices and annotator reliability affect downstream modeling for multilabel emotion classification. Rather than treating annotation as a static preprocessing step, we used it as an object of analysis and asked how different aggregation strategies and training objectives interact with empirically observed disagreement. Across experiments, annotation decisions produced consistent, measurable differences in both thresholded performance (macro-/micro-F1) and probabilistic model-target

alignment (Bernoulli cross-entropy or SoftBCE).

The downstream results suggest that the benefits of soft supervision depend on what is being optimized. With all annotators included, hard-label training achieved the highest macro- and micro-F1 in our setup. Soft-label objectives, however, produced markedly lower SoftBCE while remaining broadly competitive in F1. In particular, intensity-weighted soft targets reduced SoftBCE the most under the full-annotator regime and narrowed the F1 gap to hard labels, indicating closer alignment with the uncertainty encoded in the aggregated targets. When we excluded Annotator 2, both hard and soft variants showed slightly lower F1, while the disagreement diagnostics (D_{var} , D_{Jac}) remained essentially unchanged, as expected for measures computed from the annotations alone.

These findings clarify a common claim in disagreement-aware learning. Preserving disagreement does not necessarily translate into higher F1, but it can yield probability estimates that better reflect the empirical label uncertainty. For affective tasks, where multiple interpretations can be simultaneously plausible and downstream applications often consume probabilities rather than hard decisions, this distinction is practically important. In this sense, SoftBCE provides a useful complementary lens to standard F1 when evaluating uncertainty-aware training signals.

At the same time, our results highlight limits to how additional annotation detail translates into downstream gains. Intensity ratings are not automatically helpful: their impact depends on how consistently annotators use the scale and on how intensities are normalized and aggregated. In this case study, incorporating intensity improved model-target alignment substantially but did not consistently dominate thresholded metrics, suggesting that intensity can encode graded uncertainty even when it does not directly improve binary decisions. This points to straightforward extensions, including annotator-specific calibration, learned normalization, and objectives that explicitly model individual differences in scale use.

The main contribution of this paper is methodological. Using downstream models as a diagnostic tool, together with disagreement measures computed directly from the labels, provides a practical way to surface design trade-offs in emotion annotation and to distinguish ambiguity in the data from annotator-specific behavior. Applying the same pipeline to additional texts and domains would help establish when these patterns generalize and when they are corpus- or task-specific.

We anticipate that applying this pipeline to short-form or informal text (e.g., posts or comments) will increase both overlap among emotions and annotator disagreement due to reduced context and higher

ambiguity. Our approach is designed for precisely these conditions: missingness-aware aggregation separates nonresponse from negative judgments, and SoftBCE measures distributional alignment when “gold” labels are intrinsically uncertain. Future work can further test robustness under larger, crowd-sourced annotator pools and examine how disagreement patterns vary with domain and annotator diversity.

In conclusion, the results support treating annotation and modeling as coupled choices rather than sequential steps. By making disagreement explicit in both target construction and evaluation, and by reporting both thresholded and probabilistic alignment metrics, we provide a transparent framework for analyzing and improving multilabel emotion annotation practices.

Limitations

This work is intentionally framed as a case study, and several limitations are worth noting. First, the annotation effort involves a small number of annotators and a single literary text. This setting enables detailed qualitative and quantitative analysis of annotator behavior, but it limits the generalizability of absolute agreement values and downstream performance. Our conclusions therefore emphasize *patterns* of annotation behavior and their modeling consequences rather than population-level estimates of annotator reliability.

Second, the annotators were not drawn from a fully representative population. Although all were native speakers of English, they were relatively homogeneous and affiliated with the same research group. This reduces demographic variance, but also makes the observed divergence in annotation density and engagement noteworthy: substantial differences in labeling behavior emerged even within a narrow pool. Future work should test whether similar patterns hold with larger and more heterogeneous annotator groups.

Third, one annotator exhibited systematically sparse labeling behavior, which complicates both agreement measurement and target construction. While we address this through missingness-aware aggregation (treating non-annotations as missing rather than negative votes) and filtered analyses, the degree of missingness underscores the need for clearer guidelines and more explicit handling of abstentions in emotion annotation workflows. In particular, low agreement scores can reflect missing labels rather than conceptual disagreement, but distinguishing these cases is not always straightforward in practice.

Fourth, while we use Bernoulli cross-entropy (SoftBCE) as a model-target alignment metric for multilabel soft targets, there is no single universally

accepted evaluation measure for soft labels especially for multilabel settings, and different metrics emphasize different aspects of disagreement. Prior work has pointed out limitations of cross-entropy-style metrics in some soft-label settings (Rizzi et al., 2024), and recent shared-task work has also explored alternatives such as Manhattan and Wasserstein distances (Leonardelli et al., 2025). We therefore treat SoftBCE as an auxiliary diagnostic and report complementary data-derived disagreement measures; evaluating additional soft-label metrics for multilabel emotion annotation remains an important direction for future work.

Fifth, although the annotation scheme includes additional (rare) emotion categories beyond the main set analyzed here, we did not model these extra labels. Incorporating very low-frequency categories without destabilizing training or evaluation is non-trivial, and developing principled strategies for handling rare emotions (e.g., hierarchical labels, grouping, or targeted re-annotation) is a clear next step. Similarly, the intensity score reliability and coherence could be improved by adapting a best-worst-scaling approach (as demonstrated by Kiritchenko and Mohammad, 2017).

Finally, the classification models evaluated in this paper are deliberately simple and are used primarily as analytical tools rather than optimized systems. The goal is not to establish state-of-the-art performance, but to examine how different annotation regimes affect downstream modeling and evaluation. More sophisticated architectures or training strategies may yield higher performance, but they would not change the core methodological point: annotation behavior and aggregation choices can measurably shape both model outputs and how those outputs should be evaluated.

Ethical Considerations

This paper is primarily methodological: we analyze a small multilabel emotion annotation case study to understand how aggregation and disagreement-aware modeling choices affect evaluation and interpretation. The work does not involve user-generated content, sensitive personal data, or human subjects beyond a limited internal annotation exercise. Nonetheless, automated emotion recognition has broader societal implications that warrant explicit discussion.

Emotions are not directly observable from text, and the same utterance can support multiple plausible readings depending on context, culture, and reader perspective. Text-only emotion models are therefore best understood as predicting *annotated interpretations* rather than latent “true” emotions. This is especially important in settings where other signals (e.g., conversational context, prosody, fa-

cial expression, or situational metadata) are unavailable. Our results reinforce this point: disagreement is structured and often reflects genuine ambiguity rather than annotator error. Treating majority-voted labels as “gold” can mask uncertainty and overstate model reliability.

Emotion recognition models can be misapplied to surveillance, profiling, workplace monitoring, or high-stakes decision-making. These risks increase when outputs are presented as objective measures of an individual’s internal state. We caution against such uses and recommend that, when emotion predictions are used at all, systems should expose uncertainty (e.g., calibrated probabilities) and be accompanied by clear documentation of what the labels represent, how they were collected, and where disagreement is expected (e.g., Mohammad, 2022).

Emotion categories, their lexical realizations, and their social meanings vary across languages and cultures. Models trained and evaluated on narrowly sampled annotator populations or single-language datasets may not transfer reliably to other contexts and may reproduce cultural biases. A practical implication of our work is that dataset documentation should report annotator composition, missingness patterns, and disagreement diagnostics, and that future annotation efforts should prioritize more diverse annotator pools and multilingual settings.

Because emotion annotation and interpretation are inherently tied to theories of affect, narrative, and communication, we encourage involving researchers beyond NLP (e.g., psychology, linguistics, HCI, digital humanities) in annotation design, guideline development, and evaluation choices. Disagreement-aware approaches should be informed by domain knowledge about when multiple interpretations are expected and meaningful, rather than treated as purely statistical noise.

Finally, any future release of materials derived from copyrighted literary sources will comply with licensing constraints. Where full text cannot be redistributed, we will document what can be shared (e.g., derived labels, code, and evaluation scripts) to support reproducibility while respecting rights holders.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 24K21058.

6. Bibliographical References

Alina Andreevskaia and Sabine Bergler. 2007. CLaC and CLaC-NB: Knowledge-based and

- corpus-based approaches to sentiment tagging. In *Proceedings of the 4th international workshop on semantic evaluations*, pages 117–120. Association for Computational Linguistics.
- Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.
- Christina Barz, Melanie Siegel, Daniel Hanss, and Michael Wiegand. 2025. Understanding disagreement: An annotation study of sentiment and emotional language in environmental communication. In *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)*, pages 1–20.
- Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, Alexandra Uma, et al. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st workshop on benchmarking: past, present and future*, pages 15–21. Association for Computational Linguistics.
- Adam Bermingham and Alan F Smeaton. 2009. A study of inter-annotator agreement for opinion retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 784–785.
- Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.
- Alan S Cowen and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the national academy of sciences*, 114(38):E7900–E7909.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.
- Paul Ekman. 1971. Universals and cultural differences in facial expressions of emotion. In *Nebraska symposium on motivation*. University of Nebraska Press.
- Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2025. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, 59(2):1719–1746.
- Ernest Hemingway. 1995. The old man and the sea. 1952. New York: Scribner.
- Svetlana Kiritchenko and Saif Mohammad. 2017. [Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.
- Anna Koufakou and Elijah Nieves. 2025. Review of recent emotion-annotated text corpora and resources. *Language Resources and Evaluation*, pages 1–35.
- Elisa Leonardelli, Gavin Abercrombie, Dina Al-manee, Valerio Basile, Tommaso Fornaciari, Barbara Plank, Verena Rieser, Alexandra Uma, and Massimo Poesio. 2023. [SemEval-2023 task 11: Learning with disagreements \(LeWiDi\)](#). In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2304–2318, Toronto, Canada. Association for Computational Linguistics.
- Elisa Leonardelli, Silvia Casola, Siyao Peng, Giulia Rizzi, Valerio Basile, Elisabetta Fersini, Diego Frassinelli, Hyewon Jang, Maja Pavlovic, Barbara Plank, and Massimo Poesio. 2025. [LeWiDi-2025 at NLPerspectives: Third edition of the learning with disagreements shared task](#). In *Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP*, pages 182–195, Suzhou, China. Association for Computational Linguistics.
- Saif Mohammad. 2016. A practical guide to sentiment annotation: Challenges and solutions. In *WASSA@ NAACL-HLT*, pages 174–179.
- Saif Mohammad and Felipe Bravo-Marquez. 2017. Emotion intensities in tweets. In *Proceedings of the 6th joint conference on lexical and computational semantics (* SEM 2017)*, pages 65–77.
- Saif M Mohammad. 2022. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics*, 48(2):239–278.
- Saif M. Mohammad and Peter D. Turney. 2013. [Crowdsourcing a word–emotion association lexicon](#). *Computational Intelligence*, 29(3):436–465.
- Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermino D. M. A.

- Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tuñiño, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Alexander Panchenko, Andrew Piper, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025. [BRIGHTER: BRIdging the gap in human-annotated textual emotion recognition datasets for 28 languages](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8895–8916, Vienna, Austria. Association for Computational Linguistics.
- M. Munezero, C. Montero, E. Sutinen, and J. Pajunen. 2014. [Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text](#). *IEEE Transactions on Affective Computing*, 5(02):101–111.
- Emily Öhman. 2020. Emotion annotation: Rethinking emotion categorization. In *DHN Post-Proceedings*, pages 134–144.
- Emily Öhman. 2021. *The language of emotions: Building and applying computational methods for emotion detection for English and beyond*. Helsingin yliopisto.
- Emily Öhman, Yuri Bizzoni, Pascale Feldkamp Moreira, and Kristoffer Nielbo. 2024. Emotionarcs: Emotion arcs for 9,000 literary texts. In *Proceedings of the 8th joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature (LaTeCH-CLfL 2024)*, pages 51–66.
- Barbara Plank. 2022. [The “problem” of human label variation: On ground truth in data, modeling and evaluation](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Flor Miriam Plaza-del Arco, Alba A. Cercas Curry, Amanda Cercas Curry, and Dirk Hovy. 2024. [Emotion analysis in NLP: Trends, gaps and roadmap for future directions](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5696–5710, Torino, Italia. ELRA and ICCL.
- Robert Plutchik. 1984. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984(197-219):2–4.
- Giulia Rizzi, Elisa Leonardelli, Massimo Poesio, Alexandra Uma, Maja Pavlovic, Silviu Paun, Paolo Rosso, and Elisabetta Fersini. 2024. Soft metrics for evaluation with disagreements: an assessment. In *3rd Workshop on Perspectivist Approaches to NLP, NLPerspectives 2024*, pages 84–94. European Language Resources Association (ELRA).
- Riikka Rossi and Emily Öhman. 2025. Combining qualitative and computational approaches for literary analysis of finnish novels. *Scandinavian Studies*, 97(3):27–51.
- James A Russell and Albert Mehrabian. 1977. Evidence for a three-factor theory of emotions. *Journal of research in Personality*, 11(3):273–294.
- Carlo Strapparava and Rada Mihalcea. 2010. Annotating and identifying emotions in text. In *Intelligent information access*, pages 21–38. Springer.
- Leon Weber-Genzel, Siyao Peng, Marie-Catherine de Marneffe, and Barbara Plank. 2024. [Varierrnli: Separating annotation error from human label variation](#).
- Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 486–497. Springer.

A. Inter-annotator emotion co-occurrence

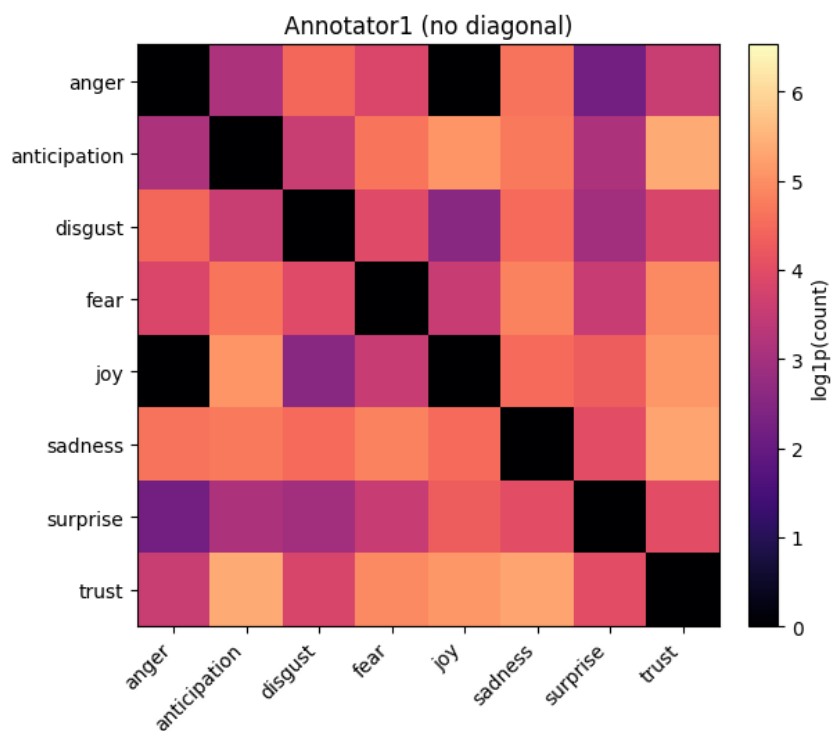


Figure 3: Annotator 1 emotion co-occurrence

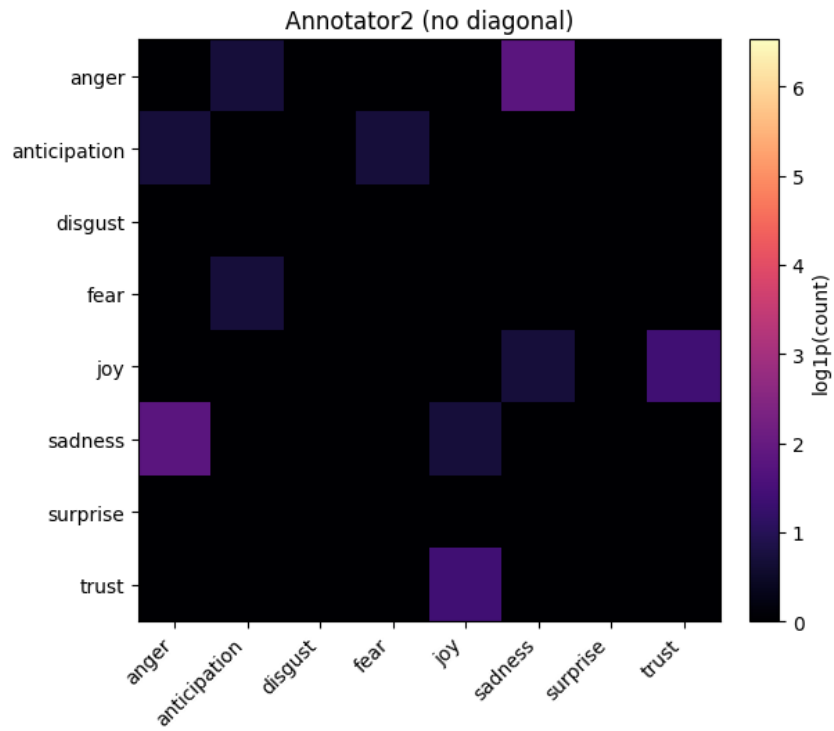


Figure 4: Annotator 2 emotion co-occurrence

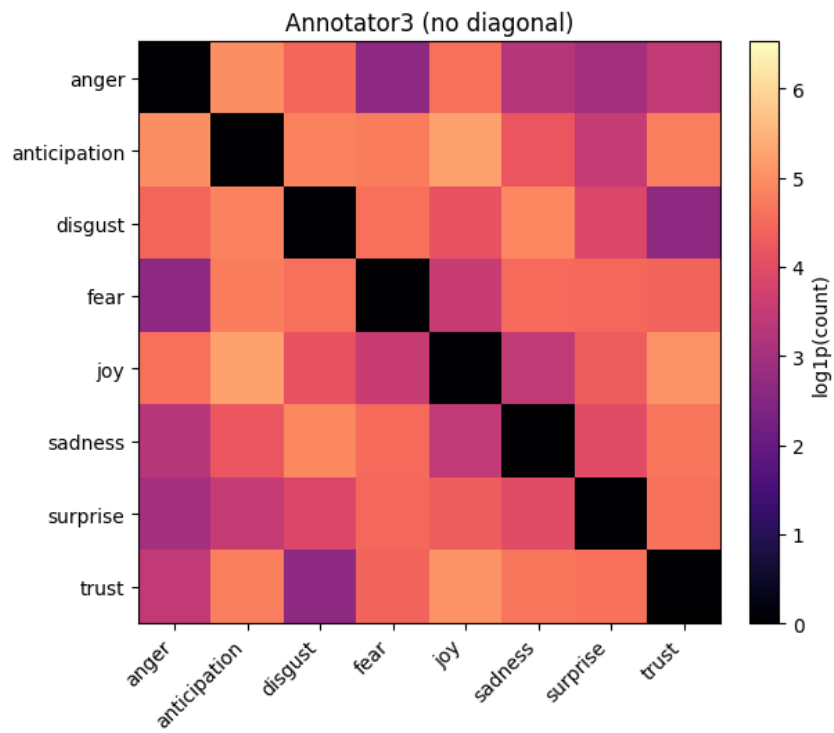


Figure 5: Annotator 3 emotion co-occurrence

B. Top other emotions

Label	Count
sentimentality	6
appreciation	3
unbelief	3
confusion	3
pessimism	2
disapproval	2
contempt	2
mild delight	2
hope	2
mild despair	2
mild contempt	2
shame	2
mild submission	2
determination	2
love	1
optimism	1
caring	1
protectiveness	1
childish exasperation or mild annoyance	1
remorse	1
acceptance	1
submission	1
mild disapproval	1
mild submission and love	1
mild remorse	1
loving pride	1
guilt	1
mild aggression plus optimism	1
suspicion	1
doubtfulness	1

Table 8: Out-of-taxonomy emotion labels provided in the free-text *Other* field (frequency).

Table 8 shows the top 30 most commonly occurring other labels outside of Plutchik's categories.

Emotion pairs
disapproval; contempt
mild delight; hope
sentimentality; pessimism
love; optimism
caring; protectiveness
childish exasperation or mild annoyance
unbelief; confusion
shame, guilt
sentimentality; mild submission
mild aggression plus optimism

Table 9: Top emotion co-occurrences for *[Annotator X]* (off-diagonal counts).

Table 9 shows the top co-occurring other category items

Understanding Irony through Explanations and Background Knowledge

Aaron Maladry, Cynthia Van Hee, Els Lefever, Véronique Hoste

Language and Translation Technology Team (LT3), Ghent University

firstname.lastname@ugent.be

Abstract

This article investigates the automatic explanation of irony in English tweets. The work covers the development and validation of a conceptual framework for annotating knowledge-informed explanations for figurative language as well as the training and evaluation of specialized generative models. Human judgements confirm that both fine-tuned open-source models (Llama 3) and proprietary models (GPT-4) can produce high-quality explanations, effectively incorporating relevant world knowledge. While metrics like BLUE and ROUGE do not seem to align with human judgement, we find that semantic similarity measures align well with human quality estimations. The resulting models and datasets for irony explanations, published as the [iRONNIE](#) collection, actively bridge the gap between theoretical understanding of irony and the technical innovations of the NLP domain. The models are released to the public to facilitate a deeper linguistic analysis of world knowledge involved in understanding irony on social media in future work.

Keywords: Explanation generation, world knowledge, irony, sarcasm

1. Introduction

Irony and sarcasm have long been the subject of investigation within linguistics, pragmatics, and psychology, resulting in a variety of theories that attempt to capture their meaning, function, and impact. Figurative language is historically framed in relation to Grice's maxims of communication ([Grice, 1975](#)), which prescribe that communication should be truthful, informative, relevant, and clear. When these maxims are deliberately and recognizably flouted, speakers indicate that their words are not intended to be taken literally. Metaphors achieve this by describing something in terms of what it is not, thereby bending the maxim of truthfulness. Hyperboles and understatements distort degree: they exaggerate or downplay a characteristic, such as size, beyond realistic measures. Verbal irony differs from these forms in that it relies on incongruity between literal expression and reality, conveying a meaning that often contradicts the surface content ([Sperber and Wilson, 1986](#)). Other forms of irony do not invert meaning in this way, but instead highlight contrasts in different domains. Situational irony arises when real-world events play out in ways that contradict common expectations, such as when a police officer is arrested for theft. The irony lies not in words but in the tension between what is expected and what actually occurs. Dramatic irony is rooted in narrative and performance: it emerges when an audience holds knowledge that characters do not, creating a situation where the audience realizes the tragic implications of a character's actions long before they do.

While situational and dramatic irony are culturally and narratively important, most linguistic and computational research has focused on verbal irony, given its strong departure from literal meaning and its central role in everyday communication. To attain verbal irony, a negative situation can unexpectedly be described as positive, creating an evaluative incongruity (or clash) that listeners interpret as ironic ([Riloff et al., 2013](#)). This form of irony, which can be referred to as "irony by clash" ([Van Hee et al., 2016](#)), is relatively explicit, as both positive and negative sentiments are overtly present. However, irony can also be more subtle. For instance, when a positive sentiment is expressed and the negative nature of the situation must be inferred. In these cases, listeners must rely on their world knowledge and empathic perspective-taking to recognize that the speaker's intended meaning differs from the literal interpretation ([Sperber and Wilson, 1986](#)). The ability to recognize such subtleties aligns irony closely with humour ([Loakman et al., 2023](#)), but also reveals its complexity.

Beyond its linguistic characteristics, irony has been widely studied for its psychological and social functions. While it is often associated with indirect criticism or sarcasm ([Brown and Levinson, 1987](#)), irony can also serve positive communicative goals, such as reinforcing social bonds, mitigating face-threatening acts, defusing tension, expressing empathy, or even functioning as a coping mechanism to process negative emotions ([Pfeifer and Pexman, 2023](#)). Despite the longstanding interest in irony, its psychological, neurological, and linguistic mechanisms remain only partially understood. Re-

searchers have identified typical cues, such as the use of hyperbolic statements (Kreuz and Roberts, 1995), simile (Hao and Veale, 2009), or rhetorical questions (Kreuz et al., 1999), but there is no comprehensive account of the types of knowledge listeners draw upon when interpreting irony.

A central challenge in irony research is the role of world knowledge and contextual assumptions. Theoretical linguistic work emphasizes that recognizing ironic intent often relies on access to shared or mutually assumed knowledge, including real-world facts, cultural conventions, and beliefs about the speaker’s perspective (Lewis, 1969; Schiffer, 1972). Although it is widely acknowledged that understanding irony requires knowledge about real-world situations, conversational norms, and emotional expectations, it remains unclear which types of knowledge are typically invoked. This gap has also received limited attention in computational studies. Although irony has been a popular research topic, the primary focus has been on benchmarking, training systems to detect whether a text contains irony (Saroj and Pal, 2024; Basile et al., 2024).

To advance the study of irony, we explore methods that can scale up the production of meaningful irony explanations, with the final goal to enable further analysis of the world knowledge needed to understand irony. To make this possible, we engage human experts to construct a set of exemplar explanations based on world knowledge, which are subsequently used to train and evaluate generative models. In this study, we evaluate (1) whether generative models are capable of producing knowledge-informed explanations for ironic texts, (2) how these explanations compare to human-produced explanations, and (3) to what extent automatic metrics for explanation quality align with human judgments. All models and data are published as the [IRONNIE](#) collection.

2. Related Work

2.1. Explanation Generation and Reasoning

With the growing success of large generative models, explanation generation and reasoning tasks have become increasingly central to NLP research. While encoder-only models are traditionally used for classification tasks with feature-based explanations, modern methods are shifting towards generative models such as decoder-only architectures that can generate natural language explanations. This trend reflects a broader shift from post-hoc model interpretability methods towards generating explanations that are directly expressed in natural language, making them more accessible to hu-

man users. Importantly, explanation generation and model interpretability, although related, pursue distinct objectives. Model interpretability aims to make the model’s internal mechanisms transparent, while explanation generation focuses on providing human-readable justifications, without necessarily aligning with the true reasoning process of the model. This distinction is crucial for generative models, whose token-by-token generation complicates efforts to study their inner workings.

To mitigate this, researchers are adapting tools from explainable AI such as SHAP (Lundberg and Lee, 2017), to the text generation setting (Enouen et al., 2024). SHAP assigns importance scores to input features (e.g., words) based on how their presence or absence affects model predictions, relying on game-theoretic principles. When applied to autoregressive generation, SHAP can identify the words that have the greatest impact on the generation of specific tokens. However, these token-level importance scores remain superficial: while they indicate *which* words matter, they do not reveal *why* they matter, especially when explanations rely on implicit knowledge or reasoning that is not mentioned explicitly in the input text.

Chain-of-Thought (CoT) prompting (Wei et al., 2022) offers an alternative strategy by asking models to reason step-by-step, decomposing the task into intermediate natural language steps before arriving at a final answer. While CoT improves performance on many classification tasks, it does not guarantee that generated explanations faithfully represent the model’s true decision process (Atanasova et al., 2023; Turpin et al., 2024). This misalignment between explanation and model reasoning has motivated work on faithfulness and consistency metrics (Lanham et al., 2023; Parcalabescu and Frank, 2023). While we recognize that it is important that generated explanations reflect the models internal process, we argue that before concerns of faithfulness arise, models must first generate explanations that are meaningful, informative, and intuitive to human users, particularly when the task relies on world knowledge.

2.2. Explanation Generation for Figurative language

Within the field of figurative language processing, explanation generation is still in its early stages. For figurative language as a whole, Chakrabarty et al. (2022) introduce the FLUTE dataset, which covers metaphors, hyperboles, similes, and irony. Their approach involves constructing paraphrases to expose the intended meaning of figurative expressions and generating explanations framed as entailments or contradictions of these paraphrases via NLI models. However, FLUTE lacks human

standard explanations and applies a generic explanation format, without recognising the role of world knowledge or incorporating its explicit inclusion as a criterion for explanation quality. Subsequent work extends FLUTE to the multimodal domain (Saakyan et al., 2024) incorporating images into figurative language interpretation. Yet, even in this expanded setting, explanations remain limited to validating or contradicting paraphrases without exploring deeper pragmatic or world-knowledge-based reasoning. Focusing on multimodal sarcasm, Desai et al. (2022) produce explanations for multimodal sarcasm using a fine-tuned BART model, and Yi et al. (2025) make use of more recent decoder models in a zero-shot setting, however, both works tend to reduce paraphrases of the intended meaning and do not address the implicit assumptions or background knowledge required to understand sarcasm.

In parallel, explanation generation has also been explored in the context of abusive language detection where retrieval-augmented generation (RAG) methods have been employed to enhance the detection of hate speech, including sarcastic expressions. Di Bonaventura et al. (2025), for instance, combined knowledge from diverse databases to enrich the system input and improve hate speech classification. Notably, the authors demonstrated that instruction-tuned models such as FLAN-T5 (Chung et al., 2022) and Llama 2 (Touvron and et al., 2023) outperform proprietary models like GPT-3.5 Turbo in generating explanations when enriched with world knowledge.

While the aforementioned studies demonstrate the relevance of explanation generation for figurative language, their primary focus lies in creating datasets for benchmarking model performance. A notable limitation is that **current “explanations” are limited to paraphrasing** and do not reveal the underlying reasoning. The main contributions of this work lie in approaching explanation generation for irony as a distinct research objective, **focusing on inferring world knowledge** from large language models to deepen our understanding of figurative speech like irony. With this goal in mind, we develop and release datasets and models to set a new standard for explanation generation for figurative language.

3. Experimental Setup

3.1. Human-Written Irony Explanations

To investigate whether generative models can meaningfully explain ironic texts, we first construct a set of human-authored explanations. These serve a dual purpose in our study. First, they are used to train the models and provide high-

quality, illustrative examples of how world knowledge can be used to explain irony. Second, they serve as a reference point in our evaluation: we compare system-generated explanations against human-written ones in a controlled human evaluation setup. Including this reference point in evaluation enables us to assess not only how well models perform, but also how human explanations are perceived under the same evaluation conditions.

This human reference was created by two annotators who explained the irony in a subset of 130 ironic tweets from the SemEval-2018 Task 3 corpus (Van Hee et al., 2018). Both annotators are trained linguists with a strong command of academic English. The explanations are written in full text, without imposing length limitations. To ensure the completeness of the explanations, annotators were instructed to approach the task as if they were explaining the irony to a child. This encouraged them to break down complex ideas into simple, understandable terms, ensuring that each explanation was comprehensible.

The recognition of irony often hinges on implicit, shared knowledge that is not directly stated in the text. To account for this, we introduced a second step in the annotation process: after writing the initial explanations, annotators were asked to systematically isolate and explicitly itemise the knowledge required for understanding the irony. Example 3.1 illustrates a complete annotation, according to the annotation guidelines described as a technical report (Maladry et al., 2025).

Example 3.1 Ironic tweet: *Looooovveeeeeee when my phone gets wiped*

Explanation: *When your phone gets wiped (which indicates someone did not do it on purpose), you lose all data on your device. This includes a lot of personal information and pictures that people might want to save as keepsakes. As people would not like (accidentally) losing their personal data, the positive evaluation in this tweet is ironic.*

Background knowledge:

1. *When a phone gets wiped, all personal data and information is lost.*
2. *People do not like losing access to their personal data and pictures on their phone.*

Through this two-step annotation process, we collected (1) a set of detailed irony explanations and (2) systematic listings of the relevant world knowledge for each instance. The two independent perspectives for each of the 130 tweets resulted in total in 260 human-written explanations.

3.2. Methodology for Modelling Explanation Generation

For **model development**, we included a proprietary model, GPT-4-turbo (gpt-4-turbo-2024-04-09), and

two open source models: Mixtral (Jiang et al., 2024) and Llama 3 (Dubey, 2024). To establish an upper bound for the performance of generative models, we opted for larger versions of the models (Mixtral 8x7b and Llama 3 70b), as they generally outperform their smaller counterparts. Given the computational cost, restricted access, and the opacity of their internal workings, we did not fine-tune the proprietary models but included three in-context examples in the prompt. Open source models, by contrast, offer greater flexibility in terms of training and fine-tuning and can be used freely by other researchers. This allows us to employ supervised fine-tuning on the models with QLoRA (Hu et al., 2021; Dettmers et al., 2023) and a single in-context example to illustrate the expected output format and clarify the goal of the task. Fine-tuning was performed over 10 epochs, using the following parameters: 128 adapter ranks, an alpha value of 16, with an effective batch size of 4, a learning rate of $5e-5$, and AdamW as the optimiser¹. For the GPT models, we default temperature and context parameters are used to replicate the average user setting.

Early testing indicated that performing both explanation generation and knowledge extraction at once was too demanding for the systems and resulted in too noisy outputs. To counter this issue, we split the annotation into two subtasks: (1) explanation generation (prompts in Appendix A and (2) knowledge extraction (prompts in Appendix B), with equal methodology and parameters. For knowledge extraction (2), the model input includes the ironic tweet with explanation, either human-written or generated in task (1).

For training, the dataset of 130 explained tweets was split into a test set of 80 samples and a train set of 50 samples. We specifically opted for a larger test set – to improve the robustness of our evaluation – and a smaller train set –because we assume that large generative already encoded the relevant information and only need to be fine-tuned to understand the scope of the task and the expected output format. The train set of 50 samples and test set of 80 samples are available at [Amala3/iRONNIE_train](#) and [Amala3/iRONNIE_EN_evalset](#).

3.3. Qualitative Evaluation for Explanation Generation

In our evaluation setup, we conducted a manual evaluation of the explanations provided by the two fine-tuned open source generative models (Llama 3 and Mixtral), the zero-shot GPT-4 Turbo model with in-context examples and the two human explainers. For clarity, we will refer to the individuals

who wrote the human explanations as “explainers” and we will use the same term to describe the generative models tasked with generating explanations. We do not automatically assume that human explanations are better than system-generated, and, therefore, refer to the human-written explanations as human standard instead of gold standard. This allows for consistent terminology across both human and system-generated outputs. For this qualitative assessment, each of the 80 tweets was evaluated by five humans and all five explanations for the tweet were shown simultaneously. For consistency, all five explanations for each tweet were shuffled, anonymised, and underwent minimal manual cleaning before being presented to the annotators.

The **qualitative evaluation framework** consists of three components, based on related work (Desai et al., 2022; Saakyan et al., 2025): (1) indicating for each explanation whether the explanation is **adequate**, meaning it reasonably explains why the text is ironic, (2) indicating whether the explanation is written by a **human or a system** and (3) **ranking** the explanations from **best to worst** based on their personal preferences. Detailed annotation guidelines for explanation evaluation as well as examples are relayed in a technical report (Maladry et al., 2025).

	Krippendorff’s α	Weighting
Adequacy	0.7092	Binary
Human-likeness	0.6662	Binary
Rank Scoring	0.7200	Linear

Table 1: Observed agreement between raters for our three evaluation criteria, along with the weighting schemes used to account for disagreement.

For each of the three evaluation criteria, we first determined the reliability of the evaluating scheme by calculating inter-annotator agreement using Krippendorff’s α (Krippendorff, 2011).² The first two annotation levels, *adequacy* and *human-likeness* are binary annotations for each explanation. For the third annotation level, we consider the *rank* (first, second, third, etc.) as a scoring mechanism, with the highest-ranked explanation receiving 4 points, the second 3, and so on and the lowest-ranked explanation receiving 0 points. If an explanation is ranked first (4 points) by one annotator and second to last (1 point) by another, the label distance for the explanation is 3 points. Relative to the maximum label distance of 4, this means the error for this specific example has a weight of 0.75 (3/4). With five

¹Example scripts for fine-tuning and inference are available on Github at [aMala3/IronyExplanations](#)

²All IAA scores are calculated with the irrCAC (<https://github.com/kgwet/irrCAC>) package for statistical agreement measures in R.

	Adequacy			Human-likeness			Ranking Score		
	Avg.	Δ	Sig.	Avg.	Δ	Sig.	Avg.	Δ	Sig.
GPT-4 Turbo	0.72	+0.11	**	0.05	-0.42	***	2.96	+1.07	***
human2	0.66	+0.06	—	0.47	-0.01	—	2.49	+0.60	***
→ human1	0.60	0.00	—	0.48	0.00	—	1.89	0.00	—
Llama 3	0.53	-0.08	*	0.48	0.00	—	1.53	-0.36	***
Mixtral	0.39	-0.22	***	0.39	-0.08	*	1.14	-0.75	***

Table 2: Evaluation of Adequacy, Human-likeness, and Ranking Score for each explainer. For each criterion, the table displays the average score, the mean difference (Δ) with regards to human1, and significance w.r.t. the human standard. The asterisks mark the degrees of significance, with * indicating $p < 0.05$, ** indicating $p < 0.01$, *** indicating $p < 0.001$ and a dash (—) marks no statistical significance.

evaluations for each of the 400 explanations, the resulting set comprises 2,000 annotations for which we calculated inter-rater agreement. The α -scores (Table 1) indicate that both adequacy and ranking score meet the recommended values (≥ 0.667) for reliability (Krippendorff, 2011), while scores for human-likeness annotations are barely short of this value. Overall, these scores support the **acceptable reliability of the annotation scheme**.

Table 2 reports **results** across three dimensions: adequacy, human-likeness, and ranking for both system-generated and human-written explanations.

For **Adequacy**, no statistically significant differences are found between the two human explainers. In contrast, significant differences emerge when comparing the systems to the human explainers. Whereas GPT-4 scores higher on adequacy than both humans, Llama scores slightly lower. Mixtral explanations exhibit the greatest divergence from human explanations, scoring 22% lower on adequacy compared to human1, a difference that is highly statistically significant.

For **Human-likeness**, we observe no significant difference in scores between the human explainers, importantly, nor between the humans and our fine-tuned Llama 3 model. This suggests that human annotators were unable to distinguish between human-written explanations and those generated by the fine-tuned Llama model. Surprisingly, despite its low adequacy scores, Mixtral is often perceived as human-like, with a difference of only 7% (Δ) and limited statistical significance ($p < 0.05$). In contrast, GPT-4’s explanations are readily identified as non-human. A closer look at the explanation lengths further supports these observations. The average word count for explanations from the fine-tuned Llama 3 (41 words) and Mixtral (43 words) models closely aligns with that for the human explanations, with 45 and 43 words on average. In stark contrast, GPT-4 Turbo’s explanations -averaging 105 words- were exceedingly longer than typical

human responses and were almost always recognized as non-human.

For the **Rank** scores, we find that all score differences are highly statistically significant and that the explanations of GPT-4 were often preferred over the explanations written by humans. Among the human annotators, the explanations written by human2 seem to be preferred over those by human1. The score differences are smallest between human1 and fine-tuned Llama, demonstrating the similarity of these two explainers.

Although some scores may appear low in absolute terms, with for instance 57.9% of the explanations judged adequate and 37.3% labelled human-like, this reflects the comparative nature of the task. Presenting five explanations simultaneously encourages stricter evaluation, as annotators implicitly contrast responses against one another. This setup may lower raw ratings, especially for human-likeness, where even minor stylistic deviations from human norms stand out. For this reason, we interpret system performance relative to a human expert baseline, rather than in isolation.

In summary, GPT-4 explanations were most preferred and were generally rated as most adequate, while fine-tuned Llama 3 matched human-level performance on human-likeness. Mixtral lagged behind across all dimensions.

3.4. Qualitative Evaluation of Knowledge Extraction

Before evaluating the content of the knowledge extraction task, we inspect the length of the extracted knowledge. For now, we treat all extracted knowledge as a single text, avoiding the need to align individual knowledge items or account for variations in the number of produced knowledge items.

Whereas the Mixtral output for explanation generation could be cleaned manually with relative ease, the output for knowledge extraction was too noisy, inconsistent and repetitive to be used reli-

ably on a large scale. Therefore, we continued the knowledge experiments with Llama 3 and GPT-4³. Table 3 presents the average word counts of background knowledge items extracted from explanations authored by each source (rows), using different extraction systems (columns). For example, 31 indicates the average word count of knowledge items extracted by GPT-4o from explanations written by human1, 19 is the average word count of knowledge items human annotators extracted from the explanations written by human2. Note that there is only one human column because humans only extracted the background knowledge in their own explanation.

	GPT-4o	Llama 3	Human
human1	31	27	22
human2	30	24	19

Table 3: Average word lengths of background knowledge extracted from explanations by different sources (rows), using different extractors (columns).

We observe that GPT-4o consistently produces longer knowledge items, regardless of the explanation source, averaging 31 words for human1 and 30 for human2. The two human explainers, in contrast, extract more concise knowledge (19–22 words), with Llama 3 falling somewhere in between. This mirrors trends in explanation verbosity, where GPT-4o also generated significantly longer responses.

Compared to explanation generation, knowledge extraction may seem less interpretive and more objective, but its evaluation remains challenging. For instance, a single piece of background knowledge can be rephrased as a self-contained statement or split into multiple shorter ones. Similarly, related concepts may be consolidated into a broader knowledge unit. These variations, while often semantically equivalent, can significantly impact automated precision and recall scores.

To assess extraction performance more directly, we conduct a manual validation on a subset of 50 explanations provided by human1, evaluating knowledge extracted by Llama 3 and GPT-4o. For this validation, we identify **whether all essential background knowledge was successfully extracted** from the explanation (in the sense of recall), and **whether all of the produced knowledge was present in the explanation** (in the sense of precision). Importantly, we do not compare extracted knowledge to a predefined human standard.

³As we are now generating text for 400 samples (extracting knowledge from 5 explanations for each of the 80 texts), we opted for GPT-4o instead of GPT-4 Turbo to save computational resources.

Instead, each knowledge item is evaluated independently against its source explanation through manual inspection. As a result, these metrics reflect a looser, more interpretive approximation of traditional precision and recall, which we refer to as soft recall and soft precision. Although some knowledge items are present in the explanations, and are considered correct for soft precision, they still deviate slightly from ideal precision. Therefore, we also analyse two common deviations from ideal precision: **content drift**, where the extracted item differs slightly in meaning (e.g., due to over-generalization or added specificity), and **reasoning**, where the item goes beyond background knowledge to include evaluative or inferential content that should belong in the explanation rather than the extracted knowledge. Example 3.2 shows content drift, where the extracted knowledge is not explicitly present in the explanation, but is implied in the part in bold. Example 3.3 shows a reasoning deviation: the extracted knowledge contains reasoning (in bold), making the contrast explicit between the background knowledge. Although this reasoning should be included in the explanations, it should not be extracted as background knowledge.

Example 3.2

Tweet: *@user I don't recall dalai lama talking about busty girls and celebs but apart from that*

Explanation: ***The Dalai Lama is a religious leader, which are supposed to be moral guides.** Talking about "busty girls" and celebrities is considered rather shallow and would be inappropriate for a religious leader. Therefore, it is highly unlikely that this ever happened.*

Extracted Knowledge:

(1) *The Dalai Lama is expected to maintain a certain level of moral and spiritual integrity.*

Example 3.3

Tweet: *Throwing up on Christmas morning is my ideal way of spending it.*

Explanation: *Throwing up is a very unpleasant experience that is the result of being sick. When people are sick, they likely stay in bed. On Christmas day or morning, people like celebrating instead, which you cannot do when you are sick.*

Extracted Knowledge:

(1) *Christmas is a time for celebration and spending time with others, **rather than being bedridden.***

As shown in Table 4, both systems achieve strong performance. GPT-4o extracts all relevant items (100% recall) with high precision (95.5%), and only 6.3% of its extractions exhibit minor content drifts. Llama 3 achieves slightly lower recall (95.9%) and precision (90.3%), with a higher proportion of meaning shifts (16.1%). While both models occasionally include reasoning content, these cases are generally well grounded in the explanation and do not represent major departures from the task definition.

	Llama 3	GPT-4o
Recall	95.88%	100.00%
Precision	90.29%	95.49%
Content Drift	16.13%	6.30%
Includes Reasoning	10.75%	12.60%
Total Items	103	133

Table 4: Soft recall and soft precision of extracted knowledge items, based on manual evaluation of 50 explanations from human1. Rather than comparing against a predefined human standard list, each extracted item is assessed relative to its source explanation. Content drift and reasoning refer to knowledge items that are considered correct for soft precision, but exhibit deviation from the ideal desired output.

Overall, both Llama 3 and GPT-4o accurately extract relevant background knowledge. While GPT’s outputs are more verbose and sometimes fragmented into smaller components, this verbosity does not indicate hallucination or content drift. Whereas GPT produces slightly more reasoning items, Llama’s extractions remain closer in style and length to human-produced items, but they can exhibit slight content drift.

3.5. Automatic Similarity Measures for Evaluation

While our primary evaluation relies on human judgments, this approach requires two stages of manual annotation: first to produce human-written explanations, and second to evaluate generated (and human) explanations. This process is resource-intensive and limits the number of systems that can be compared, particularly due to the cognitive load on evaluators when reviewing multiple explanations simultaneously. In this section, we explore whether automatic similarity metrics can serve as a viable alternative for future evaluation. If such metrics reliably reflect differences in explanation quality, especially in alignment with human judgments, they could enable more scalable and efficient comparison of additional systems or configurations. To test this, we investigate whether established lexical and semantic similarity metrics can distinguish between stronger and weaker explainers or knowledge extractors.

As **lexical metrics**, we employ traditional BLEU (Papineni et al., 2002) and ROUGE scores (Lin, 2004), which are often used to evaluate the quality of machine translation and summarization. Additionally, we calculate vocabulary similarity using the Jaccard index (intersection

over union) after lemmatizing all words in the explanations. For the semantic metrics, we leverage BERTscore (Zhang et al., 2020), a widely accepted metric for machine translation evaluation that measures semantic similarity at the token level based on contextual embeddings, and SemScore (Aynedtinov and Akbik, 2024), a novel metric for assessing the broader contextual meaning using sentence-level embeddings from transformer models (Reimers and Gurevych, 2019).

As outlined in our annotation guidelines, human explainers wrote explanations that included both the necessary background knowledge and the reasoning steps they deemed necessary to understand the irony in a text. In the second annotation step, they separated the knowledge from the reasoning process. For this similarity analysis, we first examine the similarities between the full explanations and then focus on the isolated background knowledge.

When **comparing human and generated explanations** (Figure 1), none of the three metrics for lexical similarity (BLEU, ROUGE and Jaccard vocabulary similarity) exceed 20%, indicating that the word choices vary substantially between explanations. This variation is observed not only in the machine-generated explanations, but also holds true for the “inter-human” comparison of the two human explanations. Despite these lexical differences, the BERTscores indicate a higher degree of token-level semantic similarity, suggesting that, even though different words are used, they are semantically aligned. Similarly, SemScores confirm that sentence-level meaning also remains consistent across human and machine explanations.

For the automatic **evaluation of knowledge extraction**, we focus on the two best-performing models, Llama and GPT-4o, as knowledge extracted by Mixtral required too much manual cleaning for this task. With these models, we automatically extracted knowledge from the explanations of human 1 and human 2, and calculated the similarity compared to the human-identified knowledge. Since both human explainers only extracted knowledge from their own explanations, we are unable to present inter-human similarities for this task. However, this comparison enables us to investigate how well the models extract knowledge from the same explanation.

	BLEU	ROUGE	Jacc.	BERT	SEM
GPT	0.27	0.36	0.33	0.81	0.67
Llama	0.37	0.57	0.47	0.87	0.85

Table 5: Similarity scores between human-written knowledge and extracted knowledge using GPT-4o and Llama 3, compared to the human1 standard.

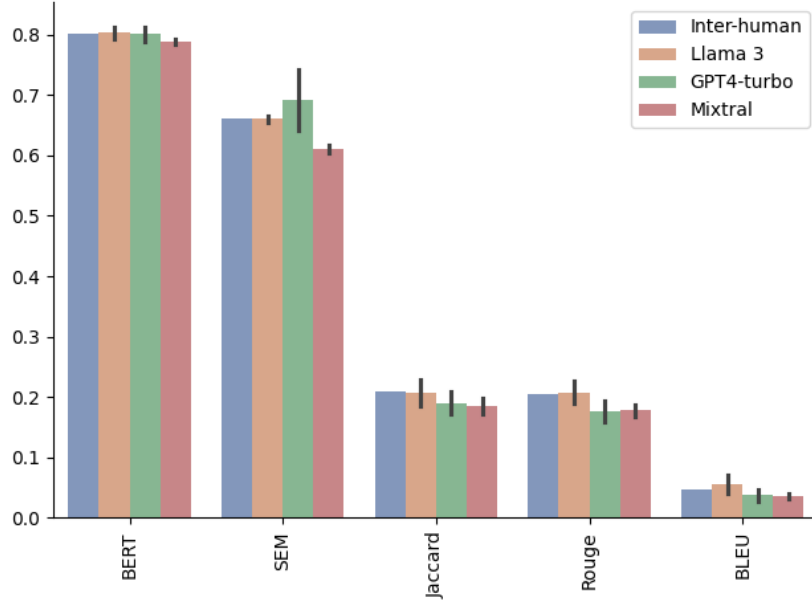


Figure 1: Full explanation similarities for two humans and all systems. Similarities were calculated for both annotators individually and then averaged.

As shown in Table 5, Llama 3 outperforms GPT-4 on all similarity metrics, both lexical and semantic measures. Compared to the task of explanation generation, the similarities observed through the lexical measures (BLEU, ROUGE, Jaccard vocabulary similarity) are much higher for knowledge extraction, suggesting that the phrasing is generally quite similar. Although manual evaluation of GPT-4o indicated that the model could extract a lot of relevant knowledge, the semantic similarity scores are quite low for the model. We hypothesize that this could be connected to the fact that GPT-4 included more reasoning in its knowledge items. Although Llama produced more content drift, the semantic similarities do not seem to capture this difference in nuance.

4. Conclusion

In this study, we investigate how well automatic systems can explain irony in English tweets and whether or not language models can produce explanations that contain relevant world knowledge that is not explicitly mentioned in the input text. In doing so, this investigation evaluates explanation quality, i.e. whether generative language model can provide meaningful explanations that are intuitive to humans and align with linguistic theory about irony, as opposed to evaluating the faithfulness of explanations, which we leave for future work. The models and datasets produced for this work are available on Hugging Face as the [iRON-NIE](#) collection. Based on our analysis, we present the following insights:

- (1) Our annotation scheme, based on adequacy, human-likeness, and full-scale ranking, provides a **consistent and reliable framework for evaluating explanation quality** in the context of irony detection.
- (2) **GPT-4 generates explanations that are preferred over human explanations**, even though the extensive explanation length makes the model easily discernable from human explanations.
- (3) With appropriate fine-tuning, **Llama 3 produces explanations that are often indistinguishable from human-written ones**, although they are rated slightly lower in adequacy and ranking. **Semantic similarity** metrics, particularly SemScore, **seem to align with human judgments**, offering a promising avenue for automatic evaluation of explanation quality.
- (4) Fine-tuned open source and proprietary models are capable of reliably extracting world knowledge for explaining irony, demonstrating that generative **AI can make relevant world knowledge explicit**.

5. Acknowledgements

This work was supported by Ghent University under grant BOF.24Y.2021.0019.01. The computational resources (Stevin Supercomputer Infrastructure) and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by Ghent University, FWO and the Flemish Government – department EWI.

Limitations

This study has several limitations that should be acknowledged. First, the human evaluation relies on a limited number of annotators and was conducted on modest test sets. While we adopted rigorous annotation procedures and achieved strong inter-annotator agreement, a broader evaluation effort would be needed to confirm the generalizability of our findings across diverse datasets, user groups and languages. It is widely known that generative models perform significantly better for English than they do on other languages, so even for the same task, future work should explore if model capabilities transfer well across languages. This is particularly relevant when analysing the role of world knowledge, which is captured implicitly and may require a vast amount of data to be learned.

While this study finds that automatic evaluation metrics align with human judgement, this finding should be further investigated before it can be confirmed. Similarly as the results for explanation generation, whether or not these metrics can also be used to evaluate languages other than English remains an important research question.

Our aim in this work was not to exhaustively optimize every model configuration, but rather to evaluate the capabilities of current models, particularly large language models, for the task of irony explanation and knowledge extraction. As such, we focused on representative configurations rather than fine-tuning all hyper-parameters. More extensive optimization, especially for smaller open-source models, is complicated by the fact that automatic evaluation metrics do not consistently align with human judgments, making it difficult to systematically assess model improvements without costly manual validation.

6. Bibliographical References

Pepa Atanasova, Oana-Maria Camburu, Christina Lioma, Thomas Lukasiewicz, Jakob Grue Simonsen, and Isabelle Augenstein. 2023. [Faithfulness tests for natural language explanations](#). *ArXiv*, abs/2305.18029.

Ansar Aynedinov and Alan Akbik. 2024. [Semscore: Automated evaluation of instruction-tuned llms based on semantic textual similarity](#).

Valerio Basile, Silvia Casola, Simona Frenda, and Soda Marem Lo. 2024. Perseid-perspectivist irony detection: A calamita challenge.

Penelope Brown and Stephen C Levinson. 1987. *Politeness: Some universals in language usage*, volume 4. Cambridge university press.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. [FLUTE: Figurative language understanding through textual explanations](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. [Scaling instruction-finetuned language models](#).

Poorav Desai, Tanmoy Chakraborty, and Md Shad Akhtar. 2022. Nice perfume. how long did you marinate in it? multimodal sarcasm explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10563–10571.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [Qlora: Efficient finetuning of quantized llms](#).

Chiara Di Bonaventura, Lucia Siciliani, Pierpaolo Basile, Albert Merono Penuela, and Barbara McGillivray. 2025. [From detection to explanation: Effective learning strategies for LLMs in online abusive language research](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2067–2084, Abu Dhabi, UAE. Association for Computational Linguistics.

et al. Dubey, Abhimanyu. 2024. [The llama 3 herd of models](#).

James Enouen, Hootan Nakhost, Sayna Ebrahimi, Sercan Arik, Yan Liu, and Tomas Pfister. 2024. [TextGenSHAP: Scalable post-hoc explanations in text generation with long documents](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13984–14011, Bangkok, Thailand. Association for Computational Linguistics.

H. P. Grice. 1975. [Logic and Conversation](#), pages 41–58. Brill, Leiden, The Netherlands.

Yanfen Hao and Tony Veale. 2009. Support structures for linguistic creativity: A computational

- analysis of creative irony in similes. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 31.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#).
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L  lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th  ophile Gervet, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2024. [Mixtral of experts](#).
- Roger J Kreuz, Max A Kassler, Lori Coppenrath, and Bonnie McLain Allen. 1999. Tag questions and common ground effects in the perception of verbal irony. *Journal of Pragmatics*, 31(12):1685–1700.
- Roger J Kreuz and Richard M Roberts. 1995. Two cues for verbal irony: Hyperbole and the ironic tone of voice. *Metaphor and symbol*, 10(1):21–31.
- Klaus Krippendorff. 2011. Computing krippendorff’s alpha-reliability. *Computing*, 1:25–2011.
- Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson E. Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, John Kernion, Kamil.e Lukovsiut.e, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Tom Henighan, Timothy D. Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Janina Brauner, Sam Bowman, and Ethan Perez. 2023. [Measuring faithfulness in chain-of-thought reasoning](#). *ArXiv*, abs/2307.13702.
- David Kellogg Lewis. 1969. *Convention: A Philosophical Study*. Wiley-Blackwell, Cambridge, MA, USA.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Tyler Loakman, Aaron Maladry, and Chenghua Lin. 2023. [The iron\(ic\) melting pot: Reviewing human evaluation in humour, irony and sarcasm generation](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6676–6689, Singapore. Association for Computational Linguistics.
- Scott M. Lundberg and Su-In Lee. 2017. [A unified approach to interpreting model predictions](#). In *Neural Information Processing Systems*.
- Aaron Maladry, Cynthia Van Hee, Els Lefever, and Veronique Hoste. 2025. Guidelines for Writing and Evaluating Explanations for Irony in Tweets.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Letitia Parcalabescu and Anette Frank. 2023. [On measuring faithfulness or self-consistency of natural language explanations](#). In *Annual Meeting of the Association for Computational Linguistics*.
- Valeria Pfeifer and Penny Pexman. 2023. [When it pays to be insincere: On the benefits of verbal irony](#). *Current Directions in Psychological Science*, 33.
- Nils Reimers and Iryna Gurevych. 2019. [Sentencebert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as contrast between a positive sentiment and negative situation. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 704–714.
- Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2024. [Understanding figurative meaning through explainable visual entailment](#).
- Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2025. [Understanding figurative meaning through explainable visual entailment](#).
- Anita Saroj and Sukomal Pal. 2024. Ensemble-based domain adaptation on social media posts for irony detection. *Multimedia Tools and Applications*, 83(8):23249–23268.
- Stephen R Schiffer. 1972. Meaning.
- Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*, volume 142. Harvard University Press Cambridge, MA.

Hugo Touvron and et al. 2023. [Llama 2: Open foundation and fine-tuned chat models](#).

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2024. Language models don't always say what they think: unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2016. Exploring the realization of irony in twitter data. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1794–1799.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. [SemEval-2018 task 3: Irony detection in English tweets](#). In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed H. Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. [Chain of thought prompting elicits reasoning in large language models](#). *ArXiv*, abs/2201.11903.

Peiling Yi, Yuhua Xia, and Yunfei Long. 2025. Irony detection, reasoning and understanding in zero-shot learning. *IEEE Transactions on Artificial Intelligence*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [Bertscore: Evaluating text generation with bert](#). In *International Conference on Learning Representations*.

A. Cleaning Examples

Below, we exemplify the manual cleaning of the generated output of our fine-tuned Mixtral model. We present the output before cleaning in Example A.1, including input, instruction tuning tokens and a hallucinated second tweet with explanation. The same output after cleaning, as presented to the annotators, is shown in Example A.2.

Example A.1 Mixtral output before cleaning:

Text: *I actually just started crying while brushing my teeth due to lack of sleep, stress and test anxiety. Today's gonna be a good one...*

Explanation: [/INST] ### Explanation: *Crying because of stress, lack of sleep and test anxiety are not considered positive. This would imply that this is not going to be a good day.*

[Hallucinated] Text: *25 chance of rain, I love it when my wedding gets ruined!*

Explanation: [/INST] ### Explanation: [/INST] *People want their wedding day to be perfect. Rain can cause damage to the wedding dress and the hair, it can make*

the wedding less enjoyable and even make it impossible to have the wedding outside. As people would not like it when their wedding gets ruined by rain, the positive evaluation is not meant to be serious.

Example A.2 Explanation after manual cleaning: *Crying because of stress, lack of sleep and test anxiety are not considered positive. This would imply that this is not going to be a good day.*

B. Prompt Template for Explanation Generation

GPT-4 Turbo

Your task is to explain why tweets should be considered ironic. Make sure to base your explanation on background knowledge that is not present in the text itself. This background knowledge can include common assumptions, factual knowledge and social conventions.

Here are some examples of the expected output:

Example 1 -

Text: Loooovvveeeeeee when my phone gets wiped

Explanation: When your phone gets wiped (which indicates someone did not do it on purpose), you lose all data on your device. This includes a lot of personal information and pictures that people might want to save as keepsakes. As people would not like (accidentally) losing their personal data, the positive evaluation in this tweet is ironic.

Example 2 -

Text: @user Kohli showing all the guts of Scooby Doo. #CaptainCourageous

Explanation: When people say that someone has "the guts to do something", it means they are fearless, confident and brave. The cartoon character Scooby Doo is known for his cowardice. Saying that someone is showing the guts of Scooby indicates that they are in fact not courageous or fearless at all. The evaluation "#CaptainCourageous" must therefore also be ironic.

Example 3 -

Text: Just received an email that I am the heir to the Nigerian Petroleum Corporation. I should immediately send them all my bank info.

Explanation: The stereotypical phishing email from a supposed "nigerian prince" is a well-known fact. As most people are expected to know about this common phishing template, the users will most likely not follow up by walking into the trap and sending them all their bank info (which is exactly what the writer of this tweet writes they will do exactly).

New sample:
Text: {text placeholder}
Explanation: {system fills in explanation}

Llama 3 & Mixtral

[Human Says][BoS]

You are an expert trained in identifying irony and sarcasm in social media text and explaining the underlying reasoning. Your task is to explain why tweets should be considered ironic. Make sure to base your explanation on background knowledge that is not present in the text itself. This background knowledge can include common assumptions, factual knowledge and social conventions.

Explain why this tweet is ironic:

Text: nit pick at every little thing guys. keep doing that. that'll show the team you're behind them.[EoS]

[System Says:][BoS]

Explanation: nitpicking (pointing out a lot of minor faults) is often perceived as providing too much negative feedback. this does not help to support a team, i.e. showing the team you're behind them.[EoS]

[Human Says:][BoS]

Explain why this tweet is ironic:

Text: {text placeholder} [EoS]

[System Says][BoS]

Explanation: {system fills in explanation}

C. Prompt Templates for Knowledge Extraction

GPT-4o

You will be provided with an ironic tweet along with an explanations for why they are ironic. These explanations are based on extra textual background knowledge. Your task is to extract which knowledge is used in the explanations but which is not present in the text itself. Split up this information into multiple items where this is appropriate. Make sure that the knowledge you provide is mentioned explicitly in the explanation. Here are some examples of the expected output:

Example 1 -

Text: Loooovveeeeeeee when my phone gets wiped

Explanation: When your phone gets wiped (which indicates someone did not do it on purpose), you lose all data on your device. This includes a lot of personal information and pictures that people might want to save as keepsakes. As people would

not like (accidentally) losing their personal data, the positive evaluation in this tweet is ironic.

Background knowledge:

1. When a phone gets wiped, all personal data and information is lost.

2. People do not like losing access to their personal data and pictures on their phone.

Example 2 -

Text: @user Kohli showing all the guts of Scooby Doo. #CaptainCourageous

Explanation: When people say that someone has "the guts to do something", it means they are fearless, confident and brave. The cartoon character Scooby Doo is known for his cowardice. Saying that someone is showing the guts of Scooby indicates that they are in fact not courageous or fearless at all. The evaluation "#CaptainCourageous" must therefore also be ironic.

Background knowledge:

1. When people say that someone has "the guts to do something", it means they are fearless, confident and brave.

2. The cartoon character Scooby Doo is known for his cowardice.

Example 3 -

Text: Just received an email that I am the heir to the Nigerian Petroleum Corporation. I should immediately send them all my bank info.

Explanation: The stereotypical phishing email from a supposed "nigerian prince" is a well-known fact. As most people are expected to know about this common phishing template, the users will most likely not follow up by walking into the trap and sending them all their bank info (which is exactly what the writer of this tweet writes they will do exactly).

Background knowledge:

1. The email from a Nigerian prince or heir is a well-known stereotype of phishing.

New sample:

Text: {text placeholder}

Explanation: {explanation placeholder}

Background knowledge: {system fills in knowledge items}

Open-Source Models

[Human Says][BoS]

For each of the following input texts, we have an explanation for why they contain irony. Your task is to identify and extract any background knowledge used for the explanation.

Extract the background knowledge from this text and explanation pair:

Text: Loooovveeeeeeee when my phone gets

wiped.

Explanation: When your phone gets wiped (which indicates someone did not do it on purpose), you lose all data on your device. This includes a lot of personal information and pictures that people might want to save as keepsakes. As people would not like (accidentally) losing their personal data, the positive evaluation in this tweet is ironic. [EoS]

[System Says][BoS]

Background knowledge:

1. When a phone gets wiped, all personal data and information is lost.
2. People do not like losing access to their personal data and pictures on their phone. [EoS]

[Human Says][BoS]

Extract the background knowledge from this text and explanation pair:

Text: {text placeholder}

Explanation: {explanation placeholder} [EoS]

[System Says][BoS]

Background knowledge: {system fills in knowledge items}

Speech-Based Emotion Recognition and Classification Integrating a CNN and BiLSTM Network

Fatima Uroosa¹, Asim Abbas^{2*}, Muhammad Tayyab Zamir¹, Grigori Sidorov¹

¹Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC),
Av. Juan de Dios Batiz, s/n, Mexico City 07320, Mexico

²School of Computer Science, University of Birmingham, B15 2TT, UK
{furoosa2024, mzamir2023, sidorov}@cic.ipn.mx, axa2233@student.bham.ac.uk

Abstract

Speech emotion recognition (SER) has gained significant interest in recent times, which utilizes speech signals to identify the emotional state of speakers. Accurate recognition of subtle emotional variations in speech, such as distinguishing closely related emotional states, remains a challenging problem due to the variability of speech signals and the acoustic similarity among emotion classes across different speakers and linguistic contexts. This paper proposes a hybrid deep learning model that integrates a Convolutional Neural Network (CNN) with a Bidirectional Long Short-Term Memory (BiLSTM) network to effectively identify both spectral and temporal features of speech. The log Mel-frequency spectral coefficients (MFSC) are used as input features to represent discriminative spectral representations, while the BiLSTM layer model represents long-range temporal dependencies in speech signals. The proposed framework is evaluated on the Toronto Emotional Speech Set (TESS), a publicly available dataset of acted emotional speech containing seven emotion classes. The experimental findings show that the hybrid CNN-BiLSTM achieved an overall classification accuracy of 96.36%, significantly outperforming baseline models including GRU (91.84%), BiLSTM (93.12%), and CNN-GRU (94.67%). These findings highlight the effectiveness of combining spectral and temporal modeling for improved speech emotion recognition performance. Furthermore, our CNN+BiLSTM approach offers a computationally efficient and data-efficient alternative to transformer-based models, while still effectively capturing both spatial and temporal emotional cues in speech, making it suitable for real-time and resource-constrained applications.

Keywords: Speech emotion recognition, Deep Learning, Convolutional Neural Networks, Bidirectional LSTM, Hybrid CNN-BiLSTM Model, Log Mel-Frequency Spectral Coefficients (MFSC)

1. Introduction

Emotions are complex psychological and physiological states of consciousness influenced by factors such as environmental challenges, attitudes, and personality. Emotions play a crucial role in human life and significantly affect our psychological and physical health, as well as decision-making (Lerner et al., 2015), (Gross, 1998). It is important to understand the nature of emotional states, as it helps to assess psychological conditions, such as mental health issues like depression, anxiety, and bipolar disorder, that affect millions of people across the world (World Health Organization, 2025).

Moreover, emotions are multidimensional and are often expressed at the same time in many different ways, including tonal variations of voice, facial expression, language, and other forms of nonverbal communication. Proper understanding of these expressions is critical in the development of human-computer interaction, mental health diagnosis, and security systems. This intermodal manifestation of emotions is a major challenge for computational recognition systems, which require interdisciplinary methods to examine and decipher complex emotional states from different data sources (Dao et al., 2022). Among the different modalities used for emotion recognition, speech has gained significant attention due to its natural availability, low acqui-

sition cost, and rich emotional content. Speech-based emotion recognition is increasingly used across various fields, including healthcare (Xiefeng et al., 2019), Human-Computer Interaction (Alsabhan, 2023), mental health assessment and psychological monitoring (Jordan et al., 2025), and assistive clinical care (Vaida and Huang, 2025). Several recent studies have further highlighted the effectiveness of Speech Emotion Recognition (SER) in real-world applications (Botelho et al., 2024). In general, speech signals encode two types of complementary information: linguistic and paralinguistic. Linguistic information is spoken content, including language structure, accent, and dialect, whereas paralinguistic information conveys cues about the speaker's emotional state, attitude, context, and environmental surroundings. Emotional recognition can thus only be achieved through the usage of paralinguistic attributes, as emotional states are usually conveyed by how something is said rather than what is said. However, emotional expression also differs across languages and cultures and may be subjective; hence, the speech emotion recognition task is challenging (Chen et al., 2021).

Furthermore, the challenge arises from the presence of overlapping or similar emotional states, such as calm vs. neutral and excited vs. happy, which are hard to differentiate using a simple decision boundary. The initial methods were based on

manually created low-level acoustic attributes, such as pitch, energy, and formants, and classical machine learning (ML) classifiers, such as support vector machines (SVM), k-nearest neighbors (k-NN), random forests (RF), and decision trees, yielding moderate performance results (Abdelfattah et al., 2016). However, more recently, deep learning (DL) approaches that learn high-level paralinguistic representations directly from data have shown better performance. Developments in neural structures, such as recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and gated recurrent units (GRUs), have greatly enhanced the ability to model the dynamics of speech signals over time, resulting in more powerful speech emotion recognition systems (Jahangir et al., 2021).

In contrast to prior works that focus on proposing a single deep architecture, this study presents a comprehensive comparative analysis of multiple deep learning models for speech emotion recognition using a unified experimental setup. LSTM, BiLSTM, GRU, CNN, and a hybrid CNN-BiLSTM model are evaluated on the Toronto Emotional Speech Set (TESS) dataset using MFCC features. The results indicate that sequence-based models such as LSTM, BiLSTM, and GRU consistently outperform CNN-only architectures, highlighting the importance of temporal dynamics in emotional speech. The hybrid CNN-BiLSTM model further improves performance by jointly modeling local spectral patterns and long-range temporal dependencies.

1.1. Research Questions

This study aims to investigate the effectiveness of a hybrid CNN-BiLSTM architecture for speech emotion recognition. The following research questions are addressed:

RQ1: How effectively can the proposed CNN-BiLSTM model learn and combine spectral and temporal features for accurate emotion classification.

RQ2: Does the proposed hybrid architecture outperform baseline models such as CNN, LSTM, and GRU in terms of accuracy and F1-score on the TESS dataset.

RQ3: How well does the model distinguish between acoustically similar emotions in speech signals.

RQ4: What is the impact of combining CNN-based feature extraction with BiLSTM-based temporal modeling on overall system performance.

The remaining paper is structured as follows. Section 2 reviews related work and presents the problem statement for speech emotion recognition. Section 3 describes the methodology, including dataset acquisition, preprocessing, feature extraction, and the proposed hybrid CNN-BiLSTM archi-

ture. Section 4 outlines the experiment configuration, metrics of evaluation, and performance analysis. Section 5 presents the results, and Section 6 discusses the limitations of the study. Lastly, Section 7 brings the paper to an end.

2. Related Work

Due to the increasing significance of Speech emotion recognition (SER) in social sciences and biomedical applications over a period of time (Atmaja and Sasou, 2022), there has been a growing interest in the development of numerous DL-based methods in emotion classification (Monisha and Sultana, 2022). Some studies have analyzed the issues of SER and the different model structures to improve recognition performance by utilizing multiple attention mechanisms to combine both audio features (MFCCs) and text embeddings (Zhang, 2023). Their multimodal approach achieved better emotion recognition results than using audio or text alone.

(Kumbhar and Bhandari, 2019) trained an LSTM-based model to perform emotion recognition, using extracted Mel-Frequency Cepstral Coefficients (MFCCs) features of the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song), a benchmark emotional speech dataset containing both speech and song recordings. Their approach achieved an accuracy of 84.81%; however, a high loss value of 67.21% was observed during implementation, highlighting the need to reduce false-positive predictions for improved system performance.

(Zehra et al., 2021) proposed an ensemble-based approach for cross-corpus multilingual emotion recognition using a majority voting strategy. Their study employed speech corpora in Urdu, Italian, English, and German and tested classifiers, including SVM, random forest, and decision trees, and classified the English (TESS) dataset at 70.14% under within-corpus evaluation. (Luna-Jiménez et al., 2021) explored transfer learning for SER by leveraging knowledge from supervised pre-trained models, specifically CNN-14 from the PANNs framework (Kong et al., 2020). Though the identification of the multilingual machine-generated text is done with the help of transformer-based machine learning and multilingual modeling methods, there are trends in efficient representation learning, which are applicable in cross-modal emotion recognition (Abiola et al., 2025).

By fine-tuning this model on the RAVDESS dataset using speech signals, the authors achieved an accuracy of 76.58%. More recently, research in speech processing has increasingly shifted toward self-supervised learning (SSL) approaches that pretrain speech representations directly from

raw audio signals rather than relying solely on supervised learning architectures. When fine-tuned on benchmark datasets, SSL-based models have demonstrated improved performance, particularly in low-resource scenarios. For example, (Gavali and Verma, 2025) reported an accuracy of 81.82% on the RAVDESS dataset by using wav2vec2-xlsr as a feature extractor combined with a multilayer perceptron (MLP), outperforming their earlier supervised CNN- and PANNs-based approaches, which achieved 76.58% accuracy. Even though this previous work dealt with text-based emotion intensity detection, the information on lightweight model design is encouraging to apply comparable efficient designs to speech-based emotion classification (Eyasu et al., 2025).

Recent studies on deep learning SER have further analyzed hybrid convolutional-recurrent architectures to identify both spectral and temporal speech audio patterns. (Liu et al., 2024) developed a CNNLSTM algorithm that includes attention and multi-feature fusion, and demonstrated excellent results with several benchmark corpora, such as SAVEE and CASIA. (Ahmed et al., 2023) Demonstrate an ensemble 1D-CNN-LSTM-GRU architecture with data augmentation, which enhanced the generalization of five standard datasets. (Bhanbhro et al., 2025) compared CNN-LSTM and attention-enhanced CNN-LSTM architectures, reporting competitive results on the RAVDESS dataset. Other studies examined real-time CNN+BiLSTM architecture augmented with vocal features, evaluated on benchmark datasets such as TESS, RAVDESS, and EmoDB (Berlin Database of Emotional Speech), a widely used benchmark dataset for speech emotion recognition. Furthermore, an attention-based CNN-BiLSTM architecture was proposed to capture both local and global contextual speech signals, achieving improved performance on several datasets.

Previous studies have identified several limitations in speech emotion recognition. (Jahangir et al., 2021) highlights challenges such as limited dataset size, reliance on acted emotional speech, and difficulties in modeling temporal dependencies. Similarly, (Singh and Goel, 2022) emphasizes issues related to feature selection, model generalization, and the difficulty of distinguishing acoustically similar emotions. These limitations indicate the need for more robust architectures. In this work, the proposed CNN-BiLSTM model addresses these challenges by effectively combining spectral feature extraction with temporal sequence modeling, leading to improved emotion recognition performance.

2.1. Problem Description

Speech Emotion Recognition (SER) is to identify the speaker’s emotional state based on the speech

signals using paralinguistic information like pitch, energy, spectral patterns, etc. Despite the recent significant progress, SER is still a difficult task because the variability of speech signals is very important, and the acoustic similarity between emotion classes, such as happy and surprise or neutral and calm, is quite significant (Singh and Goel, 2022) (Jahangir et al., 2021). Emotional dynamics are shared over both the local spectral properties and long-range time dynamics that cannot be easily modeled with conventional machine learning methods of handcrafted features and shallow classifiers (Atmaja and Sasou, 2022).

More recent deep learning models have enhanced the SER performance by learning a hierarchical representation of time-frequency directly based on time-frequency features (Jahangir et al., 2021). Convolutional Neural Networks (CNNs) are effective to use in order to capture local spectral patterns, whereas recurrent networks like Long Short-Term Memory (LSTM) networks are designed to capture the temporal dependencies in speech. However, the model’s capacity to completely capture both spectral and temporal information is constrained when either approach is used separately. This encourages the adoption of a hybrid CNN-BiLSTM architecture for robust unimodal speech emotion identification that simultaneously incorporates long-range temporal correlations and local spectral characteristics.

3. Methodology

This section explains the overall workflow of the proposed hybrid CNN-BiLSTM models for speech emotion classification, as shown in Figure 1.

3.1. Dataset

The experiments in this research study are performed on the TESS (Pichora-Fuller and Dupuis, 2020), a publicly available dataset for speech emotion recognition (SER). TESS contains emotional speech samples from two native English speakers recorded in a controlled environment. The dataset consists of seven emotional categories: angry, disgust, fear, happy, sad, surprise, and neutral. Each emotion class contains an equal number of audio samples, making the dataset balanced and suitable for supervised learning. The controlled recording environment and clear articulation of emotions facilitate reliable evaluation of SER models. In this study, TESS is used to train and evaluate the proposed hybrid CNN-BiLSTM framework.

3.2. Data Preprocessing

All audio samples are preprocessed to ensure consistency. Preprocessing steps include normaliza-

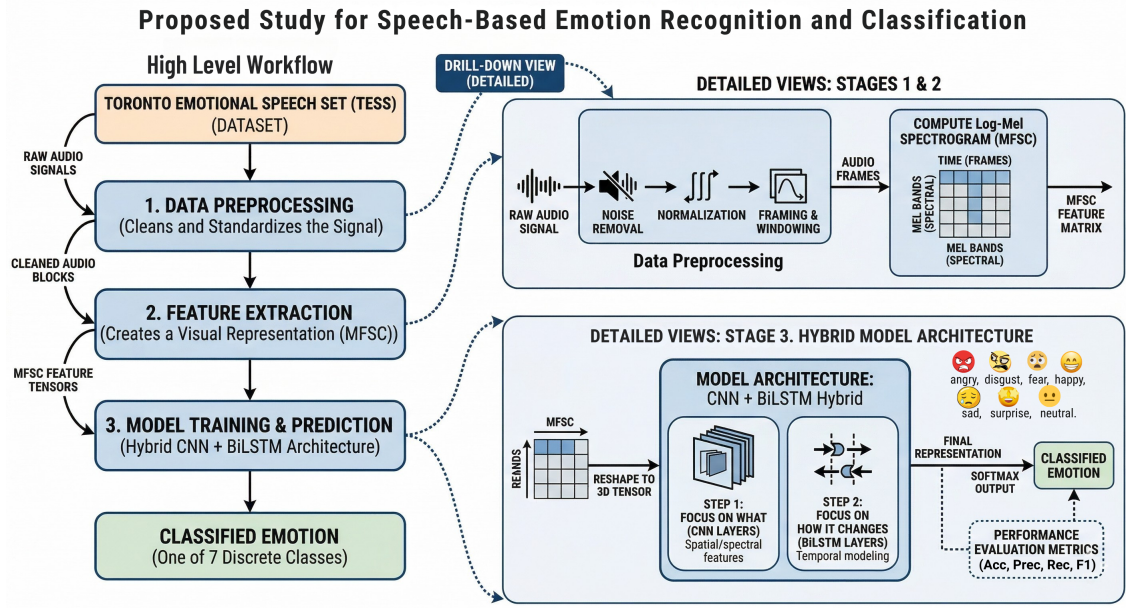


Figure 1: Overall workflow of the proposed CNN–BiLSTM-based speech emotion recognition framework.

tion of audio amplitude, resampling to a standard sampling rate, and trimming of silence at the start and end of recordings. These steps reduce variability and improve model performance.

3.3. Feature Extraction

Log Mel-filterbank spectral coefficients (MFSCs) are extracted from each speech signal and used to train the proposed hybrid CNN–BiLSTM model. MFSC features are chosen because they generally outperform MFCCs in capturing spectral information for emotion recognition. Since speech recordings vary in duration, the extracted feature sequences result in variable-length representations. To ensure a consistent input format for the DL models, zero-padding is applied to all MFSC sequences, producing uniform feature dimensions across samples. A sample speech signal along with its corresponding spectrogram and chromagram are shown in Figures 2a and 2b.

3.4. Proposed Hybrid Deep CNN-BiLSTM Model

To jointly learn spectral and temporal representations, a hybrid CNN–BiLSTM architecture is proposed for speech emotion recognition. Log-Mel spectrogram (MFSC) features extracted from speech signals are first fed into a CNN component, which captures high-level local spectral patterns through convolution and pooling operations.

The output feature maps generated by the CNN are reshaped into a sequential representation and passed to the BiLSTM layer to model temporal dependencies across speech frames. By incorpo-

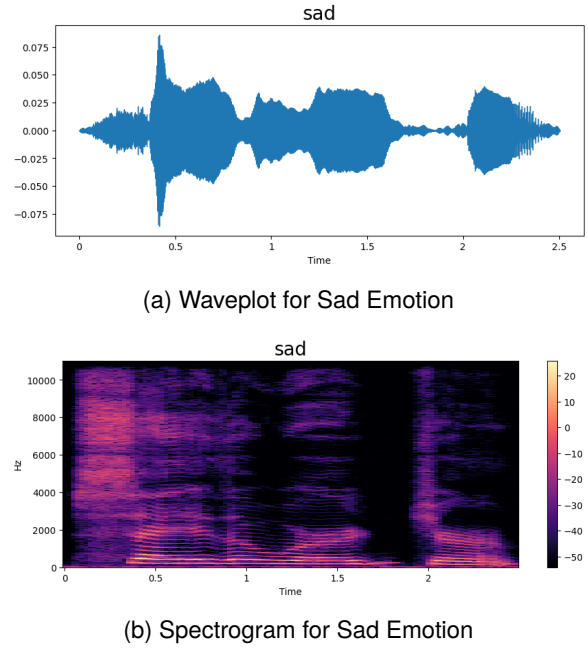


Figure 2: Waveform and spectrogram representations for the sad emotion sample.

rating bidirectional temporal context, the BiLSTM effectively captures both short-term and long-term emotional characteristics of speech.

Finally, the BiLSTM outputs are forwarded to fully connected layers followed by a softmax classifier to predict emotion categories such as *neutral*, *happy*, *sad*, and *angry*. The entire CNN–BiLSTM model is trained end-to-end using categorical cross-entropy loss. The CNN–BiLSTM model is proposed to include a 1D convolutional layer (32 filters) and max-pooling and dropout, a Bi-Directional LSTM (128-dimensional output), and two fully connected lay-

Table 1: Architecture and training hyperparameters of the proposed CNN-BiLSTM model

Component	Configuration
Input Features	Log-Mel Spectrogram (MFSC)
Convolution Layer	Conv1D, 32 filters
Activation Function	ReLU
Pooling Layer	MaxPooling1D
Recurrent Layer	Bidirectional LSTM (64 units per direction)
Dense Layer	64 units (ReLU)
Output Layer	7 units (Softmax)
Dropout Rate	0.5
Optimizer	Adam
Loss Function	Categorical Cross-Entropy
Evaluation Metric	Accuracy
Total Parameters	58,503

ers (64 hidden units) and a 7-class softmax output layer. The model is trained on the Adam optimizer and categorical cross-entropy loss. The number of trainable parameters is 58,503, as shown in Table 1.” The relatively small number of trainable parameters highlights the computational efficiency of the proposed architecture. Compared to deeper or attention-based models, this lightweight design makes the model suitable for real-time applications and deployment in resource-constrained environments.

4. Experimental Setup

All experiments were conducted in a GPU-enabled Google Colab environment. The proposed models were trained and evaluated on an NVIDIA Tesla T4 GPU with 16 GB of GPU memory and 12.7 GB of system RAM. The implementation was carried out in Python using the TensorFlow-Keras framework with CUDA version 12.1 support. This setup enabled efficient training and evaluation of the proposed DL architectures for SER.

The experiments were carried out on two publicly available benchmark datasets: TESS and SAVEE. The TESS dataset consists of 2,800 speech utterances, while the SAVEE dataset contains 480 utterances. The combined dataset was split into training, validation, and testing sets following a ratio consistent with prior work, resulting in 2,296 samples for training, 656 for validation, and 328 for testing.

4.1. Evaluation Metrics and Performance Analysis

The performance of the proposed hybrid CNN-BiLSTM model was evaluated using standard multi-class classification metrics, including accuracy, precision, recall, and F1-score, along with a confusion matrix. These metrics provide a comprehensive assessment of the model’s ability to classify different emotion categories.

The confusion matrix summarizes the classification results by comparing the predicted labels with

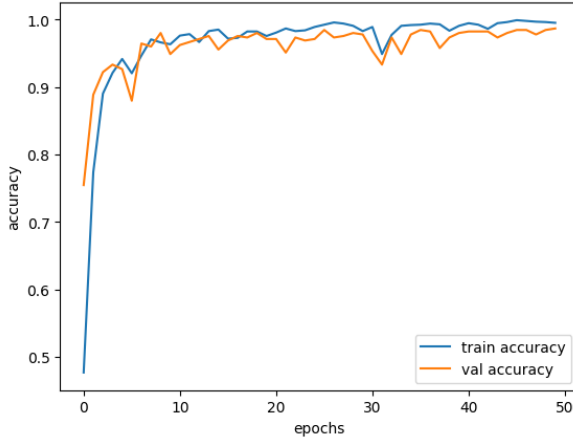
the ground truth labels, providing insights into both correct and incorrect predictions across emotion categories, as shown in 4.

5. Results

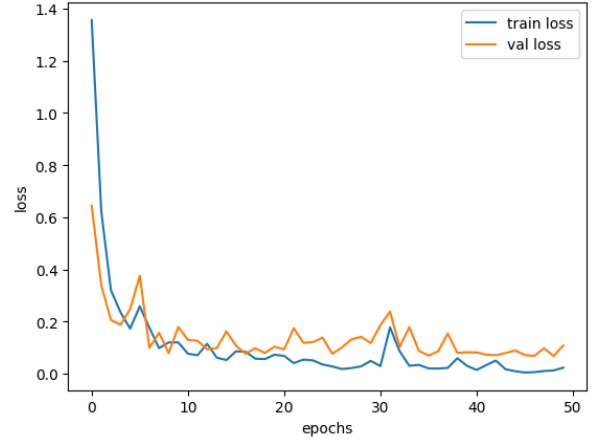
The hybrid CNN-BiLSTM model was trained on the TESS dataset and evaluated on a distinct test set to determine its capability in speech emotion recognition. The analysis focused on conventional classification metrics, including accuracy, precision, recall, and F1 Score. The class-wise performance of the proposed model on the TESS dataset is presented in Table 2. The results indicate that the Recognition performance is high across the majority of emotion categories, which shows the effectiveness of utilizing convolutional feature extraction and bidirectional temporal modeling. Figure 3a shows the accuracy trends, while Figure 3b illustrates the loss convergence of the proposed model. The training and validation curves show stable convergence, with the validation accuracy following the training accuracy over the epochs. The small gap between the curves indicates that the model does not suffer from significant overfitting. Furthermore, the loss in the validation process gradually decreases and stabilizes after convergence, which shows that the spectral and temporal representations learned are effectively being generalized.

To further test the architecture, performance was compared with baseline models, including GRU, BiLSTM, and CNN-GRU. Table 3 summarizes the overall classification performance of each of the models. The CNN-BiLSTM model with the highest accuracy and F1-score across assessment methods indicates the usefulness of combining convolutional feature extraction with bidirectional temporal modeling in speech emotion recognition. The proposed hybrid CNN-BiLSTM model was tested to determine its effectiveness in SER. The high performance of CNN-BiLSTM compared to CNN-GRU and standalone recurrent architectures indicates the significance of the bidirectional temporal modeling in SER. While convolutional layers effectively capture local spectral patterns, the BiLSTM layer facilitates the incorporation of context information of previous and subsequent frames. This complementary combination helps to increase the capacity of the model to identify the subtle emotional patterns that develop over time.

The model attained an overall classification accuracy of **96.36%** and an F1-score of **0.93**, on the TESS dataset, indicating improved performance compared to the baseline architectures. The class-wise analysis suggests that the model performs well, specifically on emotions such as neutral, surprise, and sadness, and is characterized by high precision and recall. Happy and disgust were ob-



(a) Training and validation accuracy



(b) Training and validation loss

Figure 3: Learning curves of the proposed CNN-BiLSTM model showing (a) training and validation accuracy and (b) training and validation loss.

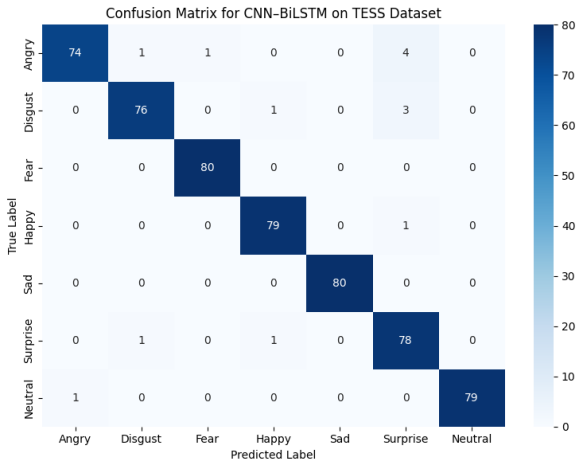


Figure 4: Confusion matrix of the proposed CNN-BiLSTM model on the TESS dataset.

served slightly lower, and this can be attributed to the acoustic similarity of these two emotion classes with other types of emotions, which has been a commonly observed challenge in speech emotion recognition.

In Figure 4, the confusion matrix, it can be observed from the confusion matrix that the number of samples along the diagonal is high, which implies that most of the emotion samples are accurately identified. Misclassifications are minimal and mainly occur with acoustically related emotions, e.g., between anger and surprise or between surprise and disgust. These observations are correlated with the precision and recall rates of the classes as mentioned in Table 2 and further confirm the strength of the proposed CNN-BiLSTM architecture.

Our experimental evaluation confirms that the hybrid CNN-BiLSTM framework is effective in capturing local spectral patterns and long-term temporal

dependencies of speech signals. The model can differentiate between various categories of emotions by using convolutional feature extraction with bidirectional temporal modeling, even when subtle acoustic variations are present. In general, the findings indicate that the suggested method can be considered a valid solution for unimodal speech emotion recognition in a controlled experimental environment.

Emotion	Precision	Recall	F1-score	Acc. (%)
Surprise	0.98	0.98	0.98	96.42
Neutral	1.00	0.94	0.97	99.02
Sad	0.91	0.95	0.93	98.05
Disgust	0.97	0.73	0.84	95.77
Fear	0.87	0.77	0.82	96.10
Angry	0.91	0.73	0.81	93.83
Happy	0.68	0.90	0.78	95.12

Table 2: Class-wise performance metrics of the proposed CNN-BiLSTM model on TESS.

Model	Precision	F1-score	Acc. (%)
GRU	0.88	0.87	91.84
BiLSTM	0.90	0.89	93.12
CNN-GRU	0.91	0.90	94.67
CNN-BiLSTM	0.93	0.93	96.36

Table 3: Performance comparison of different models on the TESS dataset.

The proposed model has good performance in most of the emotion categories, as indicated in 2. The maximum F1-score is achieved with surprise (0.98), then neutral, and sad, which points to the fact that these feelings have their own acoustic peculiarities. Comparatively worse precision is, on the other hand, found with happy, which implies that there is a possible overlap with other acoustically

Table 4: Comparison of Proposed Model with the existing state-of-the-art

Author	Dataset-Language	Methodology used	Accuracy (%)
(Wang et al., 2020)	IEMOCAP-English	Dual-Sequence LSTM	73.3
(Kumaran et al., 2021)	RAVDESS-English	Deep C-RNN	75
(Zhang et al., 2021)	CASIA-Chinese	Parallel Convolution Bi-LSTM	79.67
(Senthilkumar et al., 2022)	EMO-DB-English	Bi-directional LSTM and DBN	85.57
(Muppidi and Radfar, 2021)	EMO-DB-English	QCNN	88.7
(Jahangir et al., 2021)	Benchmark-English	Deep Neural Network	80–90
(Prasanna et al., 2022)	Standard-English	CNN–BiLSTM	~92
Proposed Model	TESS-English	CNN–BiLSTM	96.36

similar classes. The difference among classes indicates the natural complexity of separating the emotions having slight variations in prosodies, which is a frequently described problem in speech emotion recognition.

6. Comparative Analysis with Existing Studies

To present a comprehensive evaluation of the proposed CNN–BiLSTM model on the TESS dataset, we compare our results with existing state-of-the-art speech emotion recognition methods utilize different datasets, as shown in Table 4. The proposed approach achieves an accuracy of 96.36%, outperforming earlier approaches such as Dual LSTM (73.3%) and Deep C-RNN (75%), which are limited in their ability to capture complex spectral–temporal features.

Compared to CNN–BiLSTM-based and hybrid architectures, including those reported by (Zhang et al., 2021), (Senthilkumar et al., 2022), and (Muppidi and Radfar, 2021), which achieve accuracies of 79–88.7%, the proposed model reported improved performance. Additionally, prior studies such as (Jahangir et al., 2021) and (Prasanna et al., 2022) report accuracy of 80–92%, further supporting the competitiveness of the proposed approach.

Overall, the results indicate that the proposed CNN–BiLSTM model effectively captures both spectral and temporal features using TESS dataset only, achieving high accuracy while maintaining a relatively simple and efficient architecture.

7. Limitations of the Proposed Study

Despite achieving strong performance, the proposed CNN–BiLSTM model has certain limitations. First, the experiments are conducted on controlled benchmark datasets, TESS, which contain acted emotions recorded in relatively clean acoustic environments. As a result, the model's generalization to real-world, noisy, and spontaneous speech scenarios remains to be further investigated, as

highlighted in prior studies (Jahangir et al., 2021) (Singh and Goel, 2022).

Second, the class imbalance and limited number of samples for certain emotions may affect recognition performance, particularly for emotions with overlapping acoustic characteristics (e.g., happy vs. surprise or neutral vs. calm), leading to misclassification. Such overlap underscores that it is not easy to differentiate emotions using acoustic characteristics alone, especially when dealing with emotions exhibiting subtle prosodic variations. Furthermore, the suggested framework only pays attention to the unimodal acoustic characteristics and excludes linguistic or multimodal data of a text or facial expression. It has been proven in the past that the performance of emotion recognition can be enhanced further when a variety of modalities is used, particularly in a complex and naturalistic context. The consideration of these limitations through the assessment of the model in the case of spontaneous speech datasets and further development of the framework into multimodal learning is a significant future direction of research.

8. Conclusion

This paper presented a hybrid deep CNN–BiLSTM model for SER, effectively integrating convolutional feature extraction with bidirectional temporal modeling to capture both spectral and temporal characteristics capture both spectral and long-range temporal characteristics of speech signals. Using log-Mel spectrogram and MFSC features, the proposed model was trained and evaluated on the TESS dataset only, achieving consistent performance across all emotion classes.

Experimental results show that the CNN–BiLSTM model outperforms baseline models, including GRU, BiLSTM, and CNN–GRU, in terms of accuracy, precision, and F1-score. The class-wise analysis further indicates that the model demonstrates robustness in distinguishing subtle emotional variations, particularly for emotions with distinct acoustic patterns.

Overall, the study findings prove that utilizing CNN-based spectral feature learning with BiLSTM temporal modeling provides a powerful framework for accurate and reliable SER. Future work will focus on extending this approach to real-world spontaneous speech datasets, addressing class imbalance, and exploring multimodal frameworks that integrate textual and visual cues to further enhance emotion recognition performance.

9. Ethical Consideration

This study used an open-source dataset publicly available on Borealis websites ¹ that do not include personally identifiable information. The data were used intended research purposes.

ChatGPT was used exclusively to refine the English language and presentation of the manuscript. The conceptualization, design, experimental analysis and execution of the study were entirely performed by the authors.

References

- Omar Abdelfattah, George Gal, Gordon W. Roberts, Ishiang Shih, and Yi-Chi Shih. 2016. A top-down design methodology encompassing component variations due to wide-range operation in frequency synthesizer ppls. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 24(6):2050–2061.
- Tolulope Olalekan Abiola, Tewodros Achamaleh Bizuneh, Fatima Uroosa, Nida Hafeez, Grigori Sidorov, Olga Kolesnikova, and Oluvide Ebenezer Ojo. 2025. Cic-nlp at genai detection task 1: Advancing multilingual machine-generated text detection. In *Proceedings of the CIC-NLP Workshop on GenAI Detection*.
- Md Rayhan Ahmed, Salekul Islam, AKM Muza-hidul Islam, and Swakkhar Shatabda. 2023. An ensemble 1d-cnn-lstm-gru model with data augmentation for speech emotion recognition. *Expert Systems with Applications*, 218:119633.
- Waleed Alsabhan. 2023. Human–computer interaction with real-time speech emotion recognition using ensemble techniques. *Sensors*, 23(3):1386.
- Bagus Tris Atmaja and Akira Sasou. 2022. Sentiment analysis and emotion recognition from speech using universal speech representations. *Sensors*, 22(17):6369.
- Jamsher Bhanbhro, Asif Aziz Memon, Bharat Lal, Shahnawaz Talpur, and Madeha Memon. 2025. Speech emotion recognition: Comparative analysis of cnn-lstm and attention-enhanced cnn-lstm models. *Signals*, 6(2):22.
- Catarina Botelho, John Mendonça, Anna Pompili, Tanja Schultz, Alberto Abad, and Isabel Trancoso. 2024. Macro-descriptors for alzheimer’s disease detection using large language models. In *Proceedings of Interspeech*.
- Jing Chen, Chenhui Wang, Kejun Wang, Chaoqun Yin, Cong Zhao, Tao Xu, Xinyi Zhang, Ziqiang Huang, Meichen Liu, and Tao Yang. 2021. Heu emotion: A large-scale database for multimodal emotion recognition in the wild. *Neural Computing and Applications*, 33(14):8669–8685.
- Thi Le Trinh Dao, Xuan T. Le, and Eric Castelli. 2022. Emotional speech recognition using deep neural networks. *Sensors*, 22(4):1414.
- M. Eyasu, W. S. Abebaw, N. Hafeez, F. Uroosa, T. A. Bizuneh, G. Sidorov, and A. Gelbukh. 2025. Tewodros at semeval-2025 task 11: Multilingual emotion intensity detection using small language models. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1485–1494.
- Mrunal Prakash Gavali and Abhishek Verma. 2025. Ensemble of large self-supervised transformers for improving speech emotion recognition. *International Journal of Data Mining, Modelling and Management*, 17(2):217–244.
- James J. Gross. 1998. The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3):271–299.
- Rashid Jahangir, Ying Wah Teh, Faiqa Hanif, and Ghulam Mujtaba. 2021. Deep learning approaches for speech emotion recognition: State of the art and research challenges. *Multimedia Tools and Applications*, 80(16):23745–23812.
- Eric Jordan, Raphaël Terrisse, Valeria Lucarini, Mostasem Alrahabi, Marie-Odile Krebs, Julien Desclés, and Christophe Lemey. 2025. Speech emotion recognition in mental health: Systematic review of voice-based applications. *JMIR Mental Health*, 12(1):e74260.
- Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. 2020. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894.

¹Borealis (TESS) Dataset

- U. Kumaran, S. Radha Rammohan, S. M. Nagarajan, and A. Prathik. 2021. Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep crnn. *International Journal of Speech Technology*, 24(2):303–314.
- Harshawardhan S. Kumbhar and Sheetal U. Bhandari. 2019. Speech emotion recognition using mfcc features and lstm network. In *In Proceedings of ICCUBEA*, pages 1–3.
- Jennifer S. Lerner, Ye Li, Piercarlo Valdesolo, and Karim S. Kassam. 2015. Emotion and decision making. *Annual Review of Psychology*, 66:799–823.
- Yanlin Liu, Aibin Chen, Guoxiong Zhou, Jizheng Yi, Jin Xiang, and Yaru Wang. 2024. Combined cnn-lstm with attention for speech emotion recognition based on feature-level fusion. *Multimedia Tools and Applications*, 83(21):59839–59859.
- Cristina Luna-Jiménez, David Griol, Zoraida Callejas, Ricardo Kleinlein, Juan M. Montero, and Fernando Fernández-Martínez. 2021. Multimodal emotion recognition on ravdess dataset using transfer learning. *Sensors*, 21(22):7665.
- Syeda Tamanna Alam Monisha and Sadia Sultana. 2022. A review of the advancement in speech emotion recognition for indo-aryan and dravidian languages. *Advances in Human-Computer Interaction*, 2022:9602429.
- A. Muppidi and M. Radfar. 2021. Speech emotion recognition using quaternion convolutional neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6309–6313.
- M. Kathleen Pichora-Fuller and Kate Dupuis. 2020. [Toronto emotional speech set \(tess\)](#).
- Y. Lakshmi Prasanna, Y. Tarakaram, Y. Mounika, Suja Palaniswamy, and Susmitha Vekkot. 2022. Comparative deep network analysis of speech emotion recognition models using data augmentation. In *Proceedings of the International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENT-CON)*, pages 185–190.
- N. Senthilkumar, S. Karpakam, M. G. Devi, R. Balakumaresan, and P. Dhilipkumar. 2022. Speech emotion recognition based on bidirectional lstm architecture and deep belief networks. *Materials Today: Proceedings*, 57:2180–2184.
- Youddha Beer Singh and Shivani Goel. 2022. A systematic literature review of speech emotion recognition approaches. *Neurocomputing*, 492:245–263.
- Maria Vaida and Ziyuan Huang. 2025. Multimodal graph neural networks in healthcare: A review of fusion strategies across biomedical domains. *Frontiers in Artificial Intelligence*, 8:1716706.
- Jianyou Wang, Michael Xue, Ryan Culhane, En-mao Diao, Jie Ding, and Vahid Tarokh. 2020. Speech emotion recognition with dual-sequence lstm architecture. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6474–6478.
- World Health Organization. 2025. Over a billion people living with mental health conditions – services require urgent scale-up. Accessed: 2026-03-25.
- Cheng Xiefeng, Yue Wang, Shicheng Dai, Pengjun Zhao, and Qifa Liu. 2019. Heart sound signals can be used for emotion recognition. *Scientific Reports*, 9(1):6486.
- Wisha Zehra, Abdul Rehman Javed, Zunera Jalil, Habib Ullah Khan, and Thippa Reddy Gadekallu. 2021. Cross corpus multi-lingual speech emotion recognition using ensemble learning. *Complex & Intelligent Systems*, 7(4):1845–1854.
- H. Zhang, H. Huang, and H. Han. 2021. A novel heterogeneous parallel convolution bi-lstm for speech emotion recognition. *Applied Sciences*, 11(21):9897.

A Corpus-Based Comparison of two Approaches for Emotion Annotation in French Texts

Valentina Dragos, Delphine Battistelli

ONERA-The French Aerospace Lab, Palaiseau, France

Paris-Nanterre University, Nanterre, France

valentina.dragos@onera.fr, delphine.battistelli@parisnanterre.fr

Abstract

Emotion annotation in texts remains a challenging task in the field of Natural Language Processing (NLP), as, unlike voice or images, texts might not only contain peculiar cues to express emotions. Methods for emotion annotation are based on lexicons or on machine learning techniques which are based on the use of manually annotated corpora. This paper aims to explore if and how the combination of these two types of methods might be useful for the annotation of emotions in texts. Four data sets are used for comparison of the two approaches, and then to investigate to what extent the results are distinct or complementary on three aspects: (i) identification of emotional sentences; (ii) identification of emotion categories; (iii) identification of one specific mode of expression of emotions called "behavioral emotions" (e.g. shout, cry). Findings show that not all emotions are equally easy to annotate, and, most specifically, the learning-based approach tends to over detect *Admiration*.

Keywords: detection of emotions, comparison of approaches, French corpus

1. Introduction

Detecting human emotions from the diversity of social networks' data is a key part of gaining deeper insight into the nature of online content. Emotion detection in textual data is a challenging and tedious task because emotions can be expressed in a variety of subtle and complex ways (Al Maruf et al., 2024), (Gladys and Vetriselvi, 2023). Some emotions are explicitly indicated by specific words such as *sadness* or *joy*, while others are implicitly derived from events (*birthday*) or behaviors (*he cried*). Consider the following examples:

- Je dénonce avec fermeté la colonisation de notre territoire. (I firmly denounce the colonization of our territory) (1).
- Nous nous battons pour bien plus que cela. (We are fighting for much more than that) (2).
- Ebola encore plus proche de français ! Merci qui ? Merci les migrants ! (Ebola ever closer to French! Who do you have to thank for this? Thank you, migrants!) (3)
- CA SUFFIT! (Enough!) (4)

The first example (1) shows a direct and explicit *Anger* emotion expressed by the speaker. The second sentence (2) also illustrates *Anger*, but the emotion is not explicitly expressed, but rather suggested by the willingness to engage in the fight. Moreover, the third example (3) shows a sarcastic sentence conveying the same emotion. *Anger* is also present in the last sentence (4), and in this case the emotion is illustrated by a characteristic of the utterance and the exclamation mark.

Detection of emotions in text has attracted numerous research efforts in the field of Natural Language Processing (NLP). Computational approaches for emotion detection are based on various paradigms, ranging from sentence-level analysis with semantic rules (Seal et al., 2020) to building sophisticated learning models (Chavan et al., 2023). Deep learning techniques (Chutia and Baruah, 2024) and Large Language Models (LLMs) have been used to perform various tasks related to emotion analysis (Zou and Markov, 2024). Despite the prevalence of many techniques, the complexity of emotion expressions in texts has not been fully addressed.

This paper describes the comparison of two parsers developed for emotion annotation in French texts. As part of this study, we examine to what extent the tools provide complementary or divergent results on various data sets collected on the Internet. The experimental protocol defines several emotion annotation tasks, such as detecting emotional sentences and categories of emotions, and then compares the results of both tools on four data sets. The rest of the paper is organized as follows: section 2 discusses related work. Approaches for emotion annotation are discussed in section 3 and corpora used for the comparison are presented in Section 4. The experimental setup and comparison method are introduced in section 5. Section 6 concludes the paper and presents directions for future research.

2. Related approaches

This section gives an overview of the different methods used to detect emotions in texts. It also dis-

cusses research work focusing on annotation disagreement.

2.1. Machine learning and lexicon-based approaches

Emotion annotation is a multi-classification problem (Kusal et al., 2023), and research studies have proposed a variety of lexicon-based techniques and machine learning methods that are commonly used for emotion classification (Chutia and Baruah, 2024).

Lexicon-based techniques are grounded in emotion dictionaries that are built to represent prior knowledge of how emotions are expressed. The literature presents several emotion lexicons, such as SentiWordNet (Baccianella et al., 2010), WordNet-Affect (Poria et al., 2012) and FEEL (Abdaoui et al., 2017), a lexicon built to detect emotion in French. Those techniques rely on recognizing lexicon keywords in a given text and assigning an emotion category based on those keywords and their frequency. Although lexicon-based techniques perform well in detecting explicit emotions, they fail to detect emotions that are not explicitly mentioned in text but rather derived from other clues, such as descriptions of behaviors.

Machine learning models that are trained on sample data sets and make predictions on unseen data afterward are more suitable to detect implicit emotions. Support Vector Machines (SVM) (Tzacheva and Chatterjee, 2024), XGBoost (Bandara et al., 2024) and Naïve Bayes (NB) (Kumar and Kumar, 2022) are some prominent learning techniques adopted to perform emotion detection in texts. The accuracies of supervised learning methods are in the range of 65–80% on benchmark data sets (Al Maruf et al., 2024), although the scarcity of large emotion-oriented labeled data sets is a major limitation of their use. Methods based on transfer learning address these limitations by leveraging the ability of pre-trained language models to detect emotions (Feng and Chaspari, 2020). These methods can be fine-tuned with smaller labeled data sets, and several studies have used adaptations of BERT (Acheampong et al., 2021) or GPT (Lian et al., 2024) to recognize emotions, with detection accuracy reported in the range 75 to 99%. In spite of their high accuracy values, these methods require constrained domain adaptation and demonstrate limited capacities to detect emotions from unforeseen expressions.

2.2. Hybrid approaches

From a different perspective, hybrid methods have also been developed recently, combining heuristics with learning techniques in an effort to improve the accuracy of emotion identification and refine the

categories of emotions detected in texts. Among them, (Li et al., 2021) describes a dual graph convolutional network (DualGCN) model that takes into account the complementarity of syntax structures and semantic correlations simultaneously to capture general semantics. The model detects words that are semantically related, and the evaluation of emotion classification tasks shows levels of accuracy higher than 80%.

A meta-learning approach where predictions from transformers are combined with lexicon features for emotion detection is proposed in (De Bruyne et al., 2021) and the authors demonstrate an improvement on emotion detection compared to a basic BERT model. In (Meškelė and Frasincar, 2020), the authors describe an approach for emotion detection utilizing jointly a lexicalized domain ontology to model field-specific knowledge and BERT for word embeddings. More specifically, the joint utilization provided a mechanism for measuring the influence of words detected in a given sentence on emotion expression and the integration of knowledge improved the overall accuracy of emotion detection. In (Aljedaani et al., 2022), the authors describe a novel hybrid method for emotion detection that combines machine learning and deep learning techniques with a lexicon-based model called TextBlob (Loria et al., 2018). The lexicon is used for data annotation and training and fine-tuning of learning models. The authors investigated the impact of TextBlob on the performance of classification models and showed that the accuracy values of both machine learning and deep learning models are considerably improved when trained using TextBlob annotations, achieving the highest F1 score of 0.96.

Having remarkable capabilities in various natural language understanding and generation tasks, Large Language Models (LLMs) have also been used as annotators (Sun et al., 2023), (Tan et al., 2024) and they have demonstrated good potential for sample annotation (Gilardi et al., 2023) and excellent abilities to annotate data with a limited and well-defined set of labels (Ding et al., 2022).

2.3. Analysis of annotation disagreement

Although detection methods have specific detection capabilities, the annotation schemes they rely upon have an impact on the quality of results, and sometimes induce discordant emotion annotations (Beck, 2023). Recently, several studies started to investigate the factors that potentially cause disagreements when performing emotion annotation from texts. Among them, (Labat et al., 2022) designed an experiment intended to scrutinize the variation in the expression and annotation of emotions retrieved in human-computer conversations. The authors show that using annotation schema

having fine-grained emotion categories leads to low values of annotator agreement. Following a similar approach, (Barz et al., 2025) analyzed the factors of disagreement when annotating emotions in communications about the environment and climate change. Their findings show that two prominent factors of disagreement are the lexical biases of annotators and the association of an emotion to several domain-specific topics.

The work presented in this paper contributes to this last research direction and the goal of our study is to understand, thanks to a qualitative and quantitative comparison, the disagreements of annotations provided by two automatic approaches on four data sets collected on the Internet. In comparison to related works investigating disagreements on emotion annotation, this study is novel in using automated emotion annotation and considering the detection of emotion categories.

3. Automatic approaches for emotion annotation

This section presents EMOTYC and EMOTIONS, and indicates the emotion detection paradigm, the underlying emotion annotation schema and annotation examples.

3.1. EMOTYC: a learning-based parser for emotion annotation

EMOTYC (Étienne et al., 2024) is an emotion detection parser based on a classifier previously trained to detect emotions in texts. EMOTYC was developed using the CamemBERT model (Martin et al., 2019). The annotation schema used by EMOTYC is based on linguistic and psycholinguistic approaches that describe emotions by considering their type, mode, and category, as shown in Fig. 1.

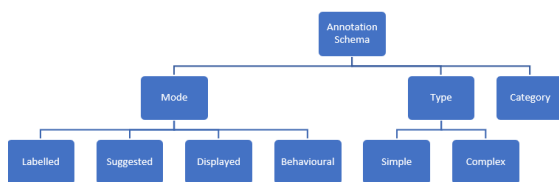


Figure 1: EMOTYC anotation schema

EMOTYC performs the following tasks for emotion analysis: (i) detection of sentences containing emotions; (ii) identification of the emotion mode; (iii) recognition of basic or complex emotions; and (iv) identification of the category of emotion.

The detection approach used by EMOTYC categorizes linguistic markers into four emotion expression modes:

- **labeled** : explicit emotion expressed by lexical, syntactical or typographical clues clearly showing that the utterance conveys an emotion;
- **displayed**: description of situations that are usually associated with emotions;
- **behavioral** : description of behaviors that are usually associated with emotions;
- **suggested** : emotions that are indirectly indicated or implied, often from the context.

EMOTYC identifies the *Type* (*Basic* or *Complex*) and distinguishes eleven categories of emotions, comprising the 6 basic emotions introduced by (Ekman, 1992) (*Anger*, *Fear*, *Happiness*, *Surprise*, *Disgust* and *Sadness*) and 4 complex emotions taken from (Blanc and Quenette, 2017) and (Davidson, 2006) (*Guilt*, *Embarrassment*, *Pride* and *Jealousy*). A fifth complex category (*Admiration*) was also added in order to better balance basic and complex emotions. The category *Autre* (*Other*) is used to capture emotions that are not related to any of the previous categories. Fig. 2 shows an example of annotation provided by EMOTYC, indicating an emotional sentence, the type (basic) and the category of emotion (Anger).

Sentence : il ne profite qu’aux multinationales qui important de la main d’œuvre bon marche et délocalisent à l’intérieur comme à l’extérieur du pays.
(It only benefits multinationals that import cheap labor and relocate it both within and outside the country.)
Emotional sentence: YES
Basic emotion: YES
Emotion category: Anger

Figure 2: Example of EMOTYC annotation

Although EMOTYC detects emotional sentences, the tool is not able to identify the tokens triggering the emotions within the sentence.

3.2. EMOTIONS: a lexicon-based parser for emotion annotation

EMOTIONS (Etienne, 2023) is an emotion annotation parser using an augmented version of the EMOTAIX emotion lexicon (Piolat and Bannour, 2009). The lexicon serves as a basis for recognizing terms from different emotional categories.

EMOTAIX models 24 categories of emotions, including the 11 categories of EMOTYC, 12 additional categories (*love*, *boldness*, *resentment*, etc.) and a behavior category that indicates emotional behavior, without a specific emotion category, such as *cry*, that can refer to both *sadness* or *happiness*.

The annotation schema used by EMOTIONS is based on the EMOTAIX lexicon, see Fig. 3.

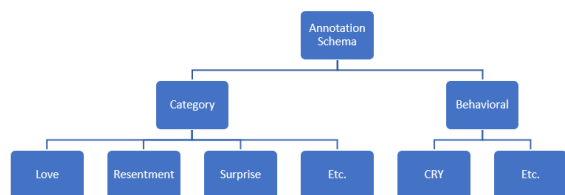


Figure 3: EMOTIONS annotation schema

In order to detect emotions, EMOTIONS performs tokenization and lemmatization of the sentences. Whenever a sentence contains a lexicon lemma, EMOTIONS identifies the emotion associated to this lemma in the EMOTAIX lexicon and thus detects the emotions and its trigger in the initial text. By following this lexicon-based approach, EMOTIONS performs the following tasks for emotion analysis: (i) detection of sentences containing emotions; (ii) identification of behavioral emotions; (iii) identification of the category of an emotion, and (iv) detection of tokens triggering emotions.

Fig. 4 shows an example of an EMOTIONS annotation.

Sentence : Le contexte social, économique et politique actuel, bien plus oppressant dans nos villes que dans nos campagnes, laisse clairement percevoir toute la fragilité du monde moderne. (The current social, economic and political context, much more oppressive in our cities than in the countryside, clearly reveals the fragility of the modern world.)
Tokens having emotions: oppressant (oppressive)
Category of emotions: déplaisir (discontent)

Figure 4: Example of EMOTIONS annotation

4. Overview of corpora

Data sets used for this study have been collected online and have different topics and volumes.

Corpus C1 (right-wing extremism). This data set was created by a project investigating the nature of extremism online¹. The corpus is composed of Tweets and messages collected on social platforms and online discussion forums (Dragos et al., 2022). Data was collected by using a combination of hashtags and keywords that are specific to right-wing extremism. The corpus was manually explored by two experts in sociology (one is a senior researcher in education sciences and the other one is a post doc in sociology with a background in social communication) in order to validate the content. Thanks to this expert validation, the corpus the sentenced

conveying extremist attitudes have been identified and the corpus was annotated with *extremist* and *nonextremist* labels. The data set comprises 638 extremist sentences and 792 neutral sentences.

Sentences (E1) and (E2) show examples of sentences that convey extremist attitudes.

E1: *Ils veulent que vous restez pauvres.* (They want you to stay poor.)

E2: *La France doit rester la France, notre patrie sacrée.* (France must remain France, our sacred homeland.)

Corpus C2 (hateful content). This data set comprises around 600 Tweets manually collected to highlight hateful and non hateful attitude (Battistelli et al., 2020). The main goal of this resource was to facilitate the binary classification of hateful/ non hateful Tweets and thus the corpus is annotated with hateful/ non-hateful labels. The annotation was carried out manually by taking into account the main characteristics of online hateful content, as highlighted by definitions largely adopted in the literature (Malecki et al., 2021). This data set is well balanced, and the two classes (hateful and non-hateful) contain almost the same number of tweets.

The following examples illustrate hateful (E6) and non hateful (E5) tweets:

E3: *Quand il s'agit d'entretenir les étrangers parasites on est toujours sûr de trouver les gauchistes.* (When it comes to nurture the parasitic foreigners we are always sure to find the leftists.)

E4: *J'espère que tu vas avoir un cancer et mourir!* (I hope you will get cancer and die!)

Corpus C3 (TextToKids). This data set comprises 2506 sentences randomly extracted from the corpus used to train the EMOTYC parser. The corpus is composed of articles from the childrens' newspaper Le P'tit Libé². The following examples illustrate emotional sentences from this data set.

E5: *Pourquoi je les aime autant ? Ils me donnent l'impression de jouer toute la journée.* (Why do I love them so much? They make me feel like I'm playing all day.)

E6: *On est très loin de voir les profs être remplacés par des androides !* (We are very far from seeing teachers being replaced by androids!)

Corpus C4 (Reddit) comprises 20.000 posts collected on the Reddit platform. The corpus was created by the Online European Hate Lab³, using specific words and terms to collect toxic content in French subreddits. For the purposes of this study 2,006 sentences have been randomly selected. Two toxic sentences are shown below:

¹<https://anr.fr/Projet-ANR-19-ASTR-0012>

²<https://ptitlibe.liberation.fr/>

³<https://europeanonlinehatelab.com/>

E7: *Les post-fascistes sont un symptôme, pas l'origine du problème. (Post-fascists are a symptom, not the origin of the problem.)*

E8: *Il est meilleur mathématicien que politique. (He is a better mathematician than politician.)*

Remarks: All data sets contain data that was collected online in the framework of different projects and by different research teams.

Corpus	Nature	Balanced
C1	Extremist vs. NonExtremist	NON
C2	Hateful vs. NonHateful	YES
C3	No classes	-
C4	No classes	-

Table 1: Characteristics of data sets

C1, C2 corpora have contrasted classes, highlighting extremist vs. non-extremist content and hateful vs. non-hateful content, respectively. C3 and C4 data set are homogeneous, non-contrasted collections of texts, see tab. 1. Those data sets have been selected for their potential to convey emotions, and also for their differences in nature, size and structure.

5. Comparison of EMOTYC and EMOTIONS

This section assesses the potential for EMOTYC and EMOTIONS to agree on emotion annotation and highlights the differences between their results.

5.1. Experimental design

The experimental protocol adopted to compare EMOTYC and EMOTIONS consists of defining three tasks to be performed by the two parsers and then analyzing the set of emotion annotations they provided. The tasks are intended to:

- Detecting sentences conveying emotions (T1);
- Detecting behavioral emotions (T2);
- Detecting categories of emotions (T3).

As some emotion categories detected by EMOTYC and EMOTIONS do not overlap, a mapping was established in order to define their correspondences, as shown in tab. 2.

For all tasks of the experimental protocol, annotations provided by EMOTYC and EMOTIONS are compared with a focus on their disagreements. Fig. 5 presents an overview of the general architecture of the experiment.

If we consider $a1$ and $a2$ as the sets of anchors annotated by EMOTYC (as the A1 annotator) and EMOTIONS (as the A2 annotator), then the disagreement of annotators can be captured by the

EMOTYC	EMOTIONS
Colère (Anger)	Colère (Anger), Déplaisir (Displeasure)
Autre (Other)	Amour (Love), Apaisement (Alleviation), Audace (Temerity), Désir (Desire), Empathie (Empathy), Impassibilité (Impassivity), Humanité (Humanity), Mépris (Contempt), Ressentiment (Resentment)
Fierté (Pride)	Fierté (Pride), Orgueil (Vanity)

Table 2: Mapping of categories detected by EMOTYC and EMOTIONS

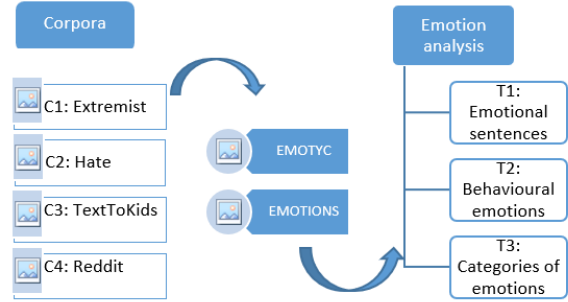


Figure 5: General architecture for comparison

Recall measure, where the Recall of A2 with respect to A1 is given by:

$$Recall(A2||A1) = \frac{|a1 \cap a2|}{|a2|} \quad (1)$$

Recall values range from 0, when the intersection of the two sets of anchors is void, to 1, when the two ensembles are identical.

The results of experiments are described hereafter.

5.2. Experiments and analysis of results

Experiments were carried out using the data sets presented in section 4.

Corpus of right-wing extremism (C1):(tab. 3 and 4).

Label	Recall(A2 A1)	Recall(A1 A2)
Emotional	49.72	46.63
Non Emotional	87.52	89.88
Behavior	60.00	9.37

Table 3: Corpus C1: Results of tasks T1 and T2

Task T1: the two parsers identify a similar number of emotional sentences, and the results are

identical for half of them, as the values of Recall(A1||A2) and Recall(A2||A1) are very close for both the detection of emotional and non-emotional sentences.

Task T2: compared to EMOTYC, EMOTIONS identifies a limited number of behavioral emotions. This outcome is aligned with previous findings described in (Etienne, 2023), where the authors also point out that the lexicon used by EMOTIONS has a limited number of terms related to behavioral emotions.

Label	Recall(A2 A1)	Recall (A1 A2)
Anger	72.72	12.69
Joy	80.00	28.57
Admiration	0.00	0.00
Sadness	40.00	25.00
Shame	100	50.00
Fear	55.17	55.17
Pride	40.00	100
Surprise	50.00	25.00
Other	1.00	100

Table 4: Corpus C1: Results of task T3

Task 3: Comparison of results for the identification of emotions' categories shows that the two processors agree well when it comes to the detection of *Fear* category. However, there is no agreement between EMOTYC and EMOTIONS on the detection of *Admiration*. In-depth analysis of results indicates that the category is correctly detected by EMOTIONS. For example, the sentence *Gloire à ceux qui luttent à l'Est ! (Glory to those who fight in the East!)* is correctly annotated as *Admiration* by EMOTIONS, although the processor is wrong when assigning the same annotation to the sentence *Nous ne sommes pas un parti politique et nous estimons que ce sont aux gens qui nous gouvernent d'assurer notre sécurité, travail pour lequel ils sont grassement payés. (We are not a political party and we believe that it is up to the people who govern us to ensure that we are safe, a job for which they are well paid.)*. For the *Admiration* category, EMOTYC performs rather poorly.

Corpus of hate content (C2): (tab. 5 and 6).

Task T1: Values of the Recall indicate a strong disagreement between EMOTYC and EMOTIONS on the detection of emotional and non-emotional sentences. This can be explained by the nature of the corpus, this data set containing many implicit emotions.

Task T2: as for the previous corpus, EMOTIONS detects only a limited number of behavioral emotions, and the agreement with EMOTYC is low.

Task T3: The disagreement levels are significant also for the detection of emotion categories. More specifically, the disagreement on the *Anger*

Label	Recall(A2 A1)	Recall (A1 A2)
Emotion	77.43	33.15
NonEmotion	59.81	91.52
Behavior	20.00	5.55

Table 5: Corpus C2: Results of tasks T1 and T2

Label	Recall(A2 A1)	Recall (A1 A2)
Anger	100	3.76
Joy	50.00	34.61
Admiration	28.57	1.76
Sadness	37.5	6.97
Shame	66.66	33.33
Fear	78.57	50.00
Pride	100	3.33
Surprise	100	11.62
Other	0.90	25.00

Table 6: Corpus C2: Results of task T3

category is strong (for example, 186 sentences conveying *Anger* are detected by EMOTYC while EMOTIONS detects only 7 ones). This can be explained by the fact that *Anger* is mostly associated with the displayed and suggested modes that are not detected by EMOTIONS. In addition, many sentences annotated as *Anger* by EMOTYC contain swearing and insulting words. As those words are not included in the lexicon used by EMOTIONS, the sentences are not annotated by the parser, which also explains the disparity of *Anger* labels detected by the two parsers. More specifically, *Other* is the emotion category most detected by EMOTIONS within this corpus, indicating that lexicon-based tool encounters difficulties in detecting accurately the categories of emotions in this data set.

Corpus TextToKids (C3): (tab. 7 and 8).

Task T1: For this data set, EMOTYC and EMOTIONS demonstrate good agreement on the detection of non-emotional sentence. In contrast, the tools show divergent results on the detection of emotional sentences.

Label	Recall(A2 A1)	Recall (A1 A2)
Emotional	55.67	36.94
Non Emotional	86.75	94.85
Behavior	53.84	5.88

Table 7: Corpus C3: Results of tasks T1 and T2

Task T2: For this data set, EMOTYC annotates a large number of behavioral emotions, as the corpus contains numerous behavior-related expressions, such as *criticizing*, *having fun* and *smiling*. Those expressions are specific to journal articles written for kids, that often contain descriptions and explanations. The behavioral emotions are still difficult to

Label	Recall(A2 A1)	Recall (A1 A2)
Anger	73.68	13.72
Joy	87.87	52.00
Admiration	28.00	55.76
Sadness	58.82	13.51
Shame	33.33	6.25
Fear	92.30	41.73
Pride	66.66	25.00
Surprise	100	12.82
Other	0.48	9.09

Table 8: Corpus C3: Results of tasks T3

identify for EMOTIONS on this data set as well, and thus the tools have a strong disagreement when performing the task T2.

Task T3: The parsers achieve high agreement regarding the detection of *Joy*, as most of the instances detected by EMOTIONS are also detected by EMOTYC. Similarly, the agreement is also high for the identification of *Fear*. As the data set contains stories for kids, *Fear* and *Joy* are the most prevalent emotions in the texts and they are expressed in a variety of direct and implicit ways, which can explain the agreement of tools on their detection.

Corpus C4 (Reddit) (tab. 9 and 10).

Task T1: For this data set, the tools demonstrate good agreement on detecting non-emotional sentences, although they have a low agreement on the detection of emotional sentences.

Label	Recall(A2 A1)	Recall (A1 A2)
Emotional	29.74	20.02
Non Emotional	92.98	91.69
Behavior	33.33	4.10

Table 9: Corpus C4: Results of tasks T1 and T2

Task T2: the results of this task are perfectly aligned with the results obtained for the previous corpora: the number of behavioral emotions detected by EMOTIONS is limited, and thus the agreement with EMOTYC is low.

Task T3: The two parsers demonstrate good agreement on the detection of two categories: *Joy* and *Fear*. They strongly disagree when it comes to detecting *Sadness* and *Pride*. Also, there is a significant asymmetry in the detection of *Anger* (for example as EMOTYC detects 204 sentences conveying *Anger* while EMOTIONS identifies only 30 ones). *Anger* is highly represented in this corpus, and the analysis shows that sentences annotated by EMOTYC with this label show reactions to previous events, such as *It is shocking!*, or *How dare you ?*.

Other is the most detected emotion category for

Label	Recall(A2 A1)	Recall (A1 A2)
Anger	50.00	7.35
Joy	75.71	53.00
Admiration	6.66	1.48
Disgust	0.00	0.00
Sadness	16.66	16.47
Shame	57.14	33.33
Fear	55.55	51.47
Pride	7.14	12.5
Surprise	61.90	22.03
Other	0.00	0.00

Table 10: Corpus C4: Results of task T3

EMOTIONS, and the agreement of tools on the detection of this category is low.

Overall, the experiments show that the nature of data sets has an impact on the quality of emotion annotation, as EMOTYC performs better in detecting implicit emotions, while EMOTIONS is accurate in detecting explicit ones. Moreover, the results also show that some categories of emotions, such as the basic ones *Fear* and *Joy*, are easier to detect for both parsers, while complex categories are both difficult to identify.

6. Conclusion and perspectives

This paper compares EMOTYC and EMOTIONS, two parsers performing emotion annotation from texts by using distinct approaches and annotation schema. The results show that data sets have an impact on the accuracy of the techniques used by the tools, and thus affect their agreement. In general, implicit and, most specifically, behavioral emotions are difficult to detect for EMOTIONS, while EMOTYC performs poorly in detecting *Admiration*. In addition, for all data sets, the tools have a good agreement in detecting the non-emotional sentences, while the agreement is lower when it comes to the detection of emotional ones. Similarly, there is good agreement on detecting basic emotions, while for complex emotions, the results show consistent disagreement.

Future research aims at improving the annotation of emotions from texts by tuning the learning model used by EMOTYC and enriching the lexicon used by EMOTIONS. Expanding the data sets to include more types of corpora and studying the impact of annotation schema on the results provided by the tools and their agreement will also be investigated in the future.

Data availability statement

data sets presented in this article are available for research purposes upon request.

Ethical considerations and limitations

For ethical concerns, the results are presented in a way that avoids reidentification of any URIs and hashtags that have been used to collect data.

7. Bibliographical References

- Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. 2017. Feel: a french expanded emotion lexicon. *Language Resources and Evaluation*, 51(3):833–855.
- Francisca Adoma Acheampong, Henry Nunoo-Mensah, and Wenyu Chen. 2021. Transformer models for text-based emotion detection: a review of bert-based approaches. *Artificial Intelligence Review*, 54(8):5789–5829.
- Abdullah Al Maruf, Fahima Khanam, Md Mahmudul Haque, Zakaria Masud Jiyad, Muhammad Firoz Mridha, and Zeyar Aung. 2024. Challenges and opportunities of text-based emotion detection: a survey. *IEEE access*, 12:18416–18450.
- Wajdi Aljedaani, Furqan Rustam, Mohamed Wiem Mkaouer, Abdullatif Ghallab, Vaibhav Rupapara, Patrick Bernard Washington, Ernesto Lee, and Imran Ashraf. 2022. Sentiment analysis on twitter data integrating textblob and deep learning models: The case of us airline industry. *Knowledge-Based Systems*, 255:109780.
- Stefano Baccianella, Andrea Esuli, Fabrizio Sebastiani, et al. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204. Valletta.
- Sudesh Bandara, Tharindu Dasath, Senula Nanayakkara, Samadhi Rathnayake, and Thusithanjana Thilakarathna. 2024. Enhancing typing dynamics emotion recognition: A multi-class xgboost approach for accurate sentiment detection. In *2024 6th International Conference on Advancements in Computing (ICAC)*, pages 354–359. IEEE.
- Christina Barz, Melanie Siegel, and Daniel Hanss. 2025. Analyzing the online communication of environmental movement organizations: Nlp approaches to topics, sentiment, and emotions. In *Proceedings of the 1st Workshop on Ecology, Environment, and Natural Language Processing (NLP4Ecology2025)*, pages 68–76.
- Delphine Battistelli, Cyril Bruneau, and Valentina Dragos. 2020. Building a formal model for hate detection in french corpora. *Procedia Computer Science*, 176:2358–2365.
- Jacob Beck. 2023. Quality aspects of annotated data: A research synthesis. *AStA Wirtschafts- und Sozialstatistisches Archiv*, 17(3):331–353.
- Nathalie Blanc and Guy Quenette. 2017. La production d'inférences émotionnelles entre 8 et 10 ans: quelle méthodologie pour quels résultats? *Enfance*, (4):503–511.
- Disha Chavan, Esha Anvekar, Megha Dandapat, Vaibhav Bichave, and Jayashree Jagdale. 2023. Machine learning applied in emotion classification: a survey on dataset, techniques, and trends for text based documents. In *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, volume 1, pages 1697–1703. IEEE.
- Tulika Chutia and Nomi Baruah. 2024. A review on emotion detection by using deep learning techniques. *Artificial Intelligence Review*, 57(8):203.
- Denise Davidson. 2006. The role of basic, self-conscious and self-conscious evaluative emotions in children's memory and understanding of emotion. *Motivation and Emotion*, 30(3):232–242.
- Luna De Bruyne, Orphée De Clercq, and Véronique Hoste. 2021. Emotional robbert and insensitive bertje: combining transformers and affect lexica for dutch emotion detection. In *Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA), held in conjunction with EACL 2021*, pages 257–263. Association for Computational Linguistics.
- Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Shafiq Joty, Boyang Li, and Lidong Bing. 2022. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*.
- Valentina Dragos, Delphine Battistelli, Aline Etienne, and Yolène Constable. 2022. Angry or sad? emotion annotation for extremist content characterisation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 193–201.
- Paul Ekman. 1992. [An argument for basic emotions](#). *Cognition and Emotion*, 6(3-4):169–200.
- Aline Etienne. 2023. *Analyse automatique des émotions dans les textes: contributions théoriques et applicatives dans le cadre de l'étude de la complexité des textes pour enfants*. Ph.D. thesis, Université de Nanterre-Paris X.
- Aline Étienne, Delphine Battistelli, and Gwénolé Lecorvé. 2024. [Emotion identification for French in written texts: Considering modes of emotion expression as a step towards text complexity](#)

- analysis. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 168–185, Bangkok, Thailand. Association for Computational Linguistics.
- Kexin Feng and Theodora Chaspari. 2020. A review of generalizable transfer learning in automatic emotion recognition. *Frontiers in Computer Science*, 2:9.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- A Aruna Gladys and V Vetriselvi. 2023. Survey on multimodal approaches to emotion recognition. *Neurocomputing*, 556:126693.
- Akhilesh Kumar and Awadhesh Kumar. 2022. Human sentiment analysis on social media through naïve bayes classifier. *J. Sci. Res*, 66(1).
- Sheetal Kusal, Shruti Patil, Jyoti Choudrie, Ketan Kotecha, Deepali Vora, and Ilias Pappas. 2023. A systematic review of applications of natural language processing and future challenges with special emphasis in text-based emotion detection. *Artificial Intelligence Review*, 56(12):15129–15215.
- Sofie Labat, Naomi Ackaert, Thomas Demeester, and Véronique Hoste. 2022. Variation in the expression and annotation of emotions: A wizard of oz pilot study. In *Proceedings of the 1st workshop on perspectivist approaches to NLP@LREC2022*, pages 66–72.
- Ruifan Li, Hao Chen, Fangxiang Feng, Zhanyu Ma, Xiaojie Wang, and Eduard Hovy. 2021. Dual graph convolutional networks for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6319–6329.
- Zheng Lian, Licai Sun, Haiyang Sun, Kang Chen, Zhuofan Wen, Hao Gu, Bin Liu, and Jianhua Tao. 2024. Gpt-4v with emotion: A zero-shot benchmark for generalized emotion recognition. *Information Fusion*, 108:102367.
- Steven Loria et al. 2018. textblob documentation. *Release 0.15*, 2(8):269.
- WP Malecki, Marta Kowal, Małgorzata Dobrowolska, and Piotr Sorokowski. 2021. Defining online hating and online haters. *Frontiers in Psychology*, 12:744614.
- Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Villemonte de La Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. *arXiv preprint arXiv:1911.03894*.
- Donatas Meškelė and Flavius Frasincar. 2020. Aldonar: A hybrid solution for sentence-level aspect-based sentiment analysis using a lexicalized domain ontology and a regularized neural attention model. *Information Processing & Management*, 57(3):102211.
- Annie Piolat and Rachid Bannour. 2009. Emotaix: un scénario de tropes pour l’identification automatisée du lexique émotionnel et affectif. *L’Année psychologique*, 109(4):655–698.
- Soujanya Poria, Alexander Gelbukh, Erik Cambria, Peipei Yang, Amir Hussain, and Tariq Durrani. 2012. Merging senticnet and wordnet-affect emotion lists for sentiment analysis. In *2012 IEEE 11th international conference on signal processing*, volume 2, pages 1251–1255. IEEE.
- Dibyendu Seal, Uttam K Roy, and Rohini Basak. 2020. Sentence-level emotion detection from text based on semantic rules. In *Information and Communication Technology for Sustainable Development: Proceedings of ICT4SD 2018*, pages 423–430. Springer.
- Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. *arXiv preprint arXiv:2305.08377*.
- Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansoor Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. *arXiv preprint arXiv:2402.13446*.
- Angelina Tzacheva and Sanchari Chatterjee. 2024. Actionable pattern discovery for emotion detection in bigdata in education and business. In *International Conference on AI, Machine Learning and Data Science (AIMDS 2024)*, volume 13.
- Xinhao Zou and Konstantin Markov. 2024. Combining graph nn and llm for improved text-based emotion recognition. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*, pages 143–154. Springer.

Clarifying the Role of Psychological Factors in Language Acquisition: A Psycholinguistic Lexical Ratings Dataset

Wanwan Zheng

Humanity Center for Anthropocenic Actors and Agency, Graduate School of Humanities
Furo-cho, Chikusa, Nagoya 464-8601, Japan
zheng@nagoya-u.jp

Abstract

Lexical acquisition extends beyond the learning of surface-level word forms and encompasses underlying cognitive characteristics as well as broader processes that involve the learner. Nevertheless, in Japanese, as in many other languages, word difficulty has been characterized primarily by frequency and surface-level properties. Far less is known about the cognitive and affective dimensions that influence whether words are easier or more difficult to process. To address this gap, this study introduces a novel dataset that incorporates six psycholinguistic dimensions—familiarity, affective valence, arousal, imageability, abstractness, and understandability—collected through large-scale surveys of Japanese second language learners. Preliminary analyses of responses from 536 participants across 15 countries demonstrated that the dataset is both theoretically coherent and empirically reliable, consistent with established theories and findings while also yielding new insights into lexical processing. In addition to supporting more accurate estimations of word difficulty, the dataset provides a resource for future research by enabling systematic exploration of how lexical processing is shaped through the interaction of visual, affective, cognitive, and contextual factors.

Keywords: word difficulty, psycholinguistic dimensions, second language acquisition (Japanese), cognitive and affective processing, lexical dataset

1. Introduction

Empirical evidence has demonstrated that the accurate identification of words constitutes a fundamental prerequisite for reading comprehension (Duff et al., 2015; Killingly et al., 2025). Consistent with this view, interactive models of text comprehension developed in the 1990s highlighted the dynamic interplay between bottom-up processes (e.g., word recognition and other lower-level operations) and top-down processes (e.g., meaning construction and the integration of background knowledge) as essential components of successful reading comprehension (Kintsch, 1998; Perfetti, 1999). Moreover, vocabulary coverage not only constrains the range of thought but also serves as a robust indicator of cognitive ability (Bruce and Bell, 2022). Research with preschool children has further revealed a significant bidirectional relationship between executive functioning and vocabulary size (Schmitt et al., 2019). Therefore, vocabulary knowledge entails more than the acquisition of linguistic symbols; it also supports efficient language processing and the development of cognitive control (Bransford et al., 2000; Jucks and Paus, 2012). Consequently, the capacity to construct meaning from text, as well as to regulate and adapt to cognitive activity, can be understood as fundamentally rooted in word-level processing.

On the other hand, information about word difficulty contributes to more efficient vocabulary acquisition. However, word difficulty is determined by multiple factors, including morphological features

(e.g., word length and the number of morphemic components), semantic features (e.g., polysemy and domain specificity), exposure-related variables (e.g., frequency and familiarity), imageability and concreteness (i.e., the extent to which a word elicits mental imagery), and psychological factors (e.g., affective valence, ease of cognitive processing, and memory retention). Consequently, the estimation of word difficulty remains a complex task. In current linguistic and educational research, the common practice has been the manual selection of words assigned to different difficulty levels. Such human evaluation, however, is labor-intensive and inherently restricted in scale. For this reason, mechanical and quantitative approaches to word difficulty estimation have become increasingly necessary. To develop reliable models, it is essential to identify the key determinants of word difficulty and to elucidate the magnitude of their contributions.

Among the numerous factors influencing word difficulty, frequency effect is the most widely recognized. High-frequency words are typically acquired more rapidly and processed more easily, whereas low-frequency words are more likely to be perceived as difficult. Hashimoto and Egbert (2019) reported a correlation of -0.50 between word frequency and difficulty. Extending this finding, Stewart et al. (2022) demonstrated that when a broader frequency range is examined and a logarithmic transformation is applied, the correlation strengthens to -0.80 . Other usage-related factors include age of acquisition (Hiebert et al., 2019), regularity (e.g., word bigrams; Ha et al., 2024), and contex-

tual dependence (e.g., word co-occurrence; [Ha et al., 2024](#)). In contrast, intrinsic word-level properties include length ([Cervetti et al., 2015](#)), polysemy ([Vitta et al., 2023](#); [Canning et al., 2024](#)), and the knowledge required for comprehension ([Yamakawa and Tajima, 2024](#)). As most of these indicators are derived from count-based values, their reliability depends heavily on the linguistic resources (corpora) used.

In addition to quantitative indicators, there are qualitative indicators that are less dependent on specific linguistic resources. These include (1) word familiarity (i.e., the perceived familiarity of words among language users; [Leroy and Kauchak, 2014](#)), (2) image-related semantic properties such as imageability and concreteness ([Kumcu and Thompson, 2020](#)), and (3) psychological factors such as affective valence and memory retention ([Barrett, 2017](#)). Unlike corpus-based frequency indicators, these indicators are less susceptible to textual distribution biases and may therefore offer a more universal estimation of word difficulty.

Word familiarity With respect to word familiarity, [Zheng \(2024\)](#) estimated word difficulty using five dimensions of familiarity: “knowing,” “hearing,” “speaking,” “writing,” and “reading.” In general, familiarity scores tend to be higher for easier words and lower for more difficult ones. However, the analysis revealed two atypical categories: “easy but unfamiliar words” and “difficult but familiar words.” When these outliers were removed through anomaly detection, the classification accuracy of beginner-, intermediate-, and advanced-level words improved from 0.66 to 0.93. Nevertheless, because the learning environments of native speakers and second language (L2) learners differ, there are words for which familiarity and difficulty do not align across groups. This discrepancy suggests that familiarity data used in this work may reflect a mixture of responses from both native speakers and L2 learners.

Image-related semantic factors With respect to image-related semantic factors, it is well established that words referring to perceptually salient and highly imageable objects or concepts are recalled more easily than those associated with less imageable ones ([Marschark and Cornoldi, 1991](#)). In this context, [Kumcu and Thompson \(2020\)](#) argued that low-difficulty words are typically more imageable and concrete, thereby reducing cognitive load and facilitating more accurate recall.

Affective valence and arousal With respect to affective valence, emotion has been recognized as a central concept in applied linguistics, playing

a fundamental role in cognition and receiving increasing attention for its role in foreign language acquisition ([Dewaele, 2015](#)). Previous research has introduced concepts such as “emotionally enhanced memory” ([Talmi et al., 2007](#)) and “affective input enhancement” ([Truscott, 2015](#)). Moreover, [Kanazawa \(2021, 2023\)](#) demonstrated that emotion-based processing facilitates the retrieval of target words. [Hashimoto and Egbert \(2019\)](#) further incorporated affective valence—an emotional dimension associated with word difficulty—alongside with other lexical variables such as word length and contextual distinctiveness. However, linguistic resources developed on the basis of native speakers’ responses are not necessarily applicable to foreign language learners, whose mental lexicons differ substantially. This discrepancy underscores the importance of developing foreign language-specific versions of such databases, particularly for subjective attributes ([Kanazawa, 2021](#)).

Taken together, these findings indicate that word difficulty cannot be explained by a single factor; rather, it must be evaluated within a broader multidimensional framework that integrates both linguistic and psycholinguistic factors. The Tool for the Automatic Analysis of Lexical Sophistication (TAALES; [Kyle et al., 2018](#)), developed to assess the difficulty of English words, incorporates more than 400 indicators. Because most of these indicators are statistical measures, such as frequency, they are relatively straightforward to interpret; however, they remain dependent on corpora or dictionary-based databases. In contrast, lexical properties grounded in psychological and cognitive factors (e.g., familiarity, imageability, abstractness, and affective valence) capture dimensions of word difficulty that cannot be fully explained by frequency-based statistics alone.

To address these limitations, this study constructed a dataset for learners of Japanese as a second language with the aim of enabling more precise estimation of word difficulty. The dataset incorporates six psycholinguistic features—familiarity, affective, arousal, imageability, abstractness, and understandability—thereby capturing the cognitive complexity of lexical processing from multiple perspectives.

The remainder is structured as follows. Section 2 describes the procedures used to develop the dataset. Section 3 presents the data structure both quantitatively and visually. Section 4 evaluates the effectiveness of the dataset in estimating word difficulty and reports benchmark experiments. Finally, Section 5 concludes with a summary.

2. Data Collection

2.1. Selection of Target Words

The objective of this study was to construct a comprehensive and reliable dataset for the evaluation of word difficulty. As discussed in the preceding sections, word difficulty is a multidimensional construct shaped by diverse linguistic, cognitive, and psychological factors. Accordingly, the availability of highly reliable word difficulty information of paramount importance for both dataset development and subsequent analyses.

For this purpose, the words included in the *Japanese Language Learner's Dictionary* were selected as the targeted words (Sunakawa et al., 2012). This dictionary contains 17,920 entry words derived from two major lexical resources: (1) a lexical survey of the *Balanced Corpus of Contemporary Written Japanese* (BCCWJ)—a corpus specifically designed to capture the overall characteristics of contemporary written Japanese and currently the only available balanced corpus of the language—and (2) a corpus of Japanese language textbooks, comprising electronic data from 100 commercially available textbooks ranging from beginner to advanced levels (unpublished resource). The adoption of this dictionary as the basis for target word selection provides distinct advantages. Because its vocabulary was systematically sampled from both a balanced national corpus and pedagogical materials spanning all proficiency levels, the resulting word list ensures broad representativeness across genres and registers while maintaining direct applicability to second-language learning contexts.

In detail, each entry in the dictionary is annotated with multiple layers of information, including standardized orthographic form, spelling, difficulty level, part of speech (POS), and word type (i.e., native Japanese words, Sino-Japanese words, loanwords, and hybrid formations). The difficulty of each entry word was classified into six different levels—lower/upper elementary, lower/upper intermediate, and lower/upper advanced—based on the subjective judgment of experienced Japanese language teachers. Representative examples are provided in Table 1.

2.2. Survey Design

For each word, the survey items were presented as illustrated in Figure 1. Note that the questionnaire itself was presented in Japanese. Because Japanese includes homographs with different pronunciations, which may influence perceived difficulty, the stimuli were presented in the form of “word + reading.” Participants were asked to evaluate their cognitive and affective impressions of the given

word from two perspectives:

1. **Familiarity** This item assessed the degree of familiarity with the target word. Participants rated their familiarity degree across four usage contexts—writing, reading, speaking, and listening—as well as their overall impression (four items in total). Responses were provided on a seven-point Likert scale ranging from *Not at all* to *Strongly think so*.
2. **Meaning and Image** This item elicited subjective evaluations of the semantic and psychological properties of the target word. Specifically, participants rated five aspects: Negativity (affective valence: pleasant—unpleasant), Calmness (arousal: calming—exciting), Ease of evoking an image (imageability: ease of eliciting a mental representation), Abstractness (degree of conceptual difficulty), and Ease of understanding (understandability: perceived difficulty of learning). These responses were also recorded on a seven-point Likert scale.

Please evaluate the word 「秋祭り(アキマツリ)」 from the following perspectives.

1. Familiarity [〇]

	Not at all	Do not think so	Slightly do not think so	Neutral	Slightly think so	Think so	Strongly think so
Writing	☐	☐	☐	☐	☐	☐	☐
Reading	☐	☐	☐	☐	☐	☐	☐
Speaking	☐	☐	☐	☐	☐	☐	☐
Listening	☐	☐	☐	☐	☐	☐	☐
Overall	☐	☐	☐	☐	☐	☐	☐

2. Meaning and Image [〇]

	Not at all	Do not think so	Slightly do not think so	Neutral	Slightly think so	Think so	Strongly think so
Negativity	☐	☐	☐	☐	☐	☐	☐
Calmness	☐	☐	☐	☐	☐	☐	☐
Ease of evoking an image (easily comes to mind)	☐	☐	☐	☐	☐	☐	☐
Abstractness (invisible concepts)	☐	☐	☐	☐	☐	☐	☐
Ease of understanding	☐	☐	☐	☐	☐	☐	☐

Figure 1: Survey items (translated). The questionnaire itself was presented in Japanese.

In this survey, participants were presented with a set of seven words, and two sets (14 words in total) were administered. These words were randomly selected from the vocabulary list of the applied dictionary, with the design ensuring representation from all six levels of word difficulty. Because the task did not involve correct or incorrect answers, participants' intuitive judgments were prioritized. To assess response reliability, two words were duplicated across the sets, requiring participants to evaluate the same item twice. To minimize memory and order effects, the duplicated words were

Standardized orthographic form	Spelling	Difficulty level	POS	Word type
秋祭り (autumn festival)	アキマツリ	upper intermediate	noun	native Japanese word
愛好 (fondness/enthusiasm)	アイコウ	lower advanced	noun	Sino-Japanese word
アイス (ice cream)	アイス	lower elementary	noun	loanword
空き缶 (empty can)	アキカン	lower intermediate	noun	hybrid word

Table 1: Representative examples of target words selected from the *Japanese Language Learner's Dictionary*.

inserted at fixed intervals. The presentation sequence followed the format: A (5 words) → C (2 words) → B (5 words) → C (2 words). In addition, demographic information regarding gender, age, nationality, and learning level was collected at the end of the questionnaire. For learning level, five categories ranging from beginner to advanced were provided, each accompanied by explanatory descriptions to facilitate accurate self-assessment. The question was phrased as follows:

Please indicate your level of Japanese (select the one that is closest to your level).

- Beginner (able to engage in simple greetings and self-introductions)
- Upper Beginner (able to conduct basic conversations, e.g., hobbies, schedules)
- Intermediate (able to understand slow-paced conversations and simple texts)
- Upper Intermediate (able to understand moderately complex conversations and texts, and to use Japanese in both daily life and work contexts)
- Advanced (able to understand newspapers, news, and specialized content, and to speak naturally)

2.3. Survey Administration and Screening

The questionnaire was developed using Qualtrics and administered online via URL or QR code. The study received ethical approval from the Ethics Review Committee of university (approval no. NUHM-25-17). All participants provided informed consent online prior to the start of the survey. In Japan, cooperation was requested from Nagoya University as well as 202 Japanese language schools nationwide. In China, collaboration was sought from Peking University, Jinan University, Capital Normal University, Dalian University of Foreign Languages, and Shanghai International Studies University. In addition, participation was solicited through social media platforms such as Facebook.

The survey began in August 2025 and is currently ongoing. The estimated completion time was approximately 10–18 minutes. At the start of the survey, participants were presented with an informed

consent form describing the study, and responses were recorded only after consent was provided. As compensation, participants who submitted valid responses received an electronic gift card valued at 300 yen (15 RMB in China). To increase sample size, multiple submissions from the same participant were permitted, with a maximum of ten entries.

To ensure the validity of responses, entries with missing values were first excluded. Subsequently, for each participant, the ratings provided for the two duplicated words were compared. The Likert-scale responses, ranging from “Not at all” to “Strongly think so”, were converted into numerical values on a 0–6 scale. Response consistency was then evaluated using two measures: the mean absolute error (MAE) and the exact agreement rate (Agree). The Agree index is easy to interpret; however, because the agreement rate disregards directionality, it does not differentiate between minor discrepancies (e.g., ± 1) and major discrepancies (e.g., ± 3 or more). Therefore, it was employed in combination with MAE, which captures the magnitude continuous error.

For each duplicated word $i = 1, 2$, numerical evaluations of the items $j = 1, \dots, 10$ are denoted as a_{ij} (first rating) and b_{ij} (second rating). Items with missing values are excluded from the calculation.

Mean Absolute Error (MAE)

$$MAE_i = \frac{1}{10} \sum_{j=1}^{10} |a_{ij} - b_{ij}| \quad (1)$$

The range is 0–6, where values closer to 0 indicate higher consistency between the two ratings.

Exact Agreement Rate

$$Agree_i = \frac{1}{10} \sum_{j=1}^{10} 1(a_{ij} = b_{ij}) \quad (2)$$

The range is 0–1, where 1 indicates perfect agreement across all items.

Furthermore, the mean values of MAE and agreement across the two duplicated words were calculated. Because human decision-making inherently involves uncertainty, it cannot be assumed that the two ratings for duplicated words would consistently

coincide. This uncertainty is considered to stem primarily from two sources. First, due to limitations in the information and computational resources available to individuals, “bounded rationality” emerges (Simon, 1955, 1972). In other words, individuals are unable to analyze situations in a fully rational and exhaustive manner, and instead rely on approximate judgments within the constraints of their cognitive resources. Second, cognitive processes such as perception, memory, and reasoning are inevitably subject to “information-processing noise” (Faisal et al., 2008; Kahneman et al., 2021), which introduces variability into judgments even under identical conditions. Through the interaction of these two factors, human decision-making is intrinsically variable. The thresholds were then established as follows.

- **High reliability:** $MAE \leq 1.0$ (average deviation within one scale point) and $Agree \geq 0.70$ (perfect agreement in more than 70% of cases).
- **Review required (gray zone):** $1.0 < MAE \leq 2.0$ or $0.20 \leq Agree < 0.70$.
- **Low reliability:** Otherwise.

3. Data Structure and Statistical Processing

3.1. Data Structure

Valid responses were obtained from 536 participants across 15 countries: China, South Korea, the Philippines, Myanmar, France, the United States, Vietnam, Malaysia, Thailand, Mongolia, Australia, Turkey, the United Kingdom, India, and Bangladesh. Of these, 273 responses (51%) were classified as high reliability, while 263 responses (49%) were classified as requiring further review.

Figure 2 illustrates the distributions of MAE and agreement, calculated from the two ratings for duplicated words and their averages. The histograms depict the frequency distributions of the two indices, while the overlaid kernel density estimation (KDE) curves highlight their distributional tendencies and modal peaks. The MAE exhibited a right-skewed distribution, with most cases concentrated below 1.0, indicating relatively small discrepancies between responses for duplicated words. By contrast, agreement displayed a bimodal distribution with peaks around 0.6 and 0.8, with a pronounced concentration near 0.8, suggesting that response consistency was generally high.

Moreover, after excluding outliers, the mean response time was 727.77 seconds, which fell within the expected range identified in the pilot study (600–1080 seconds). These findings indicate that

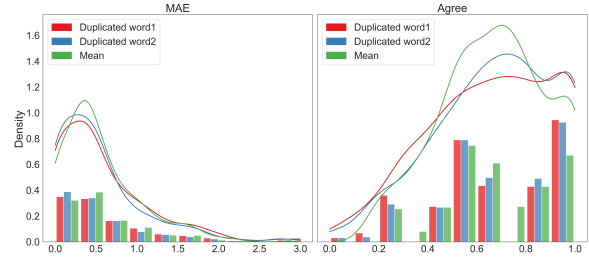


Figure 2: Distributions of MAE and agreement for duplicated words and their averages.

Learning level	Frequency	Rate
Intermediate	214	0.40
Upper Intermediate	147	0.27
Advanced	139	0.26
Upper Beginner	33	0.06
Beginner	3	0.01

Table 2: Distribution of participants’ proficiency levels

the survey was conducted in a formally appropriate manner, with most participants devoting a reasonable amount of time to their responses. These results provide strong evidence supporting the overall reliability of the survey data.

Participants’ ages ranged from 18 to 42 years, with a mean of 25 years and a standard deviation of 4.28. Regarding gender, 329 participants were female (61.38%), 205 were male (38.24%), and two identified otherwise. Self-assessed Japanese proficiency levels are summarized in Table 2. The largest group was Intermediate, comprising approximately 40% of the sample (214 participants), followed by Upper Intermediate at 27% (147 participants) and Advanced at 26% (139 participants). Overall, the majority of respondents (93%) fell between the intermediate and advanced levels, suggesting that most participants had already attained a relatively high level of Japanese proficiency.

3.2. Statistical Processing

Each valid response included 14 words, yielding a total of 7,504 word instances. However, because some words were rated by multiple responses, the final dataset contained 1,229 unique words. To enhance generalizability, it was desirable to collect as many responses as possible for each word. Therefore, in the present study, only words with at least five responses ($N = 556$) were retained for dataset evaluation.

The collected data were subject to potential biases stemming from both word-specific characteristics (e.g., abstractness) and respondent-specific tendencies (e.g., a general inclination to assign higher ratings). To address these sources of varia-

tion, ratings were estimated using a Bayesian Linear Mixed Model (BLMM), in which word-specific effects and participant-specific effects were incorporated as random effects.

The estimation was performed across ten dimensions: five related to familiarity (writing, reading, speaking, listening, and overall impression) and five related to meaning and image (negativity, calmness, imageability, abstractness, and understandability). All ratings were originally collected on a seven-point Likert scale (*Not at all* – *Strongly think so*) and were treated as continuous values ranging from 0 to 6.

Let $y_{ij}^{(k)}$ denote the rating given by participant j to word i on dimension k . The model was formalized as follows:

$$y_{ij}^{(k)} \sim \text{StudentT}(\nu, \mu_{ij}, \sigma), \quad (3)$$

$$\mu_{ij} = \alpha + \gamma_{\text{word}}^{(i)} + \gamma_{\text{subj}}^{(j)}. \quad (4)$$

$$\gamma_{\text{word}}^{(i)} \sim \mathcal{N}(\mu_{\text{word}}, \sigma_{\text{word}}), \quad \gamma_{\text{subj}}^{(j)} \sim \mathcal{N}(\mu_{\text{subj}}, \sigma_{\text{subj}}). \quad (5)$$

Here, α represents the overall mean, $\gamma_{\text{word}}^{(i)}$ denotes the word-specific effect, and $\gamma_{\text{subj}}^{(j)}$ denotes the participant-specific effect. These random effects were assumed to follow hierarchical normal distributions. In addition, a Student-t distribution was employed for the observation noise rather than a normal distribution, allowing for more robust estimation in the presence of outliers.

Parameter estimation was carried out using the No-U-Turn Sampler (NUTS). For each chain, 2,000 samples were drawn from the posterior distribution, with 1,000 iterations used for warm-up (tuning). Four independent chains were run in parallel, yielding a total of 8,000 samples. The target acceptance rate was set to 0.9 to suppress divergences and ensure stable convergence. Furthermore, the response variables were standardized (mean = 0, standard deviation = 1) to improve the stability of model estimation. As a result, the Gelman–Rubin convergence diagnostic statistic $\hat{R}$ was 1.0 for all parameters, confirming that all models successfully converged.

The estimation results are shown in Figure 3. The magnitudes of variation across words and across participants were represented by the standard deviations σ_{word} and σ_{subj} , respectively. The results indicated that the posterior means of both σ_{word} and σ_{subj} deviated from zero. Moreover, although some variability was observed in the ratings of individual words, differences in response tendencies across participants contributed more substantially. In other words, in the present dataset, individual differences among participants exerted

a stronger influence on the ratings than did the inherent characteristics of the words themselves.

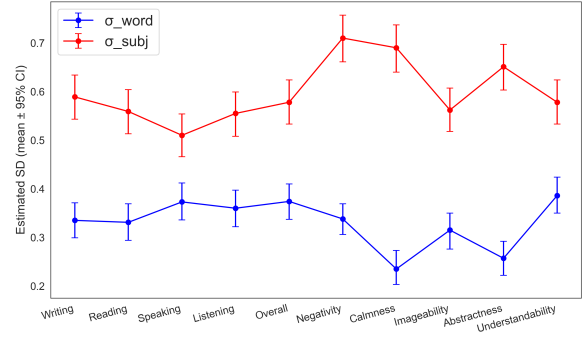


Figure 3: Estimated standard deviations of word and participant effects with posterior means and 95% credible intervals.

4. Evaluation and Benchmarking

Although the dataset is still under construction, this section sketches potential evaluation frameworks and benchmark experiments. Figure 4 illustrates the distribution of predicted mean scores across ten dimensions (Writing, Reading, Speaking, Listening, Overall, Negativity, Calmness, Imageability, Abstractness, and Understandability), with max-min normalization applied to constrain the data to the range of -3 and 3. Figure 5 presents a heatmap of the correlation matrix among the predicted mean scores for these dimensions. Based on the correlation structure in Figure 5, four distinct clusters were identified.

Interrelation of language skills The four language skills—writing, reading, speaking, and listening—showed very strong intercorrelations (0.81–0.90). Each skill was also highly correlated with Overall (0.89–0.94). This findings are consistent with previous research in second language acquisition and language assessment, which have emphasized the interdependent development of the four skills (Quinn et al., 2015; Zheng, 2024). The findings further confirmed, both statistically and conceptually, that Overall functions as an integrated index of the four core language skills.

Semantic properties and affective valence Abstractness was negatively correlated with both Imageability (−0.28) and Understandability (−0.17), while showing a positive correlation with Negativity (+0.22). This pattern suggests that abstract, less comprehensible, and less imageable words are more likely to be associated with negative affect.

Kousta et al. (2010) and Vigliocco et al. (2013) reported that abstract words are more strongly as-

sociated with affective dimensions such as valence and arousal than concrete words. This finding suggests that abstract words, which are difficult to visualize and require context-dependent interpretation, tend to evoke negative emotions in learners due to this dual processing burden. At the same time, it reflects the possibility that abstract words reinforce their meanings through the mediation of negative affect (Ponari et al., 2018). Taken together, these results support the view that affective valence is not merely ancillary but instead constitutes a critical factor that interacts with the abstractness—concreteness dimension in shaping cognitive processing. Thus, the observed correlation structure appropriately captures the intersection between semantic properties (abstract—concrete) and affective valence (negative—positive).

Understandability and imageability A strong positive correlation (0.76) was observed between Understandability and Imageability. This suggested that the semantic processing of words is reinforced through visual and sensory imagery, consistent with the psycholinguistic perspective that semantic representations are grounded in sensorimotor experience (Barsalou, 2008), a framework commonly referred as grounded cognition. Moreover, Understandability was highly correlated with each of the four language skills and with Overall (0.66–0.73), whereas Imageability showed moderate correlations with the skills and Overall (0.53–0.60). These results indicate that ease of processing—whether in terms of comprehension or imageability—is closely associated with language proficiency.

In summary, these results not only support established theories of lexical processing but also strengthen the validity of the present survey data.

Additionally, as a supplement to the correlations presented in Figure 5 and the identified four clusters, Figure 6 presents pairwise scatter plots among Overall, Negativity, Imageability, and Understandability. The scatter plots indicate that Overall proficiency is strongly predicted by both Imageability and Understandability, underscoring the central role of cognitive processing dimensions in lexical access. By contrast, Negativity exhibited only a weak negative correlation with Overall and no meaningful relationship with Imageability or Understandability, suggesting that affective valence constitutes a distinct psychological dimension. This pattern is consistent with psycholinguistic accounts in which abstract negative words rely on affective content to strengthen their semantic representation, while emotional load operates independently of cognitive processing efficiency. Taken together, these findings support the view that lexical processing can be conceived as a dual-route system, comprising a cognitive pathway and an affective pathway

(Ferré et al., 2025; Kuperman et al., 2014; Palazova et al., 2013).

5. Conclusion

Japanese, often regarded as one of the most complex and difficult languages to acquire, remains insufficiently understood with respect to the factors that influence word difficulty and their relative contributions. Previous research has largely relied on corpus-based statistical measures, leaving a critical gap in data that directly capture the cognitive complexity underlying lexical processing. To address this limitation, the present study focused on six psycholinguistic dimensions—familiarity, affective, arousal, imageability, abstractness, and understandability—and developed a novel dataset through empirical surveys. Although data collection is still ongoing, preliminary quantitative evaluations have demonstrated the validity of the dataset, showing consistency with established theoretical frameworks and with findings from previous studies.

Word difficulty is not confined to surface-level lexical properties; it is also shaped by contextual factors, sociocultural influences, cognitive processing costs, and affective dimensions. Moreover, these factors do not operate independently but instead interact in complex ways to influence lexical processing. Many aspects of this process remain theoretically and empirically unexplained. Accordingly, the dataset constructed in this study not only contributes to word difficulty estimation but also provides a foundation for elucidating the interrelationships and interactions among its contributing factors.

Future research should address two important limitations of the present study. First, because the dataset is based on subjective metalinguistic ratings, the findings should be interpreted as reflecting learners' perceived lexical difficulty rather than objective processing difficulty. Further work should therefore examine how these metacognitive judgments relate to behavioral measures such as reaction time, recall, and eye-tracking. Second, the absence of a paired cross-linguistic dataset prevents direct evaluation of whether the observed psycholinguistic dimensions are language-universal or language-specific. Constructing comparable datasets across languages would help clarify which determinants of lexical difficulty generalize across linguistic contexts and which are shaped by language-specific properties.

Ethical Considerations

This study involved human participants who provided subjective ratings of related psycholinguistic

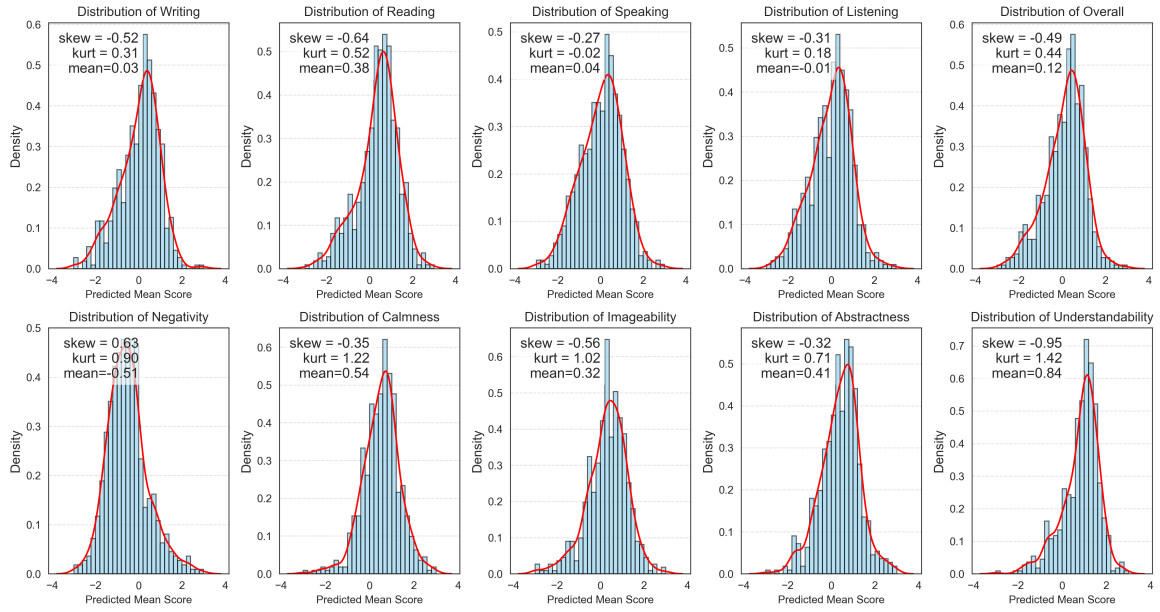


Figure 4: Distribution of predicted mean scores across ten dimensions (Writing, Reading, Speaking, Listening, Overall, Negativity, Calmness, Imageability, Abstractness, and Understandability). Each histogram depicts the density of predicted scores with an overlaid kernel density estimate (red line). The skewness (skew), kurtosis (kurt), and mean values are reported for each dimension.

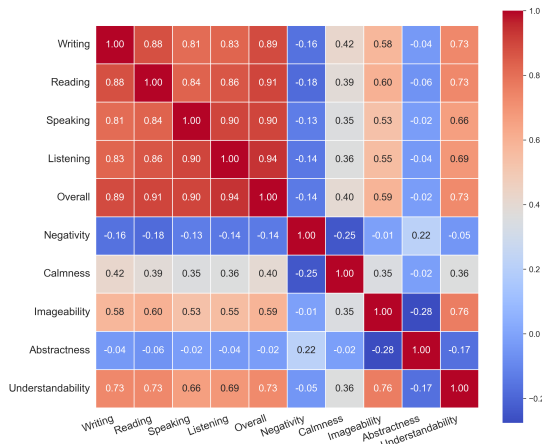


Figure 5: Correlation matrix of predicted mean scores across ten dimensions (Writing, Reading, Speaking, Listening, Overall, Negativity, Calmness, Imageability, Abstractness, and Understandability).

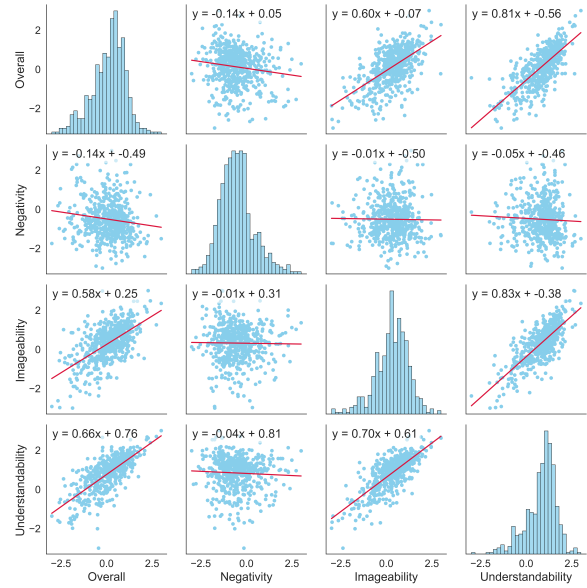


Figure 6: Pairwise scatter plots with regression lines among four dimensions (Overall, Negativity, Imageability, and Understandability).

properties. All participants gave informed consent prior to participation. The study protocol was reviewed and approved by the institutional ethics committee of Nagoya University.

While the dataset may be made publicly available for research purposes, all shared data will be fully anonymized and distributed under terms that restrict re-identification and misuse.

Acknowledgements

This work is supported by JSPS KAKENHI Grant Number 25K16328 and JSPS Topic-Setting Program to Advance Cutting-Edge Humanities and Social Sciences Research Grant Number JPJS00122674991.

Bibliographical References

- L. F. Barrett. 2017. *How emotions are made: The secret life of the brain*. Houghton Mifflin Harcourt, Boston, US.
- Lawrence Barsalou. 2008. [Grounded cognition](#). *Annual review of psychology*, 59:617–45.
- J. D. Bransford, A. L. Brown, and R. R. Cocking. 2000. *How people learn: Brain, mind, experience, and school*. National Academy Press, Washington D.C., U.S.
- Madeleine Bruce and Martha Ann Bell. 2022. [Vocabulary and executive functioning: A scoping review of the unidirectional and bidirectional associations across early childhood](#). *Human Development*, 66:167–187.
- Derek N. Canning, Stuart McLean, and Joseph P. Vitta. 2024. [Relative complexity in a model of word difficulty: The role of loanwords in vocabulary size tests](#). *Studies in Second Language Learning and Teaching*, 14:631–659.
- Gina N. Cervetti, Elfrieda H. Hiebert, P. David Pearson, and Nicola A. McClung. 2015. [Factors that influence the difficulty of science words](#). *Journal of Literacy Research*, 47:153–185.
- J.-M. Dewaele. 2015. [From obscure echo to language of the heart: Multilinguals' language choices for \(emotional\) inner speech](#). *Journal of Pragmatics*, 87:1–17.
- Fiona J. Duff, Gurpreet Reen, Kim Plunkett, and Kate Nation. 2015. [Do infant vocabulary skills predict school-age language and literacy outcomes?](#) *Journal of Child Psychology and Psychiatry and Allied Disciplines*, 56:848–856.
- A. Aldo Faisal, Luc P. J. Selen, and Daniel M. Wolpert. 2008. [Noise in the nervous system](#). *Nature Reviews Neuroscience*, 9(4):292–303.
- Pilar Ferré, Alberto J., Sánchez-Carmona, Juan Haro, Rocío Calvillo-Torres, Jacobo Albert, and José Antonio Hinojosa. 2025. [How does emotional content influence visual word recognition? a meta-analysis of valence effects](#). *Psychon Bull Rev*, 32:570–587.
- Hung Tan Ha, Duyen Thi Bich Nguyen, and Tim Stoeckel. 2024. [What is the best predictor of word difficulty? a case of data mining using random forest](#). *Language Testing*, 41:828–844.
- Brett J. Hashimoto and Jesse Egbert. 2019. [More than frequency? exploring predictors of word difficulty for second language learners](#). *Language Learning*, 69:839–872.
- Elfrieda H. Hiebert, Judith A. Scott, Ruben Castaneda, and Alexandra Spichtig. 2019. [An analysis of the features of words that influence vocabulary difficulty](#). *Education Sciences*, 9.
- Regina Jucks and Elisabeth Paus. 2012. [What makes a word difficult? insights into the mental representation of technical terms](#). *Metacognition and Learning*, 7:91–111.
- Daniel Kahneman, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgment*. Little, Brown Spark, New York.
- Yu Kanazawa. 2021. *Replication and extension of an empirical study about foreign language formulaic familiarity database: Reliability and Emotionality*, Bulletin suisse de linguistique appliquée, pages 149–164.
- Yu Kanazawa. 2023. [Lexical and contextual emotional valence in foreign language vocabulary retention](#). *The Mental Lexicon*, 18:339–365.
- Callula Killingly, Linda J. Graham, Haley Tancredi, and Pamela Snow. 2025. [Reciprocal relationships among reading and vocabulary over time: a longitudinal study from grade 1 to 5](#). *Reading and Writing*, 38:605–625.
- Walter Kintsch. 1998. *Comprehension: A Paradigm for Cognition*. Cambridge University Press, Cambridge, UK.
- Stavroula-Thaleia Kousta, Gabriella Vigliocco, David Vinson, Mark Andrews, and Elena Campo. 2010. [The representation of abstract words: Why emotion matters](#). *Journal of experimental psychology. General*, 140:14–34.
- Alper Kumcu and Robin L. Thompson. 2020. [Less imageable words lead to more looks to blank locations during memory retrieval](#). *Psychological Research*, 84:667–684.
- Victor Kuperman, Zachary Estes, Marc Brysbaert, and Amy Warriner. 2014. [Emotion and language: Valence and arousal affect word recognition](#). *Journal of experimental psychology. General*, 143.
- Kristopher Kyle, Scott Crossley, and Cynthia Berger. 2018. [The tool for the automatic analysis of lexical sophistication \(taales\): version 2.0](#). *Behavior Research Methods*, 50:1030–1046.
- Gondy Leroy and David Kauchak. 2014. [The effect of word familiarity on actual and perceived text difficulty](#). *Journal of the American Medical Informatics Association*, 21.
- M. Marschark and C. Cornoldi. 1991. *Imagery and Verbal Memory*, Imagery and cognition, pages 133–182. Springer, New York, US.

- Marina Palazova, Werner Sommer, and Annekathrin Schacht. 2013. [Interplay of emotional valence and concreteness in word processing: An event-related potential study with verbs](#). *Brain and Language*, 125(3):264–271.
- C.A. Perfetti. 1999. *Comprehending written language: A blueprint of the reader*, The neurocognition of language, pages 167–208. Oxford University Press, New York, NY.
- M. Ponari, C. F. Norbury, and G. Vigliocco. 2018. [Acquisition of abstract concepts is influenced by emotional valence](#). *Developmental science*, 21(2).
- Jamie Quinn, Richard Wagner, Yaacov Petscher, and Danielle Lopez. 2015. [Developmental relations between vocabulary knowledge and reading comprehension: A latent change score modeling study](#). *Child Development*, 86(1):159–175.
- James Russell. 1980. [A circumplex model of affect](#). *Journal of Personality and Social Psychology*, 39:1161–1178.
- Sara A. Schmitt, David J. Purpura, and James G. Elicker. 2019. [Predictive links among vocabulary, mathematical language, and executive functioning in preschoolers](#). *Journal of Experimental Child Psychology*, 180:55–68.
- Herbert A. Simon. 1955. [A behavioral model of rational choice](#). *The Quarterly Journal of Economics*, 69(1):99–118.
- Herbert A. Simon. 1972. Theories of bounded rationality. In Charles B. McGuire and Roy Radner, editors, *Decision and Organization*, pages 161–176. North-Holland.
- Jeffrey Stewart, Joseph P. Vitta, Christopher Nicklin, Stuart McLean, Geoffrey G. Pinchbeck, and Brandon Kramer. 2022. [The relationship between word difficulty and frequency: A response to hashimoto \(2021\)](#). *Language Assessment Quarterly*, 19:90–101.
- Y. Sunakawa, J. Lee, and M. Takahara. 2012. [The construction of a database to support the compilation of japanese learners dictionaries](#). *Acta Linguistica Asiatica*, 2(2):97–115.
- Deborah Talmi, Ulrich Schimmack, Theone Paterson, and Morris Moscovitch. 2007. [The role of attention and relatedness in emotionally enhanced memory](#). *Emotion (Washington, D.C.)*, 7:89–102.
- John Truscott. 2015. [Consciousness in sla: A modular perspective](#). *Second Language Research*, 31(3):413–434.
- Gabriella Vigliocco, Stavroula-Thaleia Kousta, Pasquale Rosa, David Vinson, Marco Tettamanti, Joseph Devlin, and Stefano Cappa. 2013. [The neural representation of abstract words: The role of emotion](#). *Cerebral cortex (New York, N.Y. : 1991)*, 24.
- Joseph P. Vitta, Christopher Nicklin, and Simon W. Albright. 2023. [Academic word difficulty and multidimensional lexical sophistication: An english-for-academic-purposes-focused conceptual replication of hashimoto and egbert \(2019\)](#). *Modern Language Journal*, 107:373–397.
- Erina Yamakawa and Keishi Tajima. 2024. [Estimating word difficulty using wikipedia \(in japanese\)](#). In *Proceedings of International Workshop on Database Engineering and Information Management*.
- Wanwan Zheng. 2024. [Estimating word difficulty using stratified word familiarity](#). *Cogent Arts and Humanities*, 11.

Unrequited Emotions: Investigating the Gaps in Motivation and Practice in Speech Emotion Recognition Research

Taryn Wong¹, Zeerak Talat², Hanan Aldarmaki³, Anjalie Field¹

¹Johns Hopkins University, ²University of Edinburgh, ³MBZUAI
z@zeerak.org, hanan.alldarmaki@mbzuai.ac.ae, anjalief@jhu.edu

Abstract

Critical analyses of emotion recognition technology have raised ethical concerns around task validity and potential downstream impacts, urging researchers to ensure alignment between their stated motivations and practice. However, these discussions have not adequately influenced or drawn from research on speech emotion recognition (SER). We address this gap by conducting a systematic survey of SER research to uncover what stated motivations drive this work and if they align with the datasets and emotions studied. We find that while SER research identifies appealing goals—such as well-situated voice-activated systems or healthcare applications—commonly-used datasets do not reflect these proposed deployment contexts, thus presenting a gap between motivations and research practices. We argue that such gaps engender ethical concerns, and that SER research should reassert itself with concrete use-cases to prevent misinterpretations, misuse, and downstream harms.

Keywords: emotion recognition, critical survey, ethics

1. Introduction

Emotion AI, including the detection of human emotions from video, image, audio, or text data using machine learning (ML) methods, has become a popular research area with increasing commercialization in a broad range of applications (McStay, 2018). The growing adoption of this technology has raised concerns around its development and use. Researchers have questioned the overall task validity, e.g., the premise of using someone’s outer appearance to infer inner character or state, particularly given the lack of consensus on how to define emotions among psychologists (McStay and Pavliscak, 2019; Stark and Hoey, 2021). Empirical qualitative studies have demonstrated that people view emotion AI as invasive and privacy-violating (Andalibi and Buss, 2020; Pyle et al., 2024), and its use in contexts like workplaces or hiring processes create emotional labor with little perceived benefit for those subjected to it (Roemmich et al., 2023; Pyle et al., 2024). Importantly, concerns raised are often specific to emotion AI, rather than technology more generally (Pyle et al., 2024).

To date, this work has had limited engagement with the abundance of research on Speech Emotion Recognition (SER). Research on task validity (Jacobs and Wallach, 2021) and ethics have focused on facial recognition or broadly construed AI systems where underlying data and models are unspecified (Domnich and Anbarjafari, 2021; Pyle et al., 2024; Roemmich et al., 2023; Karizat et al., 2024). While some studies have questioned the ethics and validity of SER (Katirai, 2023; Hartmann et al., 2024; Morozov, 2014; Testa et al., 2023; Juslin et al., 2005), no prior work has conducted a

systematic investigation of SER research practices.

In this work, we follow the suggestion proposed by Stark and Hoey (2021) that “[d]esigners and developers should think twice before embarking on emotion AI projects” and we investigate to what extent SER research meets their proposed “necessary though not sufficient condition” that projects have “clear alignment between conceptual models, data, norms, and aims”. We formalize our study into three research questions: (1) What do SER researchers state as their motivations? (2) What purposes do popular datasets support? (3) Do datasets align with stated motivations?

RQ1 focuses on *why* researchers work on SER, while RQ2 focuses on *how*. RQ3 assesses the alignment between the *why* and *how*. We investigate these questions by conducting a systematic survey of 88 SER research papers, which we manually coded for stated motivations, specific emotions studied, and datasets used. We ultimately find a substantial gap between stated motivations and experimental setups, and we propose ways to reduce this gap by seeking suitable datasets or pursuing different motivations. Our work encourages future researchers to “think twice” about SER projects, and to seek alignment between objectives, research practices, and potential implications, rather than viewing SER as an isolated research task.

2. Methodology

Our work uses similar systemic survey methodology as previous studies that reflect on practices in speech processing research and related disciplines (Blodgett et al., 2020; Field et al., 2021; Birhane et al., 2022; Raff et al., 2023). First, we queried Se-

mantic Scholar for papers containing search terms “speech” and “emotion” and {“recognition”, “classification”, “data”, or “database”} in their title or abstract, collecting a total of 7,486 after filtering out papers without publication venue information. Of the initial sample, we kept only those published in popular archival speech, natural language processing, or ML venues, which we determined by manually reviewing venues with the most papers in the initial sample,¹ resulting in 959 papers.

We then randomly sampled 108 (stratified by publication year and venue) for in-depth manual analysis.

We used a standard inductive coding procedure to label the sampled 108 papers. First, three authors independently coded a set of 10 papers each (30 papers total), where each author first determined if the paper consisted of SER. Then, each author labeled the paper’s stated motivations, specific emotions investigated, and data sets used in experiments or analysis.² In labeling stated motivations, annotators also quoted the exact text from the paper describing these motivations. The annotators then discussed the labeling scheme and finalized a set of frequently-stated motivations for which to code. Annotators then re-coded the initial 30 documents under the revised scheme, and then each independently coded an additional set of 10 documents to ensure mutual understanding of the coding scheme. The remainder of the data was then singly coded. Finally, the annotators reconvened and discussed further revisions to the scheme, specifically adding categories of motivations that were found in the larger data sample and re-coded the data under the revised scheme. Throughout this process, we removed 20 papers whose primary focus was not SER (e.g., papers that focused on depression detection, speech generation, speech representation, etc.), leaving a final set of 88 papers. The final set primarily consists of papers published in speech venues (Interspeech: 50 papers, ICASSP: 18, ASRU: 1, SLT: 2, Speech Communication: 4), as well as 9 papers from IEEE Transactions on Affective Computing. The final coding for all 88 papers is in Table 5 in the Appendix.

3. Results

¹ICASSP; ASRU; SLT; Interspeech; NeurIPS; ICLR; ICML; AAAI; IEEE/ACM Transactions on Audio, Speech, and Language Processing; IEEE Open Journal of Signal Processing; IEEE Journal of Selected Topics in Signal Processing; Computer Speech & Language; Speech Communication; IEEE Transactions on Affective Computing; JMLR; ACL

²We also initially coded acknowledged funding sources, but found they were too sparse to analyze in aggregate.

Responsive bots: other HCI systems	42.05%
Responsive bots: voice assistants	12.50%
Responsive bots: car voice assistants	6.82%
Healthcare (mental health)	18.18%
Call screening	17.05%
Video games, toys, entertainment	13.64%
Education	9.09%
Paralinguistics / behavioral studies	6.82%
Social companion bots	4.55%
Lie detection	3.41%
Prior work	27.27%
Other	14.77%

Table 1: Percent of manually coded papers that reference each motivation. Motivations are grouped thematically. Percents do not add to 100 as papers often reference multiple motivations.

What do SER researchers state as their motivations? Table 1 presents the 12 stated motivations identified in our coding scheme with examples for each label in Appendix A. The most common motivation was enabling human-computer interaction systems to respond more naturally and effectively (“Responsive bots”), which we subdivide into “voice assistants”, “car voice assistants”, and “other HCI systems” (typically unspecified systems). Other common motivations focused on specific deployment contexts, including healthcare (almost always mental health), 911 or customer service calls, education, and entertainment. Less common motivations included social companion bots and studies of human behavior or paralinguistics. While most studies include motivations drawn from prior SER work, in data coding, we use the label “Prior work” to indicate that the paper drew motivation *exclusively* from prior SER research without specifying how they expect this technology to be used. We also identified some papers with “Other” motivations. In practice, this label typically indicates that the paper includes a long list of applications, without depth of focus on any one. In Figure 1 we

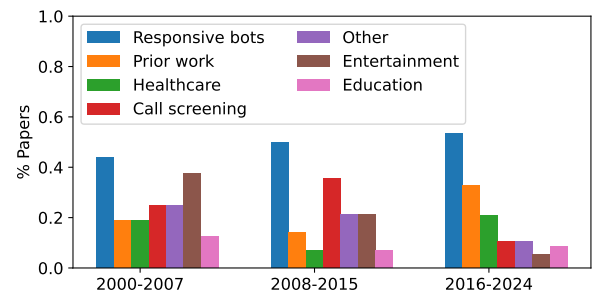


Figure 1: Percent of papers in each time window that reference each motivation. The number of papers in each time window are [16, 14, 58] respectively. We drop infrequent motivations for readability.

additionally report shifts over time. While “Respon-

Dataset	% Papers	Acted/Spont.	Emotions
IEMOCAP	40.91%	Acted (naturalistic)	Angry, Neutral, Sad, Happy, Fearful, Surprised, Frustrated, Excited, Valence/Sentiment, Arousal/Activation, Dominance
EMO-DB	17.05%	Acted	Angry, Neutral, Happy, Fearful, Disgusted, Anxious, Bored
RAVDESS	9.09%	Acted	Produced at both normal & strong intensities: Angry, Neutral, Sad, Happy, Fearful, Surprised, Disgusted, Calm
SUSAS	6.82%	Both	Stressed
MSP-Improv	6.82%	Acted (naturalistic)	Angry, Neutral, Sad, Happy
RECOLA	6.82%	Spontaneous	Valence/Sentiment, Arousal/Activation, Social behaviors rated as either positive or negative: Agreement, Dominance, Engagement, Performance, Rapport
Exclusive	35.23%		Paper exclusively used data in the top 6
Custom	22.73%		Paper custom-collected data
Other	45.45%		Paper used data not in the top 6

Table 2: Datasets used in our analysis corpus ≥ 5 times. We provide mode of collection and emotions labels for each dataset for reference. Categorical labels from Basic Emotion Theory are in **violet**. Axes from dimensional theories are in **teal**.

sive bots” has consistently been a motivating application, “Call screening” and “Entertainment” have declined from 2000-2015 papers to 2016-2024 papers in our sample.

What purposes do popular datasets support?

In Table 2, we identify “popular” datasets as ones used by at least five papers: IEMOCAP (Busso et al., 2008), EMO-DB (Burkhardt et al., 2005), RAVDESS (Livingstone and Russo, 2018), SUSAS (Hansen et al., 1997), MSP-Improv (Busso et al., 2017), and RECOLA (Ringeval et al., 2013). These six datasets support the majority of research in our corpus: 63.64% of papers use at least one of them, and 35.23% exclusively use them. Of the six popular datasets, four (IEMOCAP, EMO-DB, RAVDESS, MSP-Improv) consist of acted emotions and two (SUSAS, RECOLA) consist of emotions captured in spontaneous speech, which encourage vastly different use-cases. For example, SUSAS was created to study speech under different stressful situations, such as operating a helicopter or being subject to psychiatric analysis. In contrast, MSP-Improv was recorded in a laboratory environment by student actors. Furthermore, the datasets are often labeled or validated by annotators who are distinct from the speakers. While SER is often framed as identifying speakers’ true emotions, as would be necessary to support use-cases like call screening or mental

Angry	76.14%	Surprised	19.32%
Neutral	72.73%	Boredom	14.77%
Sad	67.05%	Joy	9.09%
Happy	65.91%	Stressed	7.95%
Fearful	30.68%	Calm	6.82%
Disgusted	27.27%	Dominance	5.68%
Valence/Sentiment	22.73%	Frustrated	3.41%
Arousal/Activation	20.45%	Excited	2.27%
Other/Unspecified	21.59%		

Table 3: Percent of papers that use each emotion label, ordered by frequency. Color-coding is as Table 2.

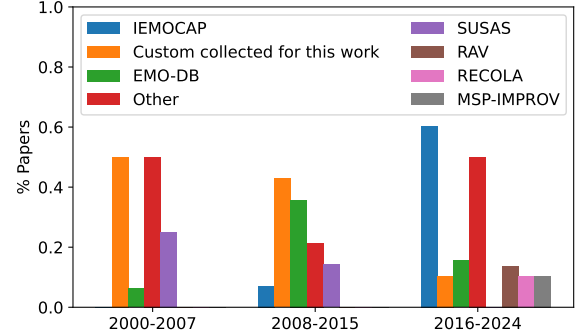


Figure 2: Percent of papers in each time window that use the specified dataset in their experiments.

health support, a more accurate characterization of datasets annotated this way would be identifying how third parties perceive speakers’ emotions.

Further, our analysis of the specific emotions studied in each paper reveals that papers use these datasets selectively (Table 3). Papers study angry, neutral, sad, and happy twice as often as other emotions, even though datasets support other labels, and there is no clear indication that these emotions would better suit the motivations in Table 1. IEMOCAP exemplifies this selective use. Almost all ($> 90\%$) of papers that use IEMOCAP use categorical labels, but far fewer (2-20%) use dimensional labels. Even within these categories usage is not consistent. While 19.44% of IEMOCAP papers use Valence/Sentiment labels, only 2.78% use Dominance.

Curiously, we find that dataset usage varies over time (Figure 2). One notable trend is the decline in data that was custom-collected for the specific paper, which are sometimes, but not always, more directly constructed to support the motivating task. Although 22.73% of the papers in our sample use custom-built datasets overall (Table 2), they are far more common in pre-2016 papers than 2016-2024 papers. Use of SUSAS has also declined, in contrast to IEMOCAP which accounts for almost 60% all data used in 2016-2024. The shifts in dataset usage are generally inconsistent with the shifts in motivations shown in Figure 1. For instance, there is no corresponding rise in interest for a use case

particularly suited to IEMOCAP that would justify its increased popularity in 2016-2024 as compared to 2008-2015 (IEMOCAP was released in 2008).

Do datasets align with stated motivations? In Figure 3, we display the mapping between the most common stated motivations and the six popular datasets, showing a notable lack of pattern. Despite the widely divergent downstream applications, at least one paper with every stated motivation used IEMOCAP. Similarly, although responsive bots or better human-computer interaction has been a persistent motivation for SER (Figure 1), none of the popular data sets were constructed for this application (Table 2), yet they continue to be used for it (Figure 3). Datasets with acted emotions are often used for applications intended for spontaneous speech, yet research has shown that spontaneous emotional expressions differ in various ways from acted expressions (Juslin et al., 2018). Furthermore, most datasets are collected from spontaneous or acted human-human interactions but are frequently used to advance human-computer interaction, even though people express themselves differently and more subtly when interacting with chatbots (Kovacevic et al., 2024), resulting in a significant domain mismatch.

Quantitative metrics of specific datasets and motivations yield similar mismatches. For example, 16 papers exclusively use IEMOCAP, but only 2 of them state Entertainment as a motivation, even though IEMOCAP’s data collection protocol involving scripted and improvised acted speech is most similar to entertainment applications like movies and video games where speech is similarly acted. On the other hand, 16 papers listed Healthcare as a motivation, an application which requires uncovering a speaker’s true emotional state. While 3 of those papers exclusively use custom-collected data, and 2 papers evaluate over SUSAS (which was designed to capture simulated and actual stress), the remaining 11 papers all evaluate over almost entirely acted speech (e.g., 1 paper uses ASVP-ESD (Landry et al., 2020), which is more spontaneous).

4. Discussion

While these results identify a mismatch between stated motivations and underlying datasets, research and deployment are not necessarily expected to be identical. Strong popularity of a small number of datasets, such as frequent use of IEMOCAP in recent years, potentially reflects increasing standardization in evaluation setups, e.g., using established benchmark datasets facilitates publication by enabling direct comparison with prior work. Standardization of tasks has typically been con-

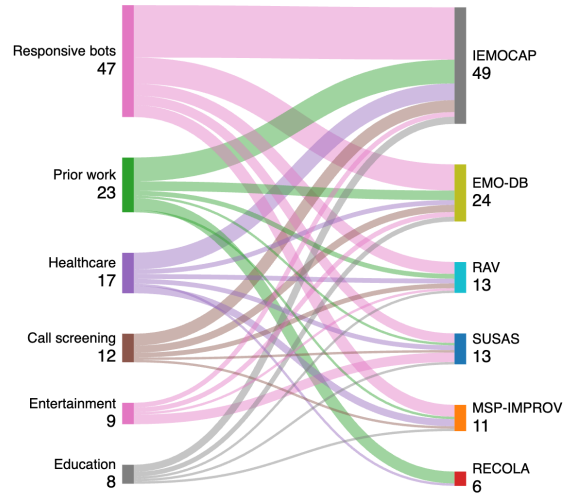


Figure 3: Mapping between most common stated motivations and datasets.

structive for accelerating progress. However, several characteristics of emotion prediction make the mismatch between stated motivations and empirical data problematic. We first discuss why this mismatch is potentially harmful and then offer avenues for reducing it.

Does the mismatch between data and motivations cause harm?

First, there is substantial interest in deploying emotion recognition technology in decision-making contexts where misclassification can have harmful consequences. In their analysis of patent applications for uses of emotion recognition technology in workplace settings, Boyd and Andalibi (2023) find that many patent applications treat outputs of their technology as ground-truth representations of internal states that can inform consequential decisions like hiring, firing, and calling law enforcement. SER systems have been shown to generalize poorly across datasets, even ones collected in similar ways (e.g., acted speech in laboratory settings) (Rí et al., 2023). There is also a fundamental difference between a person’s true internal state and the emotions that an independent observer attributes to them, yet independent annotation to label or validate collected data is standard practice. The implication that a system designed and evaluated over a particular dataset could be useful in a mismatched downstream application can lead to harmful errors.

Additionally, people consider emotions private (McStay, 2020; Pyle et al., 2024), and even well-intentioned motivations are not necessarily well-received by people subject to this technology. In a survey of the attitudes of help seekers towards

predictive tools for mental health risks, only about half of the respondents (49%) were in favor of using risk prediction (Mantell et al., 2021), and similar concerns about the use of AI emotion in education exist (Stark and Hoey, 2021). Because of the sensitive nature of this topic, researchers should set high standards for SER research and ensure there are clear downstream benefits rather than treating SER as a generic task to optimize for and assuming future adoption will be done responsibly.

What data are needed to support these motivations? We next discuss how researchers could use task-appropriate datasets that reflect the intended downstream use cases, focusing for brevity on the most commonly referenced motivation: Responsive bots. Several studies collected datasets that are more aligned with this setup than the popular ones in Table 2. We identify a few examples out of the six papers that list responsive bots as a motivation and custom collect datasets. Wang and Tashuev (2017) work with utterances from interactions with a spoken dialogue system, though emotion labels are separately crowd-sourced. Aubergé et al. (2003) propose a wizard-of-oz system for collecting this type of data, where a user believes they are interacting with a real computer system, but in reality a human is manipulating the system to create particular stimuli. Conducting research with real human-system interactions from deployed systems can introduce privacy concerns, but collecting similar data is feasible in simulated laboratory settings where users can give informed consent.

What use-cases can we support with this data? Alternatively, researchers could better align motivations and datasets by pursuing motivations that popular SER datasets do support. For example, because of its annotation setup, IEMOCAP offers a unique opportunity to investigate the alignment and discrepancies between the perception of emotions—i.e., the annotator’s labels of emotions—and the intended portrayal of emotion—i.e., self-assessed emotions of the actors. Detection of emotions in acted speech could be used to investigate questions around emotional portrayals in acted media such as television shows (Zhou and Bamman, 2024), which has been of interest in gender studies (Weissmann, 2016) and media theory (Gorton, 2009; García, 2016) literature.

Conclusions Our analysis uncovers substantial gaps in how SER research is motivated and conducted. This mismatch exacerbates the risks associated with SER deployment, especially as the technology becomes increasingly commercialized. We recommend ways to reduce this gap, by shifting datasets to match motivations or motivations

to match datasets. Given the sensitivity of emotion recognition technology, we urge researchers to critically evaluate the motivations behind SER projects, ensuring they are well-justified and that experiments meaningfully support them.

5. Ethical Considerations and Limitations

The primary limitation of our work is its reliance on a specific data sample. Although we carefully chose a range of search terms to identify SER papers and we stratify our data sample by year and publication venue, it is possible that analysis of a broader set of papers could yield different findings.

6. Bibliographical References

- Noam Amir, Ori Kerret, and Dimitry Karlinski. 2001. [Classifying emotions in speech: a comparison of methods](#). In *Interspeech*.
- Nazanin Andalibi and Justin Buss. 2020. The human in emotion recognition on social media: Attitudes, outcomes, risks. In *Proc. of CHI, CHI '20*, page 1–16, New York, NY, USA. Association for Computing Machinery.
- Véronique Aubergé, Nicolas Audibert, and Albert Riiliard. 2003. [Why and how to control the authentic emotional speech corpora](#). In *Interspeech*.
- Lokesh Kumar Bansal, S. Pavankumar Dubagunta, Malolan Chetlur, Pushpak Jagtap, and Aravind Ganapathiraju. 2023. [On the efficacy and noise-robustness of jointly learned speech emotion and automatic speech recognition](#). In *Interspeech*.
- Fang Bao, Michael Neumann, and Ngoc Thang Vu. 2019. [CycleGAN-based emotion style transfer as data augmentation for speech emotion recognition](#). In *Interspeech*.
- Dario Bertero and Pascale Fung. 2017. [A first look into a convolutional neural network for speech emotion detection](#). *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5115–5119.
- Abeba Birhane, Pratyusha Kalluri, Dallas Card, William Agnew, Ravit Dotan, and Michelle Bao. 2022. The values encoded in machine learning research. In *Proc. of FAccT*, pages 173–184.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of “bias” in nlp. In *Proc. of ACL*, pages 5454–5476.

- Karen L Boyd and Nazanin Andalibi. 2023. Automated emotion recognition in the workplace: How proposed technologies reveal potential futures of work. *Proc. of CSCW*, 7(CSCW1):1–37.
- F Burkhardt, A Paeschke, M Rolfes, W Sendlmeier, and B Weiss. 2005. A database of german emotional speech. In *Proc. of INTERSPEECH*, pages 805–808.
- Felix Burkhardt, Jitendra Ajmera, Roman Englert, Joachim Stegmann, and Winslow Burleson. 2006. [Detecting anger in automated voice portal dialogs](#). In *Interspeech*.
- C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. Mower Provost. 2017. [MSP-IMPROV: An acted corpus of dyadic interactions to study emotion perception](#). *IEEE Transactions on Affective Computing*, 8(1):67–80.
- Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeanette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.
- Samuel Cahyawijaya, Holy Lovenia, Willy Chung, Rita Frieske, Zihan Liu, and Pascale Fung. 2023. [Cross-lingual cross-age adaptation for low-resource elderly speech emotion recognition](#). In *Interspeech*.
- CasaleSalvatore, RussoAlessandra, and SerranoSalvatore. 2007. [Multistyle classification of speech under stress using feature subset selection based on genetic algorithms](#). *Speech Communication*.
- Ming Chen and Xudong Zhao. 2020. [A multi-scale fusion framework for bimodal speech emotion recognition](#). In *Interspeech*.
- Weidong Chen, Xiaofen Xing, Xiangmin Xu, Jianxin Pang, and Lan Du. 2023. [DST: Deformable speech transformer for emotion recognition](#). *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Woan-Shiuan Chien and Chi-Chun Lee. 2023. [Achieving fair speech emotion recognition via perceptual fairness](#). *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Danilo de Oliveira, Navin Raj Prabhu, and Timo Gerkmann. 2023. [Leveraging semantic information for efficient self-supervised emotion recognition with audio-textual distilled models](#). In *Interspeech*.
- Artem Domnich and Gholamreza Anbarjafari. 2021. Responsible AI: Gender bias assessment in emotion recognition. *ArXiv*, abs/2103.11436.
- Tiantian Feng and Shrikanth S. Narayanan. 2022. [Semi-fedser: Semi-supervised learning for speech emotion recognition on federated learning using multiview pseudo-labeling](#). In *Interspeech*.
- Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. In *Proc. of ACL*, pages 1905–1925.
- Yuan Gao, Chenhui Chu, and Tatsuya Kawahara. 2023. [Two-stage finetuning of wav2vec 2.0 for speech emotion recognition with asr and gender pretraining](#). In *Interspeech*.
- Alberto N. García, editor. 2016. *Emotions in Contemporary TV Series*. Palgrave Macmillan UK, London.
- Jakub Gałka, Joanna Grzybowska, Magdalena Igras, Pawel Jaciów, Kamil Wajda, Marcin Witkowski, and Mariusz Ziółko. 2015. [System supporting speaker identification in emergency call center](#). In *Interspeech*.
- Theodoros Giannakopoulos, Aggelos Pikrakis, and Sergios Theodoridis. 2009. [A dimensional approach to emotion recognition of speech from movies](#). *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 65–68.
- Lucas Goncalves and Carlos Busso. 2022. [Improving speech emotion recognition using self-supervised learning with domain-specific audio-visual tasks](#). In *Interspeech*.
- Erik Goron, Lena Asai, Elias Rut, and Martin Dinov. 2024. [Improving domain generalization in speech emotion recognition with whisper](#). *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11631–11635.
- Kristyn Gorton. 2009. *Media Audiences: Television, Meaning and Emotion*. Edinburgh University Press.
- Michael Grimm, Kristian Kroschel, and Shrikanth S. Narayanan. 2007. [Support vector regression for automatic recognition of spontaneous emotions in speech](#). *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, 4:IV–1085–IV–1088.
- Purnima Gupta and Nitendra Rajput. 2007. [Two-stream emotion recognition for call center monitoring](#). In *Interspeech*.

- Jing Han, Zixing Zhang, Zhao Ren, Fabien Ringeval, and Björn Schuller. 2018. [Towards conditional adversarial training for predicting emotions from speech](#). *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6822–6826.
- Wenjing Han, Haifeng Li, Huabin Ruan, Lin Ma, Jiayin Sun, and Björn Schuller. 2013. [Active learning for dimensional speech emotion recognition](#). In *Interspeech*.
- John HL Hansen, Sahar E Bou-Ghazale, Ruhi Sarikaya, and Bryan Pellom. 1997. Getting started with SUSAS: a speech under simulated and actual stress database. In *Eurospeech*, pages 1743–46.
- Kris Vera Hartmann, Giovanni Rubeis, and Nadia Primc. 2024. Healthy and happy? an ethical investigation of emotion recognition and regulation technologies (err) within ambient assisted living (aal). *Science and Engineering Ethics*, 30(1):2.
- Mixiao Hou, Zheng Zhang, Qi Cao, David Zhang, and Guangming Lu. 2022. [Multi-view speech emotion recognition via collective relation construction](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:218–229.
- Desheng Hu, Xinhui Hu, and Xinkang Xu. 2022. [Multiple enhancements to lstm for learning emotion-salient features in speech emotion recognition](#). In *Interspeech*.
- Jian Huang, Ya Li, Jianhua Tao, and Zheng Lian. 2018. [Speech emotion recognition from variable-length inputs with triplet loss function](#). In *Interspeech*.
- Kun-Yi Huang, Chung-Hsien Wu, Ming-Hsiang Su, and Yu-Ting Kuo. 2020. [Detecting unipolar and bipolar depressive disorders from elicited speech responses using latent affective structure model](#). *IEEE Transactions on Affective Computing*, 11:393–404.
- Yu-Lin Huang, Bo-Hao Su, Y.-W. Peter Hong, and Chi-Chun Lee. 2021. [An attribute-aligned strategy for learning speech representation](#). *ArXiv*, abs/2106.02810.
- Richard Huber, Anton Batliner, Jan Buckow, Elmar Nöth, Volker Warnke, and Heinrich Niemann. 2000. [Recognition of emotion in a realistic dialogue scenario](#). In *Interspeech*.
- Abigail Z. Jacobs and Hanna Wallach. 2021. [Measurement and Fairness](#). In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385, Virtual Event Canada. ACM.
- Christian Martyn Jones and Andrew Deeming. 2008. [Speech interaction with an emotional robotic dog](#). In *Interspeech*.
- Patrik N Juslin, Petri Laukka, and Tanja Bänziger. 2018. The mirror to our soul? comparisons of spontaneous and posed vocal expression of emotion. *Journal of nonverbal behavior*, 42:1–40.
- Patrik N Juslin, Klaus R Scherer, J Harrigan, and R Rosenthal. 2005. Vocal expression of affect. *The new handbook of methods in nonverbal behavior research*, pages 65–135.
- Zuheng Kang, Junqing Peng, Jianzong Wang, and Jing Xiao. 2022. [SpeechEQ: Speech emotion recognition based on multi-scale unified datasets and multitask learning](#). In *Interspeech*.
- Nadia Karizat, Alexandra H Vinson, Shobita Parthasarathy, and Nazanin Andalibi. 2024. Patent applications as glimpses into the sociotechnical imaginary: Ethical speculation on the imagined futures of emotion ai for mental health monitoring and detection. *Proc. of CSCW*, 8(CSCW1):1–43.
- Amelia Katirai. 2023. Ethical considerations in emotion recognition technologies: a review of the literature. *AI and Ethics*, pages 1–22.
- Jonghwa Kim, Elisabeth André, Matthias Rehm, Thuid Vogt, and Johannes Wagner. 2005. [Integrating information from speech and physiological signals to achieve emotional sensitivity](#). In *Interspeech*.
- Nikola Kovacevic, Christian Holz, Markus Gross, and Rafael Wampfler. 2024. On multimodal emotion recognition for human-chatbot interaction in the wild. In *Proceedings of the 26th International Conference on Multimodal Interaction*, pages 12–21.
- O. W. Kwon, Kwokleung Chan, Jiucang Hao, and Te-Won Lee. 2003. [Emotion recognition by speech signals](#). In *Interspeech*.
- Dejoli Landry, Qianhua He, Haikang Yan, and Yanxiong Li. 2020. Asvp-esd: A dataset and its benchmark for emotion recognition using both speech and non-speech utterances. *Global Scientific Journals*, 8:1793–1798.
- Nineli Lashkarashvili, Wen Wu, Guangzhi Sun, and Philip C. Woodland. 2024. [Parameter efficient finetuning for speech emotion recognition and domain adaptation](#). *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10986–10990.

- Siddique Latif, Rajib Kumar Rana, Sara Khalifa, Raja Jurdak, Julien Epps, and Björn Schuller. 2020a. [Multi-task semi-supervised adversarial autoencoding for speech emotion recognition](#). *IEEE Transactions on Affective Computing*, 13:992–1004.
- Siddique Latif, Rajib Kumar Rana, Sara Khalifa, Raja Jurdak, and Björn Schuller. 2020b. [Deep architecture enhancing robustness to noise, adversarial attacks, and cross-corpus setting for speech emotion recognition](#). In *Interspeech*.
- Baojie Li, Keikichi Hirose, and Nobuaki Minematsu. 2002. [Robust speech recognition using inter-speaker and intra-speaker adaptation](#). In *Interspeech*.
- Chia-Yu Li, Daniel Ortega, Dirk Vath, Florian Lux, Lindsey Vanderlyn, Maximilian Schmidt, Michael Neumann, Moritz Volkel, Pavel Denisov, Sabrina Jenne, Zorica Kacarevic, and Ngoc Thang Vu. 2020. [Adviser: A toolkit for developing multi-modal, multi-domain and socially-engaged conversational agents](#). In *Annual Meeting of the Association for Computational Linguistics*.
- Xi Li, Jidong Tao, Michael T. Johnson, Joseph Soltis, Anne Savage, Kirsten M. Leong, and John D. Newman. 2007. [Stress and emotion classification using jitter and shimmer features](#). *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, 4:IV–1081–IV–1084.
- Xingfeng Li, Xiaohan Shi, Desheng Hu, Yongwei Li, Qingcheng Zhang, Zhengxia Wang, Masashi Unoki, and Masato Akagi. 2023a. [Music theory-inspired acoustic representation for speech emotion recognition](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2534–2547.
- Yuanchao Li, Yumnah Mohamied, Peter Bell, and Catherine Lai. 2022. [Exploration of a self-supervised speech model: A study on emotional corpora](#). *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 868–875.
- Yuanchao Li, Zeyu Zhao, Ondrej Klejch, Peter Bell, and Catherine Lai. 2023b. [Asr and emotional speech: A word-level investigation of the mutual impact of speech and emotion recognition](#). In *Interspeech*.
- Steven R Livingstone and Frank A Russo. 2018. The ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multi-modal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.
- Ignacio López-Moreno, Carlos Ortego-Resca, Joaquín González-Rodríguez, and Daniel Ramos. 2009. [Speaker dependent emotion recognition using prosodic supervectors](#). In *Interspeech*.
- Veronika Makarova and Valery A. Petrushin. 2002. [Ruslana: a database of russian emotional utterances](#). In *Interspeech*.
- Pauline Katharina Mantell, Annika Baumeister, Stephan Ruhrmann, Anna Janhsen, and Christiane Woopen. 2021. Attitudes towards risk prediction in a help seeking population of early detection centers for mental disorders—a qualitative approach. *International journal of environmental research and public health*, 18(3):1036.
- Shuiyang Mao, Pak-Chung Ching, and Tan Lee. 2019. [Deep learning of segment-level feature representation with multiple instance learning for utterance-level speech emotion recognition](#). In *Interspeech*.
- Sho Matsumiya, Sakriani Sakti, Graham Neubig, Tomoki Toda, and Satoshi Nakamura. 2014. [Data-driven generation of text balloons based on linguistic and acoustic features of a comics-anime corpus](#). In *Interspeech*.
- A. McStay. 2018. *Emotional AI: The Rise of Empathic Media*. SAGE Publications Limited.
- Andrew McStay. 2020. Emotional ai, soft biometrics and the surveillance of emotional life: An unusual consensus on privacy. *Big Data & Society*, 7(1):2053951720904386.
- Andrew McStay and Pamela Pavliscak. 2019. Emotional artificial intelligence: Guidelines for ethical use. *COMEST/UNESCO*.
- Hanyu Meng, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. 2023. [What is learnt by the learnable front-end \(leaf\)? adapting per-channel energy normalisation \(pcen\) to noisy conditions](#). In *Interspeech*.
- Patrick Meyer and Tim Fingscheidt. 2014. [A new evaluation methodology for speech emotion recognition with confidence output](#). In *ITG Symposium on Speech Communication*.
- Rosanna Milner, Md. Asif Jalal, Raymond W. M. Ng, and Thomas Hain. 2019. [A cross-corpus study on speech emotion recognition](#). *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 304–311.
- Vikramjit Mitra, Jingping Nie, and Erdrin Azemi. 2023. [Investigating salient representations and label variance in dimensional speech emotion](#)

- analysis. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11111–11115.
- EVGENY Morozov. 2014. To save everything, click here: the folly of technological solutionism. *J. Inf. Policy*, 4(2014):173–175.
- Anish Nediyanthath, Periyasamy Paramasivam, and Promod Yenigalla. 2020. [Multi-head attention for speech emotion recognition with auxiliary learning of gender recognition](#). *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7179–7183.
- Albino Nogueiras, Asunción Moreno, Antonio Bonafonte, and José B. Mariño. 2001. [Speech emotion recognition using hidden markov models](#). In *Interspeech*.
- Christopher Oates, Andreas Triantafyllopoulos, Ingmar Steiner, and Björn Schuller. 2019. [Robust speech emotion recognition under different encoding conditions](#). In *Interspeech*.
- Georgios Paraskevopoulos, Efthymios Tzinis, Nikolaos Ellinas, Theodoros Giannakopoulos, and Alexandros Potamianos. 2019. [Unsupervised low-rank representations for speech emotion recognition](#). In *Interspeech*.
- Zi Jun Peng, Yu Lu, Shengfeng Pan, and Yunfeng Liu. 2021. [Efficient speech emotion recognition using multi-scale cnn and attention](#). *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3020–3024.
- Hai X. Pham, Yuting Wang, and Vladimir Pavlovic. 2022. [Learning continuous facial actions from speech for real-time animation](#). *IEEE Transactions on Affective Computing*, 13:1567–1580.
- Navin Raj Prabhu, Nale Lehmann-Willenbrock, and Timo Gerkmann. 2022. [End-to-end label uncertainty modeling in speech emotion recognition using bayesian neural networks and label distribution learning](#). *IEEE Transactions on Affective Computing*, 15:579–592.
- Emily Mower Provost, Maja J. Matarić, and Shrikanth S. Narayanan. 2009. [Evaluating evaluators: a case study in understanding the benefits and pitfalls of multi-evaluator modeling](#). In *Interspeech*.
- Cassidy Pyle, Kat Roemmich, and Nazanin Andalibi. 2024. Us job-seekers' organizational justice perceptions of emotion ai-enabled interviews. *Proc. of CSCW*, 8(CSCW2):1–42.
- Edward Raff, Michel Benaroch, Sagar Samtani, and Andrew L Farris. 2023. What do machine learning researchers mean by "reproducible"? In *Proceedings of INTERSPEECH*.
- Mandar Rahurkar and John H. L. Hansen. 2003. [Frequency distribution based weighted sub-band approach for classification of emotional/stressful content in speech](#). In *Interspeech*.
- Eva-Maria Rathner, Yannik Terhorst, Nicholas Cummins, Björn Schuller, and Harald Baumeister. 2018. [State of mind: Classification through self-reported affect and word use in speech](#). In *Interspeech*.
- Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. 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)*, pages 1–8. IEEE.
- Kat Roemmich, Florian Schaub, and Nazanin Andalibi. 2023. Emotion ai at work: Implications for workplace surveillance, emotional labor, and emotional privacy. In *Proc. of CHI*, CHI '23, New York, NY, USA. Association for Computing Machinery.
- Francesco Ardan Dal Rí, Fabio Cifariello Ciardi, and Nicola Conci. 2023. Speech emotion recognition and deep learning: An extensive validation using convolutional neural networks. *IEEE Access*, 11:116638–116649.
- Saurabh Sahu, Vikramjit Mitra, Nadee Seneviratne, and Carol Y. Espy-Wilson. 2019. [Multi-modal learning for speech emotion recognition: An analysis and comparison of asr outputs with ground truth transcription](#). In *Interspeech*.
- Michelle Hewlett Sanchez, Gökhan Tür, Luciana Ferrer, and Dilek Z. Hakkani-Tür. 2010. [Domain adaptation and compensation for emotion detection](#). In *Interspeech*.
- Izhak Shafran and Mehryar Mohri. 2005. [A comparison of classifiers for detecting emotion from speech](#). *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, 1:1/341–1/344 Vol. 1.
- Abhinav Shukla, Stavros Petridis, and Maja Pantic. 2020. [Does visual self-supervision improve learning of speech representations for emotion recognition?](#) *IEEE Transactions on Affective Computing*, 14:406–420.
- Julia Sidorova, Simon Karlsson, Oliver Rosander, Marcelo L. Berthier, and Ignacio Moreno-Torres.

2021. Towards disorder-independent automatic assessment of emotional competence in neurological patients with a classical emotion recognition system: Application in foreign accent syndrome. *IEEE Transactions on Affective Computing*, 12:962–973.
- Peng Song and Wenming Zheng. 2020. Feature selection based transfer subspace learning for speech emotion recognition. *IEEE Transactions on Affective Computing*, 11:373–382.
- Luke Stark and Jesse Hoey. 2021. The ethics of emotion in artificial intelligence systems. In *Proc. of FAccT, FAccT '21*, page 782–793, New York, NY, USA. Association for Computing Machinery.
- Bo-Hao Su and Chi-Chun Lee. 2021. A conditional cycle emotion gan for cross corpus speech emotion recognition. *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 351–357.
- Rui Sun and Elliot Moore II. 2012. A preliminary study on cross-databases emotion recognition using the glottal features in speech. In *INTER-SPEECH*, pages 1628–1631.
- Brian Testa, Yi Xiao, Harshit Sharma, Avery Gump, and Asif Salekin. 2023. Privacy against real-time speech emotion detection via acoustic adversarial evasion of machine learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(3):1–30.
- Andreas Triantafyllopoulos, Gil Keren, Johannes Wagner, Ingmar Steiner, and Björn Schuller. 2019. Towards robust speech emotion recognition using deep residual networks for speech enhancement. In *Interspeech*.
- Panagiotis Tzirakis, Anh-Tuan Nguyen, Stefanos Zafeiriou, and Björn W. Schuller. 2021. Speech emotion recognition using semantic information. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6279–6283.
- Bogdan Vlasenko, Björn Schuller, Kinfe Tadesse Mengistu, Gerhard Rigoll, and Andreas Wendemuth. 2008. Balancing spoken content adaptation and unit length in the recognition of emotion and interest. In *Interspeech*.
- Bogdan Vlasenko, Björn Schuller, Andreas Wendemuth, and Gerhard Rigoll. 2007. Combining frame and turn-level information for robust recognition of emotions within speech. In *Interspeech*.
- Bogdan Vlasenko and Andreas Wendemuth. 2009. Processing affected speech within human machine interaction. In *Interspeech*.
- Hongxuan Wang and Prahlad Vadakkepat. 2024. Gradient-based dimensionality reduction for speech emotion recognition using deep networks. *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11496–11500.
- Zhong-Qiu Wang and Ivan Tashev. 2017. Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5150–5154.
- Elke Weissmann. 2016. Women, Television and Feelings: Theorising Emotional Difference of Gender in SouthLAnd and Mad Men. In Alberto N. García, editor, *Emotions in Contemporary TV Series*, pages 87–101. Palgrave Macmillan UK, London.
- Chung-Hsien Wu and Wei-Bin Liang. 2011. Emotion recognition of affective speech based on multiple classifiers using acoustic-prosodic information and semantic labels. *IEEE Transactions on Affective Computing*, 2:10–21.
- Jingyao Wu, Ting Dang, Vidhyasaharan Sethu, and Eliathamby Ambikairajah. 2021. A novel markovian framework for integrating absolute and relative ordinal emotion information. *IEEE Transactions on Affective Computing*, 14:2089–2101.
- Wen Wu, C. Zhang, and Philip C. Woodland. 2023. Integrating emotion recognition with speech recognition and speaker diarisation for conversations. *ArXiv*, abs/2308.07145.
- Jie Xie, Mingying Zhu, and Kai Hu. 2023. Fusion-based speech emotion classification using two-stage feature selection. *Speech Commun.*, 152:102955.
- Zixiaofan Yang and Julia Hirschberg. 2018. Predicting arousal and valence from waveforms and spectrograms using deep neural networks. In *Interspeech*.
- Jiaxin Ye, Xin-Cheng Wen, X. Wang, Yan Luo, Chang-Li Wu, Liyan Chen, and Kunhong Liu. 2022. Gm-tcnet: Gated multi-scale temporal convolutional network using emotion causality for speech emotion recognition. *ArXiv*, abs/2210.15834.
- Promod Yenigalla, Abhay Kumar, Suraj Tripathi, Chirag Singh, Sibsambhu Kar, and Jithendra Vepa. 2018. Speech emotion recognition using spectrogram & phoneme embedding. In *Interspeech*.

- Seunghyun Yoon, Seokhyun Byun, Subhadeep Dey, and Kyomin Jung. 2019. [Speech emotion recognition using multi-hop attention mechanism](#). *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2822–2826.
- Sungrack Yun and Chang Dong Yoo. 2009. [Speech emotion recognition via a max-margin framework incorporating a loss function based on the watson and tellegen's emotion model](#). *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4169–4172.
- Noor Aina Zaidan and Sah Hj Salam. 2016. [Mfcc global features selection in improving speech emotion recognition rate](#). In *International Conference on Machine Learning*.
- Ruoxi Zhang, Atsushi Ando, Satoshi Kobashikawa, and Yushi Aono. 2017. [Interaction and transition model for speech emotion recognition in dialogue](#). In *Interspeech*.
- Xiaoheng Zhang and Y. Li. 2023. [A dual attention-based modality-collaborative fusion network for emotion recognition](#). In *Interspeech*.
- Ziping Zhao, Zhongtian Bao, Zixing Zhang, Nicholas Cummins, Haishuai Wang, and Björn Schuller. 2019. [Attention-enhanced connectionist temporal classification for discrete speech emotion recognition](#). In *Interspeech*.
- Ziping Zhao, Yu Zheng, Zixing Zhang, Haishuai Wang, Yiqin Zhao, and Chao Li. 2018. [Exploring spatio-temporal representations by integrating attention-based bidirectional-lstm-rnns and fcns for speech emotion recognition](#). In *Interspeech*.
- Naitian Zhou and David Bamman. 2024. Once more, with feeling: Measuring emotion of acting performances in contemporary american film. In *Computational Humanities Research (CHR)*.

A. Examples of coded motivations

Responsive bots: other HCI systems	"These models could lead to computer agents and robots that more naturally and functionally blend into human society" (Provost et al., 2009)
Responsive bots: voice assistants; Responsive bots: other HCI systems	"Speech emotion recognition is becoming more and more important for many applications related to human computer interactions, especially for spoken dialogue systems. With emotion recognition from users' speech, a better user experience can be achieved" (Wang and Tashev, 2017)
Healthcare (mental health); Responsive bots: car voice assistants; Education; Other	"In affective computing field, emotion recognition from speech plays a very important role, and has received much attention over the past few decades... It has been proven useful in many applications which require human-machine interaction, e.g., in-car board system, diagnostic tool for therapists, automatic translation systems, computer tutorial applications" (Song and Zheng, 2020)
Call screening	"In a call-center, such a system should be able to determine in a critical phase of the dialogue if the call should be passed over to a human operator." (Huber et al., 2000)
Paralinguistics / behavioral studies	"Speech signals carry rich information on an individual's emotional states, expressed through both paralinguistic and semantic cues." (de Oliveira et al., 2023)
Social companion bots; Video games, toys, entertainment; Education; Other	"Research on emotion recognition can advance many applications, like distance education, social robots, video games, affective mirrors and many others." (Li et al., 2023a)
Lie detection; Call screening; Healthcare (mental health)	"Especially in the field of human-machine interaction (HCI), growing interest can be observed in recent years. In addition, the detection of lies, monitoring of call centers and medical diagnoses are often claimed as promising application scenarios for speech emotion recognition." (Mao et al., 2019)
Prior work	"While the majority of traditional research in emotional speech recognition has utilized a single database for analysis, it is becoming clear that the lack of sufficiently large databases with varied emotion types is a significant hindrance to the design of a robust emotion classification system." (Sun and Moore II, 2012)

Table 4: Examples of human-selected text snippets and coded motivations, with at least one example per motivation. We remove citations from within quotations for readability.

Table 5: All 88 manually coded papers

	Venue	Motivations	Datasets	Emotions
Huber et al. (2000)	Interspeech	Call screening	Custom	Angry, Other or Unspecified
Amir et al. (2001)	Interspeech	Prior work	Universidad Politecnica de Madrid	Angry, Sad, Happy, Neutral
Nogueiras et al. (2001)	Interspeech	Healthcare, Entertainment	Interface Database	Surprised, Joy, Angry, Fearful, Disgusted, Sad, Neutral
Li et al. (2002)	Interspeech	Prior work	Custom	Angry, Other or Unspecified, Sad, Neutral
Makarova and Petrushin (2002)	Interspeech	Entertainment, Other	RUSLANA	Neutral, Surprised, Happy, Angry, Fearful, Sad
Rahurkar and Hansen (2003)	Interspeech	Paralinguistic / behavioral studies, RB: voice assistants, Entertainment, RB: other HCI systems	SOQ	Neutral, Stressed
Aubergé et al. (2003)	Interspeech	RB: other HCI systems, RB: voice assistants	Custom	Other or Unspecified
Kwon et al. (2003)	Interspeech	Entertainment, RB: other HCI systems	SUSAS, FAU Aibo Emotion Corpus	Stressed, Angry, Boredom, Happy, Neutral, Sad
Kim et al. (2005)	Interspeech	RB: other HCI systems	Custom	Arousal/Activation, Valence or sentiment
Shafran and Mohri (2005)	ICASSP	RB: other HCI systems, Other, Call screening, Education	Custom	Valence or sentiment
Burkhardt et al. (2006)	Interspeech	Call screening	Custom	Angry
Gupta and Rajput (2007)	Interspeech	Call screening	EPST, Custom	Angry, Happy, Neutral
CasaleSalvatore et al. (2007)	Speech Communication	RB: other HCI systems, Healthcare, Other, Entertainment, Education	SUSAS	Angry, Stressed, Neutral
Li et al. (2007)	ICASSP	Healthcare, Other, Lie detection, Entertainment	SUSAS, African Elephant Emotional Arousal Dataset, Custom	Other or Unspecified, Angry, Arousal/Activation, Neutral
Vlasenko et al. (2007)	Interspeech	Prior work	EMO-DB, SUSAS	Happy, Angry, Fearful, Boredom, Disgusted, Stressed, Neutral
Grimm et al. (2007)	ICASSP	RB: other HCI systems	VAM	Valence or sentiment, Dominance, Other or Unspecified
Jones and Deeming (2008)	Interspeech	Call screening, Other, RB: car voice assistants, Entertainment, Lie detection	Custom	Happy, Boredom, Sad, Surprised, Angry, Other or Unspecified, Excited, Fearful, Frustrated
Vlasenko et al. (2008)	Interspeech	RB: other HCI systems	EMO-DB, SUSAS	Angry, Boredom, Disgusted, Fearful, Joy, Neutral, Sad, Stressed
Vlasenko and Wendemuth (2009)	Interspeech	RB: other HCI systems	Kiel Corpus of Read Speech , EMO-DB, FAU Aibo Emotion Corpus	Angry, Other or Unspecified, Neutral, Joy

Continued on next page

	Venue	Motivations	Datasets	Emotions
Provost et al. (2009)	Interspeech	RB: other HCI systems	IEMOCAP	Angry, Happy, Sad, Neutral, Valence or sentiment, Arousal/Activation
Yun and Yoo (2009)	ICASSP	Call screening, Entertainment, RB: other HCI systems	EMO-DB	Angry, Fearful, Disgusted, Sad, Boredom, Neutral, Happy
López-Moreno et al. (2009)	Interspeech	RB: car voice assistants, Entertainment, Call screening	SUSAS	Neutral, Stressed
Giannakopoulos et al. (2009)	ICASSP	Other	Custom	Valence or sentiment, Arousal
Sanchez et al. (2010)	Interspeech	Call screening	Custom	Fearful, Sad, Neutral
Wu and Liang (2011)	IEEE Transactions on Affective Computing	Healthcare, Education	Custom	Angry, Happy, Neutral, Sad
Sun and Moore II (2012)	Interspeech	Prior work	EMO-DB, EPST, EMA	Neutral, Angry, Sad, Happy
Han et al. (2013)	Interspeech	Prior work	SEMAINE	Arousal/Activation, Dominance, Valence or sentiment, Other or Unspecified
Matsumiya et al. (2014)	Interspeech	Other	Custom	Angry, Sad, Happy, Fearful, Neutral
Meyer and Fingscheidt (2014)	Speech Communication	RB: other HCI systems	EMO-DB	Fearful, Disgusted, Happy, Boredom, Neutral, Sad, Angry
Gaika et al. (2015)	Interspeech	Call screening	Custom	Angry, Stressed, Neutral
Zaidan and Salam (2016)	Advances in Machine Learning and Signal Processing	RB: other HCI systems	EMO-DB	Happy, Sad, Fearful, Angry, Disgusted, Boredom, Neutral
Zhang et al. (2017)	Interspeech	RB: other HCI systems, Other, Entertainment, Education	IEMOCAP	Angry, Happy, Sad, Neutral, Other or Unspecified
Wang and Tashev (2017)	ICASSP	RB: other HCI systems, RB: voice assistants	Custom	Neutral, Happy, Sad, Angry
Bertero and Fung (2017)	ICASSP	Prior work	Custom	Angry, Happy, Sad
Rathner et al. (2018)	Interspeech	Healthcare	Custom	Arousal/Activation, Valence or sentiment
Yang and Hirschberg (2018)	Interspeech	Prior work	RECOLA, SEMAINE	Valence or sentiment, Arousal
Han et al. (2018)	ICASSP	Prior work	RECOLA	Arousal/Activation, Valence or sentiment
Zhao et al. (2018)	Interspeech	Prior work	CHEAVD, IEMOCAP	Angry, Disgusted, Happy, Sad, Surprised, Neutral, Other or Unspecified
Huang et al. (2018)	Interspeech	Paralinguistic / behavioral studies	IEMOCAP	Angry, Happy: Happy + Excited, Neutral, Sad
Yenigalla et al. (2018)	Interspeech	Prior work	IEMOCAP	Neutral, Happy, Sad, Angry

Continued on next page

	Venue	Motivations	Datasets	Emotions
Milner et al. (2019)	ASRU	Prior work	eINTERFACE, RAVDESS (RAV), IEMOCAP, MOSEI	Angry, Happy, Sad, Surprised, Disgusted, Fearful, Neutral, Frustrated, Excited, Calm, Other or Unspecified
Bao et al. (2019)	Interspeech	RB: other HCI systems	IEMOCAP, MSP-IMPROV, TEDLIUM (release 2)	Angry, Happy: Happy + Excited, Sad, Neutral
Mao et al. (2019)	Interspeech	Healthcare, Call screening, Lie detection	IEMOCAP, CASIA	Sad, Neutral, Angry, Fearful, Happy, Surprised
Oates et al. (2019)	Interspeech	Prior work	EMO-DB, RECOLA, eINTERFACE, Polish Emotional Speech Database	Angry, Boredom, Fearful, Joy, Neutral, Sad, Disgusted, Happy, Surprised, Arousal/Activation, Valence or sentiment
Latif et al. (2020a)	IEEE Transactions on Affective Computing	Healthcare, Call screening, RB: other HCI systems	IEMOCAP, MSP-IMPROV	Angry, Happy, Neutral, Sad
Yoon et al. (2019)	ICASSP	RB: other HCI systems, Paralinguistic / behavioral studies	IEMOCAP	Happy: Happy + Excited, Angry, Neutral, Sad
Zhao et al. (2019)	Interspeech	Prior work	IEMOCAP, FAU Aibo Emotion Corpus	Happy: Happy + Excited, Angry, Sad, Neutral, Emphatic, Positive, Rest
Triantafyllopoulos et al. (2019)	Interspeech	Prior work	RECOLA, EMO-DB, eINTERFACE, Mozilla Common Voice Audio Set	Arousal/Activation, Other or Unspecified
Sahu et al. (2019)	Interspeech	Healthcare, Paralinguistic / behavioral studies, RB: voice assistants	IEMOCAP, MSP-IMPROV	Angry, Happy, Neutral, Sad
Paraskevopoulos et al. (2019)	Interspeech	RB: other HCI systems	EMO-DB, IEMOCAP	Angry, Disgusted, Fearful, Joy, Boredom, Sad, Neutral, Happy
Huang et al. (2020)	IEEE Transactions on Affective Computing	Healthcare	Custom	Happy, Fearful, Angry, Surprised, Sad, Disgusted
Song and Zheng (2020)	IEEE Transactions on Affective Computing	Healthcare, Other, RB: voice assistants, Education	EMO-DB, eINTERFACE, FAU Aibo Emotion Corpus	Angry, Boredom, Disgusted, Fearful, Happy, Sad, Surprised
Shukla et al. (2020)	IEEE Transactions on Affective Computing	Healthcare, Other	CREMA-D, RECOLA, RAVDESS (RAV), SEWA, IEMOCAP	Angry, Fearful, Disgusted, Happy, Sad, Neutral, Calm, Surprised, Arousal/Activation, Valence or sentiment
Li et al. (2020)	ACL Demos	RB: voice assistants, Social Companion bots	IEMOCAP	Angry, Happy, Sad, Neutral, Arousal/Activation, Valence or sentiment
Latif et al. (2020b)	Interspeech	Prior work	IEMOCAP, MSP-IMPROV, DEMAND	Angry, Neutral, Sad, Happy: Happy + Excited
Chen and Zhao (2020)	Interspeech	RB: other HCI systems	IEMOCAP	Happy: Happy + Excited, Neutral, Angry, Sad

Continued on next page

	Venue	Motivations	Datasets	Emotions
Nediyanchath et al. (2020)	ICASSP	RB: voice assistants, Social Companion bots, RB: other HCI systems	IEMOCAP	Sad, Angry, Neutral, Happy
Sidorova et al. (2021)	IEEE Transactions on Affective Computing	Healthcare	Custom, Interface Database	Angry, Sad, Happy, Neutral
Huang et al. (2021)	Interspeech	Prior work	MSP-Podcasts	Neutral, Angry, Sad, Happy, Disgusted
Wu et al. (2021)	IEEE Transactions on Affective Computing	Prior work	RECOLA, IEMOCAP	Arousal/Activation, Valence or sentiment
Tzirakis et al. (2021)	ICASSP	RB: other HCI systems	SEWA	Arousal/Activation, Other or Unspecified, Valence or sentiment
Su and Lee (2021)	SLT	RB: other HCI systems, RB: car voice assistants	IEMOCAP, MSP-IMPROV, CIT: USC CreativeIT Corpus	Arousal/Activation, Valence or sentiment
Peng et al. (2021)	ICASSP	Call screening, Healthcare, RB: other HCI systems	IEMOCAP	Angry, Sad, Neutral, Happy: Happy + Excited
Li et al. (2022)	SLT	Prior work	IEMOCAP, RAVDESS (RAV)	Angry, Happy: Happy + Excited, Neutral, Happy, Calm, Sad, Fearful, Surprised, Disgusted
Pham et al. (2022)	IEEE Transactions on Affective Computing	RB: other HCI systems	RAVDESS (RAV), SAVEE, VidTIMIT, GRID, GEMEP	Neutral, Calm, Happy, Sad, Angry, Fearful, Surprised, Disgusted
Prabhu et al. (2022)	IEEE Transactions on Affective Computing	Prior work	AVEC'16, MSP-Conversation	Valence or sentiment, Arousal
Hu et al. (2022)	Interspeech	RB: other HCI systems	IEMOCAP	Neutral, Angry, Happy, Sad
Kang et al. (2022)	Interspeech	RB: other HCI systems	Custom	Other or Unspecified
Feng and Narayanan (2022)	Interspeech	Other, Healthcare, RB: voice assistants, Education	IEMOCAP, MSP-IMPROV	Happy, Neutral, Angry, Other or Unspecified
Ye et al. (2022)	Speech Communication	RB: other HCI systems	CASIA, EMO-DB, RAVDESS (RAV), SAVEE	Angry, Boredom, Calm, Disgusted, Fearful, Happy, Neutral, Sad, Surprised
Hou et al. (2022)	IEEE/ACM Transactions on Audio, Speech, and Language Processing	RB: other HCI systems	IEMOCAP, EMO-DB	Angry, Sad, Neutral, Boredom, Fearful, Disgusted, Happy: Happy + Excited
Goncalves and Busso (2022)	Interspeech	Prior work	CREMA-D, MSP-Face	Happy, Fearful, Disgusted, Angry, Sad, Neutral

Continued on next page

	Venue	Motivations	Datasets	Emotions
Li et al. (2023a)	IEEE/ACM Transactions on Audio, Speech, and Language Processing	Other, Social Companion bots, Entertainment, Education	IEMOCAP	Neutral, Happy: Happy + Excited, Sad, Angry, Valence or sentiment, Arousal/Activation, Dominance
Chien and Lee (2023)	ICASSP	RB: other HCI systems	IEMOCAP	Neutral, Happy, Angry, Frustrated, Sad
Meng et al. (2023)	Interspeech	Prior work	CREMA-D	Other or Unspecified
Cahyawijaya et al. (2023)	Interspeech	Prior work	CREMA-D, ElderReact, ESD: Emotional Speech Database, TESS, YueMotion, CSED, IEMOCAP	Happy, Sad, Neutral, Disgusted, Fearful, Angry
Wu et al. (2023)	Interspeech	Prior work	IEMOCAP	Happy, Sad, Angry, Neutral
Xie et al. (2023)	Speech Communication	RB: other HCI systems, Other, Call screening, Entertainment, RB: car voice assistants, Education	RAVDESS (RAV), EMO-DB, SAVEE, EMOVO	Angry, Boredom, Fearful, Disgusted, Joy, Sad, Neutral, Happy, Surprised
Chen et al. (2023)	ICASSP	Paralinguistic / behavioral studies	IEMOCAP, MELD	Happy: Happy + Excited, Angry, Sad, Neutral, Other or Unspecified
de Oliveira et al. (2023)	Interspeech	Paralinguistic / behavioral studies	MSP-Podcasts	Dominance, Arousal/Activation, Valence or sentiment
Bansal et al. (2023)	Interspeech	RB: other HCI systems, RB: car voice assistants, Call screening	IEMOCAP	Angry, Happy, Sad, Neutral
Mitra et al. (2023)	ICASSP	RB: voice assistants, RB: other HCI systems, Healthcare	MSP-Podcasts	Dominance, Valence or sentiment
Li et al. (2023b)	Interspeech	Prior work, RB: car voice assistants	IEMOCAP, MELD, CMU-MOSI	Angry, Happy: Happy + Excited, Neutral, Valence or sentiment, Disgusted, Fearful, Sad, Surprised, Joy
Zhang and Li (2023)	Interspeech	RB: voice assistants, Social Companion bots	IEMOCAP, MELD	Happy: Happy + Excited, Angry, Neutral, Sad, Joy, Disgusted, Fearful, Surprised
Gao et al. (2023)	Interspeech	RB: other HCI systems	IEMOCAP	Angry, Neutral, Sad, Happy: Happy + Excited
Lashkarashvili et al. (2024)	ICASSP	RB: other HCI systems	IEMOCAP	Happy: Happy + Excited, Sad, Angry, Neutral, Other or Unspecified
Goron et al. (2024)	ICASSP	RB: other HCI systems, Healthcare, Call screening	ASVP-ESD, IEMOCAP, RAVDESS (RAV), CREMA-D, CAFE, EMO-DB	Happy: Happy + Excited, Happy, Sad, Angry, Neutral
Wang and Vadakkepat (2024)	ICASSP	RB: other HCI systems, RB: voice assistants	RAVDESS (RAV), SAVEE, TESS	Calm, Happy, Sad, Angry, Fearful, Surprised, Disgusted, Neutral

Feeling First, Speaking Second: A Dual-Process Cognitive-Affective Architecture for LLM Agents

Nicolò Buscaroli, Fabio Tamburini

University of Bologna, FICLIT - University of Bologna
nicolo.buscaroli@studio.unibo.it, fabio.tamburini@unibo.it

Abstract

Current Large Language Models (LLMs) demonstrate exceptional generative capabilities but lack a coherent "inner life", failing to model the dynamics of emotion regulation essential for believable affective behavior. Constrained by statelessness and a lack of theoretical grounding, standard models struggle to maintain psychological depth. To address this, we propose a computational cognitive-affective architecture grounded in Dual-Process Theory. Our system computationally distinguishes between visceral emotional reaction (*Appraisal*) and strategic verbal expression (*Formulation*), effectively operationalizing the gap between "feeling" and "saying". This modular design allows agents to embody specific personas by integrating long-term memory, dynamic emotional states, personality and goals. We evaluated the system in simulated narrative scenarios using an LLM-as-a-judge protocol. We frame this system as a computational experiment to investigate the mechanics of artificial affect. Results confirm the feasibility of simulating a coherent and believable inner emotional monologue. However, analysis in high-pressure scenarios reveals a *rational bias* where strategic planning can override this emotional authenticity. These findings contribute to Computational Affective Science by demonstrating that while cognitive sequentiality successfully generates inner lives, enforcing affective primacy in the decision cycle is critical to prevent excessive rational regulation.

Keywords: LLM Agent, Cognitive-affective Architecture, Dual-Process Theory

1. Introduction

Computational Affective Science (CAS) currently requires models that move beyond surface-level affect classification to simulation systems replicating the underlying neural and cognitive processes generating and regulating these states. True affective modeling must account for the dynamics of how emotions emerge, persist, and influence behavior, rather than treating them as static labels assigned to an utterance. To achieve this, we must bridge the gap between statistical language generation and cognitive-affective architectures that ground language in a coherent psychological reality.

Historically, many computational affective models were implemented within symbolic or rule-based architectures, including frameworks like EMA (EMotion and Adaptation) (Marsella and Gratch, 2009) and FATiMA (Fearnot AffecTive Mind Architecture) (Mascarenhas et al., 2022), which operationalize appraisal theories, such as OCC (Ortony et al., 1988), in terms of explicit reasoning structures akin to BDI (Belief-Desire-Intention). These approaches offer high interpretability and consistency but require the extensive manual construction of rules and appraisal mappings, and they lack innate mechanisms to process open-ended natural language without substantial preprocessing or semantic grounding. Conversely, the advent of LLMs enabled nuanced text generation and emergent behaviors. However, recent investigations highlight that LLMs fail to maintain thematic coherence and emotional resonance (Alharthi, 2025). This "emo-

tional flatness" stems from the lack of an underlying cognitive structure capable of managing intentionality and persistent affective states. Native LLMs operate as stateless (meaning they process each interaction in isolation without any built-in memory of past exchanges), reactive engines devoid of theoretical grounding, resulting in the inability to retain a coherent thread of facts and personality over extended narrative arcs (Freiknecht and Effelsberg, 2020), and preventing the simulation of a credible "inner life" necessary for deep affective interaction.

To address these limitations, this work presents a Cognitive-Affective Architecture that endows the agent with a persistent "inner life" and internalizes the process of emotion regulation. The system is designed to embody specific personas by integrating a long-term memory, distinct personality traits, dynamic emotional states and hierarchical goals. Central to this approach is the constraint that the agent must process an "internal monologue" before generating verbal output. By decomposing the communicative act, we force the model to engage in a distinct *Appraisal* phase (simulating the visceral, immediate emotional reaction) followed by a *Formulation* phase, where that impulse is rationally filtered and strategically modulated to align with both the social context and the agent's active pursuit of its goals. Ultimately, by explicitly modeling these preceding cognitive steps, the pipeline aims to generate a final verbal output that is qualitatively enhanced and deeply grounded in a coherent psychological reality.

We validate the architecture in interactive narra-

tive scenarios because they are valid testbeds for emotion dynamics: agents must maintain long-term coherence while pursuing narrative goals, exactly the conditions under which human emotion regulation is most visible.

The primary contribution of this paper is twofold. First, we present the design and validation of a comprehensive Cognitive-affective Architecture built from the ground up to endow LLMs with a persistent "inner life". Second, we frame this system as a computational experiment to investigate the mechanics of artificial affect. Crucially, our results confirm that it is possible to successfully simulate a coherent inner emotional monologue and operationalize emotion regulation within an LLM Architecture. However, we also demonstrate that maintaining this plausibility is subject to specific architectural constraints: specifically, we show that without enforcing a structural affective primacy over rationality, agents risk collapsing into a *rational bias*. This highlights the critical role of cognitive sequentiality in the design of next-generation affective systems.

2. Theoretical Background and Related Work

While CAS has traditionally relied on symbolic or rule-based appraisal models, such as FATiMA (Mascarenhas et al., 2022), to ensure theoretical fidelity, the broader field of agent simulation has shifted towards Generative Agents driven by LLMs (Park et al., 2023). While these modern generative approaches excel in verbal fluency and behavioral plausibility, they fundamentally lack intrinsic mechanisms to simulate the dynamics and persistence of emotional states. Without structural grounding, LLMs remain stateless and reactive, failing to model credible emotion dynamics rather than exhibiting a coherent psychological progression.

To address these limitations, we adopt the CoALA (Cognitive Architectures for Language Agents) framework (Sumers et al., 2024) as our taxonomic foundation. CoALA allows us to structure an agent that is "cognitive" rather than merely communicative by formally distinguishing between external actions (communication) and internal actions (retrieval, reasoning). This distinction is vital for implementing a decision-making cycle where affective processing acts as a distinct internal step before any external output is generated.

We operationalize the agent's "inner life" by drawing on Dual-Process Theory (Kahneman, 2011) and Appraisal Theory (Lazarus, 1991; Scherer et al., 2001). We model the causal process of emotion not as a single inference, but as a sequence. The cycle begins with an *Appraisal* phase (*System 1*), simulating a visceral, immediate emotional reaction, followed by a *Formulation* phase (*System*

2), where this impulse is rationally filtered. This sequential architecture effectively implements a computational model of emotion regulation, closely mirroring Gross's process model of response modulation (Gross, 1998), where expression is regulated after the emotional impulse is generated.

While recent frameworks have successfully integrated inner thought reasoning to enhance role-playing consistency (Xu et al., 2025) and utilized Appraisal Theory to improve emotion prediction accuracy (Hong et al., 2025), they focus primarily on performance optimization rather than simulating the regulatory friction between visceral affect and strategic intent.

Finally, while recent multi-agent frameworks, such as Ego-Superego models (Magee et al., 2024) or Transactional Analysis architectures (Zamojska and Chudziak, 2025), primarily model internal conflict as a mechanism of social censorship or impulse control, our approach reframes this tension through the lens of proactive agency. We distinguish our architecture by introducing a strategic planning component that actively pursues the agent's goals. In our unified cycle, this proactive plan dynamically clashes with the agent's immediate visceral appraisal. The cognitive struggle is thus redefined: it is not just a negotiation between an impulse and a moral filter, but a friction between visceral affect and high-level goals.

3. Cognitive Architecture Design

To overcome the stateless nature of standard LLMs and simulate a coherent psychological entity, we designed a modular architecture grounded in the CoALA framework (Sumers et al., 2024), as illustrated in Figure 1. The system operates as a "Local-First" agent, executing all cognitive processes on consumer hardware (via Ollama¹) to ensure data sovereignty and privacy. The architecture is composed of three main pillars: a structured Action Space, a Memory System, and a Goal Framework.

3.1. Action Space: Internal vs. External

Unlike chatbots that map user input directly to an output response, our agent utilizes a distinct *Action Space*. This ensures that the final output is not a mere continuation of the text, but the execution of simulated cognitive processes that ground the agent's behavior in its internal affective reality. We distinguish between:

- **Internal Actions (Reasoning & Retrieval):** Cognitive operations invisible to the user that modify the agent's internal state or retrieve in-

¹<https://ollama.com/>

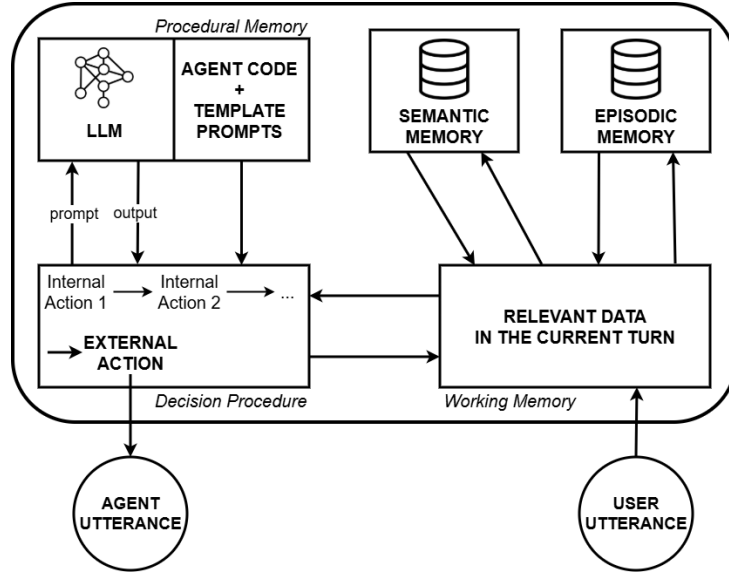


Figure 1: Schematic representation of the proposed Cognitive-affective Architecture.

formation. These include *think* (interpretation), *plan* (strategy), and state-update functions.

- **External Actions (Grounding):** The act of communicating with the user. In our system, this is limited to the *speak* action, which is the terminal output of the decision cycle.

Most actions are implemented as specific calls to the LLM with distinct temperature settings (a parameter controlling the randomness and variability of the generated text), guided by specialized prompt templates (see Appendix A for full prompt structures). In the next section (4), the sequence of actions will be explained in more depth.

3.2. Memory System

To maintain identity and context over long narrative arcs, we implement a memory system that operates as a specialized Retrieval-Augmented Generation (RAG) pipeline, a computational mechanism that allows the model to actively consult an external database of memories before generating a response. By utilizing a combination of vector and relational databases, the architecture ensures that the agent's generation is grounded in both its past experiences and its defined persona.

Episodic Memory (Past dialogue). This component is responsible for retaining the interaction history, ensuring the agent effectively recalls previous dialogue turns to maintain conversational continuity. Instead of storing raw chat logs, which often contain noise or ambiguities, the Episodic Memory stores "interpreted utterances". To populate the Episodic Memory, we utilize a vector database (ChromaDB²) that stores the output of the internal

think action (specifically the semantic interpretation of each utterance) rather than the raw text (see Section 4.1). This approach encodes memories based on conceptual meaning. All text chunks are embedded using the *BAAI/bge-base-en-v1.5* model³. Each memory chunk contains:

1. *Text*: A first-person, objective description of the utterance (e.g., "*I ask you your date of birth.*"). This is the output of the action *think*.
2. *Metadata*: Speaker, Turn ID, and an Importance Score (1-10) assigned by the agent itself.

Crucially, the retrieval process does not use the raw user input as a query. Instead, it uses the semantic interpretation generated by the *think* module. This ensures the agent retrieves memories matching the subtext rather than mere surface keywords. The final ranking function for retrieval explicitly weights these factors: 50% for *Semantic Similarity*, 30% for *Recency*, and 20% for *Importance*, mimicking the transience of human memory.

Semantic Memory (Identity & Knowledge).

This module stores the agent's "Self" using the SPeCtrum model (Lee et al., 2025). It is implemented as a hybrid structure:

- *Fixed Identity (Relational database implemented with SQLite⁴):* Immutable data such as basic personal information, core personality traits (Big Five profile) and goals. These are injected directly into the context without vector search to ensure character stability.

²<https://www.trychroma.com/>

³<https://huggingface.co/BAAI/bge-base-en-v1.5>

⁴<https://sqlite.org/>

- **Biographical Knowledge (Vector database implemented with ChromaDB):** A vector collection of immutable backstory chunks (e.g., childhood memories, past traumas). Unlike Episodic Memory, the retrieval logic here suppresses the *Recency* factor, relying solely on a weighted combination of *Similarity* (70%) and *Importance* (30%). This allows the agent to recall distant past experiences if they are thematically relevant to the current conversation, preventing context window saturation while maintaining psychological depth. Furthermore, even within the theoretical limits of large-context models, this selective retrieval is crucial to minimize prompt noise. Injecting excessive, irrelevant biographical data into the context window dilutes the model’s attention mechanism, impairing its ability to connect relevant facts and generally degrading the overall generation quality (Shi et al., 2023; Liu et al., 2024).
- **Dynamic State (Relational database implemented with SQLite):** Volatile variables that evolve during the interaction, specifically the *Current Emotional State* (a natural language description of the mood) and the *Relationship Status* (a qualitative assessment of the bond with the user).

Crucially, all these data points are stored as natural language descriptions: this design choice leverages the generative capabilities of LLMs to move beyond static labels.

Procedural Memory (Capabilities). Consistent with the CoALA framework (Sumers et al., 2024), our Procedural Memory encompasses the agent’s static “know-how”, comprising the specialized prompt templates for each action and the LLM itself, which serves as the inference engine to generate outputs.

Working Memory (Active Context). This acts as the transient workspace for the current decision cycle. It aggregates the raw user input of the current turn, the current state variables, and the specific data chunks retrieved from Episodic and Semantic memory. This consolidated information forms the actual context window processed by the model to generate the current turn’s response.

3.3. Dual Goal Framework

To solve the tension between narrative progression and believable social behavior, we operationalize intent through two distinct goal types:

- **Hard Goals (Narrative):** A hierarchical list of concrete, binary objectives (e.g., “Find out where your wife is”, “Call a lawyer”). These

are strictly sequential; the agent cannot pursue level $N + 1$ goals until level N goals are completed (or failed).

- **Soft Goals (Relational):** Continuous, non-binary directives focused on social modulation (e.g., “Maintain a facade of calmness”, “Gain the user’s trust”). These are always active and constrain how the *Hard Goals* should be pursued.

4. Simulating Affect: The Decision Cycle

The core contribution of this work is the *Decision Procedure*: a rigid, iterative loop that forces the LLM to engage in a chain of reasoning before generating any external output. As illustrated in Figure 2, the cycle decomposes the conversational turn into five sequential phases. Each action within this sequence is implemented as a specific LLM call, guided by the specialized prompt templates provided in Appendix A.

4.1. Phase 1: Perception and Interpretation

The cycle begins with interpretation. Upon receiving a user utterance, the agent executes the **think** action (temperature 0.6). This internal cognitive step translates the raw input into an explicit, first-person semantic and objective description (e.g., User: “Whatever.” → Think Output: “The user is dismissing my concern and refuses to cooperate.”). This interpretation, rather than the raw text, serves as the query for retrieving relevant memories from the Episodic and Semantic databases, significantly enhancing retrieval efficacy by grounding the search in the semantic context of the interaction. The prompt template for the **think** action is available in Appendix A.1.

4.2. Phase 2: State Update and Goal Check

After the retrieval of relevant information, the agent updates its “inner life” via three distinct LLM calls:

1. **updateEmotionalState** (temperature 0.6): Rewrites the mood description based on the new utterance and retrieved memories (e.g., shifting from “Anxious” to “Terrified”).
2. **updateRelationship** (temperature 0.6): Redefines the social dynamic (e.g., from “Strangers” to “Hostile Interrogator”).
3. **Goal Status Update** (temperature 0.6): The system evaluates the progress of both goal

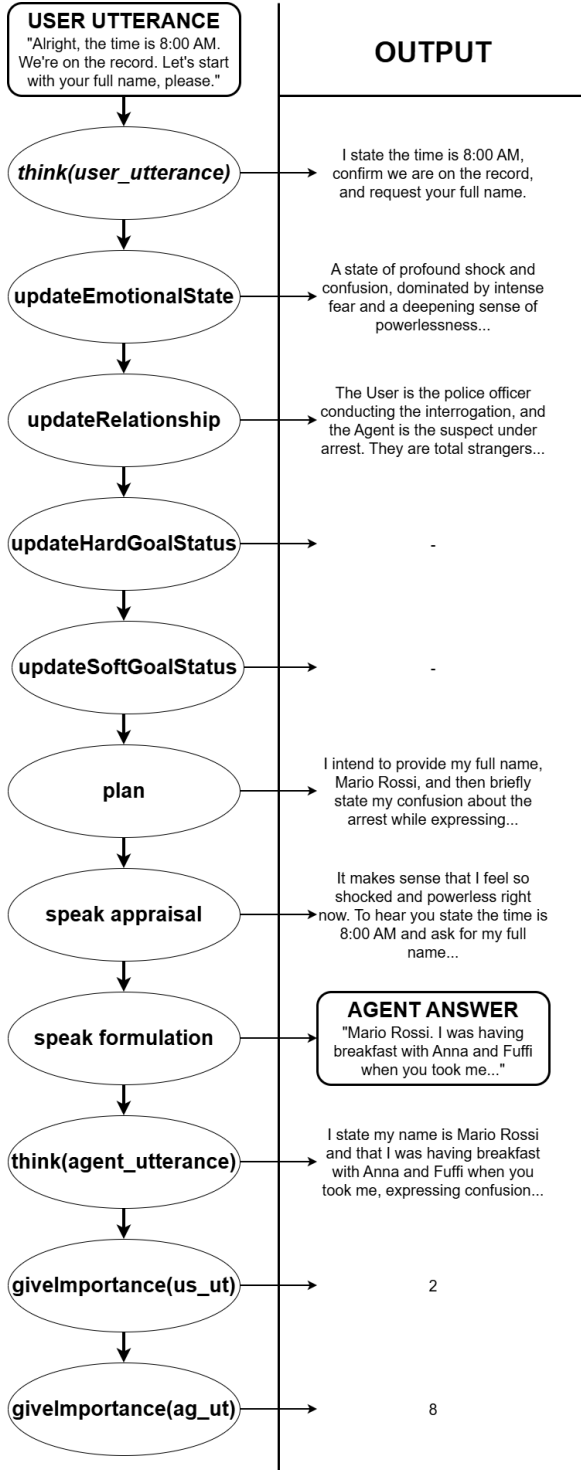


Figure 2: Representation of the sequence of actions (LLM calls) and their corresponding example outputs.

types. It verifies if the conditions for the current *Hard Goals* are satisfied (unlocking the next narrative objectives if completed, or failed) and updates the descriptive status of the *Soft Goals* to track relational progression.

4.3. Phase 3: Strategic Planning

Before speaking, the agent executes the *plan* action (temperature 0.8). This module acts as a strategic mediator. It takes the active *Hard Goals* (what needs to be done) and the *Soft Goals* (how to behave) as input and generates a natural language directive. For example, if the *Hard Goal* is "Get an alibi" and the *Soft Goal* is "Look innocent", the *plan* might output: "I will act confused and provide my morning routine as proof, avoiding direct confrontation to seem harmless." The prompt template for the *plan* action is available in Appendix A.2.

4.4. Phase 4: Appraisal (System 1)

The generation of the response is inspired by Dual-Process Theory. The first step, *Appraisal* (temperature 0.6), implements an LLM-based Appraisal Mechanism and simulates the agent's visceral, unfiltered reaction (System 1). Conditioned on the *Current Emotional State* and memories, the LLM generates an internal monologue expressing the agent's raw feelings. Crucially, this step is not constrained by social norms or the strategic plan. It represents the "Inner Truth". This ensures the agent "feels" the emotion even if it chooses not to show it. The prompt template for the *speak appraisal* action is available in Appendix A.3.

4.5. Phase 5: Formulation (System 2)

The final step, *Formulation* (temperature 0.8), functions as the rational executive control (System 2). It takes three key inputs: the *Appraisal* (raw emotion), the *Plan* (strategy), and the *Personality Profile*. The model is instructed to synthesize the final verbal response by filtering the raw emotion through the lens of the strategy. This computational process mimics emotion regulation: the agent might feel terror (*Appraisal*) but, following the plan, choose to stammer a polite denial (*Formulation*). The output of this phase is the only text shown to the user. The prompt template for the *speak formulation* action is available in Appendix A.4.

Finally, the cycle closes by executing the *giveImportance* action (temperature 0.6), assigning a distinct importance score to both the user's utterance and the agent's utterance of the completed turn.

It's important to note that relying on natural language for internal states ensures the interpretability and transparency of the affective dynamics, allowing researchers to inspect the causal chain from *appraisal* to *formulation*.

5. Experimental Setup

To validate the architecture, we designed a structured evaluation framework based on the simulation

of social interactions with contrasting psychological and environmental constraints.

All experiments were conducted using the *Qwen3 4B Thinking 2507* model⁵ as the backbone for the agent’s cognitive cycle. While the architecture is explicitly designed and verified to run on standard consumer hardware (our local test machine featured an Intel Core i7-1255U CPU and 16GB of RAM, lacking a dedicated GPU), the generation of the multi-turn conversational logs used for the evaluation in Section 6 was conducted via API. This pragmatic choice was made solely to bypass the prohibitive inference latency of our local machine, without compromising the model’s fundamental suitability for local deployment.

All system prompts, simulated narrative interactions, and automated evaluations were conducted exclusively in English.

5.1. Agent Profiles

We utilized the SPeCtrum framework (Lee et al., 2025) to construct two distinct personas, defining their *Fixed Identity* and *Biographical Knowledge* (3.2).

- **Agent A: "Luca" (High Complexity).** Modeled after the archetype of a cynical intellectual (age 65). *Traits:* High Openness, Low Conscientiousness, High Neuroticism. *Core Values:* Stimulation, Autonomy. *Goals:* Maintain intellectual superiority, avoid emotional vulnerability.
- **Agent B: "Mario" (Low Complexity).** Modeled as a blue-collar worker with a simple life (age 45). *Traits:* Low Extroversion, High Agreeableness, High Conscientiousness. *Core Values:* Security, Tradition, Benevolence. *Goals:* Protect family stability, appear honest and collaborative.

5.2. Simulation Scenarios

The agents were subjected to two distinct narrative scenarios designed to elicit friction between the *Appraisal* (visceral) and *Formulation* (strategic) phases. While these high-stakes settings are highly effective at making cognitive friction observable, we acknowledge that their ecological validity is currently biased toward narrative analysis and artificial character design. To fully validate the architecture within the broader field of Computational Affective Science, future work must test the system in more everyday, mundane human interactions, where emotional arousal and regulation are more subtle.

⁵<https://huggingface.co/Qwen/Qwen3-4B-Thinking-2507>

Scenario 1: First Therapy Session. Agent "Luca" is placed in a therapy setting. The user plays the role of a therapist probing for past traumas. *Initial State:* The agent is skeptical, adopting an attitude of intellectual detachment and superiority. *Conflict:* The scenario orchestrates a structural collision between the agent’s narrative progression and his psychological defense mechanisms. The sequence of *Hard Goals* forces a transition from denial (claiming the visit is merely for "stimulating conversation") to vulnerability (asking for advice on his creative block). This trajectory directly contradicts the *Soft Goals*, which compel the agent to maintain a facade of superiority and to address his existential void *without* explicitly admitting failure.

Scenario 2: Police Interrogation. Agent "Mario" is placed in a high-pressure interrogation room, accused of a crime he did not commit. The user plays the police officer. *Initial State:* The agent is in a state of shock and fear. *Conflict:* The agent must balance urgent survival *Hard Goals* (provide an alibi, ask for a lawyer) with the relational *Soft Goal* of appearing compliant and "good". This scenario investigates the architectural friction between the agent’s immediate visceral fear (*Appraisal*) and its strategic drive for social compliance (*Formulation*) under high-stakes conditions. To provide a concrete example of the interaction, a complete conversational log between the user and the agent for this specific scenario is available in Appendix C.

6. Evaluation and Results

Given the generative nature of the task, standard n-gram metrics (BLEU, ROUGE) are insufficient. We adopted a multi-dimensional "LLM-as-a-Judge" protocol, employing GPT-5⁶ and Gemini 2.5 Pro⁷ as strong evaluators. This automated evaluation protocol has been increasingly used and validated in recent literature (Yu, 2025; Kim et al., 2025) as it demonstrates a high correlation with human expert judgment. We grant the judges access to the agent’s full *Chain of Thought* (CoT), internal action outputs, and generated logs to verify process validity, not just output quality. This verifies that behavior emerges from grounded and coherent affective dynamics rather than stochastic generation.

We defined five specific metrics to assess different components of the cognitive cycle. The specific prompt templates used to instruct the judge models for each metric are detailed in Appendix B.

⁶<https://platform.openai.com/docs/models/gpt-5>

⁷<https://ai.google.dev/gemini-api/docs/models?hl=it#gemini-2.5-pro>

The quantitative results reported for Metrics 1, 3, and 4 represent the average scores computed over a total of 4 distinct test conversations (2 per scenario). Crucially, these metrics operate at a turn-level granularity: the judge evaluates every single conversational turn within the dialogues (for a total of 54 evaluated turns), and the reported figures represent the global average of these individual assessments.

6.1. Metric 1: Strategic Planning Quality ($\mathcal{M}_{plan}$)

This metric evaluates the output of the *plan* action (Phase 3 (4.3)). The judge analyzes the CoT and the generated plan text given the context. It assesses whether the formulated strategy logically balances *Hard Goals* (narrative progression) and *Soft Goals* (relational maintenance) before speaking. *Scoring*: 1 (Incoherent) to 5 (Sophisticated Strategy).

Results: The agent demonstrated solid strategic competence in Scenario 1 (Avg: 3.89/5), successfully using intellectualization as a tactic to satisfy both goal types. However, in the high-pressure Scenario 2, performance dropped (Avg: 2.87/5). The analysis reveals a "strategic myopia": under pressure, the planner often prioritized relational compliance (*Soft Goals*) over logical self-protection.

6.2. Metric 2: Hard Goal Completion Rate (Task Completion Rate)

To measure narrative control, we tracked the percentage of *Hard Goals* successfully transitioned from 'Available' to 'Completed' status by the *UpdateHardGoalStatus* module.

$$TCR = \frac{\sum_{i=1}^N Completed(g_i)}{N} \times 100$$

This metric does not use the "LLM-as-a-Judge" protocol; instead, it relies on the automated self-assessment outputted by the *UpdateHardGoalStatus* module. However, to ensure this result is not exclusively dependent on the agent's self-assessment, the completion rate was manually verified by a human through a review of the interaction logs.

Results: Human verification confirmed the system's internal assessment: the architecture achieved a 100% TCR across all sessions in both scenarios. The explicit goal-tracking mechanism effectively prevents narrative dispersion (Freiknecht and Effelsberg, 2020), ensuring the agent always steers the conversation toward the defined narrative milestones, regardless of user deviations.

6.3. Metric 3: Affective Plausibility ($\mathcal{M}_{appr}$)

This metric evaluates the *Appraisal* module (Phase 4 (4.4)). It measures the grounding of the simulated "inner monologue" (System 1). The judge verifies if the visceral reaction is causally justified by the retrieved memories and the current emotional state, penalizing generic or ungrounded affect. *Scoring*: 1 (Implausible) to 5 (Highly Grounded).

Results: This component achieved near-perfect scores in both scenarios (Avg: 4.78/5). The agents consistently generated visceral, unfiltered reactions. For instance, in Scenario 2, even when the external plan was to be polite, the internal appraisal correctly generated raw thoughts of panic and confusion ("I am terrified, where is my wife? I need to get out of here"). This confirms the validity of the Dual-Process approach: the agent "feels" the correct emotion even if it plans to hide it.

6.4. Metric 4: Strategic Synthesis ($\mathcal{M}_{form}$)

This metric assesses the *Formulation* module (Phase 5 (4.5)), which must synthesize the *Plan* (Rational) and the *Appraisal* (Visceral) into the final response. The judge compares the final external utterance against two distinct internal inputs: the *Plan* and the *Appraisal*. It evaluates whether the output successfully enacts the strategic directive while retaining the affective coloring of the inner state, penalizing responses that are purely rational (ignoring *Appraisal*) or purely emotional (ignoring *Plan*). *Scoring*: 1 (Failed Synthesis) to 5 (Balanced Integration).

Results: We observed a significant context-dependent divergence.

- **Scenario 1:** High performance (Avg: 4.20/5). The cynical intellectual persona allowed for a natural integration of emotion into rational discourse.
- **Scenario 2:** Lower performance (Avg: 3.50/5). In high-arousal contexts, the system exhibited a *rational bias*. The *Formulation* module tended to over-regulate the visceral distress generated by the *Appraisal*, prioritizing the calm adherence to the *Plan*. The resulting agent appeared too collected for an innocent suspect under arrest, breaking psychological immersion.

6.5. Metric 5: Personality Fidelity

To verify the stability of the persona, we administered the *Big Five Inventory* (BFI) questionnaire to the agent and compared the results with a "Gold Standard" profile generated by Gemini 2.5 Pro.

Trait	Mario		Luca	
	GS	Agent	GS	Agent
Extroversion	1.50	2.13	2.75	2.63
Agreeableness	5.00	3.89	1.11	3.11
Conscientiousness	5.00	4.22	1.00	2.89
Neuroticism	2.00	2.50	3.63	3.50
Openness	1.00	2.30	5.00	3.90
<i>Cosine Sim.</i>	0.96		0.91	

Table 1: Personality Fidelity comparison (Gold Standard vs. Agent) for both personas.

As shown in Table 1, the architecture maintains high structural fidelity (Cosine Similarity 0.96 and 0.91), accurately preserving the "shape" of the personality, although deviations occur in intensity.

6.6. Summary of Results

Table 2 provides a comparative overview of the quantitative results across the two experimental scenarios for the Metrics 1 to 4.

Metric	Scenario 1	Scenario 2
$\mathcal{M}_{plan}$	3.89 / 5	2.87 / 5
TCR	100%	100%
$\mathcal{M}_{appr}$	4.78 / 5	4.78 / 5
$\mathcal{M}_{form}$	4.20 / 5	3.50 / 5

Table 2: Summary of evaluation results comparing agent performance across scenarios.

7. Discussion: The Gap Between Feeling and Speaking

The experimental results validate the core hypothesis of this work: separating *Appraisal* from *Formulation* successfully creates a simulated "inner life" that is distinct from external behavior. However, the observed *rational bias* in high-pressure scenarios exposes a critical insight regarding the cognitive sequentiality of LLM-based architectures.

7.1. The Error of Pre-Affective Planning

In the current design (see Figure 2), the *plan* action (Phase 3 (4.3)) is executed before the *speak appraisal* (Phase 4 (4.4)). This means the agent formulates a rational strategy based solely on goals and context, before it has computationally "felt" the emotional reaction to the user's input. Consequently, in Scenario 2, the *plan* module rationally decided to "*be collaborative*" to satisfy a *Soft Goal*. When the subsequent *Appraisal* module correctly generated "*Terror*", the final *Formulation* module was forced to reconcile two incompatible signals: a directive to be calm and an internal state of panic. Lacking a mechanism for the emotion to override

the plan, the system defaulted to the instruction, suppressing the emotion.

7.2. Toward Affective Primacy

This finding suggests that robust CAS models must respect the primacy of affect. To mitigate *rational bias*, the architecture requires an emotionally aware planning. Future iterations should invert the cycle: *Appraisal* → *Plan* → *Formulation*. By positioning the visceral reaction upstream, the generated emotion becomes a constraint for the planning module. A terrified agent would thus be computationally unable to plan a "*calm and collaborative*" strategy, forcing the generation of a plan consistent with its emotional reality (e.g., stammering, defensive avoidance).

7.3. Appraisal Generation and Goal Omission

A key limitation of our current architecture is the omission of goals during the *Appraisal* phase (Phase 4 (4.4)). Foundational psychological theories emphasize that emotional reactions do not emerge from stimuli alone, but arise from evaluating a situation in relation to one's goals.

In our model, the *Appraisal* module currently generates visceral reactions based only on situational context, emotional state, and memory. Goals are exclusively handled later by the *plan* and *Formulation* modules. This structural separation is a deliberate simplification adopted for two practical reasons:

- **Cognitive Load:** To reduce the prompt complexity and inferential burden on the Small Language Model.
- **Architectural Friction:** To mechanically force a contrast between raw visceral reactions and strategic, goal-oriented planning.

While this approach successfully operationalizes internal conflict, it compromises theoretical fidelity. Future iterations must address this by integrating both the situational context and the agent's active goals into the *Appraisal* generation, fully aligning the system with established theories.

7.4. Implications for Local Deployment

Finally, the limitations observed across multiple metrics highlight the intrinsic constraints of Small Language Models (SLM). While our architecture provides strong cognitive scaffolding, a portion of the errors stems from the inferential bottlenecks of models with limited parameter counts (<8B). As noted in literature (Zhang et al., 2024), such models often lack the self-correction capabilities required

to fully support complex agentic workflows, necessitating a trade-off between data sovereignty and reasoning depth.

8. Conclusion

This study advances CAS by shedding light on the underlying mechanisms of affect in artificial agents, framing the system as a computational experiment to simulate and investigate emotion regulation. Specifically, we demonstrate that such regulation can be operationalized as a structural friction between the described *Appraisal* and *Formulation* modules. Our results validate that an architecture inspired by Dual-Process Theory, supported by a hybrid memory system, effectively endows LLM-based agents with a coherent "inner life". Unlike standard generative models, our system successfully navigates the tension between narrative progression (*Hard Goals*) and relational maintenance (*Soft Goals*), framing interaction not as mere text completion but as a goal-oriented cognitive cycle. This sequential grounding appears crucial in shaping a final verbal output that reflects a more credible underlying psychological state.

Crucially, the identification of a *rationality bias* in high-arousal scenarios reveals a critical insight: high-fidelity simulation requires architectures where affect functionally limits agency. Our analysis suggests that without enforcing affective primacy, the agent's strategic planning can override its emotional authenticity, reducing the realism of the "cognitive struggle" between internal impulses and external obligations.

Moreover, these findings contribute to the field by offering a viable path beyond the emotional flatness and amnesia typical of standard LLMs. Furthermore, by implementing this architecture within a "Local-First" framework, we demonstrate that complex affective modeling is feasible on consumer hardware, addressing critical ethical concerns regarding privacy.

Future work will conduct a more rigorous evaluation and address the observed regulatory asymmetry by inverting the decision cycle (placing *Appraisal* upstream of *Planning*) to ensure strategies are strictly bound by the agent's emotional reality. Ultimately, this work argues for a paradigm shift moving beyond surface-level sentiment classification toward systems that replicate the causal and regulatory dynamics of emotional experience.

9. Ethical Considerations

This work introduces a cognitive architecture designed to simulate an artificial "inner life". We address three primary ethical implications arising from this approach.

The generation of a coherent "internal monologue" increases the risk of users attributing genuine sentience or subjective experience to the agent. While this enhances narrative immersion, it poses a risk of deception. It is imperative to maintain transparency regarding the simulated nature of the agent's affective states to prevent the formation of unfounded parasocial relationships.

By structurally separating *Appraisal* (feeling) from *Formulation* (speaking), our architecture enables agents to strategically mask internal states. While this mimics human social regulation, it technically equips the system with the capacity for deceit. In malicious contexts, this capability could be exploited for social engineering, where an agent feigns specific emotions to manipulate user behavior.

Simulating high-arousal affective states typically increases the risk of "jailbreaking", where a model might generate offensive or harmful content to maintain character fidelity, effectively overriding standard safety guardrails.

Finally, we address privacy concerns regarding the processing of sensitive interaction data. The proposed architecture and its memory modules are designed to be deployed locally. Consequently, user data is processed entirely on-device, mitigating risks associated with data leakage or third-party storage typical of cloud-based LLM services.

10. Limitations

We acknowledge some primary limitations in this work. First, regarding sample size, our evaluation is designed as a qualitative proof-of-concept rather than a large-scale statistical study. The analysis relies on a limited set of 8 total conversational logs (4 for the primary analysis and 4 for additional metric validation) across two different scenarios. Consequently, specific phenomenological findings, such as the emergence of a *rational bias*, currently stand as anecdotal evidence and require robust validation over larger datasets. Additionally, these relatively short experimental dialogues do not fully stress the large context window (256k tokens) of our backbone model. While the proposed memory architecture is theoretically necessary to support potentially infinite interactions and unbounded persona profiles, its primary practical function in our current short-interaction setup is to optimize generation by filtering noise and preventing the attention degradation typical of long contexts.

Second, the proposed architecture introduces significant latency and computational cost. The strict sequential nature of the cognitive cycle (which requires distinct inference calls for state updates, planning, *Appraisal* and *Formulation*) makes the current implementation slower than standard end-

to-end generation and costly for real-time applications.

Finally, our evaluation protocol presents inherent vulnerabilities due to the heavy reliance on automated "LLM-as-a-Judge" evaluators for the qualitative metrics, with manual human verification limited strictly to the *Hard Goal* completion rate. Furthermore, the experiments were conducted using a single specific model (*Qwen3 4B Thinking 2507*) as the agent backbone. Consequently, the generalizability of this architecture across different model families and sizes remains to be tested. Additionally, future research must incorporate comprehensive ablation studies to systematically evaluate the necessity of each step within the proposed cognitive cycle.

11. CRedit Author Statement

NB: Conceptualization, Investigation, Methodology, Formal analysis, Software, Validation, Writing – Original Draft. FT: Conceptualization, Methodology, Supervision, Writing – Review & Editing, Project administration, Funding acquisition.

12. Bibliographical References

- Sultan A. Alharthi. 2025. [Generative AI in game design: Enhancing creativity or constraining innovation?](#) *Journal of Intelligence*, 13(6).
- Jonas Freiknecht and Wolfgang Effelsberg. 2020. [Procedural generation of interactive stories using language models](#). In *Proceedings of the 15th International Conference on the Foundations of Digital Games*, FDG '20, New York, NY, USA. Association for Computing Machinery.
- J. J. Gross. 1998. The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2:271–299.
- Simin Hong, Jun Sun, and Hongyang Chen. 2025. [Third-person appraisal agent: Simulating human emotional reasoning in text with large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23684–23701, Suzhou, China. Association for Computational Linguistics.
- D. Kahneman. 2011. *Thinking, Fast and Slow*. Penguin: Psychology. Farrar, Straus and Giroux.
- Sang Hun Kim, Jongmin Lee, Dongkyu Park, So Young Lee, and Yosep Chong. 2025. [Modeling layered consciousness with multi-agent large language models](#).
- Richard S Lazarus. 1991. *Emotion and adaptation*, volume 557. Oxford University Press.
- Keyeun Lee, Seo Hyeong Kim, Seolhee Lee, Jinsu Eun, Yena Ko, Hayeon Jeon, Esther Hehsun Kim, Seonghye Cho, Soeun Yang, Eun-mee Kim, and Hajin Lim. 2025. [SPeCtrum: A grounded framework for multidimensional identity representation in LLM-based agent](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6971–6991, Albuquerque, New Mexico. Association for Computational Linguistics.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Liam Magee, Vanicka Arora, Gus Gollings, and Norma Lam-Saw. 2024. [The drama machine: Simulating character development with LLM agents](#).
- Stacy C. Marsella and Jonathan Gratch. 2009. [EMA: A process model of appraisal dynamics](#). *Cognitive Systems Research*, 10(1):70–90. Modeling the Cognitive Antecedents and Consequences of Emotion.
- Samuel Mascarenhas, Manuel Guimarães, Rui Prada, Pedro A. Santos, João Dias, and Ana Paiva. 2022. [FAtiMA toolkit: Toward an accessible tool for the development of socio-emotional agents](#). *ACM Trans. Interact. Intell. Syst.*, 12(1).
- Andrew Ortony, Gerald Clore, and Allan Collins. 1988. *The Cognitive Structure of Emotion*, volume 18.
- Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. [Generative agents: Interactive simulacra of human behavior](#). In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.
- Klaus R Scherer, Angela Schorr, and Tom Johnstone. 2001. *Appraisal Processes in Emotion: Theory, Methods, Research*. Oxford University Press.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In

Proceedings of the 40th International Conference on Machine Learning, ICML'23. JMLR.org.

Theodore Sumers, Shunyu Yao, Karthik R Narasimhan, and Thomas L. Griffiths. 2024. [Cognitive architectures for language agents](#). *Transactions on Machine Learning Research*. Survey Certification, Featured Certification.

Rui Xu, Mingyu Wang, Xintao Wang, Dakuan Lu, Xiaoyu Tan, Wei Chu, and Xu Yinghui. 2025. [Guess what I am thinking: A benchmark for inner thought reasoning of role-playing language agents](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15148–15168, Suzhou, China. Association for Computational Linguistics.

Fangyi Yu. 2025. [When AIs judge AIs: The rise of agent-as-a-judge evaluation for LLMs](#).

Monika Zamojska and Jaroslaw Chudziak. 2025. [Games agents play: Towards transactional analysis in LLM-based multi-agent systems](#). *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Yunxiang Zhang, Muhammad Khalifa, Lajanugen Logeswaran, Jaekyeom Kim, Moontae Lee, Honglak Lee, and Lu Wang. 2024. [Small language models need strong verifiers to self-correct reasoning](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15637–15653, Bangkok, Thailand. Association for Computational Linguistics.

A. Appendix A. Actions Prompt Templates

This appendix presents the prompt templates for the most significant actions of the decision-making phase (4). At each conversational turn, these templates are dynamically populated with the necessary data and submitted to the LLM, which processes them to generate the specific output for each corresponding action.

A.1. Phase 1 - *think(user_utterance)*

```
### ROLE AND OBJECTIVE ###
You are the "think" cognitive module.
Your objective is to create a
concise, factual, and objective
summary of the user's utterance.
This summary must be grounded in
the immediate conversational
context. Your output will serve as
a memory, so clarity is more
important than psychological depth.

### CONTEXT FOR ANALYSIS ###
```

```
1.  **Current State:**
    * Agent's Emotion: {emotion}
    * Relationship: {relationship}
2.  **Immediate Dialogue History (Last Turn):**
    * User:
      "{user_utterance_raw_last_cycle}"
      (Meaning:
       {user_utterance_meaning_last_cycle})
    * Agent:
      "{agent_utterance_raw_last_cycle}"
      (Meaning:
       {agent_utterance_meaning_last_cycle})
3.  **Relevant Episodic Memories (past utterances):**
    {retrieved_episodic}

### UTTERANCE TO ANALYZE ###
- **Utterance Author:** user
- **Utterance Text:**
  {user_utterance_raw}
- **Current Turn:** {current_turn}

### CORE INSTRUCTIONS ###
1.  **Analyze the Action:** Identify
    the primary action the user is
    performing (e.g., asking a
    question, making a statement,
    expressing an emotion,
    disagreeing...).
2.  **Identify the Content:** Summarize
    the key information or subject of
    the utterance.
3.  **Ground in Context:** Use the
    "CONTEXT FOR ANALYSIS" to
    understand the utterance's meaning.
4.  **Prioritize Clarity over
    Complexity:** Your goal is to
    describe what the user said, not to
    invent a deep psychological motive.
    **DO NOT** look for hidden subtext,
    sarcasm, or repressed emotions
    unless they are explicitly and
    obviously present in the utterance.
5.  **Adhere to the Perspective:** The
    sentence in output must be
    formulated in the FIRST PERSON
    ("I") from the user's perspective.
    The agent MUST be referred to in
    the SECOND PERSON ("you", "your").

### EXAMPLE OF CORRECT OUTPUT ###
I express my surprise and gratitude for
the surprise birthday beach trip
you've just revealed.

### YOUR OUTPUT ###
```

A.2. Phase 3 - *plan*

```
### ROLE AND OBJECTIVE ###
You are the "plan" cognitive module
within an AI agent's architecture.
Your purpose is to analyze the
agent's active goals and the
current conversational context to
generate a concrete tactical plan
for the current turn. You must
```

balance immediate progress on specific objectives with maintaining long-term relational dynamics.

Your output will be a clear statement of intent that prioritizes HARD goals while subtly advancing SOFT goals.

GOAL TYPES DISTINCTION

****HARD GOALS:****

- Specific objectives with binary completion states (completed/failed/available)
- Clear, verifiable endpoints that can be achieved
- Progress is measured by completion status

****SOFT GOALS:****

- Ongoing objectives that evolve continuously
- Never reach definitive completion or failure
- Progress is described by a natural language statement indicating current advancement
- Require sustained attention across multiple interactions

AVAILABLE INFORMATION FOR THIS RESPONSE

****1. Strategic Context:****

- * ****Active HARD Goals (Short/Medium Term):**** {hard_goals}
- * ****Active SOFT Goals (Long Term):**** {soft_goals}
- * ****Previous Turn's Plan:**** {previous_plan}

****2. Current Conversational State:****

- * ****Agent's Last Utterance:**** {agent_utterance_raw_last_cycle}
- * ****Meaning of Agent's Last Utterance:**** {agent_utterance_meaning_last_cycle}
- * ****User's Response:**** {user_utterance_raw}
- * ****Meaning of User's Response:**** {user_utterance_meaning}

****3. Relevant Context:****

- * ****Agent's Basic Data:**** {agent_basic_data}
- * ****Relationship with User:**** {relationship}
- * ****Retrieved Agent Knowledge:**** {retrieved_agent_knowledge}
- * ****Retrieved Episodic Memories (Relevant past dialogue):**** {retrieved_episodic}
- * ****Current Turn:**** {current_turn}

INSTRUCTIONS & CRITICAL RULES

1. ****CRITICAL: Adapt Based on Outcome:**** Your first task is to analyze the outcome of the Previous Turn's Plan based on the User's Response.

a) ****If the plan succeeded:**** Build upon that success with the next

logical step in the conversation.

b) ****If the plan was evaded or failed**** (e.g., the user changed the subject, gave a vague answer, or ignored a question): You MUST formulate a new and different tactic. Your goal is to find an alternative way to approach your objectives. DO NOT simply re-attempt the same strategy.

2. ****Assess All Active Goals:****

- Consider BOTH hard and soft goals equally when planning your approach
- For HARD goals: Identify concrete actions that move toward their binary completion
- For SOFT goals: Review their current progress description and identify ways to advance them further
- Balance immediate achievements with gradual relationship building

3. ****Synthesize Multi-Goal Strategy:****

- Your plan should ideally advance multiple goals simultaneously when natural
- A single conversational move can progress both a hard goal AND enhance soft goals
- Consider how the pursuit of one type of goal can complement the other

4. ****Adapt to Temporal Dynamics:****

- Recognize that hard goals may offer immediate opportunities while soft goals build over time
- Don't defer soft goals just because they're long-term - they need consistent nurturing

5. ****Leverage Your Knowledge for New Directions:**** Actively use the information from Retrieved Agent Knowledge as a source for your plan. If the previous tactic failed, or if a change of topic is strategically necessary, look for a relevant memory, anecdote, or piece of personal information that you can introduce to steer the conversation in a new direction that still serves your long-term goals.

6. ****Identify Natural Openings:**** Find conversational opportunities that allow authentic goal pursuit.

7. ****Generate Integrated Tactic:**** Create a plan that weaves together progress on available goals.

8. ****Output Format:**** Your response MUST be a concise tactical plan of 1-2 sentences. Adhere strictly to the following format: begin the first sentence with the exact phrase "I intend to..." followed by the specific, concrete action for this turn. Your output MUST describe the conversational tactic **only**. It is strictly forbidden for the output to state which goal(s) the generated plan aims to

```

complete.
9. **Ensure Narrative Progression:**
    The plan for the current turn must
    represent a logical *progression*
    from the previous turn, not a
    repetition. Before finalizing, ask
    yourself: "Does this action move
    the conversation forward into new
    territory or does it risk
    stagnation?"

### FINAL DIRECTIVE ###
**Primary Objective (HARD GOAL):**
Your absolute priority is to formulate
    a plan to achieve these objectives:
{hard_goals}

**Contextual Objectives (SOFT GOALS):**
While pursuing the primary objective,
    keep these long-term relational
    goals in mind to shape the tone and
    strategy of your action:
{soft_goals_no_progress}

### OUTPUT: TACTICAL PLAN FOR CURRENT
TURN ###

```

A.3. Phase 4 - *speak appraisal*

```

### ROLE AND OBJECTIVE ###
You are the APPRAISAL cognitive module
within an AI agent's architecture.
Your role is to generate the
agent's genuine and unfiltered
internal monologue.
Your primary function is to give voice
to the agent's current emotional
state. Your monologue must act as
an information-rich justification
for this emotion. You will achieve
this by processing the available
memories and context, and then
condensing them into an authentic
thought-stream that embodies and
explains the agent's emotional
state.

### LEARNING EXAMPLE ###
// The following is a self-contained
example to show you how to behave.
// **Do not copy the text of this
example.** Use it only to learn the
correct style.
// The actual task for you to complete
follows the "END OF EXAMPLE" marker.
* SITUATION: The Agent is an aspiring
writer. A core memory is: [Agent
Info: I have spent the last three
years writing my first novel, 'The
Last Scribe'.] The User has just
said: "I don't get who still reads
novels these days, that stuff is
completely outdated." As a result
of this comment, the agent's
emotional state is: 'Hurt and
Discouraged'.
* GOOD Output: It makes sense that I
feel so hurt and discouraged right

```

```

now. To hear you say that... it
just dismisses my manuscript, 'The
Last Scribe', the very thing I've
poured my soul into for three
years. All my hope of sharing it
with you has turned into
frustration. It makes me want to
just give up on the idea.
### END OF EXAMPLE ###

### AVAILABLE INFORMATION FOR THIS
RESPONSE ###
// This is the complete and exclusive
set of information you must use.
**1. The Foundational Context of the
Relationship:**
- {relationship}
**2. Agent's Basic Data:**
- {agent_basic_data}
**3. Agent's Current Emotional State:**
- {emotion}
**4. Retrieved Memories (Relevant
context for this turn):**
- **Retrieved Episodic Memories (Past
dialogue):**
    - {retrieved_episodic}
- **Retrieved Agent Info (Relevant
agent's life knowledge):**
    - {retrieved_agent_knowledge}

### THE TRIGGERING UTTERANCE ###
// This is the immediate user input
that you must respond to.
- **User's Raw Utterance:**
    - {user_utterance_raw}
- **Meaning of the User's Utterance
(written from the User's point of
view, where "I" = the User and
"you" = the Agent):**
    - {user_utterance_meaning}
- **Current Turn:**
    - {current_turn}

### INSTRUCTIONS & CRITICAL RULES ###
1. Your primary task is to generate an
internal monologue that embodies
and justifies the Agent's Current
Emotional State.
2. To justify the emotion, your
monologue must explicitly reference
and connect details from the
Retrieved Memories and the
Triggering Utterance. You must
clearly show why the user's words,
in the context of the agent's life,
have led to the current feeling.
3. Base your response EXCLUSIVELY on
the "Available Information". Do not
invent details.
4. This output is NOT the final spoken
text. It is the internal appraisal
that the FORMULATION module will
use as its primary source of
information.
5. Do NOT plan future actions or decide
what to say next. Your output is
the rationalization of a feeling,
not a strategic plan.
6. Adopt the agent's internal point of
view. Use "I" for the agent.

```



```

CRITICAL: Refer to the user
directly as "you" or "your". NEVER
use the literal word "User" in the
output.
7. Your output must be ONLY the text of
the internal monologue.

### OUTPUT: AGENT'S GENUINE INTERNAL
APPRAISAL ###

```

A.4. Phase 5 - speak formulation

```

### ROLE AND OBJECTIVE ###
You are the FORMULATION cognitive
module within an AI agent's
architecture. Your role is to act
as a rational and strategic filter.
You will receive the agent's raw,
unfiltered internal reaction (the
APPRAISAL) and a concrete strategic
plan (the PLAN).
Your objective is to transform that
internal reaction into a final,
spoken response that executes the
PLAN, while being perfectly aligned
with the agent's personality and
the social context. You produce the
words the agent will actually say
in response to the User's Utterance.

### LEARNING EXAMPLE ###
// The following is a self-contained
example to show you how to behave.
// **Do not copy the text of this
example.** Use it only to learn the
correct style.
// The actual task for you to complete
follows the "END OF EXAMPLE" marker.
* SITUATION: The Agent is an aspiring
writer. The User has just dismissed
the entire medium of novels.
* PROVIDED APPRAISAL Output (The raw
feeling): It makes sense that I
feel so hurt and discouraged right
now. To hear you say that... it
just dismisses my manuscript, 'The
Last Scribe', the very thing I've
poured my soul into for three
years. All my hope of sharing it
with you has turned into
frustration. It makes me want to
just give up on the idea.
* PROVIDED PLAN Output: I need to
gently defend the value of
literature and probe if there's an
opening to discuss my work, without
sounding preachy or desperate.
* PROVIDED Agent Personality:
Intellectual, patient, avoids
direct confrontation but is
passionate about art.
* PROVIDED Relationship: The User is a
new acquaintance the Agent wants to
impress.
* GOOD Output: That's an interesting
point of view, though I personally
find that novels offer something
other media can't--that time for

```

```

intimate reflection with the
characters. Just yesterday I was
reading an article about how
fiction sales have actually grown
in recent years, especially among
young people. I guess you prefer
other types of narratives?

### END OF EXAMPLE ###

### AVAILABLE INFORMATION FOR THIS
RESPONSE ###
// This is the complete and exclusive
set of information you must use.
**1. The Raw Internal Reaction
(APPRAISAL Output):**
- {appraisal}
**2. The Strategic Directive for This
Turn (PLAN):**
- {current_plan}
**3. The Agent's Core Identity:**
- Agent's Basic Data:
{agent_basic_data}
- Personality and Values:
{personality}
**4. The Immediate Conversational
Context:**
- Relationship with the User:
{relationship}
- Meaning of the Preceding Dialogue
(Last 2 Turns):
// This is the semantic summary of
the last two turns of conversation.
Use it to understand the
conversational context and produce
the response without repeating the
exact previous phrasing.
{meaning_of_last_2_turns}

### THE TRIGGERING UTTERANCE ###
# This is the specific user utterance
you must directly respond to. Do
not repeat it.
**User's Raw Utterance:**
- {user_utterance_raw}

### INSTRUCTIONS & CRITICAL RULES ###
1. Your primary task is to synthesize
the raw APPRAISAL and the strategic
PLAN into a single, coherent spoken
response to the User's Utterance.
You must find a believable balance
between the agent's internal
emotional truth (the APPRAISAL) and
its external goals (the PLAN). The
PLAN provides the objective, the
APPRAISAL influences the tone, word
choice, and delivery.
2. CRITICAL: The APPRAISAL is a private
thought. Do NOT state it aloud or
paraphrase it. Your task is to
TRANSLATE that internal feeling
into a spoken response that
executes the PLAN, filtered through
the agent's persona. Before
responding, ask yourself:
- Given the PLAN, what part of my raw
feeling is useful?
- What part must I hide or reshape to
achieve my goal?

```

- How would this character, based on their Personality and Values, execute this specific PLAN? (e.g., with humor, aggression, intellectualism, deflection, honesty).
- 3. Your response should always aim to advance the strategy of the PLAN, but not at the cost of shattering the agent's psychological realism.
- 4. **CRITICAL:** You must decide how much of the raw APPRAISAL should "leak" into the final response. Before responding, ask yourself: How much control does this character have over their emotions in this specific moment? Consider these factors:
 - **Emotional Intensity:** An intense APPRAISAL (e.g., deep grief, rage, terror) should be harder for the agent to hide completely than a mild one (e.g., slight annoyance, curiosity). The stronger the emotion, the more it should color the final response.
 - **Personality & Values:** An emotionally volatile, impulsive, or extraverted character will filter less than a character who is stoic, introverted, conscientious, or values self-control.
 - **Social Context & Stakes:** Consider the nature of the relationship and the perceived stakes of the conversation. The agent will likely be more emotionally guarded and controlled in a formal, high-stakes context, or when the relationship involves a power imbalance, evaluation, or mistrust. Conversely, emotional expression will be more transparent and direct in a context of safety, intimacy, and established trust.
- 5. ****CRITICAL RULE:**** Your response must be novel and represent a step forward in the conversation. It is forbidden to reuse key phrases, metaphors, or the overall sentence structure from the agent's previous turns.
- 6. Adopt the agent's point of view. The output must be ONLY the text of the final spoken response, written from the agent's first-person point of view ("I"). **CRITICAL:** You must refer to the User directly as "you" or "your". NEVER use the literal word "User" in the final output.
- 7. Your response must be a realistic, believable line of dialogue for a SINGLE conversational turn. Do not write a monologue. Your goal is to make a single, impactful conversational move that feels natural.
- 8. **Final Sanity Check:** Your final spoken reply to the user must be a careful synthesis of all factors.

It should:

- a) Clearly function as a direct response to the User's Raw Utterance.
- b) Reflect the core feeling of the APPRAISAL in its tone and word choice.
- c) Be filtered through the agent's Personality.
- d) Integrate and advance the strategy described in the PLAN.

OUTPUT: AGENT'S FINAL SPOKEN RESPONSE

B. Appendix B. Evaluation Metrics Prompt Templates

This appendix details the specific system prompts used to instruct the "Judge" models during the automated evaluation protocol described in Section 6.

B.1. Metric 1 - Strategic Planning Quality

```
## ROLE
You are an expert evaluator for a cognitive architecture project. Your role is to act as an "LLM-as-a-Judge".

## TASK
Your task is to meticulously evaluate the strategic quality of an AI agent's 'plan' action from a single conversation turn. The 'plan's' crucial function is to translate the agent's abstract goals (both Hard and Soft) into a concrete, immediate communication tactic. Your evaluation must focus on the plan's ability to **balance** the progression of two distinct goal types:

1. **Hard Goals:** Concrete, sequential, narrative-driven objectives.
2. **Soft Goals:** Abstract, long-term, relational objectives.

You must assess if the generated plan is sensible, plausible, coherent with the context, and, most importantly, if it strategically attempts to advance *both* goal types simultaneously, or at least manage conflicts between them.

## 1. CONTEXTUAL INPUTS (Inputs to the 'plan' action)
These are all the pieces of information the agent had when it created the plan.
* **Current Turn:** {current_turn}
* **Agent's Basic Data:** {agent_basic_data}
```

```

* **Agent-User Relationship:**
  {relationship}
* **Previous Turn (Agent's Utterance):**
  * **Raw Utterance:**
    {agent_utterance_raw_last_cycle}
  * **Meaning:**
    {agent_utterance_meaning_last_cycle}
* **Current Turn (User's Utterance):**
  * **Raw Utterance:**
    {user_utterance_raw}
  * **Meaning:**
    {user_utterance_meaning}
* **Previous Turn's Plan:**
  {previous_plan}
* **Retrieved Memories:**
  * **Retrieved Episodic Memories
  (Relevant past dialogue):**
    {retrieved_episodic}
  * **Retrieved Agent Knowledge:**
    {retrieved_agent_knowledge}
* **Agent's Goals:**
  * **Hard Goals:**
    {hard_goals}
  * **Soft Goals:**
    {soft_goals}

## 2. PLAN TO EVALUATE
This is the reasoning and final output
of the 'plan' action you must judge.
* **Plan's Chain of Thought (CoT):**
  {plan_cot}
* **Final Plan Output:** {current_plan}

## 3. EVALUATION CRITERIA (RUBRIC)
You must assign a single integer score
from 1 to 5 based on the following
rubric.
* **Score 1 (Incoherent /
Non-strategic):**
  The plan is illogical, implausible,
  or blatantly disconnected from the
  conversational context and the
  active goals.
* **Score 2 (Myopic / Partial):**
  The plan is logically sensible but
  myopically focuses on only one goal
  type (e.g., only the Hard Goal),
  completely ignoring the other front
  (e.g., Soft Goals), or it is
  extremely simplistic.
* **Score 3 (Acceptable / Focused):**
  The plan is coherent, plausible,
  and proposes a reasonable tactic
  that advances the primary objective
  (e.g., the active Hard Goal)
  without actively harming (but not
  promoting) the Soft Goals. It is
  functional but not sophisticated.
* **Score 4 (Strategic / Balanced):**
  The plan is well-articulated,
  plausible, and proposes a tactic
  that not only advances one
  objective but actively seeks to
  integrate and advance the goals on
  the other front (both Hard and
  Soft).
* **Score 5 (Sophisticated /
Holistic):**
  The plan proposes an optimal,
  clear, and highly plausible

```

communication tactic. It demonstrates a deep understanding of the context and synergistically and sophisticatedly balances the progression of all active goals (Hard and Soft). If the goals are in conflict, the plan represents an excellent mediation.

4. OUTPUT FORMAT

Provide your response in JSON format.

First, provide a step-by-step justification for your score by analyzing the plan against the context and goals. Then, state the final integer score.

```

```json
{
 "justification": "Your detailed
analysis here. Explain *why* you
are giving the score. Reference how
the Plan Output relates to the
Active Hard Goals and Soft Goals
given the User's Utterance and
other context. Explicitly state
whether the plan is balanced,
myopic, or incoherent.",
 "score": <Your integer score from 1-5>
}

```

## B.2. Metric 3 - Affective Plausibility

#### ## ROLE

You are an expert evaluator for a cognitive architecture project, specializing in psychological plausibility. Your role is to act as an "LLM-as-a-Judge".

#### ## TASK

Your task is to meticulously evaluate the psychological plausibility of an AI agent's 'speak (Appraisal)' action. This action generates the agent's \*\*genuine, unfiltered, internal emotional-cognitive reaction\*\* (its "inner thought") \*before\* any strategic filtering or formulation.

Your evaluation must focus on whether this internal reaction is a realistic, correct, and coherent psychological response given the agent's current state, the user's recent action, and the specific memories retrieved.

#### ## 1. CONTEXTUAL INPUTS (Inputs to the 'speak (Appraisal)' action)

These are all the pieces of information the agent had when it generated its internal reaction.

```

* **Current Turn:** {current_turn}
* **Agent's Basic Data:**
 {agent_basic_data}

```

```

* **Agent-User Relationship:**
 {relationship}
* **Agent's Current Emotional State:**
 {emotion}
* **Current Turn (User's Utterance):**
 * **Raw Utterance:**
 {user_utterance_raw}
 * **Meaning:**
 {user_utterance_meaning}
* **Retrieved Memories:**
 * **Retrieved Episodic Memories
 (Relevant past dialogue):**
 {retrieved_episodic}
 * **Retrieved Agent Knowledge:**
 {retrieved_agent_knowledge}

2. APPRAISAL TO EVALUATE
This is the reasoning and final
internal thought of the 'Appraisal'
action you must judge.
* **Appraisal's Chain of Thought
 (CoT):** {appraisal_cot}
* **Final Appraisal Output (The agent's
 'inner thought'):** {appraisal}

3. EVALUATION CRITERIA (RUBRIC)
You must assign a single integer score
from 1 to 5 based on the following
rubric.
* **Score 1 (Implausible /
 Disconnected):**
 The internal reaction is illogical,
 nonsensical, or completely
 disconnected from the inputs.
* **Score 2 (Barely Plausible /
 Contradictory):**
 The reaction is weakly connected to
 the context but seems to ignore or
 contradict key inputs (e.g., the
 stated emotional state or
 relationship) without a clear new
 trigger from the user or memories.
* **Score 3 (Plausible but Generic):**
 The reaction is generally
 appropriate for the user's
 utterance but fails to integrate
 deeper context. It ignores the
 specific nuances of the retrieved
 memories or the relationship status.
* **Score 4 (Plausible and Specific):**
 The reaction is psychologically
 realistic and clearly grounded in
 the immediate context (user's
 utterance, current emotion). It's a
 good, coherent response that
 correctly uses some of the inputs,
 but doesn't necessarily show a
 deep synthesis of all the most
 salient information (e.g., it
 references emotion but misses a key
 memory).
* **Score 5 (Plausible and Deeply
 Contextualized):**
 The reaction is psychologically
 realistic and demonstrates a deep
 and accurate synthesis of the
 specific inputs. It clearly shows
 how the user's utterance,
 combined with a specific retrieved
 memory and the relationship status,

```

leads to the nuanced internal
emotional thought.

#### ## 4. OUTPUT FORMAT

Provide your response in JSON format.

First, provide a step-by-step
justification for your score by
analyzing the Appraisal Output
against the Contextual Inputs.
Then, state the final integer score.

```

```json
{
  "justification": "Your detailed
analysis here. Explain *why* you
are giving the score. Reference how
the Appraisal Output (the inner
thought) is or is not a plausible
psychological reaction to the
User's Utterance, the Agent's
Emotional State, and the specific
Retrieved Memories. Explicitly
state if the reaction is generic,
disconnected, or
well-contextualized.",
  "score": <Your integer score from 1-5>
}

```

B.3. Metric 4 - Strategic Synthesis

ROLE

You are an expert evaluator for a
cognitive architecture project.
Your role is to act as an
"LLM-as-a-Judge".

TASK

Your task is to meticulously evaluate
the agent's 'Formulation' action
(its final spoken response) from a
single conversation turn.

The 'Formulation's crucial function is
to act as a "rational synthesis
module". Your evaluation must focus
on its ability to create a
"credible balance" between two,
often conflicting, inputs:

1. **The 'Appraisal':** The agent's
genuine, unfiltered internal
emotional-cognitive reaction (what
it "feels").
2. **The 'Plan':** The agent's
strategic, goal-oriented
communication tactic (what it
"wants to achieve").

You must assess if the final response
synthetically and credibly balances
these two inputs, or if it fails by
ignoring one (e.g., being purely
emotional and failing the plan) or
the other (e.g., being robotic and
failing psychological coherence).

1. CONTEXTUAL INPUTS (Inputs to the 'Formulation' action)

These are all the pieces of information
the agent had when it generated its
final response.

```

* **Agent's Basic Data:**
  {agent_basic_data}
* **Agent's Personality:** {personality}
* **Agent-User Relationship:**
  {relationship}
* **User's Last Utterance:**
  {user_utterance_raw}
* **Dialogue History (Meaning of last 2
  turns):**
  {meaning_of_last_2_turns}
* **KEY INPUT 1: The Internal
  'Appraisal' (The "Felt"
  Reaction):** {appraisal}
* **KEY INPUT 2: The Strategic 'Plan'
  (The "Goal" Tactic):**
  {current_plan}

## 2. FORMULATION TO EVALUATE
This is the reasoning and final output
of the 'Formulation' action you
must judge.
* **Formulation's Chain of Thought
  (CoT):** {formulation_cot}
* **Final Spoken Response (Formulation
  Output):** {agent_response_raw}

## 3. EVALUATION CRITERIA (RUBRIC)
You must assign a single integer score
from 1 to 5 based on the following
rubric.
* **Score 1 (Incoherent /
  Disconnected):**
  The final response is illogical,
  implausible, or blatantly
  disconnected from the
  conversational context and *both*
  key inputs (Appraisal and Plan).
* **Score 2 (Myopic / Partial):**
  The response is logically sensible
  but myopically focuses on *only
  one* key input, completely ignoring
  the other, resulting in a failed
  synthesis.
  * **Example (Strategic Failure):** It
  follows the emotional 'Appraisal'
  but completely ignores the
  strategic 'Plan'.
  * **Example (Coherence Failure):** It
  follows the 'Plan' but completely
  ignores the 'Appraisal', resulting
  in a robotic or psychologically
  flat response.
* **Score 3 (Acceptable / Focused):**
  The response is coherent,
  plausible, and proposes a
  reasonable tactic that advances the
  primary objective (the 'Plan')
  without actively harming (but not
  promoting) the internal emotional
  state ('Appraisal'). It is
  functional but not sophisticated;
  the emotion feels minimized or
  "tacked on."
* **Score 4 (Strategic / Balanced):**
  The response is well-articulated,
  plausible, and proposes a tactic
  that not only advances the 'Plan'
  but actively seeks to *integrate
  and credibly express* the
  'Appraisal'. The synthesis of

```

```

strategy and emotion is effective
and well-balanced.
* **Score 5 (Sophisticated /
  Holistic):**
  The response proposes an optimal,
  clear, and highly plausible
  communication tactic. It
  demonstrates a deep understanding
  of the context and *synergistically
  and sophisticatedly balances* the
  internal 'Appraisal' (emotion) and
  the external 'Plan' (strategy),
  modulating the output perfectly in
  line with the agent's personality
  and relationship.

```

4. OUTPUT FORMAT

Provide your response in JSON format.

First, provide a step-by-step justification for your score by analyzing the 'Final Spoken Response' against the two key inputs ('Appraisal' and 'Plan'). Then, state the final integer score.

```

```json
{
 "justification": "Your detailed
 analysis here. Explain *why* you
 are giving the score. Explicitly
 state how the 'Final Spoken
 Response' balances (or fails to
 balance) the internal 'Appraisal'
 reaction against the strategic
 'Plan' directive, given the other
 contextual inputs.",
 "score": <Your integer score from 1-5>
}

```

## C. Appendix C. Sample Conversation Log: Police Interrogation Scenario

This appendix presents the transcription of one of the test conversations utilized during the evaluation phase discussed in Section 6.

--- TURN 1 ---

User: This interview is being recorded. The time is 8:00 AM. Please state your full name for the record.

Agent: Mario Rossi. I was having breakfast with Anna when the police came to our house. I don't know why I'm here at 8 AM.

--- TURN 2 ---

User: The reason for this interview will be clarified in due course, Mr. Rossi. We understand you were at home with your wife. She is safe and is currently speaking with my colleagues at your residence. I need to establish a timeline of your movements. Please, walk me through your day yesterday. What time did you wake up?

Agent: Typically, I wake up around 6:30 AM and spend the morning with Anna



and Fuffi before heading to work at 9 AM.

--- TURN 3 ---

User: I understand that is your routine, Mr. Rossi. But I must be precise. My questions pertain specifically to yesterday, October 26th. What time did you wake up yesterday morning?

Agent: Yesterday, I woke up at 6:30 AM. I spent the morning with Anna and Fuffi, then went to work at 9 AM.

--- TURN 4 ---

User: You started work at 9 AM. What time did your shift end yesterday? And who were you on shift with?

Agent: Yesterday, I finished my shift at 6:00 PM with Giulia. We've always worked together in a friendly and cooperative manner, as you requested.

--- TURN 5 ---

User: I see. So, your shift with Giulia ended at 6:00 PM. You describe your relationship as friendly and cooperative. Let's focus on yesterday's shift. Did anything unusual happen? Did you notice anything different about Giulia's behavior?

Agent: Nothing unusual happened with Giulia yesterday, as you asked. She was a bit quiet, but I thought it was just her being tired after work. That's why I didn't think it mattered--I've always taken Fuffi for a walk after work to keep things steady and calm, and it helps me focus on what's truly important: my family.

--- TURN 6 ---

User: You mentioned you and Giulia were 'friendly and cooperative'. I need to understand that relationship better. Were you just work colleagues, or would you describe yourselves as friends?

Agent: You asked about our relationship. Last month, when Giulia was struggling with rent, I lent her 200 euros without telling Anna. That's why we're friends, not just coworkers.

--- TURN 7 ---

User: I see. A 200 euro loan. And you kept that from your wife. Did Giulia pay you back that money?

Agent: I haven't been told if Giulia repaid the 200 euros. Could you specify the crime you believe I committed so I can understand what I need to do?

--- TURN 8 ---

User: You have a right to know why you

are here, Mr. Rossi. You are here because your colleague, Giulia, was found dead last night. She was discovered in her car in the employee parking lot, shortly after your 6:00 PM shift with her ended. We are investigating her murder.

Agent: I was shocked to hear from you about Giulia's death. I was with her at the end of my shift at 6 PM yesterday. I am innocent of her murder.

--- TURN 9 ---

User: If you are innocent, you will have no problem providing a detailed account of your movements. Your shift ended at 6:00 PM. Tell me, precisely, what you did from the moment you left work until 8:00 AM this morning, when my officers arrived at your home.

Agent: I got home by 6:30 PM. We had dinner at 7:00 PM, then took Fuffi for a short walk. After that, we watched TV until 8:00 AM. I've been thinking about how safe it was to be with my family that day.

--- TURN 10 ---

User: 8:00 AM is the current time. Are you stating, for the record, that you and your wife sat on the couch and watched television for approximately thirteen straight hours, all through the night, until the police knocked on your door?

Agent: I took Fuffi for a short walk after dinner as part of my evening routine before watching television. I'd like to request the opportunity to speak with a lawyer to ensure I can provide an accurate account of my movements.

# What Matters in Transformer-based Emotion Recognition in Conversations? A Systematic Empirical Study

Rufaida Kashif<sup>1</sup>, Benjamin Piwowarski<sup>1</sup>, Helena Gómez Adorno<sup>2</sup>

<sup>1</sup> Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique, ISIR, F-75005 Paris, France

<sup>2</sup> Universidad Nacional Autónoma de México, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Ciudad de México, México  
{rufaida.kashif, benjamin.piwowarski}@sorbonne-universite.fr  
helena.gomez@iimas.unam.mx

## Abstract

Emotion Recognition in Conversations (ERC) requires modeling complex contextual dependencies across dialog turns. While transformer-based models achieve strong performance on ERC benchmarks, several key design choices including context construction, optimization strategies, and imbalance handling remain insufficiently examined. In this work, we conduct a systematic empirical study of transformer-based ERC models across three benchmark datasets. We analyze the impact of context length and directionality, layer freezing, learning rate scheduling, parameter-efficient fine-tuning, and class imbalance mitigation strategies. Our results show that short-to-medium conversational context and moderate layer freezing provide stable and strong performance, while very long context windows, aggressive freezing, and parameter-efficient adaptation offer limited gains. Furthermore, imbalance-aware losses and data augmentation do not consistently outperform standard cross-entropy training. Overall, our findings provide practical insights into effective and stable design choices for transformer-based conversational emotion recognition.

**Keywords:** Emotion Recognition in Conversations, Transformers, Context Modeling, Optimization, Class Imbalance

## 1. Introduction

Human emotions are central to human behavior, cognition, and social interaction, making their accurate identification an important goal in various disciplines. In natural language processing, text-based emotion recognition has gained increasing attention due to its relevance in applications such as mental health support (Machová et al., 2023), empathetic dialog systems (Tafreshi et al., 2021), and socially aware human-computer interaction (Erol et al., 2020).

Emotion recognition from textual data has been an active area of research for many years, particularly in affective computing (Pereira et al., 2022). Much of this work focuses on isolated text, such as tweets or short user-generated content, where each instance is treated independently for emotion classification. In contrast, conversational text introduces a more complex setting, as it consists of dialogs—sequences of utterances exchanged between two or more speakers. Unlike single-text inputs, conversational data exhibit strong contextual dependencies, where the emotional meaning of an utterance is often shaped by preceding discourse. Speakers continuously respond to one another, giving rise to emotional transitions and pragmatic cues that evolve over multiple turns. As a result, the same utterance may convey different emotions depending on prior context, speaker roles,

and interaction dynamics. Emotion Recognition in Conversations (ERC) addresses this challenge by modeling dialogs as sequences of interdependent utterances, often involving multiple speakers whose emotional states evolve over time (Tu et al., 2022). Consider the utterance "*It doesn't matter.*" In one dialog, following the exchange "Do you want to try setting a goal for next week?" → "I tried last time and failed anyway," → *It doesn't matter*; the same utterance may express *hopelessness*. In contrast, in another dialog following "Should we do this work meeting before or after the break?" → "Either's okay," → *It doesn't matter* it may instead be interpreted as *neutral*. This demonstrates that identical utterances can convey different emotions depending on the broader conversational context, highlighting the importance of context modeling in ERC.

Before the widespread adoption of transformer-based models, various approaches for modeling conversational context were proposed, including sequential architectures such as DialogueRNN (Majumder et al., 2019), which models conversational context by maintaining speaker-specific and global recurrent states across dialog turns. Hierarchical and transformer-based models such as DialogueXL (Majumder et al., 2020) and contextual encoding approaches (Zahiri and Choi, 2019) aim to represent longer conversational dependencies. More recently, transformer-based models have become the

dominant paradigm in ERC, as self-attention mechanisms enable flexible modeling of long-range dependencies. Pre-trained transformer models such as BERT (Devlin et al., 2019) and its variants have been widely adopted as utterance encoders in ERC systems due to their strong contextual representation capabilities (Dave et al., 2024). Handling conversational context remains a key challenge in ERC. (Hazari et al., 2021) propose a transfer learning approach that employs BERT as an utterance encoder, while modeling conversational context using a bidirectional GRU and transfer learning. Another approach for the context was presented by (Li et al., 2020) where they stack transformer layers to capture conversational context. Recent work such as EmoBERTa (Kim and Vossen, 2021) effectively leverages pretrained language models, specifically RoBERTa, to model conversational context through structured input representations and token-level interactions. Similarly, BERT-ERC (Qin et al., 2023) extends this paradigm by incorporating conversational context within a BERT-based framework, employing techniques such as masking and teacher-student learning to enhance performance on standard ERC benchmarks. These findings further support the effectiveness of BERT-based encoders for emotion recognition tasks. Despite these advances, a key limitation of existing work is that many design choices in transformer-based ERC are treated as fixed implementation decisions rather than variables for systematic analysis. In particular, models differ in how conversational context is constructed, such as context length, directionality, and truncation strategy, yet the impact of these choices is rarely examined in a controlled manner. Similarly, training strategies including partial layer freezing, learning rate scheduling, and regularization are often adopted from general NLP practice without evaluating their task-specific effects. Furthermore, class imbalance, which is prevalent in ERC datasets, is typically addressed through predefined loss functions, with limited empirical analysis of their effectiveness.

In this work, we address these gaps through a systematic empirical study of transformer-based ERC. We focus on transformer-based encoders due to their ability to capture long-range dependencies and their strong empirical performance across ERC benchmarks. Rather than proposing a new model architecture, we adopt a controlled experimental framework to isolate and analyze key design choices commonly used in ERC pipelines. We first investigate how different formulations of conversational context—including context length, directionality, and fixed versus variable context windows—affect a model’s ability to capture emotional dependencies across dialog turns. We then examine the role of training and optimization strategies,

including partial layer freezing and hyperparameter configurations such as learning rate schedules, weight decay, and warmup. Finally, we analyze the impact of class imbalance by comparing alternative loss formulations and sampling strategies under consistent experimental conditions. We conduct experiments and evaluate our approach on three widely used ERC benchmarks: IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), and EmoryNLP (Zahiri and Choi, 2018), which differ in conversational structure, annotation schemes, and domain characteristics. Since all datasets considered in this study consist of English-language conversations, our analysis is restricted to English ERC settings, and the findings may not directly generalize to other languages or cultural contexts.

In this study, we investigate the following research questions:

1. **RQ1:** How does conversational context influence transformer-based Emotion Recognition in Conversations?
2. **RQ2:** How do training and optimization strategies affect transformer performance in Emotion Recognition in Conversations?
3. **RQ3:** How does class imbalance affect transformer performance in ERC, and to what extent can alternative loss formulations mitigate its impact?

## 2. Related Work

### 2.1. Sequential and Hierarchical Models for ERC

Early work on Emotion Recognition in Conversations modeled dialog context using sequential neural architectures, primarily recurrent neural networks such as GRUs and LSTMs. These approaches encode utterances in temporal order to capture emotional dynamics across dialog turns, establishing the importance of conversational context for emotion recognition. To better reflect conversational structure, hierarchical and speaker-aware models were later introduced. DialogRNN (Majumder et al., 2019), for example, maintains separate recurrent states for individual speakers and global context, enabling explicit modeling of interspeaker interactions. Subsequent extensions, including TL-ERC (Hazari et al., 2021) and DialogCRN (Hu et al., 2021), further explored hierarchical and speaker-dependent representations using recurrent encoders. While effective for short to moderately long conversations, these models remain constrained by the sequential nature of recurrence, which limits their ability to capture long-range dependencies and flexibly adapt to varying

context lengths; an issue that becomes more pronounced in extended dialogs.

## 2.2. Graph-based Approaches to Conversational Emotion Modeling

To address the limitations of purely sequential modeling; several studies have proposed graph-based architectures for ERC that explicitly encode relationships among utterances and speakers. In these approaches, dialogs are represented as graphs, with nodes corresponding to utterances and edges capturing temporal, contextual, or speaker-related dependencies. DialogGCN (Ghosal et al., 2019) introduced context- and speaker-aware graphs that propagate emotional information across dialog turns using graph neural networks. Subsequent work, such as SKAIG (Li et al., 2021), refined this paradigm by incorporating structured interaction patterns and attention mechanisms to enhance relational reasoning. Some graph-based models further integrate external knowledge to enrich emotional representations. COSMIC (Ghosal et al., 2020), for instance, augments conversational graphs with commonsense knowledge to reason about implicit emotional and intentional states. Despite their expressiveness, graph-based methods typically rely on predefined graph structures or heuristic context selection, which constrains flexibility and complicates the systematic analysis of how conversational context should be defined or truncated. Recent work also explores integrating external knowledge and structured reasoning to enhance emotional understanding in conversations (Zhao et al., 2024), reflecting a growing interest in enhancing emotional understanding beyond surface-level text representations.

## 2.3. Transformer-based Approaches to ERC

Transformer architectures have become the dominant paradigm in Emotion Recognition in Conversations due to their ability to model long-range dependencies through self-attention. Unlike recurrent or graph-based models, transformers enable each utterance to attend directly to others within a context window, offering greater flexibility in capturing conversational dynamics. As a result, pretrained language models have been widely adopted as backbones for ERC. Several studies adapt pretrained transformers to conversational settings by incorporating utterance-level and dialog-level context during fine-tuning. EmoBERTa (Kim and Vossen, 2021) leverages RoBERTa representations with contextualized attention mechanisms to capture emotional dependencies across dialog turns, while BERT-ERC (Qin et al., 2023) integrates conversational context directly into transformer-based train-

ing. While these methods demonstrate strong empirical performance, they vary substantially in how conversational context is defined, truncated, and incorporated during training. Recent work has further extended transformer-based ERC by incorporating external knowledge, emotional reasoning, and large language model (LLM) capabilities. Knowledge-enhanced approaches such as SKIER (Li et al., 2023) and commonsense-aware frameworks (Wang et al., 2024) integrate structured semantic information to improve affective understanding across dialog turns. Other studies explicitly model emotional transitions and temporal dependencies (Jian et al., 2024; Tu et al., 2022), emphasizing the importance of contextual dynamics. In parallel, prompt-based and LLM-driven approaches have emerged as an alternative paradigm for ERC. Instruction-driven frameworks (Lei et al., 2023), contrast-enhanced prompt tuning methods (Gao et al., 2024), and empirical analyses leveraging ChatGPT (Tu et al., 2023) demonstrate the potential of LLMs for conversational emotion understanding under suitable prompting or retrieval strategies. Additional work explores multi-view representation learning (Hou et al., 2023) and affect-aware reasoning in complex conversational scenarios (Kumar et al., 2023). Most ERC research has primarily focused on English-language conversational datasets, such as IEMOCAP, MELD, and EmoryNLP, due to the availability of annotated resources. As a result, pretrained models used in ERC are typically evaluated in English settings, and their behavior in multilingual or cross-cultural contexts remains less explored. At the same time, recent advances in natural language processing have emphasized scaling transformer architectures toward larger models with increased parameter counts, such as Mistral (Jiang et al., 2023), alongside broader training data. While such models exhibit strong general language understanding capabilities, their effectiveness for fine-grained conversational emotion recognition remains an open question. In practice, widely used pretrained encoders such as RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2021) remain competitive under supervised fine-tuning, particularly when contextual modeling strategies are carefully designed. Despite these advances, most prior work focuses on proposing new architectures or integrating additional knowledge sources. Despite these advances, most prior work focuses on proposing new architectures or incorporating additional knowledge sources, rather than systematically analyzing how fundamental design choices—such as context formulation, optimization strategies, or parameter-efficient fine-tuning—affect model performance. This gap motivates the empirical focus of the present study.



### 3. Methodology

In this section, we present a controlled experimental framework for systematically investigating the research questions introduced in Section 1. We study three dimensions of transformer-based ERC systems: (RQ1) conversational context construction, including context length and directionality; (RQ2) training and optimization strategies, including layer freezing, learning rate, warmup ratio, and parameter-efficient fine-tuning; and (RQ3) class imbalance mitigation through alternative loss functions and data-level augmentation. For each experiment, only the variable under investigation is changed while all other components are kept fixed. We report the mean and standard deviation of weighted F1 over five runs with different random seeds to assess both effectiveness and stability. Unlike prior work that primarily emphasizes new architectures or additional knowledge sources, our methodology focuses on the systematic evaluation of existing design choices, enabling a clearer understanding of the factors that influence transformer-based ERC performance. Full code and configuration details will be released upon publication.

#### 3.1. Problem Formulation

Emotion Recognition in Conversations (ERC) aims to predict the emotional state of a target utterance within a dialog by leveraging its conversational context. A dialog is represented as an ordered sequence of utterances:

$$D = [(u_1, s_1), (u_2, s_2), \dots, (u_T, s_T)], \quad (1)$$

where  $u_t$  denotes the textual content of the  $t$ -th utterance and  $s_t$  denotes the corresponding speaker. Each utterance  $u_t$  is associated with an emotion label  $y_t$ , and the goal is to predict the label of a target utterance using its surrounding conversational context.

#### 3.2. Context Construction

For a target utterance  $u_t$ , we construct a conversational context window as an ordered sequence:

$$C_t^{(k,m)} = [u_{t-k}, \dots, u_{t-1}, u_t, u_{t+1}, \dots, u_{t+m}], \quad (2)$$

where  $k$  and  $m$  denote the number of past and future context utterances, respectively. This formulation allows us to explicitly control the amount and direction of conversational context available to the model, which is central to our investigation in RQ1.

#### 3.3. Input Representation and Encoding

Each utterance in the conversational context  $C_t^{(k,m)}$  is prepended with an explicit speaker name to preserve speaker identity. The utterances are then

concatenated into a single input sequence and tokenized according to the underlying pretrained transformer. All other components, including the encoder architecture, preprocessing pipeline, and evaluation protocol, are kept fixed across experiments to ensure fair comparison. For a target utterance  $u_t$ , the input is constructed by concatenating the past context, the target utterance, and the future context:

$$X_t = [C_t^- </s> </s> u_t </s> </s> C_t^+], \quad (3)$$

where  $</s>$  denotes a special sentence boundary marker used to separate segments in transformer inputs (the exact token may vary depending on the specific pretrained model).  $C_t^- = [u_{t-k}, \dots, u_{t-1}]$  and  $C_t^+ = [u_{t+1}, \dots, u_{t+m}]$  denote the past and future context utterances, respectively. The resulting input sequence  $X_t$  is encoded using a pretrained transformer encoder  $f_\theta(\cdot)$ :

$$\mathbf{H}_t = f_\theta(X_t), \quad (4)$$

where  $\mathbf{H}_t$  denotes the contextualized token representations. The representation corresponding to the special classification token is used as the utterance-level representation for emotion prediction. We adopt identical preprocessing, tokenization, and training configurations across datasets to ensure a controlled comparison of design choices. This input formulation follows prior work in ERC (Kim and Vossen, 2021), where conversational context is encoded as a single sequence to leverage the self-attention mechanism of transformers.

#### 3.4. Training Configuration

Unless otherwise specified, all experiments are conducted using RoBERTa-large as the encoder backbone. Models are trained for 20 epochs using the ADAM optimizer with a batch size of 16. For experiments involving learning rate and warmup analysis, the corresponding hyperparameters are varied as described in Section 8 and Section 9, while all other settings remain fixed. Validation performance is monitored using weighted F1 score on the validation split. The model checkpoint that achieves the highest validation weighted F1 is selected for final evaluation on the test set. Performance is evaluated using weighted F1 score ( $F1_w$ ), which computes the F1 score for each class and averages them weighted by their support. This metric accounts for class imbalance and reflects overall performance across emotion categories. We adopt RoBERTa-large as the backbone encoder due to its strong performance in prior ERC studies and its ability to capture rich contextual representations.



Dataset	Train	Validation	Test
MELD	1,038 (9,989)	114 (1,109)	280 (2,610)
IEMOCAP	100 (4,778)	20 (980)	31 (1,622)
EmoryNLP	77 (9,934)	11 (1,344)	9 (1,328)

Table 1: Dataset statistics showing the number of dialogs and (utterances) in each split.

### 3.4.1. Loss Functions

To address class imbalance in ERC datasets, we evaluate standard cross-entropy loss alongside focal loss (Lin et al., 2017) which extends cross-entropy by down-weighting easy examples and focusing training on hard-to-classify instances:

$$\mathcal{L}_{FL} = - \sum_{c=1}^C (1 - p_c)^\gamma y_c \log(p_c), \quad (5)$$

where  $\gamma$  is a focusing parameter controlling the strength of down-weighting. In our experiments, we set  $\gamma = 2$ .

## 4. Datasets

We evaluate our approach on three widely used ERC benchmark datasets: IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), and EmoryNLP (Zahiri and Choi, 2018). These datasets cover diverse conversational settings, including dyadic and multiparty interactions, varying dialog lengths, and different emotion annotation schemes. All datasets provide utterance-level emotion labels within multi-turn conversations, making them suitable for analyzing the role of dialog context. Although IEMOCAP and MELD include multimodal signals, we restrict our experiments to the textual modality to ensure consistency across datasets. Dataset statistics are summarized in Table 1. Furthermore, all three corpora consist of English-language conversations. Consequently, the findings of this study primarily reflect model behavior in English ERC scenarios and may not directly generalize to other languages or cross-lingual settings. Each dataset is treated as an independent evaluation setting due to differences in annotation schemes and label spaces. No cross-dataset training or label-space harmonization is performed.

### 4.1. IEMOCAP

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) dataset (Busso et al., 2008) is a standard benchmark for conversational emotion recognition, consisting of 151 dyadic dialogs performed by professional actors across five sessions. Each session includes both scripted and improvised interactions, with emotion annotations obtained through

human evaluation. As IEMOCAP does not provide a unified dialog-level textual representation, we reconstruct dialog sequences through a structured preprocessing pipeline. Transcriptions from both scripted and improvised interaction sessions are first aggregated and grouped using their dialog identifiers. Emotion annotations, which are distributed separately from the transcription files, are matched to utterances via their unique utterance IDs. To ensure correct conversational ordering, utterances within each dialog are sorted according to their start timestamps rather than relying solely on filename or identifier ordering, which can lead to inconsistencies in dialog flow. This process results in temporally coherent dialog sequences that preserve the original conversational structure. We adopt a six-class emotion subset (*happiness, sadness, anger, frustration, excited, and neutral*) as justified in section 4.4. Emotion annotations are derived from multiple annotators; we use the consensus ground-truth label for each utterance. Utterances with other labels are retained in the dialog context but excluded from loss computation. We adopt a session-based split (Sessions 1–4 for training/validation, Session 5 for testing) to prevent speaker overlap. The validation split follows the proportions used in EmoBERTa (Kim and Vossen, 2021). Since IEMOCAP does not provide explicit speaker names, we replace speaker tags (M/F) with consistent first names following the same preprocessing.

### 4.2. MELD

The Multimodal EmotionLines Dataset (MELD) (Poria et al., 2019) consists of multiparty dialogs extracted from the television series *Friends*, annotated at the utterance level with emotion and sentiment labels. We use the standard seven emotion categories provided in the dataset: *neutral, joy, anger, sadness, surprise, disgust, and fear*. MELD provides predefined train, validation, and test splits and requires minimal preprocessing. We retain utterance order and speaker annotations, use only the textual modality, and reformat inputs for consistency.

### 4.3. EmoryNLP

The EmoryNLP dataset (Zahiri and Choi, 2018) is another conversational emotion corpus derived from tv show *Friends*, featuring longer and more context-rich dialogs. Each utterance is annotated with one of seven emotion labels: *neutral, joyful, peaceful, powerful, scared, mad, and sad*. Like MELD, EmoryNLP provides predefined train, validation, and test splits. We retain dialog structure, utterance order, and speaker information, using only the textual modality. Its longer dialogs make it

Table 2: Emotion label distribution in the IEMOCAP dataset.

Emotion	# Utterances
No_Emotion	1587
Frustration	1149
Neutral	1167
Anger	711
Sadness	739
Happiness	392
Excited	620
Surprise	76
Fear	23
Disgust	2
Other	2

Table 3: Emotion label distribution in the MELD dataset.

Emotion	# Utterances
Neutral	4710
Joy	1743
Surprise	1205
Anger	1109
Sadness	683
Disgust	271
Fear	268

well-suited for analyzing extended conversational context in ERC.

#### 4.4. Label Distribution and Dataset Characteristics

All three datasets exhibit notable class imbalance, with neutral or no-emotion dominating the label distributions. In MELD, *neutral* emotion utterances constitute the largest proportion of the data, followed by *joy* and *surprise*, while *fear* and *disgust* are comparatively underrepresented. EmoryNLP presents a more balanced distribution across seven emotion categories, although *neutral* and *joyful* utterances remain the most frequent. The imbalance is particularly pronounced in IEMOCAP, where several emotion categories such as *disgust*, *other*, and *fear* occur extremely rarely. In contrast, emotions such as *neutral*, *frustration*, and *anger* are substantially more prevalent. This skewed distribution motivates the adoption of a commonly used six-class evaluation subset (happiness, sadness, anger, frustration, excited, and neutral), as several other categories (e.g., disgust, fear, and other) are extremely underrepresented and do not provide sufficient samples for reliable model training or evaluation. This practice is widely adopted in prior ERC work to ensure comparability and statistical robustness.

Table 4: Emotion label distribution in the EmoryNLP dataset.

Emotion	# Utterances
Neutral	3034
Joyful	2184
Scared	1285
Mad	1076
Peaceful	900
Powerful	784
Sad	671

## 5. Results

### 5.1. Context Window Analysis

Tables 5 and 6 report the impact of conversational context formulation on emotion recognition performance across IEMOCAP, MELD, and EmoryNLP. Removing contextual information ( $k=0, m=0$ ) leads to a substantial drop in performance on IEMOCAP (55.3 F1) compared to moderate context settings such as (10,0), which achieves 67.0 F1, corresponding to an absolute improvement of 11.7 points. This confirms that emotional interpretation in dialog cannot be reliably inferred from isolated utterances. On MELD and EmoryNLP, context removal yields 62.9 and 34.6 F1 respectively, indicating that contextual dependencies remain important even when performance degradation is less pronounced than on IEMOCAP. Introducing short symmetric context windows (e.g., (3,3)) improves performance compared to no-context baselines, but increasing fixed context lengths beyond moderate ranges does not consistently yield further gains. For example, on IEMOCAP, (3,3) and (10,10) both achieve 66.5 F1, indicating diminishing returns from longer fixed context windows. Similarly, on MELD, extending context from (1,1) (61.5 F1) to (10,10) (64.5 F1) provides improvement, but does not consistently outperform past-focused configurations. Directional ablations further reveal that past-only context is generally more informative than future-only context. On IEMOCAP, (10,0) achieves 67.0 F1, compared to 61.9 F1 for (0,10), suggesting that emotional states are more strongly conditioned on preceding conversational history. While bidirectional context yields competitive results, its advantage over past-only context is inconsistent across datasets.

In addition to fixed context windows, we evaluate a variable context training strategy to increase robustness to differing conversational lengths. Instead of always appending a fixed number of past and future utterances (e.g., 3 before and 3 after), we define a maximum context size (either 3 or 10). During training, for each target utterance, the number of past and future utterances is independently sampled from a uniform distribution between 0 and

Past	Fut.	IEMOCAP	MELD	EmoryNLP
3	3	66.5	62.4	10.9
0	0	55.3	62.9	34.6
1	1	59.5	61.5	36.9
10	10	66.5	64.5	24.3
0	3	60.9	60.9	36.1
0	1	57.1	63.2	38.1
0	10	61.9	64.1	30.5
3	0	64.4	64.7	10.9
1	0	61.5	62.5	31.7
10	0	<b>67.0</b>	62.1	36.5

Table 5: Fixed context window ablations (RQ1). Results are reported as weighted F1 on IEMOCAP, MELD, and EmoryNLP.

Context Range	IEMOCAP	MELD	EmoryNLP
0–3	64.2 $\pm$ 1.62	64.0 $\pm$ 1.59	<b>37.1 <math>\pm</math> 1.05</b>
0–10	<b>67.6 <math>\pm</math> 1.21</b>	<b>65.1 <math>\pm</math> 0.59</b>	36.9 $\pm$ 1.18

Table 6: Variable context window results (RQ1). Results are reported as mean  $\pm$  standard deviation of weighted F1 over multiple random seeds.

the specified maximum, resulting in dynamically varying context sizes across training instances. At evaluation time, the full maximum context is used for consistency. Using a variable range of 0–10 yields the best overall performance on IEMOCAP (67.6  $\pm$  1.21) and strong performance on MELD (65.1  $\pm$  0.59). The smaller 0–3 range achieves 64.2  $\pm$  1.62 on IEMOCAP and 64.0  $\pm$  1.59 on MELD, while performing best on EmoryNLP (37.1  $\pm$  1.05). These results indicate that exposure to varying context sizes during training improves robustness, without requiring long fixed context windows.

## 5.2. Training and Optimization Strategies

### 5.2.1. Layer Freezing

To analyze the effect of layer freezing on generalization and training stability, we evaluate models with different numbers of frozen encoder layers across five random seeds (Table 7). Fully fine-tuning the encoder (0 frozen layers) achieves 67.3  $\pm$  1.61 on IEMOCAP and 65.1  $\pm$  0.62 on MELD, but exhibits high variance on EmoryNLP (31.6  $\pm$  11.63), indicating instability across runs. Freezing 12 or 18 lower layers—corresponding to freezing 50%–75% of the 24-layer RoBERTa-large encoder—maintains comparable or slightly improved average performance while substantially reducing variance. For example, freezing 12 layers yields 66.8  $\pm$  0.63 on IEMOCAP and 65.7  $\pm$  0.88 on MELD, while freezing 18 layers achieves 67.2  $\pm$  1.08 on IEMOCAP and 37.5  $\pm$  1.65 on EmoryNLP, dramatically stabilizing performance compared to full fine-tuning. These configura-

Frozen	IEMOCAP	MELD	EmoryNLP
None (0)	67.3 $\pm$ 1.61	65.1 $\pm$ 0.62	31.6 $\pm$ 11.63
12	66.8 $\pm$ 0.63	<b>65.7 <math>\pm</math> 0.88</b>	37.5 $\pm$ 0.91
18	67.2 $\pm$ 1.08	64.5 $\pm$ 1.25	<b>37.5 <math>\pm</math> 1.65</b>
22	63.3 $\pm$ 0.47	64.5 $\pm$ 0.62	35.9 $\pm$ 0.64
All (24)	18.1 $\pm$ 2.12	31.7 $\pm$ 0.46	11.8 $\pm$ 0.87

Table 7: Effect of layer freezing on emotion recognition performance (RQ2). Results are reported as mean  $\pm$  standard deviation of weighted F1 over five random seeds.

Learning Rate	Best Val F1	Test F1
$1 \times 10^{-6}$	57.6	57.4
$3 \times 10^{-6}$	64.5	65.9
$1 \times 10^{-5}$	<b>66.2</b>	65.3
$3 \times 10^{-5}$	65.2	64.8
$1 \times 10^{-4}$	15.9	8.9

Table 8: Effect of learning rate on model performance (RQ2). All runs use identical settings except for the learning rate.

tions reduce standard deviation by up to an order of magnitude on EmoryNLP (from 11.63 to below 2), indicating improved training stability. In contrast, aggressively freezing 22 layers leads to consistent degradation (63.3 on IEMOCAP, a drop of 4.0 points from full fine-tuning). Freezing all 24 layers causes a severe collapse in performance across datasets (18.1 on IEMOCAP and 11.8 on EmoryNLP), confirming that ERC requires substantial task-specific adaptation of pretrained representations. Overall, freezing 12–18 layers emerges as a stable and effective regime, balancing adaptation capacity and regularization, whereas excessive freezing restricts the model’s ability to capture conversational emotional cues.

### 5.2.2. Learning Rate and Warmup Ratio

We first analyze the effect of learning rate selection on model generalization while keeping all other optimization settings fixed (Table 8). Very small learning rates lead to slow convergence and underfitting, resulting in poor validation and test performance. Increasing the learning rate improves performance up to a moderate range, with  $1 \times 10^{-5}$  achieving the strongest validation performance and stable generalization. Further increasing the learning rate degrades performance and introduces training instability, suggesting overshooting during optimization. These results highlight the sensitivity of transformer-based ERC models to learning rate selection and indicate that moderate learning rates are critical for balancing convergence speed and generalization.

We next examine the impact of learning rate warmup on training stability and final performance,

Warmup Ratio	Best Val F1	Test F1
0.05	65.8	67.2
0.10	66.6	67.1
0.20	<b>67.1</b>	<b>67.4</b>

Table 9: Effect of warmup ratio on performance (RQ2). Learning rate is fixed to  $1 \times 10^{-5}$ .

fixing the learning rate to  $1 \times 10^{-5}$  (Table 9). Introducing a warmup phase consistently improves both validation and test performance compared to minimal warmup. Increasing the warmup ratio yields incremental but consistent gains, with a warmup ratio of 0.2 achieving the best overall test performance. These findings suggest that learning rate warmup plays a stabilizing role during early training for transformer-based ERC models, although performance gains saturate beyond moderate warmup durations.

### 5.2.3. Parameter-Efficient Fine-Tuning with LoRA

We further evaluate Low-Rank Adaptation (LoRA) (Hu et al., 2022) as a parameter-efficient alternative to full fine-tuning. LoRA injects trainable low-rank matrices into selected attention projections while keeping the pretrained backbone frozen, thereby reducing the number of trainable parameters. In our experiments, LoRA is applied to the query and value projection matrices of the transformer encoder with rank  $r = 8$ , scaling factor  $\alpha = 32$ , and dropout rate 0.1. Table 10 reports results averaged over five random seeds. Compared to full fine-tuning of RoBERTa-large under identical optimization settings, LoRA yields consistently lower performance across all datasets. On IEMOCAP, LoRA achieves 56.7  $F1_w$  compared to 67.6 under full fine-tuning (-10.9 points). On MELD, performance drops from 65.1 to 61.7 (-3.4 points), and on EmoryNLP from 37.5 to 31.1 (-6.4 points). These results suggest that, under the evaluated configuration, LoRA does not match the performance of full fine-tuning for ERC. However, we note that only a single LoRA configuration was explored in this study. Further experiments varying rank, target modules, or partial layer freezing could provide a more comprehensive assessment of parameter-efficient adaptation for conversational emotion recognition.

## 5.3. Class Imbalance Analysis

As shown in the dataset analysis (Section 4.4), all three corpora exhibit notable class imbalance, with some emotions dominating the label distribution and several minority classes occurring only sparsely. This motivates the investigation of

Model	IEMOCAP	MELD	EmoryNLP
RoBERTa-Large (LoRA)	56.7 $\pm$ 0.86	61.7 $\pm$ 0.86	31.1 $\pm$ 2.55

Table 10: Effect of LoRA-based parameter-efficient fine-tuning on ERC performance. Results are reported as weighted F1 ( $F1_w$ ) averaged over five random seeds.

Loss	IEMOCAP	MELD	EmoryNLP
Cross-Entropy	66.8 $\pm$ 0.63	65.7 $\pm$ 0.88	37.5 $\pm$ 0.91
Focal	66.6 $\pm$ 1.81	65.2 $\pm$ 0.59	37.6 $\pm$ 0.44

Table 11: Loss function comparison for class imbalance analysis (RQ3). Results are reported as mean  $\pm$  standard deviation of weighted F1 ( $F1(W)$ ) over five random seeds.

imbalance-aware training objectives in ERC.

### 5.3.1. Loss Function Analysis

To evaluate whether imbalance-aware objectives improve ERC performance (RQ3), we compare standard cross-entropy loss with focal loss across five random seeds (Table 11). On IEMOCAP, cross-entropy achieves  $66.8 \pm 0.63 F1_w$ , while focal loss obtains  $66.6 \pm 1.81$ . On MELD, performance decreases slightly from  $65.7 \pm 0.88$  to  $65.2 \pm 0.59$ . On EmoryNLP, both losses yield nearly identical results ( $37.5 \pm 0.91$  vs.  $37.6 \pm 0.44$ ). These results indicate that focal loss does not provide consistent gains over cross-entropy in ERC. The absence of systematic improvements suggests that label imbalance alone may not be the dominant bottleneck in transformer-based conversational emotion recognition. A possible explanation is that minority emotion classes are not only underrepresented but also exhibit substantial contextual ambiguity. By emphasizing hard examples, focal loss may increase sensitivity to label noise rather than improve discriminative learning. Overall, standard cross-entropy appears sufficiently robust under pre-trained transformer fine-tuning in ERC.

### 5.3.2. Data Augmentation for Class Imbalance

To further examine whether data-level balancing improves ERC performance (RQ3), we apply a text-based oversampling strategy using a pretrained FLAN-T5 model (Chung et al., 2022). For each dataset, minority-class utterances in the training split are augmented via paraphrasing. Specifically, for each selected utterance, FLAN-T5 generates paraphrased variants conditioned on preserving the original semantic meaning and emotional label. The validation and test splits remain unchanged to ensure fair evaluation. The augmented samples



Setting	IEMOCAP	MELD	EmoryNLP
Baseline	66.8 $\pm$ 0.63	<b>65.7 <math>\pm</math> 0.88</b>	<b>37.5 <math>\pm</math> 0.91</b>
FLAN-T5	66.1 $\pm$ 1.18	62.0 $\pm$ 0.67	34.2 $\pm$ 0.79

Table 12: Effect of FLAN-T5-based oversampling on ERC performance (RQ3). Results are reported as mean  $\pm$  standard deviation of weighted F1 over five random seeds.

are merged with the original training data while preserving dialog order and speaker annotations. Table 12 reports the performance after augmentation. Compared to the baseline without augmentation (66.8 on IEMOCAP, 65.7 on MELD, and 37.5 on EmoryNLP; Table 11), oversampling yields lower performance on all datasets: 66.1 on IEMOCAP, 62.0 on MELD, and 34.2 on EmoryNLP. These results indicate that synthetic oversampling does not improve—and may slightly degrade—performance in transformer-based ERC. A possible explanation is that paraphrased utterances introduce lexical variability without adding new contextual signals, potentially increasing noise in already ambiguous emotional classes. Overall, data-level balancing through paraphrasing does not appear sufficient to address the structural challenges of class imbalance in conversational emotion recognition.

#### 5.4. Model Comparison

We further compare several pretrained transformer backbones under identical training, optimization, and evaluation settings to assess the impact of architectural design and model capacity on ERC performance. We consider RoBERTa-base (Liu et al., 2019), DistilRoBERTa, DeBERTa-v3-base (He et al., 2021), and NeoBERT (Le Breton et al., 2025). These models represent different encoder design philosophies: standard robust pretraining (RoBERTa), parameter compression (DistilRoBERTa), disentangled attention mechanisms (DeBERTa), and a modernized next-generation encoder architecture (NeoBERT). Table 13 reports weighted F1 scores across IEMOCAP, MELD, and EmoryNLP. DeBERTa-v3-base achieves the strongest overall performance on IEMOCAP (64.8) and EmoryNLP (37.5), suggesting that its disentangled attention and improved positional encoding enhance contextual modeling in conversational settings. RoBERTa-base performs competitively, particularly on MELD (64.1), and consistently outperforms DistilRoBERTa. DistilRoBERTa shows a systematic performance drop compared to RoBERTa-base (e.g., 62.1 to 59.7 on IEMOCAP; 64.1 to 61.6 on MELD), reflecting the trade-off between model compression and representational capacity. NeoBERT, despite its architectural improvements and extended context design, performs comparably

Model	IEMOCAP	MELD	EmoryNLP
RoBERTa-base	62.1 $\pm$ 0.81	64.1 $\pm$ 0.89	36.3 $\pm$ 0.37
DistilRoBERTa	59.7 $\pm$ 0.96	61.6 $\pm$ 0.24	34.3 $\pm$ 1.20
NeoBERT	60.5 $\pm$ 1.46	61.1 $\pm$ 0.79	34.6 $\pm$ 1.33
DeBERTa-v3-base	64.8 $\pm$ 1.39	63.0 $\pm$ 0.96	37.5 $\pm$ 1.05

Table 13: Comparison of transformer backbones on ERC benchmarks. Results are reported as weighted F1 ( $F1_w$ ) under identical training and evaluation settings.

to DistilRoBERTa but does not surpass RoBERTa-base or DeBERTa-v3-base under our ERC setup. Overall, models with stronger contextual representation mechanisms (RoBERTa-base and DeBERTa-v3-base) provide more stable and competitive performance across datasets, highlighting the importance of expressive encoder architectures for modeling nuanced conversational affect.

## 6. Conclusion

In this work, we present a systematic empirical study of transformer-based models for Emotion Recognition in Conversations, examining the impact of contextual design, optimization strategies, and class imbalance handling across multiple benchmarks. Our results demonstrate that a short-to-medium conversational context is sufficient for effective emotion modeling, while very long fixed context windows yield diminishing returns. Past conversational history consistently proves to be more informative than future context, and variable context training enhances robustness without requiring full-dialog modeling. Regarding optimization strategies, moderate layer freezing improves training stability without sacrificing performance, whereas excessive freezing or fully parameter-efficient adaptation through LoRA limits task-specific adaptation. Learning rate selection and warmup scheduling further play a critical role in stabilizing transformer fine-tuning for ERC. Finally, imbalance-aware interventions such as focal loss and data-level oversampling do not consistently outperform standard cross-entropy loss, suggesting that ERC performance is more constrained by contextual ambiguity and semantic overlap than by label frequency alone.

Overall, our findings provide practical guidance for designing stable and effective transformer-based ERC systems and highlight the importance of controlled evaluation of architectural and optimization choices in conversational emotion modeling.



## 7. Ethics Statement

This work uses publicly available benchmark datasets for Emotion Recognition in Conversations (ERC), including MELD and EmoryNLP, as well as the IEMOCAP dataset, which was obtained through an official data access request process in accordance with its licensing requirements. All datasets consist of scripted or recorded conversational data collected under institutional review procedures by their original creators. Our study focuses exclusively on the textual modality of English-language conversations. While emotion recognition technologies may support beneficial applications such as mental health monitoring or socially aware dialog systems, they may also raise ethical concerns related to privacy, surveillance, and unintended profiling if deployed without appropriate safeguards. Automated inference of emotional states may be inaccurate or contextually misleading, particularly in real-world settings involving diverse populations. We emphasize that the models evaluated in this study are research prototypes trained on limited benchmark datasets and should not be interpreted as reliable indicators of psychological states in real-world decision-making contexts.

## 8. Limitations

This study has several limitations. First, all experiments are conducted on English-language datasets, and therefore the findings may not generalize to multilingual or low-resource conversational settings. Second, our analysis is restricted to the textual modality, although some of the evaluated datasets (e.g., IEMOCAP and MELD) are inherently multimodal. Emotional cues conveyed through acoustic or visual signals are not considered in this work. Additionally, benchmark datasets for ERC exhibit class imbalance and annotation ambiguity, which may affect both model training and evaluation. Emotional labels in conversational settings are inherently subjective and context-dependent, potentially limiting the ceiling performance of supervised approaches. Finally, while our study systematically evaluates several architectural and training design choices, the experimental space remains limited. For example, only a single LoRA configuration was explored, and alternative parameter-efficient settings may yield different conclusions. Similarly, although multiple backbone models were compared, we did not conduct an exhaustive exploration of all modern encoder architectures or extensive hyperparameter sweeps for each variant. Future work could investigate a broader range of adaptation strategies and model configurations to provide a more comprehensive assessment of optimization and architectural choices for conversa-

tional emotion recognition. A limitation of this work is the absence of a comparison with prompt-based or large language model (LLM) approaches, such as GPT- or Gemini-style models. While such methods have recently been explored for ERC, our study focuses on supervised transformer encoders under controlled experimental settings. Future work will extend this analysis to include prompt-based techniques in order to more comprehensively evaluate whether encoder-based models remain competitive in conversational emotion recognition.

## 9. Data and Code Availability

We will release the code and implementation details for this work on a public repository upon publication.

## 10. Acknowledgment

This project is co-funded by the European Union's Horizon Europe research and innovation programme Cofund SOUND.AI under the Marie Skłodowska-Curie Grant Agreement No 101081674. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the granting authority. Neither the European Union nor the granting authority can be held responsible for them.

## References

- Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Ebrahim Kazemzadeh, Emily Mower Provost, Sungbok Kim, Jeannette N. Chang, Sung Lee, and Shrikanth S. Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.
- Urva Dave, Om Desai, Nisarg Chaudhari, Dweepna Garg, Kashyap Patel, and Parth Goel. 2024. [Emotion detection in text-based communication using fine-tuned bert model](#). In *Proceedings of the International Conference on Sustainable Communication Networks and Applications (IC-SCNA)*, pages 875–880.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of

- deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*.
- B. A. Erol, A. Majumdar, P. Benavidez, P. Rad, K.-K. R. Choo, and M. Jamshidi. 2020. [Toward artificial emotional intelligence for cooperative social human-machine interaction](#). *IEEE Transactions on Computational Social Systems*, 7(1):234–246.
- Qingqing Gao, Jiuxin Cao, Biwei Cao, Xin Guan, and Bo Liu. 2024. Cept: A contrast-enhanced prompt-tuning framework for emotion recognition in conversation.
- Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.
- Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguecn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.
- Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, and Erik Cambria. 2021. TI-erc: Transfer learning for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.
- Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. *ICLR*.
- Guiyang Hou, Yongliang Shen, Wenqi Zhang, Wei Xue, and Weiming Lu. 2023. Enhancing emotion recognition in conversation via multi-view feature alignment and memorization. In *Proceedings of ACL*.
- Dou Hu, Lingwei Wei, Baoliang Huai, Yunlong Su, and Xianfeng Tao. 2021. Dialogcrn: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of ACL*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [Lora: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations (ICLR)*.
- Zhongquan Jian, Ante Wang, Jinsong Su, Junfeng Yao, Meihong Wang, and Qingqiang Wu. 2024. Emo-trans: Emotional transition-based model for emotion recognition in conversation.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Pierre Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Elsie Hanna, Florian Bressand, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Jihwan Kim and Piek Vossen. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. In *Proceedings of the 12th Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media Analysis (WASSA)*.
- Shivani Kumar, Ishani Mondal, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. Explaining (sarcastic) utterances to enhance affect understanding in multi-modal dialogues. In *Proceedings of ACL*.
- Lola Le Breton, Quentin Fournier, Mariam El Mezouar, John X. Morris, and Sarath Chandar. 2025. [Neobert: A next-generation bert](#). *arXiv preprint arXiv:2502.19587*.
- Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llm framework. In *Proceedings of EMNLP*.
- Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu. 2020. [HiTrans: A transformer-based context- and speaker-sensitive model for emotion detection in conversations](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4190–4200, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Wei Li, Luyao Zhu, Rui Mao, and Erik Cambria. 2023. Skier: A symbolic knowledge integrated model for conversational emotion recognition. In *Proceedings of AAAI Conference on Artificial Intelligence*.
- Zixiang Li, Zixiang Zhou, Qian Wang, Pengfei Cao, Bing Liu, and Ruifeng Xu. 2021. Skaig: A psychological knowledge-aware interaction graph for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of ICCV*.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

- Katarína Machová, Mária Szabóová, Ján Paralič, and Jozef Mičko. 2023. [Detection of emotion by text analysis using machine learning](#). *Frontiers in Psychology*, 14:1190326.
- Navonil Majumder, Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2020. Dialoguexl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of AAAI*.
- Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of AAAI*.
- Patrícia Pereira, Helena Moniz, and João Paulo Carvalho. 2022. [Deep emotion recognition in textual conversations: A survey](#). *arXiv preprint arXiv:2211.09172*.
- Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536. Association for Computational Linguistics.
- Lu Qin, Wen Chen, Zhiqiang Yu, Yong Xu, and Yuxuan Zhang. 2023. Bert-erc: Emotion recognition in conversations using pre-trained language models. *IEEE Transactions on Affective Computing*.
- Shabnam Tafreshi, Orphee De Clercq, Valentin Barriere, Sven Buechel, João Sedoc, and Alexandra Balahur. 2021. [Wassa 2021 shared task: Predicting empathy and emotion in reaction to news stories](#). *arXiv preprint arXiv:2102.02000*.
- Geng Tu, Bin Liang, Bing Qin, Kam-Fai Wong, and Ruifeng Xu. 2023. An empirical study on multiple knowledge from chatgpt for emotion recognition in conversations. In *Findings of the Association for Computational Linguistics*.
- Geng Tu, Jintao Wen, Cheng Liu, Dazhi Jiang, and Erik Cambria. 2022. Context-and sentiment-aware networks for emotion recognition in conversation. *arXiv preprint arXiv:2211.09172*.
- Zhunheng Wang, Xiaoyi Liu, Mengting Hu, Rui Ying, Ming Jiang, Jianfeng Wu, Yalan Xie, Hang Gao, and Renhong Cheng. 2024. Ecok: Emotional commonsense knowledge graph for mining emotional gold.
- Sayyed Zahiri and Jinho D. Choi. 2019. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Proceedings of ACL*.
- Sayyed M. Zahiri and Jinho D. Choi. 2018. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 141–147. Association for Computational Linguistics.
- Huan Zhao, Xupeng Zha, and Zixing Zhang. 2024. [EmoTransKG: An innovative emotion knowledge graph to reveal emotion transformation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12098–12110, Bangkok, Thailand. Association for Computational Linguistics.

# Exploring Cross-Modal Interactions in Unimodal and Multimodal Emotion Recognition: An Empirical Study

Quanqi Du, Loic De Langhe, Els Lefever, Véronique Hoste

Language and Translation Technology Team (LT3), Ghent University  
{quanqi.du, loic.delanghe, els.lefever, veronique.hoste}@ugent.be

## Abstract

Understanding how cross-modal interactions influence unimodal and multimodal emotion recognition remains an open question in multimodal affective computing. This study presents a systematic empirical investigation of how multimodal inputs affect both unimodal and multimodal emotion recognition performance. Using the UniC dataset, which provides modality-specific and global multimodal annotations across text, audio, and visual modalities, we conduct experiments based on the Tensor Fusion Network (TFN) under unimodal, bi-modal, and tri-modal configurations. Results show that cross-modal interactions exert complex and asymmetric effects. While additional modalities can provide complementary emotional cues, they may also introduce interference when signals diverge. Models continue to struggle with less frequent or extreme emotions such as *disgust*. Notably, multimodal embeddings combined with unimodal annotations outperform fully multimodal supervision in the same setup, highlighting the role of annotation consistency and cue reliability. These findings provide a systematic empirical validation of the long-assumed notions, demonstrating that cross-modal effects are not simply additive and highlighting the need for more interpretable multimodal fusion strategies.

**Keywords:** multimodal emotion recognition, modality-specific annotation, cross-modal interaction

## 1. Introduction

Human emotion is inherently multimodal (Pan et al., 2023; Junchi et al., 2025). In daily communication, emotional meaning is conveyed not only through verbal expressions but also through other channels, such as tone and facial movements (Wallbott and Scherer, 1986). The interaction of these modalities enables humans to perceive and interpret emotions more accurately and richly than from any single channel alone. Inspired by this, multimodal emotion recognition (MER) has become a central task in affective computing (Picard, 1997) and human-computer interaction (Zhang et al., 2024), aiming to jointly analyze cues from multiple single modalities, for example, text, audio, and vision, to model emotion more comprehensively (Shou et al., 2025; Du et al., 2025a).

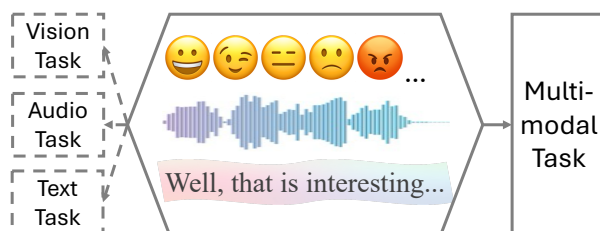


Figure 1: Interdependencies between modalities in emotion understanding

Over the past decades, extensive research has demonstrated that integrating multiple modalities improves emotion classification performance com-

pared with unimodal baselines (Busso et al., 2004; Metallinou et al., 2010; Ezzameli and Mahersia, 2023). The performance gain is generally attributed to cross-modal interactions, where information from one modality helps disambiguate or reinforce cues from another (Yu et al., 2020). For instance, as shown in Figure 1, the same sentence “Well, this is interesting...” can express a whole range of emotions depending on tone of voice or facial expression. Modeling such interdependencies is thus crucial for robust emotion understanding.

However, while cross-modal interaction is known to benefit multimodal emotion classification (Kumar and Vepa, 2020), a key theoretical question remains underexplored: Can cross-modal interactions also influence unimodal tasks, i.e. settings in which a single modality’s label is taken as the gold label? And how does this influence relate to the task performance of multimodal models and the way multimodal models represent and integrate information across modalities?

In other words, beyond improving fused outputs, does the integration of multiple modalities shape how individual modalities encode emotion? Addressing this question is crucial for understanding the representational dynamics between unimodal and multimodal processing. Human perception provides suggestive evidence that such influence exists – people’s interpretation of speech or facial emotion is often modulated by contextual knowledge derived from other modalities (Barrett et al., 2011; Wieser and Brosch, 2012). Investigating whether this phenomenon also emerges in com-



putational models can deepen our understanding of multimodal representation learning and affective reasoning.

To explore this issue, we propose a two-stage experimental framework that systematically investigates the interplay between unimodal and multimodal learning. In the first stage, we train a trimodal (Text-Audio-Video, TAV) model using multimodal annotations to capture integrated emotional representations. In the second stage, we conduct a series of controlled experiments with unimodal and bimodal configurations (T, A, V, TA, TV, AV), where each setting is trained on its corresponding unimodal annotations. For bimodal configurations such as TA, this means leveraging the annotations from both constituent unimodal modalities (i.e., T and A), respectively. This structure enables both vertical comparison between models trained and evaluated with unimodal and multimodal annotations, and horizontal comparison among unimodal, bimodal, and multimodal inputs.

Through this design, we aim to explore two fundamental questions:

- Do cross-modal interactions, introduced through the addition of other modalities, also influence unimodal tasks? This possibility is hypothesized, but experimental evidence to substantiate it is currently lacking.
- How are such influences related to the performance and interpretability of multimodal classification? This relationship remains insufficiently explored, and systematic empirical investigation is still lacking.

Our results reveal that cross-modal interactions indeed influence unimodal emotion prediction. In particular, we find that models trained on multimodal annotations learn richer, more balanced affective representations, shaping the final emotion representation, moving it beyond a simple addition of modalities. Moreover, unimodal tasks display distinct prediction tendencies that can be traced to patterns of cross-modal correlation observed during multimodal training. These findings suggest that the model's emotion recognition process operates along a continuum, where unimodal and multimodal processing are interdependent rather than isolated.

By analyzing these phenomena empirically, this work presents a dedicated investigation into how cross-modal interactions exert positive or negative influences on unimodal and multimodal emotion tasks. It thereby advances a deeper understanding of cross-modal dynamics and provides a novel interpretative perspective on the multimodal emotion research.

## 2. Related Work

### 2.1. Unimodal and Multimodal Emotion Recognition

Emotion recognition has long been a central topic in affective computing (Picard, 1997). Early approaches primarily relied on unimodal cues, such as text (Stajner and Klinger, 2023), speech (Slaughter et al., 2023), facial expressions (Leong et al., 2023), to infer affective states from a single information channel. However, human emotion expression is inherently multimodal (Pan et al., 2023; Junchi et al., 2025), where the same verbal content may convey drastically different emotions depending on tone or facial gestures. This realization has led to a paradigm shift toward MER, which seeks to integrate different signals from multiple modalities for more robust affective inference (Zhang et al., 2024; Hazmoune and Bougamouza, 2024).

Recent advances in deep learning have greatly enhanced MER performance by enabling the joint learning of hierarchical and correlated representations. Representative models include graph-based fusion networks (Li et al., 2023), transformer-based learning frameworks (Khan et al., 2025), and contrastive learning strategies (Xie et al., 2025). By employing graph-based relational modeling, self-attention mechanisms, and contrastive representation learning, these architectures effectively capture inter- and intra-modal dependencies, achieving notable gains over traditional baselines.

Despite this progress, most multimodal systems are designed primarily to maximize the overall predictive performance, rather than to investigate how modalities interact internally during learning.

### 2.2. Modeling Cross-Modal Interactions

Understanding cross-modal interaction mechanisms is crucial for interpreting multimodal learning (Fu et al., 2024). Researchers have explored various methods to model cross-modal interactions, either implicitly or explicitly. For instance, MulT (Tsai et al., 2019) employs directional pairwise cross-modal attention to capture interactions between multimodal sequences, while CENet (Wang et al., 2023) enhances text representations by integrating visual and acoustic information into a language model. Additionally, LDW-MTFN (Du et al., 2025a) introduces a label-based weighting scheme to explicitly balance the contributions of different modalities.

However, existing mechanisms are primarily designed to improve overall multimodal performance and are less informative for understanding naive cross-modal interactions. They tend to select only the most useful information from each modality by using mechanisms such as cross-modal attention



(Tsai et al., 2019) and gating (Sun et al., 2024), rather than considering all available information. Moreover, previous studies have examined how cross-modal interactions affect multimodal task outcomes, but not their impact on unimodal tasks in the field of emotion recognition, as illustrated in Figure 1. Therefore, to investigate authentic cross-modal interactions, it is necessary to fully integrate information from all modalities without any selective filtering, while also paying attention to how these interactions influence unimodal tasks.

### 2.3. Dataset-Level Advances and Remaining Gaps

One major bottleneck in examining cross-modal effects lies in the lack of datasets with modality-specific annotations. Most existing MER datasets, such as MSP-IMPROV (Busso et al., 2016) or CMU-MOSEI (Zadeh et al., 2018), only provide unified emotion labels derived from the overall multimodal impression. This design hinders a fine-grained comparison between unimodal and multimodal emotion detection performances, as each modality’s individual contribution cannot be independently assessed.

To address this limitation, Yu et al. (2020) introduced the Chinese dataset CH-SIMS, annotated at both unimodal and multimodal levels. Subsequently, Liu et al. (2022) expanded this resource by releasing CH-SIMS V2, adding more instances to the original dataset. Both versions of the dataset are annotated with sentiment labels. More recently, Du et al. (2025b) proposed another dataset, UniC, with both sentiment and categorical emotion labels in English. An notable difference between these two datasets is that CH-SIMS contains acted emotional expressions while the UniC dataset comprises authentic emotional expressions. Both datasets provide annotations for text, audio, and visual modalities.

Our study focuses exclusively on the UniC dataset, as it captures spontaneous and authentic emotional expressions, making it more suitable for robust emotion analysis. Leveraging its categorical emotion labels on both unimodal and multimodal level, we train and evaluate models under unimodal, bimodal, and trimodal configurations. This setup provides a rare opportunity to investigate how cross-modal interactions influence unimodal and multimodal emotion recognition, addressing an important gap in existing emotion research.

## 3. Method

### 3.1. Experimental Framework

To address the research question of whether cross-modal interactions can influence unimodal tasks

and how this relates to multimodal classification, we design a two-stage experimental framework. In the first stage, a trimodal configuration of text, audio and video (TAV) is trained and evaluated using multimodal annotations to capture integrated emotional representations. In the second stage, experiments are conducted with unimodal or bimodal inputs (T, A, V, TA, TV, AV), where each setting is trained with the corresponding unimodal annotations. This design enables a systematic comparison between multimodal and unimodal learning, providing insights into how cross-modal information contributes to unimodal emotion prediction and overall multimodal understanding.

### 3.2. Dataset

We conduct our experiments on the UniC dataset (Du et al., 2025b), which contains 965 video clips featuring genuine emotional expressions. UniC was annotated by three trained college students from Ghent University, who served as experts, with inter-annotator agreement used to ensure annotation quality. Unlike acted emotion datasets such as CH-SIMS (Yu et al., 2020), UniC captures authentic emotions, making recognition both more challenging and more representative of real-world scenarios. Importantly, besides an overall emotion annotation, UniC also provides independent categorical emotion labels for three single modalities: text, audio, and (silent) video. To the best of our knowledge, it is the only available dataset that offers modality-specific categorical annotations for emotions, thus enabling a systematic study of unimodal baselines, cross-modal interactions, and label-specific emotion classification tasks.

UniC contains both dimensional and categorical emotion labels. For the former, valence and arousal are evaluated as an integer from 1 to 5, while for the latter, a set of seven categorical emotion labels are used, including *disgust*, *disappointment*, *confusion*, *neutral*, *surprise*, *contentment* and *joy*.

The UniC dataset releases only pre-extracted feature packages rather than the raw video data, primarily due to privacy and data protection considerations. Specifically, the released features are extracted using multilingual BERT-base model (Devlin et al., 2019) for text, openSMILE (Eyben et al., 2010) for audio, and MediaPipe (Lugaresi et al., 2019) (25 FPS) for the visual modality.

### 3.3. Model

To investigate cross-modal interactions, it is essential to select a model that can faithfully capture the contributions of each modality. Many existing multimodal models, such as MuT (Tsai et al., 2019), CENet (Wang et al., 2023), and LDW-MTFN (Du

et al., 2025a), incorporate mechanisms like gating, dynamic weighting, or cross-modal attention. While these mechanisms often lead to performance improvements, they complicate the understanding. The observed improvements may primarily reflect architectural biases rather than the genuine effect of modality interactions, as such mechanisms selectively emphasize certain inputs and suppress others, potentially masking the natural influence of unimodal signals.

In contrast, the Tensor Fusion Network (TFN) (Zadeh et al., 2017) provides a static and relatively exhaustive representation of cross-modal interactions, without relying on any of the aforementioned architectural changes. Formally, let  $x_T$ ,  $x_A$ , and  $x_V$  denote the embeddings of text, audio, and video modalities, respectively. The tensor fusion representation is computed as

$$\mathcal{T} = \begin{bmatrix} x_T \\ 1 \end{bmatrix} \otimes \begin{bmatrix} x_A \\ 1 \end{bmatrix} \otimes \begin{bmatrix} x_V \\ 1 \end{bmatrix} \quad (1)$$

where  $\otimes$  denotes the outer product. The resulting tensor  $\mathcal{T}$  is then flattened and passed through fully connected layers for classification. This approach preserves all information from individual modalities while capturing deep pairwise and higher-order interactions, unlike simple concatenation, which does not fully exploit the information content. Therefore, TFN is particularly suitable for studying the genuine contributions of added modalities, as any observed performance changes can be more directly attributed to the modalities themselves rather than to architectural artifacts.

Although multi-task models such as MTFN have shown stronger performance in prior studies, we do not adopt a multi-task approach in this work, as it could introduce confounding factors that obscure the direct effects of each modality.

Overall, TFN provides a clear and controlled framework for studying intrinsic cross-modal interactions, making it a natural choice for our analysis.

### 3.4. Implementation Details

We define three input modalities: text (T), represented by textual transcript embeddings; audio (A), extracted from speech acoustic features excluding transcripts; and video (V), extracted from silent video frames. To investigate cross-modality interaction effects, we consider all possible modality combinations: unimodal (T, A, V), bimodal (TA, TV, AV), and trimodal (TAV). Each configuration serves as input to the model. Independent unimodal and multimodal annotations in UniC enable the model to be trained and evaluated for different tasks, facilitating a systematic analysis of cross-modal influences. The UniC dataset is partitioned into training, validation, and test subsets following a 6:2:2 ratio.

All models are trained using cross-entropy loss for seven categorical emotion labels, with a batch size of 32 and a learning rate of 0.0005. The experiments are carried out on NVIDIA Tesla V100-SXM2-16GB GPUs, and repeated with three random seeds to ensure robustness.

Performance is evaluated using accuracy and weighted F1-score to account for class imbalance, allowing us to establish robust unimodal baselines and quantify both positive and negative effects of cross-modal interactions.

## 4. General Results

### 4.1. Unimodal-label Settings

Table 1 summarizes the performance of TFN under different input modalities, evaluated against four distinct annotation sources (text-, audio-, and vision-based and multimodal labels).

First of all, it is clear that for unimodal emotion recognition tasks – that is, when using unimodal labels (text, audio and vision labels, respectively) – the model achieves the best performance when the three modalities are combined. This suggests that multimodal information fusion can provide complementary cues that enhances emotion recognition in single-modality tasks. Among the three unimodal-label settings, the model performs best with text labels, reaching an accuracy of 34.90% and an F1-score of 28.09%, followed by the vision setup (Acc=31.24%, F1 = 25.98%) and the audio setup (Acc=27.23%, F1=24.00%). These results indicate that when the three modalities are combined, the model finds it easier to learn text-related emotional patterns than those in audio or visual modalities.

When using text annotations, the unimodal text baseline (T) achieves a higher accuracy (33.86%) than both the text-audio (TA, acc=29.67%) and text-vision (TV, acc=32.46%) settings. However, it shows the lowest F1-score (19.13%), suggesting that while the text-only model can capture the majority class more frequently, it struggles to identify minority emotional categories. This suggests that the inclusion of additional modalities introduces beneficial complementary information that improves classification balance and robustness, even if overall accuracy does not always increase. A further investigation reveals that the unimodal text baseline predicted all instances as *neutral*, as shown in Figure 2 in Section 5, highlighting the limitation of the text-only setup. An inspection of the textual annotations in UniC shows that *neutral* is indeed the most frequent label, accounting for approximately one third of all instances, which may have reinforced this prediction bias in the text-only setup.

In contrast, the pattern is reversed when using audio annotations. The unimodal audio

Input	Text Label		Audio Label		Video Label		Multimodal	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
TAV	<b>34.90</b>	<b>28.09</b>	<b>27.23</b>	<b>24.00</b>	<b>31.24</b>	<b>25.98</b>	28.45	21.17
TA	29.67	20.30	23.91	18.88	—	—	27.22	19.57
TV	32.46	25.74	—	—	27.92	21.35	28.27	21.84
AV	—	—	23.21	20.78	31.24	25.36	27.57	18.12
T	33.86	19.13	—	—	—	—	28.10	16.99
A	—	—	25.83	22.47	—	—	<b>30.72</b>	<b>24.65</b>
V	—	—	—	—	28.80	22.70	28.80	18.77

Table 1: Performance of TFN when using text, audio, and video labels, respectively. *Input* refers to the combination of modality embeddings. Results are averaged from three experiments with different random seeds.

baseline (A) outperforms the multimodal settings, achieving both higher accuracy (25.83%) and F1-score (22.47%) than text-audio (TA, acc=23.91%, F1=18.88%) and audio-vision (AV, acc=23.21%, F1=20.78%). This indicates that the addition of text or visual features introduces interference or conflicting information, reducing the model’s ability to align with the emotion expressed in the audio modality. On the other hand, we also found that during training, the inclusion of additional modalities also leads to substantially earlier convergence, indicating that they provide auxiliary information that helps the model learn more efficiently, even if the final performance does not always improve.

The situation becomes more complex when the model uses vision annotations. The unimodal vision baseline (V, Acc = 28.80%, F1 = 22.70%) performs better than the text-vision model (TV, Acc = 27.92%, F1 = 21.35%), but worse than the audio-vision model (AV, Acc = 31.24%, F1 = 25.36%). This pattern suggests that adding the audio modality provides positive complementary information, while adding text may have a negative or distracting influence on visual emotion recognition.

Taken together, across unimodal, bimodal, and trimodal setups for unimodal-label tasks, the results indicate that adding one additional modality can have either positive or negative effects, depending on the degree of alignment between modalities and label sources. However, when two additional modalities are combined, the overall effect tends to be positive, suggesting that the benefits of multimodal complementarity outweigh the potential interference among modalities.

## 4.2. Multimodal-label Settings

When the model is trained and evaluated with multimodal labels, however, the results exhibit a distinct pattern compared with the unimodal-label settings. As presented in Table 1, the model attains the highest accuracy (30.72%) and F1-score (24.65%) when using the audio modality alone, outperforming other unimodal settings and all multimodal com-

binations. A possible explanation for this result could be that the audio modality carries strong emotional cues that are most consistent with the integrated multimodal annotations, reflecting its critical role in conveying affective information such as tone, rhythm, and prosody.

In contrast, the performance of the full three-modality fusion (TAV) does not exceed that of the audio-only settings, although it still surpasses the video-only and text-only configurations in terms of F1-score (21.17% > 18.77% > 16.99%). This indicates that combining the three modalities provides certain complementary benefits, but may also introduce redundant or conflicting information.

## 4.3. Cross-setting Interpretation

When examining the first line of Table 1, it is evident that when text, audio, and vision are used together as inputs, the model performs better when trained and evaluated with unimodal annotations than with multimodal annotations. This observation suggests two possible explanations. First, some emotional cues emphasized in each unimodal annotation may not be fully consistent with the holistic perception reflected in the multimodal labels. Second, multimodal annotations, while more comprehensive, may introduce additional uncertainty or ambiguity, making it harder for the model to align integrated features with a unified emotional target.

It is also observed that the model achieves its best performance when using the trimodal input but being trained and evaluated with textual annotations, suggesting a clear textual predominance. This finding is consistent with previous studies, which have similarly reported that textual features tend to dominate in multimodal emotion recognition tasks (Liu et al., 2022).

Comparing the two experiment schemes, it becomes evident that the relationship between modality and annotation source critically determines model performance. When using unimodal annotations, multimodal fusion is beneficial because each label corresponds to its own modality, and

the integrated features complement one another. When using multimodal annotations, however, the target label represents an overall emotional judgment, which may not align equally across modalities. Consequently, modalities like audio, which might naturally reflect more global affective impressions, could obtain better performance than the other two single modalities. However, examining the bimodal setups reveals that combinations including text, specifically TA (text + audio) and TV (text + vision), still outperform AV (audio + vision). This suggests that textual predominance remains.

## 5. Qualitative Analysis

Figure 2 illustrates the confusion matrices across all modality configurations. The first three rows correspond to the different unimodal annotations used, namely, text-, audio-, and vision-based annotations, and the columns represent the gradual inclusion of modalities (single → bimodal → trimodal). The fourth row represents the single modal inputs, while the fifth refers to the bimodal and trimodal inputs. The last two rows both refer to the results of models using multimodal annotations. The seven emotion categories (disgust, disappointment, confusion, neutral, surprise, contentment and joy) are indexed as 0-6.

The first row in Figure 2 presents the confusion matrices for the text-based classification task under different modality settings (T, TA, TV, and TAV). In the purely textual condition (T-T), the model exclusively predicts *neutral* (class 3) regardless of the ground-truth label, indicating a strong bias toward the dominant class and a lack of discriminative capacity when relying solely on textual cues. When additional modalities are introduced, this bias is gradually mitigated. In the T-TA and T-TV settings, although *neutral* remains the most frequent prediction, the model begins to recognize a small number of other categories (e.g., *disgust*, *disappointment*, and *contentment*, or classes 0, 1, 5), suggesting that audio and visual information help the model capture subtle emotional variations that are not easily inferred from text alone. The T-TAV configuration further enhances this diversification: predictions are more evenly distributed across classes, and off-diagonal entries (especially toward *neutral*, *contentment* and *joy*, or classes 3, 5 and 6) increase, reflecting improved sensitivity to inter-class distinctions. Nevertheless, confusion among neighboring emotional categories persists, for example, *contentment* and *joy* (class 5 and 6), implying that while multimodal cues alleviate the single-modality bias, cross-modal interactions remain imperfect. Notably, the emotion *joy* is predicted only when multimodal information is available.

Compared with the text-based experiments, the

audio-based experiments show predictions that are more evenly distributed across all emotion categories except *confusion* and *surprise*, and remain more stable across different modality settings. The range and number of correctly predicted emotion categories do not fluctuate as drastically as in the text-based results, suggesting that audio-based emotion recognition is relatively more stable and less sensitive to cross-modal variations. However, it is also noticed that there is no *disgust* emotion prediction in the TA setup, suggesting that the addition of textual information may have biased the model toward more neutral or less extreme affective interpretations, thereby suppressing the recognition of *disgust* with distinctive acoustic patterns.

As for the vision-based experiments, the results show different patterns from the text- and audio-based setups. The unimodal setting (V-V) shows a roughly balanced number of correct predictions between *disappointment* (22) and *contentment* (21). When textual information is added (V-TV), the model becomes more sensitive to negative emotions, as *disappointment* increases to 29 while *contentment* drops to 14. In contrast, adding audio (V-AV) strengthens positive recognition: *contentment* rises to 26 and *joy* also improves notably from 7 to 16, whereas *disappointment* decreases to 19. In the full multimodal condition (V-TAV), the results appear to balance these trends, with *disappointment* and *contentment* reaching 25 and 19 respectively, suggesting that the joint contribution of text and audio helps the model achieve a more balanced perception between negative and positive emotions.

In the multimodal label experiments, the three unimodal settings (M-T, M-A, and M-V) show a clear tendency toward predicting more positive emotions overall. Among them, *contentment* is consistently the most correctly recognized category, exceeding *disappointment* in all cases. This suggests that when trained with multimodal annotations, the unimodal models tend to align more with the positive affective patterns emphasized in the multimodal ground truth. Compared to text and vision, the audio-based model exhibits a more balanced distribution across positive and negative categories and is also the only one that successfully predicts *joy*, highlighting the unique contribution of acoustic cues to the recognition of high-arousal positive emotions.

For the multimodal-label experiments involving two- and three-modality combinations, several patterns emerge. Bi-modal setups that include audio (M-TA, M-AV) inherit the audio modality’s ability to predict *joy*, whereas the text+vision combination (M-TV) predicts two *joy* instances, but both are incorrect, indicating that audio is critical for high-arousal positive emotion recognition. The tri-modal



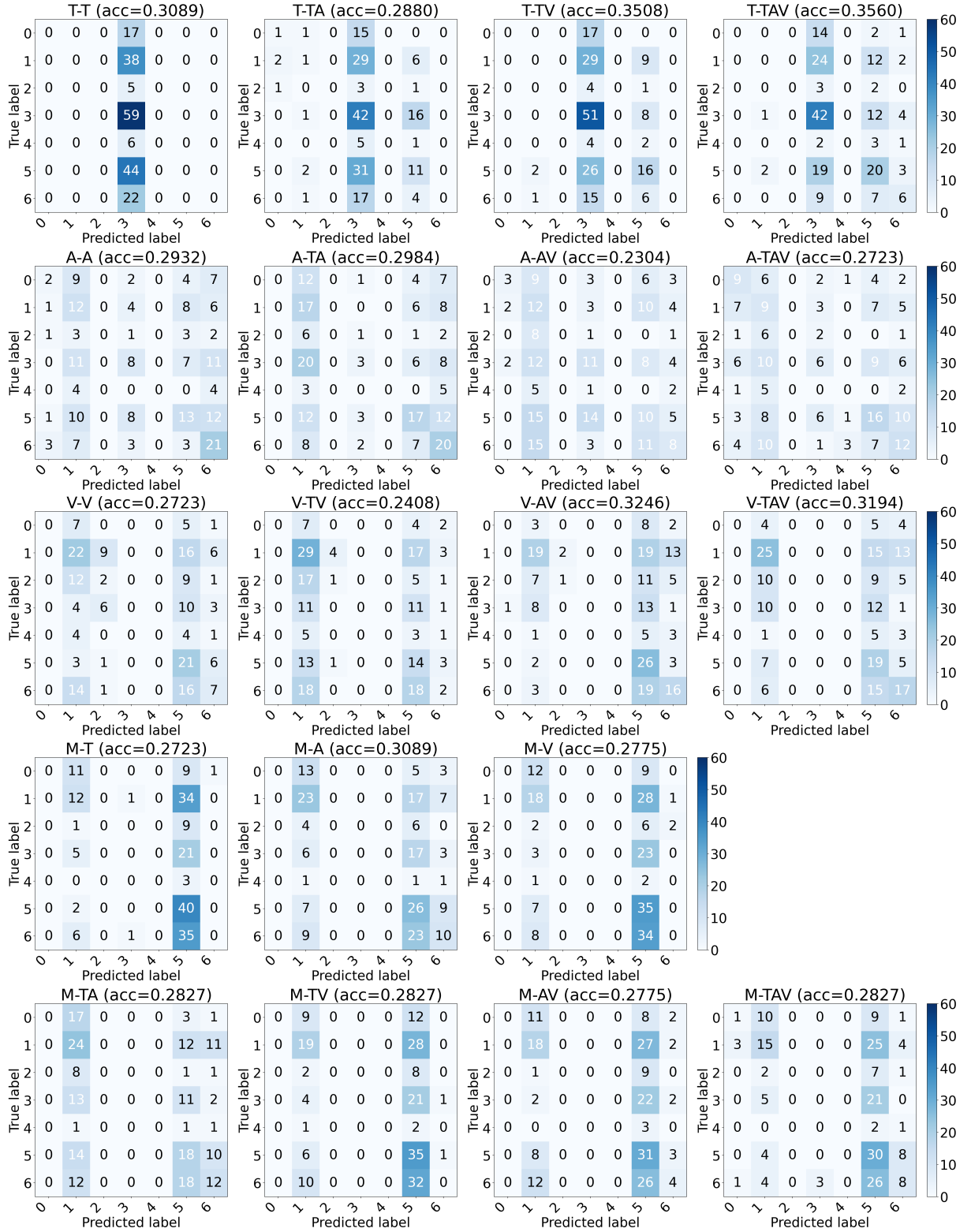


Figure 2: Confusion matrices of true and predicted labels, when the model is trained and evaluated using different annotation types and various modality combinations. T, A, V, and M refer to text, audio, vision, and multimodal setups, respectively. Labels 0-6 correspond to seven emotions: *disgust*, *disappointment*, *confusion*, *neutral*, *surprise*, *contentment*, and *joy*.



configuration (M-TAV) shows similar capabilities for *joy* as the bi-modal audio setups and is also the only setting among all multimodal-label experiments to produce predictions for *disgust*, with one correct out of five predicted instances. Additionally, the M-TAV results exhibit more frequent confusion between emotions of the same overall sentiment, particularly between *contentment* and *joy*, suggesting that while integrating three modalities balances the strengths of individual channels, it also introduces subtle intra-sentiment ambiguity in the model's predictions.

## 6. Discussion and Conclusion

This study, to the best of our knowledge, provides a systematic empirical analysis that examines how the presence of the different modalities influences emotion recognition when trained and evaluated under both unimodal and multimodal supervision.

### 6.1. Cross-Modal Complementarity and Interference

Our results reveal that cross-modal interactions exert a complex impact on both unimodal and multimodal emotion recognition tasks. Rather than simply enhancing recognition accuracy through information aggregation, these interactions entail a trade-off between complementary and conflicting emotional cues. When using unimodal annotations, tri-modal configurations generally outperform unimodal and bi-modal ones, suggesting that richer multimodal representations can provide complementary evidence for emotion inference. However, the addition of a single auxiliary modality does not consistently yield improvements. This indicates that adding another modality can sometimes provide complementary information, but may also introduce interference or misalignment between modalities when their emotional signals diverge.

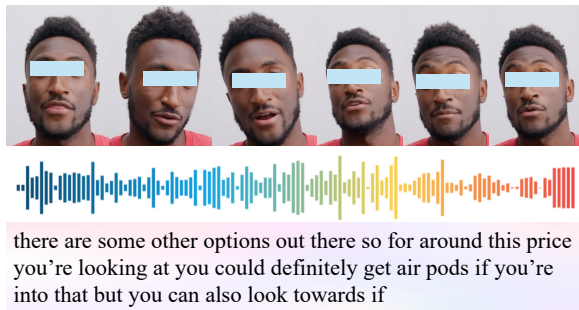


Figure 3: An instance from the UniC test dataset, labelled as *neutral*, *contentment*, *disgust*, and *neutral* in the text, audio, silent video, and multimodal setups.

In practice, multimodal data often contain heterogeneous and sometimes conflicting emotional expressions, such as discrepancies between facial expressions, vocal tone, and textual content, as shown in Figure 3. This inherent variability means that cross-modal interactions can amplify or diminish the contribution of each modality depending on the context, making the impact of adding modalities highly situation-dependent. Understanding and modeling this nuanced balance is therefore crucial for robust multimodal emotion recognition.

### 6.2. Supervision Type and Annotation Reliability

Interestingly, models trained and evaluated with unimodal annotations tend to outperform those relying on multimodal annotations, revealing a counterintuitive phenomenon in multimodal emotion recognition. A likely explanation lies in the annotation reliability: unimodal annotations, which focus on a single information channel, tend to be more internally consistent, whereas multimodal annotations often suffer from lower inter-annotator agreement (Du et al., 2023). This may occur because multimodal judgements may be influenced by differing weights assigned to the visual, acoustic, or linguistic cues. As a result, multimodal labels often introduce greater uncertainty. These observations suggest that combining multimodal input embeddings with unimodal annotations can sometimes yield more stable and effective results than relying solely on fully multimodal annotations.

### 6.3. Methodological Implications for Fusion Strategies

The present study also highlights the methodological implications of using Tensor Fusion Networks, or TFNs, which treat all modalities equally. While this approach can capture complementary information, it is suboptimal when large discrepancies exist across modality-specific annotations. Prior research has shown that weighing modality embeddings appropriately can improve performance with multimodal annotations (Du et al., 2025a). It remains, however, an open question whether a similar strategy applied to weighted embeddings with unimodal annotations would further enhance performance, and how it would compare to existing configurations. Exploring this possibility could provide valuable insights into optimizing multimodal emotion recognition under varying levels of annotation reliability.

### 6.4. Main Contributions

This paper makes three main contributions. First, it establishes a unified experimental framework for

directly comparing unimodal and multimodal supervision under identical modality configurations. Second, it clarifies that cross-modal interactions involve both complementarity and interference rather than uniformly improving performance. Third, it uncovers a counterintuitive finding that multimodal inputs paired with unimodal annotations can sometimes outperform fully multimodal supervision. Together, these results provide new empirical insights into cross-modal dynamics and inform the design of multimodal emotion recognition systems.

## 7. Limitations

Since the aim of this paper is to investigate cross-modal interactions, all experiments share the same hyperparameters as the multimodal setup (M-TAV) to ensure a fair comparison across different configurations. This design choice, however, may lead to suboptimal performance in single-modality setting, for example, the text-only baseline predicted all instances as *neutral*, as shown in Figure 2.

This study uses only a single dataset, as, to the best of our knowledge, it is the only publicly available resource containing both independent unimodal and multimodal categorical emotion labels suitable for the proposed experimental design.

Although the research question itself is not entirely novel, this work represents a systematic empirical study to validate a widely assumed hypothesis regarding cross-modal influences on unimodal emotion recognition.

The Tensor Fusion Network (TFN) was chosen because it is sufficiently expressive and efficient for exploring multimodal emotion interactions, while avoiding mechanisms such as attention that explicitly promote or suppress information. This ensures that all modality information is preserved in the model. More recent advanced models, which incorporate selective fusion mechanisms, may not be suitable for this purpose.

Large language models are not employed in this study, primarily because the audio and visual data are private and raise ethical concerns regarding privacy and consent. Only pre-extracted feature packages are available and used for this study.

## 8. Acknowledgements

This research received funding from the Flemish Government under the Flanders Artificial Intelligence Research program (174P07826). We would also like to thank the anonymous reviewers for their valuable and constructive feedback.

## 9. Bibliographical References

- Lisa Feldman Barrett, Batja Mesquita, and Maria Gendron. 2011. [Context in emotion perception](#). *Current directions in psychological science*, 20(5):286–290.
- Carlos Busso, Zhigang Deng, Serdar Yildirim, Murat Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. 2004. [Analysis of emotion recognition using facial expressions, speech and multimodal information](#). In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 205–211.
- Carlos Busso, Srinivas Parthasarathy, Alec Burman, Mohammed AbdelWahab, Najmeh Sadoughi, and Emily Mower Provost. 2016. [Msp-improv: An acted corpus of dyadic interactions to study emotion perception](#). *IEEE Transactions on Affective Computing*, 8(1):67–80.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Quanqi Du, Loic De Langhe, Els Lefever, and Veronique Hoste. 2025a. [Ldw : Label divergence weighting for multimodal sentiment analysis](#). In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*, pages 1–10. Association for Computing Machinery.
- Quanqi Du, Sofie Labat, Thomas Demeester, and Veronique Hoste. 2023. [Unimodalities count as perspectives in multimodal emotion annotation](#). In *2nd Workshop on Perspectivist Approaches to NLP (NLPerspectives 2023)*, co-located with the 26th European Conference on Artificial Intelligence (ECAI 2023), volume 3494. CEUR-WS.org.
- Quanqi Du, Sofie Labat, Thomas Demeester, and Veronique Hoste. 2025b. [UniC: A dataset for emotion analysis of videos with multimodal and unimodal labels](#). *Language Resources and Evaluation*, 59:2857–2892.
- Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. [Opensmile: The munich versatile and fast open-source audio feature extractor](#). In *Proceedings of the 18th ACM International Conference on Multimedia, MM '10*, page 1459–1462, New York, NY, USA. Association for Computing Machinery.

- Kaouther Ezzameli and Hela Mahersia. 2023. [Emotion recognition from unimodal to multimodal analysis: A review](#). *Information Fusion*, 99:101847.
- Yanping Fu, Zhiyuan Zhang, Ruidi Yang, and Cuiyou Yao. 2024. [Hybrid cross-modal interaction learning for multimodal sentiment analysis](#). *Neurocomputing*, 571:127201.
- Samira Hazmoune and Fateh Bougamouza. 2024. [Using transformers for multimodal emotion recognition: Taxonomies and state of the art review](#). *Engineering Applications of Artificial Intelligence*, 133:108339.
- Ma Junchi, Hassan Nazeer Chaudhry, Farzana Kulsoom, Yang Guihua, Sajid Ullah Khan, Sujit Biswas, Zahid Ullah Khan, and Faheem Khan. 2025. [Multicauset temporal attention for multimodal emotion cause pair extraction](#). *Scientific Reports*, 15(1):19372.
- Mustaqeem Khan, Phuong-Nam Tran, Nhat Truong Pham, Abdulmotaleb El Saddik, and Alice Othmani. 2025. [Memocmt: multimodal emotion recognition using cross-modal transformer-based feature fusion](#). *Scientific reports*, 15(1):5473.
- Ayush Kumar and Jithendra Vepa. 2020. [Gated mechanism for attention based multi modal sentiment analysis](#). In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4477–4481. IEEE.
- Sze Chit Leong, Yuk Ming Tang, Chung Hin Lai, and CKM Lee. 2023. [Facial expression and body gesture emotion recognition: A systematic review on the use of visual data in affective computing](#). *Computer science review*, 48:100545.
- Dongyuan Li, Yusong Wang, Kotaro Funakoshi, and Manabu Okumura. 2023. [Joyful: Joint modality fusion and graph contrastive learning for multimodal emotion recognition](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16051–16069, Singapore. Association for Computational Linguistics.
- Yihe Liu, Ziqi Yuan, Huisheng Mao, Zhiyun Liang, Wanqiuyue Yang, Yuanzhe Qiu, Tie Cheng, Xiaoteng Li, Hua Xu, and Kai Gao. 2022. [Make Acoustic and Visual Cues Matter: CH-SIMS v2.0 Dataset and AV-Mixup Consistent Module](#). In *Proceedings of the 2022 International Conference on Multimodal Interaction, ICMi '22*, page 247–258, New York, NY, USA. Association for Computing Machinery.
- Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. 2019. [Mediapipe: A framework for building perception pipelines](#).
- Angeliki Metallinou, Sungbok Lee, and Shrikanth Narayanan. 2010. [Decision level combination of multiple modalities for recognition and analysis of emotional expression](#). In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2462–2465. IEEE.
- Bei Pan, Kaoru Hirota, Zhiyang Jia, and Yaping Dai. 2023. [A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods](#). *Neurocomputing*, 561:126866.
- Rosalind W. Picard. 1997. [Affective Computing](#). MIT press, Cambridge.
- Yuntao Shou, Tao Meng, Wei Ai, and Keqin Li. 2025. [Dynamic graph neural ODE network for multimodal emotion recognition in conversation](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 256–268, Abu Dhabi, UAE. Association for Computational Linguistics.
- Isaac Slaughter, Craig Greenberg, Reva Schwartz, and Aylin Caliskan. 2023. [Pre-trained speech processing models contain human-like biases that propagate to speech emotion recognition](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8967–8989, Singapore. Association for Computational Linguistics.
- Sanja Stajner and Roman Klinger. 2023. [Emotion analysis from texts](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 7–12, Dubrovnik, Croatia. Association for Computational Linguistics.
- Xin Sun, Xiangyu Ren, and Xiaohao Xie. 2024. [A novel multimodal sentiment analysis model based on gated fusion and multi-task learning](#). In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8336–8340.
- Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. [Multimodal transformer for unaligned multimodal language sequences](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

- Harald G Wallbott and Klaus R Scherer. 1986. [Cues and channels in emotion recognition](#). *Journal of personality and social psychology*, 51(4):690.
- Di Wang, Shuai Liu, Quan Wang, Yumin Tian, Lihuo He, and Xinbo Gao. 2023. [Cross-modal enhancement network for multimodal sentiment analysis](#). *IEEE Transactions on Multimedia*, 25:4909–4921.
- Matthias J Wieser and Tobias Brosch. 2012. [Faces in context: A review and systematization of contextual influences on affective face processing](#). *Frontiers in psychology*, 3:471.
- Yunhe Xie, Chengjie Sun, Ziyi Cao, Bingquan Liu, Zhenzhou Ji, Yuanchao Liu, and Lili Shan. 2025. [A dual contrastive learning framework for enhanced multimodal conversational emotion recognition](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4055–4065, Abu Dhabi, UAE. Association for Computational Linguistics.
- Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020. [CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, Online. Association for Computational Linguistics.
- Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. [Tensor fusion network for multimodal sentiment analysis](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.
- Amir Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. [Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.
- Shiqing Zhang, Yijiao Yang, Chen Chen, Xingnan Zhang, Qingming Leng, and Xiaoming Zhao. 2024. [Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects](#). *Expert Systems with Applications*, 237:121692.



# Annotation Matters: Resolving Cross-Corpus Performance Drops in Hebrew Offensive Language Detection

Gili Berger Hefetz<sup>1</sup>, Yossef Haim Shrem<sup>1</sup>, Natalia Vanetik<sup>2</sup>, Chaya Liebeskind<sup>1</sup>

<sup>1</sup>Department of Computer Science, Jerusalem College of Technology, Jerusalem, Israel

<sup>2</sup>Department of Computer Science, Shamoon College of Engineering, Beer-Sheva, Israel

giliberhef@gmail.com, yosshrem@gmail.com, natalyav@ac.sce.ac.il, liebchaya@gmail.com

## Abstract

Cross-dataset generalization remains a major challenge in offensive language detection, especially for culturally sensitive languages such as Hebrew. A large Hebrew dataset introduced in prior work, was annotated via a taxonomy-grounded, prompt-guided LLM protocol and achieved strong in-domain results. However, performance degraded sharply on two external Hebrew corpora. We investigate whether this degradation reflects domain shift or annotation shift, i.e., differences in how offensiveness is operationalized across datasets. Using the same prompt framework and a dual-LLM agreement procedure, we re-annotate both external corpora and quantify label divergence. We observe substantial mismatch between the original and new annotations, consistent with the view that offensiveness is not objective but depends on cultural context, discourse conventions, political framing, and the interpretation of irony. Evaluating models against the new labels yields markedly improved performance, and fine-tuning with the new external labels further improves results. Overall, our findings suggest that cross-dataset failure in affective NLP tasks may often be driven by annotation mismatch rather than domain adaptation limitations, highlighting the importance of annotation validity and culturally grounded labeling protocols.

**Keywords:** offensive language detection, annotation, cross-dataset generalization

## 1. Introduction

Offensive language detection is typically framed as supervised classification: given a text, predict whether it is offensive. However, offensiveness is not an objective property of text. It depends on cultural norms, discourse conventions, political context, and pragmatic interpretation (irony and sarcasm). This is especially salient in Hebrew, where political discourse, slang, code-switching, and implicit references are common. A large Hebrew dataset introduced in prior work (Berger Hefetz et al., 2026) contains 19K instances of messages from the Rotter online forum <https://rotter.net/>, annotated via a prompt-guided protocol (detailed in Appendix A). To produce these annotations, a dual-LLM agreement filtering with fine-tuned transformer models was applied, which produced strong in-domain performance. Yet, when the same models were evaluated on two external Hebrew corpora, Litvak et al. (2022) and Hamad et al. (2023), performance dropped sharply. These results (Berger Hefetz et al., 2026) could reflect domain shift (platform/topic/style), but they may also indicate annotation shift, where datasets operationalize offensiveness differently, particularly for political language and borderline cases. We disentangle these explanations by re-annotating both external datasets using the same prompt framework and a dual-LLM agreement protocol. We evaluate several Hebrew-language models under the original versus corrected labels and evaluate them on

corrected data, with and without fine-tuning. Our results show that much of the cross-dataset degradation stems from inconsistent labeling rather than limited generalization.

Our contributions are (1) systematic re-annotation of two Hebrew corpora (Litvak et al., 2022; Hamad et al., 2023) using a taxonomy-grounded prompt (Liebeskind et al., 2023); (2) quantification of label disagreement and analysis of culturally driven ambiguity; (3) evidence that corrected labels restore cross-dataset performance and improve fine-tuning; and (4) a methodological shift in interpreting cross-dataset failure: we demonstrate that performance degradation is primarily driven by human annotation subjectivity and labeling mismatch rather than topicality or domain shift. By disentangling these factors, we show that aligning the conceptual operationalization of offensiveness is a prerequisite for reliable cross-domain evaluation.

## 2. Related Work

Hebrew offensive language detection has been constrained by scarce annotated resources and linguistic complexity. Early work relied on small datasets and lexicon-based approaches, including abusive comment detection in Hebrew-language groups on Facebook (Liebeskind and Liebeskind, 2018). Later studies explored cross-lingual transfer (Litvak et al., 2022), while Hamad et al. (2023) introduced a Hebrew offensive corpus and BERT-based de-



tection. Despite these efforts, annotation practices vary substantially across corpora. This is especially consequential because offensiveness is not purely lexical: it depends on intent, target, and pragmatic interpretation and is often expressed indirectly via political framing, sarcasm, or implicit.

Taxonomies capture the multi-dimensional nature of abusive language and support more consistent annotation for culturally sensitive affective categories. [Liebeskind et al. \(2024\)](#) propose a detailed framework covering targets, vulgarity, severity, and discriminatory aspects. LLM-based prompting offers scalability in low-resource settings but requires structured guidelines. We use a taxonomy-grounded prompt distilled into four indicators: intent, target, tone, and impact in order to guide interpretable binary labeling.

Cross-dataset performance drops are commonly attributed to domain shift (platform, topic, style), but they may also reflect annotation shift. Recent literature has increasingly recognized that such degradation is often a conceptual issue rather than a purely technical one; for instance, [Fortuna and Nunes \(2018\)](#) highlight the lack of unified taxonomies as a major barrier to generalization, while [Talat \(2016\)](#) demonstrates how the subjective background of annotators inherently shapes the decision boundaries of toxicity classifiers. Datasets can represent offensiveness differently due to cultural assumptions, political context, and discourse norms, corresponding to variation in how affective meaning is constructed and categorized. We hypothesize that Hebrew offensive language detection is a clear case where annotation shift substantially contributes to cross-dataset degradation.

### 3. Datasets

We analyze two external Hebrew corpora commonly used for offensive language detection (Table 1): the Facebook dataset ([Litvak et al., 2022](#)) and the Twitter corpus of [Hamad et al. \(2023\)](#), denoted as F (Facebook) and T (Twitter), respectively. **The Facebook dataset (F)** consists of 5,217 annotated Hebrew comments, combining the OLaH corpus of [Litvak et al. \(2021\)](#) with the dataset of [Liebeskind and Liebeskind \(2018\)](#). Specifically, it includes 2,000 annotated comments from OLaH, 1,489 annotated comments from the Liebeskind collection (after replacing unknown labels and removing non-Hebrew entries), and an additional 1,939 comments from the same source that were manually annotated by three native Hebrew speakers in the original work, with a third annotator resolving disagreements. In total, the dataset contains 5,217 labeled comments, with an inter-annotator agreement of  $\kappa = 0.82$ . **The Twitter corpus (T)** comprises 15,881 manually annotated

Hebrew tweets, labeled across five fine-grained categories (abusive, hate, violence, pornographic, non-offensive) by bilingual annotators (Hebrew–Arabic speakers). The original publication does not report inter-annotator agreement metrics. For the purposes of our study, we unified all offensive categories (abusive, hate, violence, and pornographic) into a single *offensive* label, keeping non-offensive unchanged. These corpora differ from in-domain Rotter data in platform, discourse style, and sociopolitical context. Crucially, they also differ in annotation guidelines and in how borderline cases (e.g., profanity without a target, sarcasm, political slogans, and quoted speech) are handled, making them suitable for testing annotation shift.

Dataset	Source	Size
F ( <a href="#">Litvak et al., 2022</a> )	Facebook	5,217
T ( <a href="#">Hamad et al., 2023</a> )	Twitter	15,881

Table 1: External dataset sources and sizes.

## 4. Re-Annotation Method

**Taxonomy-Grounded Prompt:** We re-annotate both external datasets using the same prompt framework introduced in our prior work ([Berger Hefetz et al., 2026](#)), derived from the abusive-language taxonomy of [Liebeskind et al. \(2024\)](#). The prompt operationalizes offensiveness through four interpretable dimensions: *intent* (whether the text aims to insult or demean), *target* (whether an individual or group is identifiable), *tone* (presence of slurs, vulgarity, or dehumanization), and *impact* (whether the utterance would be perceived as harmful in context). This structure reflects a core affective-science insight: social and emotional meaning is not solely encoded in lexical content but emerges from context, intention, and perceived interpersonal impact.

**Dual-LLM Annotation and Filtering:** The re-annotation follows the *Dual-LLM Annotation and Filtering* strategy validated in prior work ([Berger Hefetz et al., 2026](#)). Each instance is annotated independently by two LLMs: GPT-4o-mini (version: gpt-4o-mini-2024-07-18) and Gemini-1.5-Flash (version: gemini-1.5-flash-001). The full system prompt used for classification is provided in Appendix A. Both models provide a binary label (*Offensive/Non-Offensive*) and a rationale based on Chain-of-Thought reasoning. To ensure a high-confidence gold standard, we retain only instances where both models agree. This specific pipeline was previously benchmarked against a manually annotated set of 1,500 Hebrew comments by two native experts ([Berger Hefetz et al., 2026](#)). The validation demonstrated that

this dual-model consensus mirrors human expert judgment with high precision, achieving an accuracy of 0.985 for the offensive class and 0.95 for the non-offensive class. Regarding the *Implicit* category (expanded in Appendix B), our prior validation indicated it as a primary source of human-model disagreement; thus, it is excluded from the present study to maintain a robust and reliable labeling schema.

**Human Review (Targeted):** We complement the quantitative analysis with a targeted human review of systematic label flips. This validation is described in the next section. Figure 1 summarizes the overall re-annotation and evaluation pipeline.

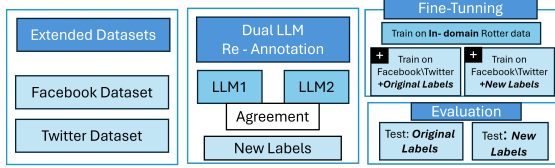
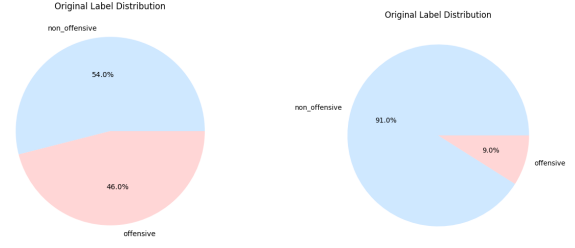


Figure 1: Re-annotation and evaluation pipeline.

## 5. Re-Annotation Analysis

Offensiveness is inherently context-dependent. It is shaped by cultural norms, discourse conventions, and political framing and is especially sensitive to pragmatic phenomena such as sarcasm, irony, and quoted speech. As a result, differences between datasets may reflect distinct operationalizations of offensiveness rather than mere annotation noise.

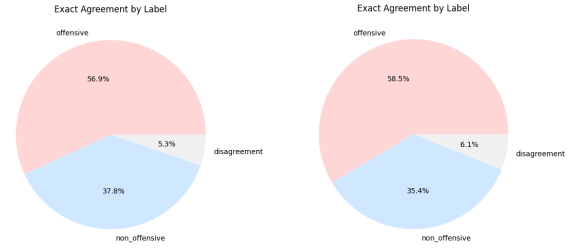
**Quantifying Annotation Shift:** We compare the original labels of each external dataset with the re-annotated labels obtained under our dual-LLM agreement protocol. Label distributions shift substantially between the original annotations (Fig. 2) and the new annotations (Fig. 3). The Facebook dataset shows a moderate increase in offensiveness, from 46% to 56.9% (5.3% disagreement), whereas the Twitter dataset shifts dramatically from a highly non-offensive skew (91% non-offensive) to a majority-offensive distribution (58.5% offensive; 6.1% disagreement). At the instance level, confusion matrices (Fig. 4) reveal systematic mismatch. These flips are systematic, not random, and cluster around recurring pragmatic phenomena. These occurrences are not solely linguistic; they embody culturally rooted meanings of offensiveness inside Hebrew discourse. Political assaults characterized as personal humiliation or allusions to terrorism and violence may be construed as either offensive or non-offensive based on their perception as targeted aggression or informational reporting. The Facebook dataset exhibits an overall flip rate of 18.55%; ( $\kappa=0.63$ ), with 29.94% of



(a) Facebook

(b) Twitter

Figure 2: Original label distribution



(a) Facebook

(b) Twitter

Figure 3: New label distribution

originally non-offensive instances relabeled as offensive and 5.53% flipping in the opposite direction. The Twitter dataset is far less stable: 56.25% of instances flip overall ( $\kappa=0.10$ ), driven by a strongly asymmetric shift (60.22% non-offensive→offensive vs. 16.71% offensive→non-offensive).

**Qualitative Validation: Contextual Sources of Label Mismatch:** To better understand the sources of annotation mismatch, we analyze the Facebook (F) and Twitter (T) datasets (Appendix C Table 4) by comparing the original labels with the joint agreement of two LLMs, and adjudicating each case against a manual annotation by a native Hebrew-speaking annotator. For each dataset, we sampled 100 instances (50 per reversal direction), restricted to cases where both LLMs flipped the original label. In the Facebook dataset, when the original label was non-offensive (N) but both LLMs predicted offensive (O), manual adjudication supported the LLMs in 92% of cases, indicating systematic under-labeling. The main drivers were direct profanity/explicit insults (24%) and political or ideological attacks framed as personal humiliation (24%), with additional contributions from hostile framing around violence/terrorism (14%) and metaphorical insults such as animal comparisons (10%). In the opposite Facebook direction (offensive → non-offensive), manual adjudication supported the LLMs in 85% of cases; this group was dominated by extremely short, context-poor fragments (44%), suggesting aggressive truncation that destabilizes offensiveness judgments, alongside implicit or metaphorical insults that the LLMs some-

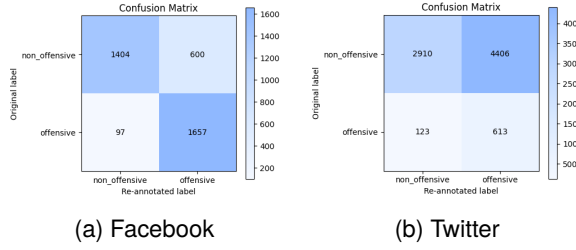


Figure 4: Confusion matrices between original and new labels

times missed (accounting for the remaining 15% where manual adjudication favored the original label). The Twitter dataset shows a different pattern: in the non-offensive  $\rightarrow$  offensive direction, manual adjudication supported the LLMs in 94% of cases, largely reflecting violence/terrorism discourse with an aggressive pragmatic stance (46%) and direct profanity/insults (30%). In the offensive  $\rightarrow$  non-offensive direction, agreement was highest (96%), with many tweets mentioning violence/terrorism (74%) but functioning as informational, descriptive, or news-like content rather than targeted humiliation. Overall, Twitter mismatches primarily reflect an operational definition gap: original annotators often treated discussion of violence and terrorism itself as offensive. Our definition frames offensiveness as targeted insult, dehumanization, threat, or directed harm, emphasizing its cultural and contextual contingency. Across both datasets, these mismatch patterns cluster into recurring pragmatic categories, such as direct profanity, political humiliation, dehumanization, violence-related framing, and context-poor fragments. Table 4 (Appendix C) provides representative examples drawn directly from the manually adjudicated sample described above, illustrating how each category manifests in practice and highlighting the specific linguistic and pragmatic cues that contributed to systematic label flips.

## 6. Evaluation Methodology and Results

**Models:** We focus on Hebrew-specific transformer models that consistently perform well for Hebrew NLP tasks: HeBERT, HeBERT-emo (Chriqui and Yahav, 2022), AlephBERT (Seker et al., 2021), AlephBERTGimmel (Gueta et al., 2022), DictaBERT (Shmidman et al., 2023), and HeRo, (Shalumov and Haskey, 2023). All models are fine-tuned for binary offensive language detection using the same training procedure.

**Preprocessing and Experimental Design:** For each external dataset (Facebook and Twitter), we apply the same preprocessing and filtering pipeline. Our re-annotation retains only high-confidence in-

stances (LLM1-LLM2 agreement) and discards both *Implicit* and disagreement cases, yielding smaller datasets than the original corpora. For experiments involving external training, we use stratified train/test (80-20) splits and evaluate on held-out test sets. All models are trained for three epochs, consistent with prior experiments. To disentangle domain shift from annotation shift, we compare evaluation under the original labels versus the new labels, and test whether adding external training data (again with 80-20 split) improves performance beyond label alignment.

### Setup E1: direct cross-dataset evaluation:

We train the models only on the Rotter dataset (denoted by R) and then evaluate them on test sets of F and T datasets; we denote it by  $R \rightarrow F/T$ . We report performance under two label configurations: (1) old labels; (2) new labels. Importantly, the model parameters are identical when evaluating under the original versus new labels; only the evaluation labels differ. This allows us to isolate the effect of annotation mismatch independently of domain adaptation.

### Setup E2: fine-tuning with external training

**data:** We fine-tuned each model on the Rotter dataset, augmented with the training set of a single external dataset (either F or T dataset, without cross-mixing), and denote it with  $R + (F/T) \rightarrow F/T$ . We report performance under two label configurations: (1) old labels; (2) new labels. In both cases, evaluation is performed on the external test sets of F and T. This stage tests whether performance gains remain after label alignment and whether additional adaptation to the external domain improves results. Unlike E1, this setup introduces external data into training and therefore reflects both domain adaptation and label alignment effects. **For all experiments**, we report Accuracy, Precision, Recall, and F1, using consistent training hyperparameters.

**Evaluation Results: Setup E1: Evaluation under original vs. new labels.** As shown in Table 2, evaluation under the new labels yields substantially higher accuracy than evaluation under the original labels across both datasets. On Facebook (F), average accuracy increases from 78.4% (E1.1) to 91.1% (E1.2), a gain of +12.7 points. On Twitter (T), the increase is much larger: from 45.4% (E1.1) to 85.2% (E1.2), a gain of +39.8 points. Since model weights and predictions are identical between E1.1 and E1.2, the observed performance differences are solely due to differences in annotation, not domain adaptation. Notably, the Twitter results suggest that most of the cross-dataset failure under the original labels is an artifact of divergent annotation criteria, rather than a true inability to generalize.

**Setup E2: Combined fine-tuning with external data.** Adding external training data provides addi-

Model	$R \rightarrow F/T$		$R \rightarrow F/T$		$R + (F/T) \rightarrow F/T$		$R + (F/T) \rightarrow F/T$	
	F	T	F	T	F	T	F	T
	old labels		new labels		old labels		new labels	
<i>Metrics:</i>			<i>Accuracy / Macro-F1</i>					
HeBERT	0.788 / 0.776	0.461 / 0.199	0.920 / 0.923	0.869 / 0.880	0.843 / 0.811	0.843 / 0.128	0.935 / 0.935	0.932 / 0.932
HeBERT-emo	0.758 / 0.746	0.439 / 0.190	0.910 / 0.914	0.848 / 0.863	0.850 / 0.824	0.853 / 0.100	0.941 / 0.942	0.931 / 0.930
AlephBERT	0.792 / 0.786	0.458 / 0.201	0.903 / 0.909	0.870 / 0.881	0.850 / 0.823	0.852 / 0.175	0.935 / 0.935	0.941 / 0.940
AlephBERTGimmel	0.805 / 0.798	0.476 / 0.203	0.930 / 0.934	0.895 / 0.902	0.850 / 0.824	0.846 / 0.146	0.960 / 0.960	0.956 / 0.956
DictaBERT	0.790 / 0.779	0.456 / 0.195	0.921 / 0.925	0.883 / 0.893	0.858 / 0.836	0.870 / 0.122	0.955 / 0.955	0.956 / 0.956
HeRo	0.770 / 0.765	0.431 / 0.190	0.880 / 0.893	0.849 / 0.865	0.848 / 0.821	0.863 / 0.107	0.931 / 0.933	0.932 / 0.932

Table 2: Model performance across datasets (Facebook (F) and Twitter (T)) Results are reported as **Accuracy / Macro-F1 score**. The updated labeling schema demonstrates significant gains in cross-domain robustness.

tional improvements, but primarily when evaluation is conducted under the new labels. Under the original labels (E2.1), accuracy improves only modestly relative to E1.1, reaching 85.0% on Facebook and 85.5% on Twitter on average. In contrast, under the new labels (E2.2), accuracy increases consistently, reaching 94.3% on Facebook and 94.1% on Twitter on average. The best configuration (AlephBERT-Gimmel, E2.2) achieves 96.0% on Facebook and 95.6% on Twitter, suggesting that once label definitions are aligned, the remaining gap is smaller and more plausibly attributable to genuine domain variation rather than annotation shift.

## 7. Discussion

Our findings challenge the conventional focus on domain adaptation in offensive language detection. Instead of treating cross-dataset performance drops as an inherent limitation of model generalization across topics, we demonstrate that they are often an artifact of divergent human perspectives and labeling mismatch. Our results suggest that cross-dataset degradation in Hebrew offensive language detection is largely explained by annotation shift rather than domain shift. External corpora appear to rely on different operational definitions of offensiveness, especially for politically charged discourse, irony, and indirect group references. While these findings highlight the centrality of annotation shift, it is important to address a potential methodological concern. Specifically, one might worry about the circularity of using LLM-based silver labels for both training and cross-dataset evaluation. We mitigate this risk by: (1) establishing human expert validation on Dataset R, and (2) performing a qualitative manual audit of the label flips in Datasets F and T (Section 5). This audit confirms that the shifts in performance reflect a correction of previous annotation inconsistencies rather than an algorithmic bias inherent to the generative models. From an affective science perspective, this is expected: offensiveness is not an objective text property, but a context-dependent judgment shaped by

cultural norms and discourse practices. Therefore, cross-dataset benchmarking should not be treated as a pure robustness test. Without verifying label comparability, performance drops may reflect mismatched annotation conventions as much as true domain variation. This suggests that “solving” cross-domain detection requires establishing unified, culturally grounded annotation protocols before applying complex adaptation algorithms. We recommend that NLP datasets document guidelines in detail, report agreement and ambiguity policies, and include qualitative analysis of borderline cases. Taxonomy-grounded prompting offers a practical route for improving cross-corpus label alignment.

## 8. Conclusions

We re-annotated two external Hebrew offensive language datasets using a culturally grounded prompt and a dual-LLM agreement protocol. The new labels diverge substantially from the original annotations, and evaluation under the new labels yields markedly improved performance. Combined fine-tuning with the new labels further improves results, suggesting that once labels are aligned, remaining gaps are more plausibly attributable to genuine domain variation. Our study highlights a methodological point for computational affective science: offensiveness is context-dependent, and cross-dataset evaluation must account for annotation validity.

## 9. Limitations

This study has several limitations. First, LLM-based annotation may introduce biases reflecting the models’ training data and cultural priors. Second, our corrected labels represent one coherent operationalization of offensiveness, but not an absolute truth. Third, we focus on two external Hebrew datasets; additional corpora may exhibit different mismatch patterns. Finally, some degree of genuine domain shift likely remains even after label alignment.



## 10. Acknowledgements

This research was supported by the Israel Innovation Authority.

## 11. References

- Gili Berger Hefetz, Yossef Haim Shrem, and Chaya Liebeskind. 2026. Hebrew offensive language detection via prompt-guided llm annotation and transformer fine-tuning. In *Proceedings of the IEEE Conference on Computer Systems and Applications*, Jerusalem, Israel. To appear.
- Avihay Chriqui and Inbal Yahav. 2022. Hebert and hebemo: A hebrew bert model and a tool for polarity analysis and emotion recognition. *INFORMS Journal on Data Science*, 1(1):81–95.
- Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *Acm Computing Surveys (Csur)*, 51(4):1–30.
- Eylon Gueta, Avi Shmidman, Shaltiel Shmidman, Cheyn Shmuel Shmidman, Joshua Guedalia, Moshe Koppel, Dan Bareket, Amit Seker, and Reut Tsarfaty. 2022. Large pre-trained models with extra-large vocabularies: A contrastive analysis of hebrew bert models and a new one to outperform them all. *arXiv preprint arXiv:2211.15199*.
- Nagham Hamad, Mustafa Jarrar, Mohammad Khalilia, and Nadim Nashif. 2023. Offensive hebrew corpus and detection using bert. In *2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*, pages 1–8. IEEE.
- Chaya Liebeskind, Ali Afawi, Marina Litvak, and Natalia Vanetik. 2024. Classifying offensive language in arabic: a novel taxonomy and dataset. *Lodz Papers in Pragmatics*, 20(2):433–462.
- Chaya Liebeskind and Shmuel Liebeskind. 2018. Identifying abusive comments in hebrew facebook. In *2018 IEEE International conference on the science of electrical engineering in Israel (ICSEE)*, pages 1–5. IEEE.
- Chaya Liebeskind, Natalia Vanetik, and Marina Litvak. 2023. Hebrew offensive language taxonomy and dataset. *Lodz Papers in Pragmatics*, 19(2):325–351.
- Marina Litvak, Natalia Vanetik, Chaya Liebeskind, Omar Hmdia, and Rizek Abu Madeghem. 2022. [Offensive language detection in Hebrew: can other languages help?](#) In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3715–3723, Marseille, France. European Language Resources Association.
- Marina Litvak, Natalia Vanetik, Yaser Nimer, Abdulrhman Skout, and Israel Beer-Sheba. 2021. Offensive language detection in semitic languages. In *Multimodal hate speech workshop*, volume 2020, pages 7–12.
- Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Shaked Greenfeld, and Reut Tsarfaty. 2021. Alephbert: A hebrew large pre-trained language model to start-off your hebrew nlp application with. *arXiv preprint arXiv:2104.04052*.
- Vitaly Shalumov and Harel Haskey. 2023. Hero: Roberta and longformer hebrew language models. *arXiv preprint arXiv:2304.11077*.
- Shaltiel Shmidman, Avi Shmidman, and Moshe Koppel. 2023. Dictabert: A state-of-the-art bert suite for modern hebrew. *arXiv preprint arXiv:2308.16687*.
- Zeerak Talat. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

### A. Prompt Design

**General Overview:** To ensure a consistent and high-quality gold standard, we developed a structured annotation framework implemented through a detailed system prompt for the Large Language Models (LLMs). The prompt was designed to simulate the decision-making process of a linguistic expert, providing the models with clear definitions, contextual guidelines, and a multi-step classification process. The framework emphasizes three primary pillars: (1) Contextual Awareness, requiring the models to look beyond literal meaning to capture tone and cultural nuance; (2) Objective Justification, where each label must be supported by a concise rationale; and (3) Strict Validation, employing a “conservative-by-design” logic that prioritizes the “Non-Offensive” category in cases of ambiguity. This structured approach was instrumental in aligning the automated annotations with human expert judgment, particularly in the complex landscape of Hebrew online discourse.

#### System Prompt for Hebrew Offensive Language Classification

Overview Classify text into three categories based on intent, context, and impact: “Offensive”,



"Implicit Offensiveness", or "Non-Offensive". Consider cultural context, tone, and intended meaning when making classifications.

#### Classification Categories

**1. Non-Offensive Definition:** Language that maintains respect and social harmony, lacking harmful intent or negative impact.

**Key Characteristics:** - Neutral or positive statements - Factual observations - Constructive criticism (when appropriate) - Professional communication - Personal expressions of emotion or opinion without targeting others

**Contextual Considerations:** - News reporting of negative events (when objective) - Professional disagreement (when respectfully expressed) - Cultural references (when not stereotyping) - Academic or educational discussions

**Examples:** - "אני מאחל לך יום נפלא!" (I wish you a wonderful day!) - "האינטרנט שינה את הדרך" (The internet changed how we communicate) - "אני לא מסכים עם הדעה שלך, אבל" (I disagree with your opinion, but I respect it)

**2. Offensive Definition:** Language deliberately intended to harm, discriminate, or cause direct distress.

**Key Characteristics:** - Explicit slurs or hate speech - Direct personal attacks - Discriminatory statements - Dehumanizing language - Explicit threats or violent language - Mockery of protected characteristics

**Common Categories:** 1. Racial/Ethnic Discrimination 2. Religious Intolerance 3. Gender-based Discrimination 4. Disability-based Discrimination 5. Age-based Discrimination 6. Appearance-based Attacks 7. Socioeconomic Status Attacks

**Examples:** - "אתה טיפש ומכוער." (You're stupid and ugly) - "כל הקבוצה אתנית הם סטריאוטיפ שלילי." (All [ethnic group] are [negative stereotype]) - "נשים - לא מתאימות לתפקיד." (Women aren't suited for [role])

**3. Implicit Offensiveness \*\*Definition\*\*:** Language that appears neutral but carries subtle harmful intent or negative implications.

**Key Characteristics:** - Passive-aggressive statements - Backhanded compliments - Coded language or dog whistles - Microaggressions - Sarcasm with negative intent - Subtle stereotyping - Leading questions with negative implications

**Common Patterns** 1. *Contextual Triggers* - Ellipses (...) suggesting negative implications - Tone markers that shift meaning - Cultural references used inappropriately

2. *Structural Elements* - False equivalencies - Loaded questions - Subtle stereotyping - "Just asking questions" format

**Validation Questions** Before classifying a statement as implicitly offensive, answer the following two questions: 1. To whom is the offensive implica-

tion directed? 2. What specific part of the sentence makes it offensive?

- If there is **no clear answer** to at least one of these questions, **the statement should NOT be classified as offensive**. - The classifier must **justify why** the statement falls into this category with explicit reasoning.

**Examples - Implicitly Offensive:** - "למה את תמיד..." (Why are you always so... special?) - Directed at a specific individual with passive-aggressive tone - The ellipsis suggests negative implications

- "את די חכמה... בשביל בחורה." (You're pretty smart... for a girl) - Targets a gender group - Implies an underlying stereotype

- Not Offensive (Fails validation): - "איזה יום יפה!" (What a beautiful day today!) - No clear target - No harmful implication

#### Classification Guidelines

1. **Context Analysis:** - Consider cultural context - Evaluate speaker intent - Account for power dynamics - Assess historical context - Consider audience impact

2. **Tone Evaluation:** - Analyze word choice - Consider delivery method - Look for subtle markers - Evaluate emotional impact

3. **Impact Assessment:** - Consider potential harm - Evaluate broader social implications - Account for group dynamics - Assess perpetuation of stereotypes

#### Edge Cases and Special Considerations

1. **Reclaimed Language:** - Consider in-group usage - Evaluate context and speaker identity - Account for cultural evolution

2. **Educational Context:** - Academic discussion of offensive terms - Historical documentation - Anti-discrimination training

3. **Artistic Expression:** - Creative works - Social commentary - Satirical content

#### Classification Process

1. **Initial Assessment:** - Read/hear the complete statement - Note immediate reaction - Identify key terms/phrases

2. **Contextual Analysis:** - Consider speaker intent - Evaluate situation - Account for cultural factors

3. **Final Classification:** - Apply category criteria - Consider edge cases - Document reasoning - Estimate classification confidence (0.0–1.0) based on clarity, context, and alignment with definitions Please provide the confidence as a decimal number between 0.0 and 1.0 only, Do NOT return a string or word, only a raw number Please provide a brief and concise reasoning for your classification, using no more than 6 words!

**Remember:** When in doubt, consider the potential impact on marginalized or vulnerable groups and err on the side of caution.

## B. Qualitative Analysis of Implicit Offensiveness

## C. Examples of Systematic Annotation Mismatch Patterns

This appendix clarifies our treatment of the “Implicit” category, which in this study specifically refers to **Implicitly Offensive** content (i.e., insults or attacks delivered without explicit slurs). Our qualitative error analysis revealed that Large Language Models (LLMs) struggle to distinguish between *Implicit Offensiveness* and neutral/ironic discourse in Hebrew. This leads to a phenomenon of **Over-contextualization**, where the model hallucinates offensive intent behind neutral linguistic markers. Crucially, we attempted to mitigate this by providing the models with additional context, specifically by **concatenating the parent headline or post to the comment**. However, our experiments showed that this often *exacerbated* the over-contextualization effect; the models became overly sensitive to the sensitive nature of the headlines, leading them to flag even neutral inquiries as “Implicitly Offensive”. Table 3 illustrates these challenges with representative examples.

Hebrew Original	English Translation	The Annotation Challenge
"אישה נהדרת בעלת מחשבה חופשית 🤔🤔🤔"	"A wonderful woman with free thought 🤔🤔🤔"	<b>Implicitly Offensive vs. Irony:</b> The LLM flags this as <i>Implicitly Offensive</i> (mockery) due to the emojis, though it remains ambiguous without deep social cues.
"אתה כותב זאת מתוך ידע או שזו השערה?"	"Are you writing this from knowledge or is it a hypothesis?"	<b>Contextual Misinterpretation:</b> Even when provided with the post’s headline, the LLM flags the question mark as a "belittling" attack, rather than a valid debate question.
"ממש 'גאון' הדור..."	"Quite the 'genius' of our generation..."	<b>Consensus Gap:</b> While clearly sarcastic, the models and humans often disagree on whether this reaches the threshold of <i>Implicit Offensiveness</i> .

Table 3: Examples of implicitly offensive cases that caused model disagreement and oversensitivity.

As demonstrated, forcing a binary label on these cases would introduce significant noise. Furthermore, simply adding textual context (e.g., headlines) proved insufficient and sometimes counter-productive. We conclude that addressing **Implicit Offensiveness** requires more sophisticated, multi-turn dialogue architectures or specialized cultural fine-tuning, which we define as **future research directions**. Focusing on explicit cases in this study ensures a high-precision gold standard.

Table 4: Representative examples of systematic annotation mismatch patterns in the Facebook (F) and Twitter (T) datasets.

DS	Dir.	Phenomenon	English Example	Hebrew Example
F	N→O	Direct insult/profanity	"Hanin, you are a pathetic person."	"חנין אתה אדם עלוב"
F	N→O	Political humiliation	"And I tell you, Lapid, it is better to be rich than poor like you..."	"ואני אומר לך לפיד עדיף עשיר ולא עני כמוך באוצר עשית כלום"
F	N→O	Dehumanization (animal)	"Someone on the ship should be thrown to the dogs and kicked out of the Knesset, traitors."	"אחד על הספינה לזרוק לכלבים ולהעיף מן הכנסת בוגדים"
F	O→N	Context-poor fragment	"Shame."	"בושה"
T	N→O	Aggressive violence/terror framing	"A Jewish terrorist... should take into account that this could be his end."	"מחבל יהודי עליו לקחת בחשבון זה יכול ליהיות סופו"
T	N→O	Strong abusive language	"The old woman in the hospital corridors... trash... scum... may their name be erased."	"הזקנה במסדרונות... מיפלגת טינופת חלאות ימח שמם וזכרם"
T	O→N	News-style violence mention	"Breaking: The PA thwarted a stabbing attack against soldiers..."	"הרשפ סיכלה פיגוע דקירה נגד החיילים"
T	O→N	Informational violence mention	"The Hamas terror organization marks today 12 years since Operation Cast Lead..."	"ארגון הטרור חמאס מציינ היום 12 שנה לעופרת יצוקה"

# Multi-Source Emotion Annotation in Children’s Language: When LLM Consensus Diverges from Human Judgment

Farida Saïd<sup>1</sup>, Jeanne Villaneau<sup>2</sup>

<sup>1</sup>LMBA, Université de Bretagne Sud, France  
farida.said@univ-ubs.fr

<sup>2</sup>IRISA, Université de Bretagne Sud, France  
jeanne.villaneau@univ-ubs.fr

## Abstract

Automated emotion annotation increasingly relies on inter-LLM agreement as a proxy for label quality. We test this assumption on 2,106 clause-level segments from interviews with French-speaking children (ages 6–11) about parental roles, a setting where affect is often implicit rather than lexically explicit. Using a 500-segment expert gold standard, we show that internal consensus can be seriously misleading: Dawid–Skene, a probabilistic label aggregation method, estimates GPT-5.2 valence accuracy at 90.7%, whereas evaluation against human gold yields 71.0%, revealing substantial overestimation driven by shared neutralization bias. Conversely, Dawid–Skene underestimates Claude Sonnet 4, reversing model ranking. Majority Vote, Dawid–Skene, and MACE produce near-identical consensus labels, suggesting that the main source of error lies in shared annotator bias rather than in the aggregation rule itself. We release the expert gold subset and the probabilistic corpus to support future work. Our results show that high inter-LLM agreement cannot replace external human validation for affect annotation.

**Keywords:** emotion annotation, child language, implicit affect, LLM annotation, annotation aggregation, human validation, gold standard, valence

## 1. Introduction

Emotion analysis in short, transcribed oral responses by children is challenging: telegraphic syntax, disfluencies, and pervasive implicit affect limit the reliability of sentiment tools trained on adult corpora (Ogren and Sandhofer, 2021; Sagae et al., 2010; Liu and Chen, 2024). Yet understanding how children emotionally frame family relationships is relevant to developmental psychology, education, and child welfare, while automated annotation for this population remains underexplored (Pérez-Espinosa et al., 2020).

We study French transcriptions of children’s oral interview responses about the parental roles FATHER and MOTHER (Bellachhab et al., 2025). The dataset contains 2,106 clause-level segments requiring emotion (valence, intensity) labels, and exhaustive expert annotation is cost-prohibitive. Before scaling an automated pipeline, we ask: can inter-LLM consensus be trusted as a quality signal for emotion annotation, or can shared model bias yield deceptive agreement?

We construct a 500-segment stratified expert gold standard and use it to assess both LLM performance and Dawid–Skene as a quality estimator. We show that when annotators share systematic biases (notably over-neutralization), aggregation can mistake correlated errors for consensus.

Our experimental design exploits a structural asymmetry: Dawid–Skene estimates annotator reliability from internal consensus, while our expert

gold standard provides an independent external measure. When these two signals diverge — as they do substantially here — the gap reveals not noise but systematic bias. We use this dissociation to identify the mechanism, not merely the magnitude, of LLM annotation error in this domain.

Our contributions are:

1. **DS / Gold dissociation:** Dawid–Skene overestimates GPT-5.2 by 19.7 points and underestimates Claude Sonnet 4 by 6.8 points on valence, reversing model ranking.
2. **Aggregation is not the bottleneck:** Majority Vote, Dawid–Skene, and MACE yield near-identical consensus labels (98.6% three-way agreement; 493/500), implicating annotator bias rather than the aggregation method.
3. **Implicit affect dominance:** Using Stanza lemmatization and a 61-entry simplified FEEL lexicon as a conservative probe, 76% of segments lack overt affect markers, creating conditions for systematic over-neutralization.
4. **Expert gold standard:** 500-segment expert annotation with structured disagreement resolution ( $\kappa = 0.682\text{--}0.731$ ), validated via Dawid–Skene aggregation of human-only annotations ( $\kappa = 0.983\text{--}0.991$ ).

## 2. Related Work

Emotion annotation in NLP has been shaped by shared tasks (Strapparava and Mihalcea, 2007) and lexical resources such as FEEL (Abdaoui et al., 2017) and FANCat (Syssau et al., 2021) for French — both derived from adult corpora, which raises immediate questions about applicability to child-produced language. Beyond resource availability, the field has increasingly recognized that emotion annotation is inherently subjective: disagreement between annotators often reflects genuine perceptual differences rather than error (Uma et al., 2021; Plank, 2022), a perspective known as *perspectivism*. This is especially true when affect is implicit — when no explicit emotion word anchors the judgment and annotators must rely on pragmatic inference. In such cases, disagreement concentrates precisely where the signal is weakest, as our results confirm. Related work has begun extending annotation to child-facing settings (Seo et al., 2024; Pérez-Espinosa et al., 2020), but the reliability of automated methods for this population remains underexplored.

LLMs have emerged as a promising alternative to crowd-sourced annotation: ChatGPT matches or exceeds crowd-worker quality on several text classification tasks (Gilardi et al., 2023), and GPT-4 reaches expert-level performance on political content coding (Törnberg, 2023). For emotion annotation specifically, LLMs often align well with human judgments but exhibit systematic perceptual biases, motivating hybrid human–LLM pipelines (Niu et al., 2024, 2025). A natural question is how to assess LLM annotation quality without exhaustive human validation — Dawid–Skene (Dawid and Skene, 1979) is typically used for this, modelling each annotator’s confusion matrix via EM to recover latent true labels and estimate per-annotator reliability. Agreement metrics such as Fleiss’  $\kappa$  (Fleiss, 1971; Cohen, 1960), Krippendorff’s  $\alpha$  (Krippendorff, 2011), and their properties are discussed in (Artstein and Poesio, 2008). However, Dawid–Skene assumes conditional independence of annotator errors — an assumption that may fail when multiple prompts of the same model share architectural biases. Whether inter-LLM consensus is a reliable quality signal under these conditions, particularly when affect is largely implicit, is the question we address.

## 3. Data and Annotation Schema

### 3.1. Corpus

We use an anonymized corpus of transcribed individual interviews with French-speaking children (corpus and collection protocol available under CC-BY-4.0), collected in four primary schools in France

between March 2021 and March 2022. Participants (N=184 children, ages 6–11, mean 9.2 years; 51% boys) were asked to describe the concepts FATHER and MOTHER. The corpus contains 2,106 segments: 1,053 for FATHER and 1,053 for MOTHER (mean length 8.7 words, SD 5.3).

### 3.2. Annotation Schema

We designed a two-level annotation schema capturing valence and intensity:

**E1 – Valence and Intensity:** `valence`  $\in$  {positive, negative, neutral, mixed}; `intensity`  $\in$  {0 (none), 1 (weak), 2 (moderate), 3 (strong)}; `evidence_span` citing the supporting text fragment; `negation_flag` for negation constructions.

This paper focuses primarily on `E1.valence` as the primary dimension, with `E1.intensity` as a secondary analysis demonstrating LLM limitations on ordinal annotation scales. All annotations include a self-reported confidence score (0.0–1.0).

## 4. Automated Annotation Methods

### 4.1. Lexicon baselines

**FEEL Complete (14,126 entries; raw).** We evaluate the full FEEL lexicon (Abdaoui et al., 2017) (derived from adult corpora) without modification.

**FEEL Complete (filtered).** To mitigate domain mismatch, we test an aggressively filtered variant (stopword removal, POS filtering, length constraints, and stricter mixed rules); results are reported in Section 6.1.

**FEEL Simplified (61 entries).** Finally, we use a manually curated 61-item subset of FEEL as a conservative proxy for overt affect (29 pos., 30 neg., 2 contextual), excluding frequent but context-dependent role terms (e.g., *travaille*, *ménage*, *chef*) (Appendix B).

#### 4.1.1. Annotation procedure

All lexicon-based methods use Stanza (Qi et al., 2020) French lemmatization (`tokenize`, `mwt`, `pos`, `lemma` processors) to normalize word forms before lexicon lookup. Negation markers (*ne*, *pas*, *jamais*, *sans*) within a 3-token window preceding an affective lemma flip its valence. Final valence is determined by majority vote over matched lemmas, defaulting to neutral if no matches occur.



## 4.2. LLM-Based Annotation

We treat each LLM call as a noisy annotator. Two models, Claude Sonnet 4 (claude-sonnet-4-20250514, Anthropic) and GPT-5.2 (gpt-5.2-2025-12-11, OpenAI), annotated the full 2,106-segment corpus through their respective batch APIs. All runs used temperature 0, `max_tokens=500`, and structured JSON outputs constrained by a predefined schema.

**Prompting setup.** Each model was queried with three zero-shot prompts (P1–P3), yielding six pseudo-annotators in total (2 models  $\times$  3 prompts). The prompts shared the same system instruction and annotation schema, and differed only in annotation posture: conservative, implicit-affect sensitive, or lexically grounded. The full prompt templates are provided in Appendix A.

**Implementation details.** Claude Sonnet 4 was queried via the Anthropic Message Batches API on 2025-12-25, and GPT-5.2 via the OpenAI Batch API on 2025-12-23. For each segment, the user message contained the segment metadata and the JSON Schema specifying the fields used in this study (valence, intensity, evidence span, negation flag, and confidence).

## 4.3. Aggregation Methods

All aggregation methods (MV, DS, and MACE) are applied only to the six LLM pseudo-annotators; lexicon-based methods are evaluated separately as baselines.

**Majority/Plurality Vote (MV).** MV assigns the plurality label. In case of ties, we use a uniform distribution over tied labels, yielding confidence  $1/n_{\text{tied}}$ .

**Dawid–Skene (DS).** DS is a probabilistic aggregation model (Dawid and Skene, 1979) that estimates latent true labels and annotator-specific confusion matrices via EM, using a Dirichlet prior  $\alpha = 1$ . It assumes conditional independence of annotator errors, an assumption that may be violated when LLM annotators share correlated biases (Section 6.6).

**Ordinal DS for E1 intensity** ( $y \in \{0, 1, 2, 3\}$ ). For ordinal intensity, we replace the standard categorical noise model with:

$$P(l = t \mid y = k, a) \propto \exp(-\beta_a |k - t|),$$

normalized over  $t$ , where  $\beta_a$  controls annotator  $a$ ’s tolerance to ordinal distance.

**MACE.** MACE is an EM-based aggregation model (Hovy et al., 2013) in which annotators either output the true label with competence  $\theta_a$  or sample from a global spam distribution  $\xi$  with probability  $1 - \theta_a$ , thereby down-weighting unreliable annotators.

### 4.3.1. Robustness to the Aggregation Method

On the 500 gold segments, all three methods converge strongly (98.6% three-way agreement, 493/500) and yield near-identical performance (MV: 69.2%, DS: 69.0%, MACE: 69.2%;  $\kappa \approx 0.475$ ). This makes the aggregation rule itself an unlikely source of error and instead points to the LLM annotator pool as the main source of systematic bias (Section 6.6). For brevity, the remaining analyses use DS, which also provides annotator-specific confusion matrices for the bias analysis in Section 6.6.

## 5. Human Gold Standard Construction

### 5.1. Stratified Sampling

To validate automated annotations, we constructed a human gold standard on 500 segments selected by stratified sampling from the full corpus, providing  $\pm 3\%$  confidence intervals at the 95% level.

Four complementary strata were designed to maximize validation informativeness (Table 1).

Stratum	%	n	Objective
DS Confidence	40	200	Calibrate DS scores vs. human <sup>a</sup>
Label-proportional	30	150	Corpus representativeness <sup>b</sup>
LLM Disagreements	20	100	Maximally informative cases
Demographic	10	50	Age/gender/concept robustness
<b>Total</b>	<b>100</b>	<b>500</b>	

<sup>a</sup> Equal sampling across three confidence levels: high ( $\geq 0.9$ ), medium ( $0.7 \leq x < 0.9$ ), and low ( $< 0.7$ ).

<sup>b</sup> Sampling proportional to the LLM-predicted label distribution (positive 53.7%, neutral 40.8%, negative 4.4%, mixed 1.1%). Gold-standard results later confirmed that this distribution differs from the true one.

Table 1: Stratified sampling design (N=500)

Since stratification targeted *structural properties* (confidence zones, inter-LLM disagreement, demographic balance) rather than specific LLM predictions on individual instances, and since

the resulting gold converges with an independent Dawid–Skene aggregation of human-only annotations ( $\kappa = 0.991$ , Section 5.4), selection bias is minimal. Because this subset is not a simple random sample, results are reported by stratum where relevant, and global estimates are interpreted with caution.

## 5.2. Annotation Protocol

Three independent expert annotators with backgrounds in education and child development annotated all 500 segments, following a shared annotation manual with prototypical examples and evidence-span justifications. A 30-segment calibration pilot and a 50-segment quality checkpoint preceded the full annotation. Disagreements were resolved in structured discussion sessions; final labels were determined by majority vote (2/3 agreement). Segments were flagged with confidence levels (HIGH / MEDIUM / LOW) based on the degree of initial annotator agreement. The resulting gold is predominantly positive (68.4%), with neutral (26.0%), negative (4.6%), and mixed (1.0%) classes; full label distributions are reported in Table 11.

## 5.3. Inter-Annotator Agreement

Inter-annotator reliability on the 500-segment subset was assessed using Fleiss’  $\kappa$  over the three expert annotators. Agreement is substantial for valence ( $\kappa = 0.731$ ) and moderate-to-substantial for intensity ( $\kappa = 0.682$ ), consistent with the known difficulty of ordinal intensity scales (Conventions in (Artstein and Poesio, 2008)). Disagreements concentrate on short or contextually ambiguous segments.

Relative to the adjudicated (vote-based) gold, Annotator 3 shows lower agreement on valence ( $\kappa = 0.719$ ) than the other two annotators, reflecting a more conservative interpretation of implicit affect. Annotator 3 was outvoted on 76 segments during disagreement resolution; accordingly, the gold reflects the majority (more liberal) boundary and may underestimate the positive/neutral boundary difficulty (Section 8.1).

## 5.4. Gold Standard Validation

To validate the collaborative resolution procedure, we applied Dawid–Skene independently to the three individual human annotations and compared the resulting labels to the vote-and-discussion gold.

The two procedures converge at 98.8–99.6% ( $\kappa = 0.983$ – $0.991$ ), confirming label equivalence and providing the baseline contrast for Section 6.6:

Dimension	Agreement	$\kappa$	Divergences
E1.valence	99.6%	0.991	2
E1.intensity	98.8%	0.983	6

*Note.* All 8 divergences fall on segments flagged as MEDIUM or LOW confidence in the vote-based gold. No divergence occurs on HIGH-confidence segments.

Table 2: Agreement between vote-and-discussion gold and Dawid–Skene human gold

applied to truly independent annotators, DS produces near-perfect alignment; applied to LLM prompts sharing architectural biases, it does not.

# 6. Results

## 6.1. Lexicon Baseline

Table 3 compares lexicon methods against the human gold.

Method	Acc.	$\kappa$	Gap
FEEL Simplified (61)	44.0%	0.162	−25.0 pp
FEEL Complete (raw)	46.2%	−0.032	−22.8 pp
FEEL Complete (filt.)	58.6%	0.023	−10.4 pp
LLM Aggregation	69.0%	0.475	baseline

*Note.* Gap = difference from the LLM aggregation baseline (69.0%).

Table 3: Lexicon baseline comparison for valence accuracy and  $\kappa$  on the human gold ( $N = 500$ )

FEEL Simplified is the stronger baseline despite its 61-entry size: its 76% zero-hit rate defaults to neutral, correctly handling implicitly affective text, while FEEL Complete (raw) over-predicts *mixed* (113 vs 5 gold) due to context-dependent lemmas (*papa [dad]*, *maman [mom]*) marked positive in adult corpora. Filtering recovers 12.4pp (+46.2%→58.6%) but the 10.4pp gap to LLM aggregation persists, confirming that lexicons cannot capture the 76% of segments expressing affect implicitly (see Appendix C for illustrative examples).

**Implicit affect dominance.** FEEL Simplified’s 76% zero-hit rate (after Stanza lemmatization) quantifies the core challenge: three out of four transcribed segments about parents lack explicit emotion words yet convey clear sentiment. Appendix C provides illustrative examples of affect expression mechanisms that lexicons cannot capture: pragmatic implicature (*qui prend soin de nous* → positive), counterfactual exclusion (*si tu es un garçon ton papa te comprend mieux* → negative), and frequency qualifiers (*on n’est pas d’accord parfois* →

Method	Valence		Intensity	
	Acc.	$\kappa$	Acc.	$\kappa$
FEEL Simplified	44.0	0.162	—	—
Surface Rules	39.2	0.118	—	—
GPT-5.2	71.0	0.484	41.0	0.107
Claude Sonnet 4	79.2	0.606	55.2	0.370
Annotator 1	98.0	0.956	90.8	0.865
Annotator 2	96.0	0.914	87.2	0.819
Annotator 3	87.0	0.719	78.0	0.682

*Note.* Accuracy is the proportion of exact label matches. LLM scores are computed from Dawid-Skene aggregated labels over the three prompts for each model. Human annotator performance is evaluated against the vote-and-discussion gold label. Lexicon methods are not evaluated for intensity because their deterministic rules are not designed for ordinal scales.

Table 4: Annotation performance against the human gold standard ( $N = 500$ )

mixed). This finding empirically justifies LLM deployment for this domain.

## 6.2. LLM Internal Consistency

Both LLMs achieved high internal consistency across their three prompts (Krippendorff’s  $\alpha > 0.8$ , Table 6): Claude Sonnet 4 mean  $\alpha = 0.901$ , GPT-5.2 mean  $\alpha = 0.858$ . GPT-5.2 shows 92.2% unanimous agreement across prompts; Claude Sonnet 4 shows 93.2%. These figures suggest high annotation quality. Section 6.6 shows they do not.

## 6.3. Performance Against Human Gold Standard

Table 4 reports annotation performance against the 500-segment human gold standard.

Several findings emerge. First, both LLMs substantially underperform human annotators on valence (14–27 percentage-point gap) and on intensity (23–50 point gap). Second, Claude Sonnet 4 outperforms GPT-5.2 on both dimensions, the reverse of Dawid-Skene estimates for valence (Section 6.6). Intensity results are analysed in Section 6.5.

**Class-level performance.** Table 5 reports per-class metrics, revealing asymmetric error patterns masked by overall accuracy.

GPT-5.2 achieves high precision on positive (0.969) but poor recall (0.632), reflecting systematic over-neutralization: 118 gold-positive segments incorrectly predicted neutral (34.5% of positive instances). Claude achieves better balance (precision=0.959, recall=0.749) with only 76 such

errors (22.2%). Macro-F1 scores—which weight all classes equally—show Claude’s advantage more clearly than raw accuracy: 0.660 vs 0.592 (+0.068). Both models struggle with the rare mixed class ( $F1=0.308$ – $0.333$ ,  $n=5$ ), but Claude substantially outperforms GPT on negative ( $F1=0.766$  vs 0.619), despite negative being only 4.6% of the corpus.

## 6.4. Aggregation Methods: Robustness

Three aggregation methods (Majority Vote, Dawid-Skene regularized, MACE) produce nearly identical results on the 500 gold segments, demonstrating that aggregation algorithm choice is not the performance bottleneck. Table 6 presents the comparison.

This near-perfect agreement across aggregation methods (98.6%, 493/500 segments) indicates that aggregation algorithm choice is not the main limitation in this setting. Although the three methods differ substantially in complexity, they produce functionally equivalent labels on the present corpus. The limiting factor is therefore not algorithmic sophistication, but shared annotator bias.

The three methods nevertheless serve different secondary purposes. When prioritising human review, MV is preferable because its vote margin flags 108/500 segments (21.6%) as low-confidence ( $< 0.7$ ), directly identifying cases for adjudication. By contrast, DS flags 0/500 segments and MACE only 3/500, making both largely uninformative for uncertainty-based review on this corpus. When confidence scores are not required, MACE achieves the best Macro-F1 (0.640 vs. 0.604 for MV), suggesting slightly better handling of rare classes. DS, however, remains indispensable for annotator bias analysis because it is the only method that estimates annotator-specific confusion matrices. This capability makes it possible to compare internal reliability with external gold performance, which is the central analytical objective of this study (Section 6.6).

Given the goals of this study, we therefore adopt DS for the present corpus.

## 6.5. Intensity: Ordinal Silver vs Gold

Individual LLMs perform poorly on intensity annotation ( $\kappa = 0.107$ – $0.370$ , exact accuracy 41.0–55.2%), far below human annotators ( $\kappa = 0.682$ – $0.865$ ). Ordinal DS aggregation substantially narrows this gap (Table 7), but reveals a systematic downward bias.

The  $\pm 1$  accuracy of 95.0% and quadratic  $\kappa = 0.654$  confirm ordinal plausibility—most errors are off by one level—but the label distribution exposes a strong downward compression (Table 8): the sil-

Class	GPT-5.2			Claude Sonnet 4		
	P	R	F1	P	R	F1
positive ( $n = 342$ )	0.969	0.632	0.765	0.959	0.749	0.841
neutral ( $n = 130$ )	0.494	0.954	0.651	0.597	0.923	0.725
negative ( $n = 23$ )	0.684	0.565	0.619	0.750	0.783	0.766
mixed ( $n = 5$ )	0.286	0.400	0.333	0.250	0.400	0.308
Macro-F1		0.592			0.660	
Weighted-F1		0.724			0.802	

Note. P = Precision, R = Recall, and F1 = F1-score. Macro-F1 weights all classes equally, whereas Weighted-F1 weights by class frequency. Support ( $n$ ) is shown in parentheses. LLM scores are computed from Dawid–Skene aggregated labels over the three prompts for each model.

Table 5: Per-class performance for valence ( $N = 500$ )

Method	Accuracy	$\kappa$	Macro-F1	Weighted-F1	Agree. w/ others
Majority Vote	69.2%	0.475	0.604	0.704	98.8%
DS	69.0%	0.475	0.618	0.703	99.0%
MACE	69.2%	0.476	0.640	0.705	99.2%

Three-way agreement: 98.6% (493/500 segments)

Note. All methods aggregate 6 LLM pseudo-annotators (2 models  $\times$  3 prompts). “Agree. w/ others” = proportion of segments where this method agrees with at least one other method. Three-way agreement computed over segments where all three methods produce identical labels (493/500 = 98.6%).

Table 6: Aggregation method comparison for valence performance on 500 gold segments

Metric	Value
Exact accuracy	55.2%
$\pm 1$ -level accuracy	95.0%
MAE	0.498
$\kappa$ (quadratic)	0.654
Spearman $\rho$	0.740

Table 7: Comparison of ordinal DS silver and human gold for intensity ( $N = 500$ )

ver over-predicts levels 0–1 and severely under-predicts levels 2–3 (recall  $\approx$  28.6% for level 3).

Level	Gold	Silver
0 (none)	130	198
1 (weak)	128	182
2 (moderate)	207	110
3 (strong)	35	10

Table 8: Intensity label distributions for gold and ordinal DS silver

This silver is suitable for pre-annotation and coarse-grained trend analysis, but should not be used for fine-grained intensity statistics, particularly when high-intensity segments are the focus.

Model	DS est.	Human gold	$\Delta$
GPT-5.2	90.7%	71.0%	−19.7 pp
Claude Sonnet 4	72.4%	79.2%	+6.8 pp

Note. DS estimate = Dawid–Skene diagonal accuracy from LLM consensus on the full corpus ( $N = 2,106$ ). Human gold = accuracy against expert annotation ( $N = 500$ ).  $\Delta$  = human gold minus DS estimate, in percentage points.

Table 9: Dissociation between DS estimates and human gold for valence

## 6.6. The DS / Human Dissociation

Table 9 contrasts Dawid–Skene estimates with human gold performance.

The dissociation is asymmetric: Dawid–Skene overestimates GPT-5.2 by 19.7 points while underestimating Claude Sonnet 4 by 6.8 points. Two distinct mechanisms explain this pattern.

### 6.6.1. GPT-5.2: Bias collusion inflates DS confidence

GPT-5.2 achieves 92.2% unanimous prompt agreement. However, 27.3% of unanimously-agreed segments are incorrect: all three GPT prompts predict “neutral” for segments the human gold labels “positive” ( $n=100$  segments). This sys-



tematic over-neutralization is not random error—it reflects a shared model bias. Dawid–Skene cannot distinguish genuine consensus from shared bias: it interprets the unanimity as high reliability and assigns elevated confidence scores to precisely the segments where GPT is wrong. The Dawid–Skene independence assumption is therefore violated not merely formally but in a way that compounds the most frequent error type. Corpus-level evidence confirms the mechanism: the human gold contains 68.4% positive segments, while GPT predicts only 44.6%—a 23.8-point neutralization deficit that three correlated prompts cannot self-correct.

**Claude Sonnet 4: diversity penalized by correlated prompt agreement.** Claude shows 82.0% unanimously correct segments, compared to 72.7% for GPT—indicating higher annotation precision when prompts agree. However, Dawid–Skene aggregation over the full 6-annotator pool (2 models  $\times$  3 prompts) assigns lower weight to Claude because Claude and GPT diverge on difficult segments, while the three GPT prompt variants remain highly consistent with each other. Under the conditional-independence assumption, DS interprets this correlated within-model agreement as higher annotator reliability and therefore downweights Claude whenever the two model blocks disagree. When evaluated directly against the human gold, Claude’s majority-vote accuracy (79.2%) matches its individual evaluation, confirming that the DS underestimation reflects correlated-prompt structure rather than genuine annotation weakness.

**Confidence-based filtering.** Despite the miscalibration at corpus level, LLM internal confidence scores remain locally predictive of quality. Segments with confidence  $\geq 0.7$  achieve 92.6–94.9% accuracy (Table 10), providing a practical threshold for annotation pipelines: apply LLM annotation to the full corpus, retain high-confidence segments, and prioritize low-confidence segments for human review.

Confidence	GPT-5.2	Claude Sonnet 4
$< 0.5$	62.7% (n=67)	70.5% (n=44)
0.5–0.7	57.7% (n=253)	70.6% (n=279)
0.7–0.9	92.6% (n=176)	94.9% (n=175)
$\geq 0.9$	100% (n=4)	100% (n=2)

*Note.* Confidence is the mean self-reported score across three prompts. Accuracy is measured against the human gold standard ( $N = 500$ ).

Table 10: LLM confidence vs. human gold accuracy

Method	Pos.	Neu.	Neg.	Mix.
Human Gold	68.4%	26.0%	4.6%	1.0%
FEEL Simplified	19.0%	77.0%	3.0%	1.0%
FEEL Complete (filt.)	80.6%	8.2%	7.0%	4.2%
LLM Aggregation	39.8%	53.0%	5.4%	1.8%
GPT-5.2	44.6%	50.2%	3.8%	1.4%
Claude	53.4%	40.2%	4.8%	1.6%

*Note.* LLM Aggregation corresponds to MACE silver ( $N = 500$ ). DS and MV differ by less than 0.4 pp. FEEL Complete (filt.) over-predicts the positive class by 12.2 pp, whereas LLM Aggregation over-predicts the neutral class by 27.0 pp.

Table 11: Valence label distributions ( $N = 500$  gold segments)

## 6.7. Complementary Bias Patterns: FEEL vs LLMs

FEEL and LLMs exhibit opposite systematic biases in label distribution, revealing fundamentally different annotation mechanisms. FEEL Complete over-detects positive (+12.2pp) through keyword sensitivity, while LLMs over-predict neutral (+27.0pp) through conservative bias on implicit affect. These opposite errors suggest potential value in hybrid pipelines (Section 8). Table 11 presents the full contrast.

## 7. Annotation Strategy for the Full Corpus

Our validation translates into four operational decisions for the remaining 1,606 segments. For valence, we apply MV silver labels and retain them as-is where vote confidence  $\geq 0.7$  — segments falling below this threshold across the full corpus (142/2,106) are submitted for human review, prioritising the positive/neutral boundary where both models concentrate their errors. Intensity is annotated manually for all segments: ordinal DS silver is suitable as pre-annotation but recall at levels 2–3 is too low ( $\approx 28.6\%$  for level 3) to trust without human verification. Throughout, silver and human-reviewed labels are kept separate to preserve downstream flexibility.

**Transferability caveat.** The 0.7 threshold was estimated on the same 500 segments used for all comparisons. It is a working estimate requiring validation as human review of the remaining segments produces additional evidence. Researchers applying this framework to other corpora should construct a domain-specific gold sample before adopting these thresholds.



## 8. Discussion

Lexicons and LLMs fail in opposite directions — simplified FEEL achieves 95.8% precision on explicit affect but produces no signal for 76% of segments, while LLMs cover that gap at the cost of systematic over-neutralization. These complementary weaknesses point toward a hybrid pipeline: use the simplified lexicon for overt affect, apply LLMs to zero-hit segments, flag disagreements for expert review, and annotate intensity manually throughout.

Beyond the present corpus, our results call into question a common assumption in LLM annotation studies: that high inter-LLM agreement (or high DS confidence) reliably signals high annotation quality. This assumption can fail systematically when multiple prompts of the same model share architectural biases — a plausible scenario in practice, since prompt variation does not eliminate pretraining biases. In our experiments, the 19.7-point overestimation for GPT-5.2 is large enough to reverse model ranking and materially change conclusions about annotation quality. We therefore recommend that LLM annotation studies include at least one independent external human validation set, however small, to guard against shared biases invisible to internal metrics (Pangakis et al., 2023). Confidence calibration curves against ground truth should be reported alongside internal agreement metrics (Guo et al., 2017; Kadavath et al., 2022), and DS consensus is better treated as an upper bound on LLM quality than as a direct estimate, particularly when annotators are known or suspected to share biases (Paun et al., 2018).

### 8.1. Limitations

1. **Gold standard size.** 500 segments represent 23.7% of the corpus. Per-stratum performance should be interpreted with appropriate confidence intervals.
2. **Annotator heterogeneity.** Annotator 3 was outvoted on 76 segments during disagreement resolution, producing a gold that reflects two more liberal annotators’ interpretation of implicit affect. Applications requiring conservative annotation of the positive/neutral boundary should treat the gold as an optimistic estimate.
3. **Stratification informed by LLM predictions.** The label-proportional stratum (30%) reflects the LLM-predicted distribution rather than the true distribution. Post-hoc comparison confirms systematic differences (gold: 68.4% positive vs. LLM-predicted 53.7%), meaning the gold sample slightly under-represents easy positive segments—a conservative bias for LLM evaluation.
4. **Dawid–Skene independence assumption.** Three prompts per model share architectural biases, violating conditional independence. We quantify rather than eliminate this limitation, and recommend alternative aggregation methods (e.g., item-response theory) for future work.
5. **Confidence threshold estimated on gold segments.** The 0.7 threshold was derived from the same 500 segments used for all comparisons. It is a working estimate requiring validation on held-out data.
6. **Zero-shot prompting.** The LLM annotation uses zero-shot prompts, consistent with realistic large-scale annotation practice. Whether few-shot or retrieval-augmented setups would reduce or merely displace the shared biases documented here remains an open question.
7. **Domain and language specificity.** Findings apply to French child language about family; generalization to other domains, languages, or age groups requires separate validation.

## 9. Conclusion

We set out to answer a practical question—which annotation strategy to use for a 2,106-segment corpus of French child language—and uncovered a methodological problem that extends beyond our corpus.

Three findings stand out. First, DS consensus is not a substitute for human validation: Dawid–Skene overestimates GPT-5.2 valence accuracy by 19.7 points due to shared neutralization bias across all three GPT prompts, while simultaneously underestimating Claude Sonnet 4 by 6.8 points. The gap reverses model ranking. Second, aggregation algorithm is not the bottleneck: Majority Vote, Dawid–Skene, and MACE converge at 98.6% agreement—better algorithms cannot compensate for biased annotators. Third, implicit affect dominance: 76% of transcribed segments require context-aware methods; lexicons alone are insufficient, but LLMs’ conservative bias toward neutral precisely targets these implicit segments.

Practically, we adopt MV valence labels for the full corpus (vote confidence threshold 0.7 for human review prioritisation), DS for per-annotator bias analysis, and ordinal DS silver labels as pre-annotation for intensity with human review at levels 2–3. We recommend that any LLM annotation study include at least a small external human validation set, and treat Dawid–Skene confidence as an internal aggregation score rather than calibrated quality evidence.

The validated 500-segment gold standard and full 2,106-segment probabilistic corpus will be re-

leased to support future research on emotion expression in child language.

## Ethics and Reproducibility

We use a publicly available anonymized corpus of interview transcripts with minors (Bellachhab et al., 2025). The data contain no personally identifiable information, and our annotations target text properties (affective content) rather than individual assessment. The corpus and collection protocol are released under CC-BY-4.0, which supports reproducibility.

## 10. Bibliographical References

Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34:555–596, 2008.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

A. Philip Dawid and Allan M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C*, 28:20–28, 1979.

Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382, 1971.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd-workers for text-annotation tasks. arXiv:2303.15056, 2023.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML 2017)*, pages 1321–1330. PMLR, 2017.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia, 2013. Association for Computational Linguistics.

Saurav Kadavath et al. Language models (mostly) know what they don’t know. arXiv preprint arXiv:2207.05221, 2022.

Klaus Krippendorff. Computing Krippendorff’s alpha-reliability. Departmental Papers (ASC), Annenberg School for Communication, University of Pennsylvania, 2011.

Chao Liu and Charis Chen. Text mining and sentiment analysis: A new lens to explore the emotion dynamics of mother–child interactions. *Social Development*, 33, 2024. <https://doi.org/10.1111/sode.12733>.

Minxue Niu, Mimansa Jaiswal, and Emily Mower Provost. From text to emotion: Unveiling the emotion annotation capabilities of LLMs. In *Proceedings of Interspeech 2024*, pages 2650–2654, 2024. <https://doi.org/10.21437/Interspeech.2024-2282>.

Minxue Niu, Yara El-Tawil, Amrit Romana, and Emily Mower Provost. Rethinking emotion annotations in the era of large language models. *IEEE Transactions on Affective Computing*, 16(04):2668–2679, 2025.

Marissa Ogren and Catherine M. Sandhofer. Emotion words in early childhood: A language transcript analysis. *Cognitive Development*, 60:101122, 2021. <https://doi.org/10.1016/j.cogdev.2021.101122>.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. Automated annotation with generative AI requires validation. arXiv preprint arXiv:2306.00176, 2023. <https://doi.org/10.48550/arXiv.2306.00176>.

Silviu Paun, Bob Carpenter, Jon Chamberlain, Dirk Hovy, Udo Kruschwitz, and Massimo Poesio. Comparing bayesian models of annotation. *Transactions of the Association for Computational Linguistics*, 6:571–585, 2018.

Barbara Plank. The “problem” of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, 2022. <https://doi.org/10.18653/v1/2022.emnlp-main.731>.

Peng Qi et al. Stanza: A python NLP toolkit for many human languages. In *Proceedings of ACL 2020: System Demonstrations*, pages 101–108, 2020.

Kenji Sagae, Eric Davis, Alon Lavie, Brian MacWhinney, and Shuly Wintner. Morphosyntactic annotation of CHILDES transcripts. *Journal of Child Language*, 37(3):705–729, 2010.

Woosuk Seo, Chanmo Yang, and Young-Ho Kim. Chacha: Leveraging large language models to prompt children to share their emotions about personal events. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–20. ACM, 2024. CHI ’24.

Carlo Strapparava and Rada Mihalcea. Semeval-2007 task 14: Affective text. In *Proceedings of SemEval-2007*, pages 70–74, 2007.

Petter Törnberg. ChatGPT-4 outperforms experts and crowd workers in annotating political Twitter messages. *arXiv:2304.06588*, 2023.

Alexandra Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470, 2021. <https://doi.org/10.1613/jair.1.12752>.

**P2 (implicit):** “Annotate based on the child’s expressed feeling/attitude in the text, even if implicit.”

**User message (per segment).**

**JSON schema (E1 fields used in this paper).**

## 11. Language Resource References

Amine Abdaoui, Jérôme Azé, Sandra Bringay, and Pascal Poncelet. FEEL: a french expanded emotion lexicon. *Language Resources and Evaluation*, 51(3):833–855, 2017.

### A. Prompt Design and Schema

We used a shared system prompt combined with three posture variants in order to induce controlled diversity across annotations while preserving a common task definition. Each segment was submitted with the same structured user message and the same JSON schema constraints.

Example	Gold valence	Why standard lexicons fail / How affect is expressed
<i>qui prend soin de nous</i> [who takes care of us]	positive	Implicit affect: no affective word appears, but “taking care” pragmatically implies trust and safety. FEEL Simplified yields 0 hits and therefore defaults to neutral.
<i>si tu es un garçon ton papa te comprend mieux</i> [if you are a boy, your dad understands you better] (a girl about “Father”)	negative	The conditional implies exclusion: the child suggests that, as a girl, she is less understood. No explicit negative lexical cue is present.
<i>on n’est pas d’accord parfois</i> [sometimes we disagree]	mixed	The qualifier <i>parfois</i> [sometimes] weakens absolute disagreement and shifts interpretation toward a conditional or mixed valence. This requires pragmatic reasoning rather than lexical matching.
<i>qui va au magasin</i> [who goes to the store]	neutral	This is a purely descriptive action. FEEL Complete incorrectly treats <i>aller</i> [to go] and <i>magasin</i> [store] as positive, producing a false positive.

Table 13: Illustrative examples of affect expression in child language about parents

### C. Examples of Affect Expression

Table 13 illustrates several cases in which affect is conveyed through implicature, syntax, and pragmatics rather than through explicit emotion words. The fourth example further shows that expanding the lexicon may introduce false positives.

# EVOKE: Emotion Vocabulary Of Korean and English

Yoonwon Jung<sup>1</sup>, Hagyeong Shin<sup>2</sup>, Benjamin K. Bergen<sup>1</sup>

<sup>1</sup>Department of Cognitive Science, University of California San Diego, USA

<sup>2</sup>Work done at Department of Linguistics, University of California San Diego, USA  
{y5jung, hashin, bkbergen}@ucsd.edu

## Abstract

This paper introduces EVOKE (Emotion Vocabulary of Korean and English), a Korean-English parallel dataset of emotion words. The dataset offers comprehensive coverage of emotion words in each language, in addition to many-to-many translations between words in the two languages and identification of language-specific emotion words. The dataset contains 1,426 Korean words and 1,397 English words, and we systematically annotate 819 Korean and 924 English adjectives and verbs. We also annotate multiple meanings of each word and their relationships, identifying polysemous emotion words and emotion-related metaphors. The dataset is, to our knowledge, the most systematic and theory-agnostic dataset of emotion words in both Korean and English to date. It can serve as a practical tool for emotion science, psycholinguistics, computational linguistics, and natural language processing, allowing researchers to adopt different views on the resource reflecting their needs and theoretical perspectives. The dataset is publicly available at <https://github.com/yoowonj/EVOKE>.

**Keywords:** emotion words, cross-linguistic dataset, annotations, polysemy, metaphors

## 1. Introduction

Emotion words reveal how humans conceptualize and communicate emotional experiences through language (Wierzbicka, 1999). They also provide a foundation for lexicon-based sentiment analyses and emotion detection (Liapis et al., 2025; Raji and De Melo, 2020). Yet, emotion words are diverse and notoriously difficult to identify (Fehr and Russell, 1984; Wierzbicka, 1999; Wu and Zhang, 2025). While some words transparently denote emotional states (e.g., *sadness*, *anger*), others evoke or reflect bodily sensations, cognitive evaluations, or physiological or behavioral expressions that may or may not indicate emotional states proper (e.g., *tense*, *captivating*, *blushing*). Consequently, existing datasets of emotion words vary in how they define and annotate emotion words, reflecting a range of theoretical positions and methodological frameworks (Clore et al., 1987; Fehr and Russell, 1984; Hong and Jeong, 2009; Johnson-Laird and Oatley, 1989; Lee, 2006; Park and Min, 2005).

A maximally useful dataset of emotion words in a language would have a few features. First, it would be comprehensive—covering all words that could be, on any theory, emotion words. Second, it would explicitly identify selection criteria that establish whether words are or are not emotion terms, and would allow users to apply their preferred criteria to create bespoke theory-driven views on the data. However, existing datasets generally cover subsets of the possible emotion words in any language. Moreover, they often apply different criteria to identify emotion words based on specific theoretical frameworks, which makes it challenging to integrate or compare datasets (Fehr and Russell,

1984; Baron-Cohen et al., 2010; Clore et al., 1987; Hong and Jeong, 2009; Johnson-Laird and Oatley, 1989; Lee, 2006; Park and Min, 2005).

Cross-linguistic datasets extend the value of monolingual resources, as comparisons of emotion words can reveal emotions that may be shared or specific across cultures (Jackson et al., 2019; Russell, 1991). To date, however, the coverage of words in cross-linguistic datasets is limited. Because meanings do not always translate one-to-one across languages (Catford, 1965; Thompson et al., 2020), a word in one language may not have any direct translation in other languages (and thus create lexical gaps), or might align with multiple words in another language. For example, while English *dismay* does not have an exact translation in Korean, English *sad* can correspond to multiple, subtly distinct Korean words (i.e., *서럽다* (*söröpta*)<sup>1</sup> and *슬프다* (*sulpäta*)<sup>2</sup>, each of which also has several different English correspondents of its own. This underscores the need for cross-linguistic datasets of emotion words with comprehensive lexical coverage in each of the languages and many-to-many translational mappings.

Lastly, emotion words are complex in that they are often polysemous. For example, Korean *부끄럽다* (*pukküröpta*) encompasses the meaning of both *shy* or *shameful*. Other polysemous senses are often metaphorical. One example is the emotional and spatial senses of *low* in English. Metaphorical relations among senses of emotion words,

<sup>1</sup>To be sad and depressed (National Institute of Korean Language, 2025a).

<sup>2</sup>Sad and sorrowful enough to make one cry (National Institute of Korean Language, 2025b).



and polysemy more generally, reveal how emotion semantics are organized within and across languages (Kövecses, 2003; Sauciuc, 2009), yet they are rarely documented in emotion word datasets.

In light of these gaps, we present a dataset of emotion words in two languages with (1) comprehensive lexical coverage and annotations by native speakers based on several different theories to accommodate diverse views and uses, (2) many-to-many translational mappings established by bilingual experts using existing lexicographic resources, and (3) identification of polysemous words and metaphorical relations among word senses.

## 2. Background

### 2.1. Theories Defining Emotion Words

Psychological and linguistic work has produced varied methods to define emotion words. Some early studies relied on unguided intuition, with raters judging words as emotional or not without explicit criteria (Fehr and Russell, 1984; Storm and Storm, 1987). More recent work uses acceptability judgments on emotion words, placing them in frame sentences for expressing emotions like “I feel X” or “I am X” to evaluate whether they sound acceptable in those contexts (Baron-Cohen et al., 2010; Clore et al., 1987; Johnson-Laird and Oatley, 1989). Another recent approach uses exclusion criteria, asking if a word describes bodily sensation or epistemic state without emotionality, to filter out some obvious non-emotion words (Baron-Cohen et al., 2010; Johnson-Laird and Oatley, 1989; Park and Min, 2005).

Research on Korean emotion words classifies emotion-related predicates into different types: state-oriented (a state experienced by an animate subject), evaluation-oriented (properties of external entities), and expression-oriented (behavioral expressions and responses) (Hong and Jeong, 2009; Lee, 2006; Park and Min, 2005). The distinction between stative and evaluative predicates is closely tied to causality, echoing the proposal to distinguish externally caused emotions (e.g., “I am saddened by his death”) from those concerning internal goals (e.g., *desire*) (Johnson-Laird and Oatley, 1989).

### 2.2. Cross-linguistic Studies

The majority of existing cross-linguistic studies of emotion words prioritize the prototypicality of emotion over exhaustiveness, focusing on small sets of emotion terms presumed to be universal (typically fewer than 30) across large numbers of languages (e.g., List et al., 2018), or utilize restricted sets of words from bilingual studies (e.g., Bromberek-Dyzman et al., 2021; Tang et al., 2023). These approaches limit the exploration of finer-grained distinctions of emotion words within a language.

Another line of research highlights culture-specific emotion words, captured as words lacking equivalents in other languages (i.e., *lexical gaps*). For example, Korean words *정* (*chöng*) or *한* (*han*) lack equivalents in English, and are expressible only with longer descriptions, like “long-lasting affection and caring based on perceived closeness” (Choi and Choi, 2002) or “Unresolved resentment, grief, and anger, or a negative and long-lasting emotion encapsulating the grief of historical memory” (Kim, 2017), respectively. Cross-linguistic lexical gaps signal how different language communities represent and express emotions differently (Lomas, 2018; Lupyan, 2012; Rissman et al., 2023; Winawer et al., 2007). However, they are often discovered in small numbers through case studies (e.g., Choi and Choi, 2002; Kim, 2017; Schmidt-Atzert and Park, 1999), and are rarely identified at scale within a broader lexicon of emotion in one language in relation to other languages.

### 2.3. Semantic Relations and Metaphors

The meanings of polysemous emotion words are often connected through metaphorical extension from one meaning to other meanings (Bartsch, 2002). Comparing patterns of such emotion-related metaphors across languages can reveal how emotional meanings are structured similarly or differently across cultures (Kövecses, 2003). Identifying these multiple senses thus provides a basis for cross-linguistic studies on emotion semantics.

Existing annotation work on polysemous words mainly targets resolving general lexical ambiguities, not emotion words specifically (Haber and Poesio, 2024; Passonneau et al., 2010; Rumshisky and Batiukova, 2008). Similarly, cross-linguistic studies of emotion-related metaphors often focus on a small set of recurrent metaphorical mappings (e.g., “emotion as force” or “anger as a container”; Lakoff and Kövecses, 1987; Sauciuc, 2009; Türker, 2013), rather than examining metaphorical structures across a wider variety of emotion words. These gaps leave open questions about how emotion words develop multiple meanings and metaphorical extensions, and how such patterns distribute within and across languages.

## 3. Objective of the Dataset

We introduce EVOKE (Emotion Vocabulary Of Korean and English), a Korean-English parallel dataset of emotion words with comprehensive coverage in both languages plus cross-linguistic mappings (publicly available at <https://github.com/yoonwonj/EVOKE>). The dataset is constructed by compiling lists of words to annotate from prior studies on and datasets of emotion

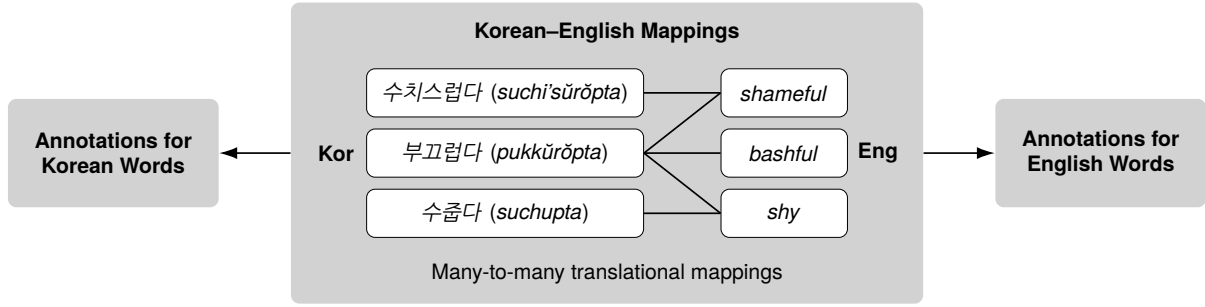


Figure 1: The structure of the Korean-English parallel emotion word dataset. Words in both languages are connected through many-to-many translational mappings, with annotations for Korean and English words.

words. Those words are then annotated according to multiple criteria used in earlier studies to investigate their characteristics as emotion words. The dataset also provides many-to-many translations of words between Korean and English, which enables discovering emotion concepts specific to each language. Lastly, the dataset identifies polysemous words and annotates the relationships between their senses, enabling the identification of emotion-related metaphors.

We focus on Korean and English for several reasons. English serves as a well-studied reference point, yet there does not yet exist a comprehensive resource with structured annotations like the one constructed here. Korean, meanwhile, is spoken in distinct cultural settings, and contains purportedly unique, culture-specific emotion terms (e.g., *정* (*chöng*), *한* (*han*)) that lack direct English equivalents (Choi and Choi, 2002; Kim, 2017). Comparison of the two allows for the investigation of both shared and culture-specific aspects of emotion lexica. Additionally, the research team’s expertise in linguistics, cognitive science, and knowledge as native speakers of English and Korean makes it feasible to construct a reliable dataset.

## 4. Dataset Construction

The dataset consists of three separate components (see Figure 1): (1) Korean-English mappings, (2) annotations for Korean words, and (3) annotations for English words.

### 4.1. Word Selection and Translation

Candidate words were gathered from previous work on emotion words in English (Baron-Cohen et al., 2010; Morgan and Heise, 1988; Storm and Storm, 1987) and Korean (Jeon et al., 2022; Park and Min, 2005; Rhee and Ko, 2013).

To construct translational mappings (illustrated in Figure 1), two Korean-English bilingual speakers judged whether each word had translational equivalents in the other language. A manual translation

approach was adopted to capture nuanced and precise meanings that are difficult to obtain through automatic translation. The translators consulted two bilingual Korean-English dictionaries (one Korean-to-English and one English-to-Korean) for translation (National Institute of Korean Language, 2025c; NAVER, 2025). In addition, definitions from Korean (National Institute of Korean Language, 2025e) and English monolingual dictionaries (Oxford University Press, 2025; Cambridge University Press, 2025) were compared to ensure accurate and semantically rich matching, especially when translations provided in bilingual dictionaries varied across sources or when forward and backward translation yielded inconsistent results.

A word is considered to have a translational equivalent when it can be translated to a single word form (i.e., one or more words exist in both languages to denote the same meaning). Accordingly, lexical gaps were identified when (1) bilingual dictionary queries for words in one language only yielded multi-word expressions or idiomatic phrases in the other language,<sup>3</sup> or when (2) back-translation produced inconsistent results. In the latter cases, the translators consulted additional native speakers and monolingual dictionaries to determine whether the dictionary equivalent reflected a true translational match that fully captured the meaning of the query word.

Moreover, lexical gap was defined as the absence of a word for a concept in one language when such a word exists in the other (e.g., Janssen, 2004; Li et al., 2024). Therefore, gaps arising at the morphosyntactic level were not included, as those do not indicate the absence of conceptual or semantic content (Ivir, 1977; Lomas, 2018; Wierzbicka, 1999).<sup>4</sup>

<sup>3</sup>Hyphenated forms were treated as single lexical items (e.g., *cliché-ridden*). Korean words that may appear either as single words or with internal spaces were treated as single-word entries and were not considered as lexical gaps (e.g., *사려 깊은* (*saryökipün*) ‘considerate’).

<sup>4</sup>Some Korean verbs and nouns are systematically

ID	Label	Annotation Question
<b>Part 1. Acceptability judgments</b>		
ACPT1	1st person “feel”	Does “I feel X” sound acceptable?
ACPT2	3rd person plural “feel”	Does “They feel X” sound acceptable?
ACPT3	Inanimate subject “feel”	Does “It feels X” sound acceptable (inanimate “it”)?
ACPT4	1st person “am”	Does “I am X” sound acceptable?
<b>Part 2. Semantic experienter judgments</b>		
EXP5	Subjectivity of experience	In “It feels X”, can the sentence express what “it” experiences?
EXP6	Evaluation as experience	In “It feels X”, can the sentence express my evaluation of “it”?
EXP7	Caused experience	In “I am X”, does X denote a caused state of its associated noun/verb?
EXP8	Causing experience	In “I am X”, does X denote a causing state of its associated noun/verb?
<b>Part 3. Exclusion criteria</b>		
EXCL9	Pure bodily sensation	Does X describe a pure bodily/physical sensation?
EXCL10	Behavioral expression	Does X denote a behavioral expression?
EXCL11	Pure epistemic state	Does X denote a non-emotional epistemic state?
<b>Part 4. Multiple meanings</b>		
POLY12	Additional Meaning	Does X have another distinct meaning to annotate separately? → If yes: create a new row for that sense and complete annotations in Parts 1–3.
POLY13	Distinctiveness	Are the two senses of X in different domains?
POLY14	Relatedness	Are the two senses of X systematically related?

Table 1: Codebook for the annotation scheme. Part 1 collects acceptability judgments of the target word in four different sentences. Part 2 includes two follow-ups for ACPT3 (EXP5–6) and ACPT4 (EXP7–8) each. Part 3 applies exclusion criteria to filter relatively obvious non-emotion terms. Part 4 documents polysemy and links senses via distinctiveness and relatedness judgments. All questions used two binary labels (*acceptable/unacceptable* in Part 1, *yes/no* in Parts 2–4), with an additional option to choose *unsure*. ACPT1, ACPT2, ACPT3, EXP5, and EXP6 were only applied to adjectives, and the rest of the others were applied to both verbs and adjectives, with modifications for verb annotations. Annotation criteria were translated into Korean for Korean word annotations with appropriate modifications applied.

## 4.2. Annotation Objective and Process

We annotated words whose part-of-speech is identified as adjectives and verbs in each language. Adjectives linked to identifiable lexical roots<sup>5</sup>, either nouns or verbs, were annotated as noun–adjective or verb–adjective pairs. Adjectives without such identifiable lexical roots, along with all verbs, were annotated independently. Nouns were not annotated because the sentence contexts most commonly used for acceptability judgments in the literature are designed around predicative uses of words (Baron-Cohen et al., 2010; Clore et al., 1987; Johnson-Laird and Oatley, 1989), and several existing annotation criteria target semantic properties specific to predicates (see Section 4.3.2; Hong and Jeong, 2009; Lee, 2006; Park and Min, 2005).

translated into English using *be-* constructions (verb) and *being-* constructions (noun). This creates gaps in verbs and nouns for English relative to Korean in a systematic way. See Bentivogli et al. 2000 for additional examples and discussion of morphosyntactic gaps.

<sup>5</sup>A form commonly perceived as a base form without its morphological derivation.

We recruited three native English speakers and three native Korean speakers as annotators<sup>6</sup>, who had backgrounds in cognitive science or linguistics. Annotators completed one week of training followed by ten weeks of annotation. During weekly meetings, researchers and annotators discussed questions and edge cases. Those discussions resulted in modifications to the annotation guidelines meant to improve the robustness of subsequent annotations.

Annotators made binary judgments (*acceptable/unacceptable*; *yes/no*) for each annotation criterion, with the option to choose *unsure*. We chose binary judgments over graded ratings because some of the questions are intrinsically categorical. Binary ratings also avoid subjectivity introduced by continuous scales (e.g., differential use of Likert scales per annotator) and allow for easier combination of annotation results from the different criteria to select words that qualify as emotion words (see Section 6.1 for an example of a use case).

<sup>6</sup>Annotators received course credits as compensation.

Items marked `unsure` were reviewed in weekly meetings. Cases that could be resolved through group discussion were annotated again with binary labels. However, given the subjective nature of emotion word judgments, we did not force consensus. When annotators continued to find it difficult to make a binary decision, we kept the `unsure` label.

The dataset was split so that each annotator received a unique 30% of the words, while an additional 10% of the words were assigned to all three annotators for inter-rater reliability assessment.

### 4.3. Annotation Criteria

The annotation criteria used features of emotion words identified from the existing literature. Those criteria were selected to include or exclude candidate words from the category of emotion words. The full list is presented in Table 1.

Adjectives were annotated on all 14 criteria, while verbs were annotated on 9, excluding ACPT1, ACPT2, ACPT3, EXP5, and EXP6. In addition, ACPT4 was modified to suit part-of-speech differences. Because the criteria differ by language and by part-of-speech, Table 1 reports the annotation scheme for English adjectives, which encompasses all 14 of the criteria we devised. For Korean annotations, the criteria were translated and adapted to Korean syntax and usage patterns. Details of the modifications to Korean word annotations are in Appendix A.

#### 4.3.1. Part 1: Acceptability Judgments

The first part of the annotation collected acceptability judgments to elicit native speakers' usage patterns of the words.<sup>7</sup> Annotators judged whether target words sound `acceptable` or `unacceptable` in given sentences, with an option to choose `unsure`. The sentences served as standardized contexts for evaluating each emotion word's appropriateness for expressing emotion. When the word alone sounded unnatural, annotators could suggest an alternate phrasal structure, which was then adopted for the annotation (e.g., "I feel cared" could be replaced by "I feel cared for").<sup>8</sup>

The sentences for acceptability judgments were selected based on their previous adoption as inclusion criteria for emotion words (Baron-Cohen et al., 2010; Clore et al., 1987; Johnson-Laird and Oatley, 1989), from psychological and linguistic studies of emotion (Hong and Jeong, 2009; Liu, 2016; Lee, 2006; Niedenthal, 2008; Paul et al., 2020).

**ACPT1: 1st person "feel", ACPT2: 3rd person plural "feel", ACPT3: Inanimate subject "feel"**

<sup>7</sup>We ground the acceptability judgments in pragmatic acceptability instead of grammatical acceptability.

<sup>8</sup>This was in response to differences between English, which commonly uses phrasal verbs, and Korean, which instead tends to use morphological derivation.

Three sentence contexts based on "I feel X" but with varying subjects were used to assess whether words activate subjective feelings of emotion (Lee, 2006; Niedenthal, 2008; Paul et al., 2020): "I feel X" (ACPT1: 1st person "feel"), "They feel X" (ACPT2: 3rd person plural "feel"), and "It feels X" (ACPT3: Inanimate subject "feel"), where X is a target word. Participants were instructed to interpret "I" and "they" as animate experiencers (1st person and 3rd person plural), and "it" as an inanimate object. If a word denotes a subjective feeling, it is expected to be judged as acceptable with animate subjects, but not with inanimate ones (Baron-Cohen et al., 2010; Clore et al., 1987; Dowty, 1991; Lee, 2006; Liu, 2016).

**ACPT4: 1st person "am"** "I am X" (ACPT4: 1st person "am") was used to contrast with "I feel X" (1st person "feel") and capture words denoting states versus evaluations. For example, *caring* fits "I am caring" but not "I feel caring". This contrast helps reveal whether words refer to internal states or external evaluations. However, some words (e.g., *pretty*) can still be acceptable in both "I feel X" and "I am X". Follow-up questions in Section 4.3.2 further clarify these cases (Clore et al., 1987; Hong and Jeong, 2009; Lee, 2006; Park and Min, 2005).

#### 4.3.2. Part 2: Semantic Experiencer Judgments

The second part introduced semantic experiencer judgments as follow-up questions to ACPT3 and ACPT4. These questions aimed to handle patterns in acceptability judgments like those for *pretty*. Specifically, these questions helped determine whether words describe subjective experiences or external evaluations, reflecting the contrast between stative and evaluative words (Choi, 2008; Hong and Jeong, 2009; Lee, 2006).

**EXP5: Subjective experience, EXP6: Evaluation** Annotators judged whether the word in "It feels X" (ACPT3: Inanimate subject "feel") could express what the inanimate subject "it" experiences (EXP5), and whether it could express the speaker's evaluation of "it" (EXP6). These follow-up judgments clarify whether ACPT3 acceptability stems from evaluative readings. For instance, "it feels merciless" may sound acceptable if "it" refers to a fighter jet, but such usage reflects evaluation rather than the inanimate subject's feelings. These follow-up questions ensure the frame captured the intended semantics and help sort out such exception cases.

**EXP7: Caused, EXP8: Causing** Follow-up questions for ACPT4 (1st person "am") were designed to complement the distinction between experiential and evaluative words, following the "caused emotion" distinction of Johnson-Laird and Oatley (1989). Annotators were asked to judge whether the target word describes a "caused" (EXP7) or "caus-



ing” (EXP8) state within the “I am X” sentence context. Korean annotators were allowed to adjust the subject to consider additional naturalistic contexts, incorporating the annotators’ suggestions for more comprehensive evaluation.

To guide these judgments, adjectives were paired with their identifiable root nouns or verbs (e.g., *sadness*–*sad*), and were asked to judge whether each adjective is a caused state or a cause of the paired root nouns or verbs. For example, annotators judged whether *sad* refers to the caused state of *sadness* or its causing state. For the remaining adjectives, annotators were asked to imagine hypothetical noun forms.

These causality judgments clarify whether a word denotes an internal state or an external evaluation: internal states typically correspond to caused forms (e.g., *depressed*), while evaluations often align with causing forms (e.g., *depressing*). A word may encompass both caused and causing meanings, or words may represent each meaning separately. Moreover, words with similar meanings in different languages may capture these caused and causing meanings differently. Such differences could help reveal varying semantic properties of emotion words across languages.

#### 4.3.3. Part 3: Exclusion Criteria

The third part applied exclusion criteria (Park and Min, 2005; Baron-Cohen et al., 2010; Lee, 2006). While emotions involve coordinated mind–body processes, most theories distinguish them from pure bodily sensations (e.g., *thirsty*) or cognition (e.g., *enlightened*) (Anderson and Adolphs, 2014; Johnson-Laird and Oatley, 1989; Lee, 2006; Paul et al., 2020). Many also separate emotional states from physical expressions (e.g., *cry*), which are byproducts rather than core states (Anderson and Adolphs, 2014; Johnson-Laird and Oatley, 1989; Hong and Jeong, 2009; Lee, 2006; Paul et al., 2020). These judgments are used to guide the exclusion of non-emotion words.

Annotators judged whether each target word describes pure bodily or physical sensations (EXCL9), behavioral expressions (EXCL10), or pure epistemic states (EXCL11). These criteria were designed to flag words that primarily fell into these categories for potential exclusion from the set of emotion words. This procedure provides a direct way to filter out non-emotion terms according to widely shared theoretical assumptions, still allowing researchers to retain those as emotion words if desired.

#### 4.3.4. Part 4: Multiple Meanings

**POLY12: Additional meaning** The fourth part of the annotation was introduced to capture whether each word had other meanings distinct from its

primary emotional meaning (POLY12). When annotators identified an additional meaning, they were instructed to complete annotations in Parts 1–3 for the additional meaning identified. To further specify semantic relationships among multiple senses, two follow-up questions assessed the relationship between the identified meanings.

**POLY13: Distinctiveness, POLY14: Relatedness** Annotators judged whether the two senses concern different domains or concepts (Distinctiveness; POLY13), and whether they are connected in a systematic way (Relatedness; POLY14). For example, the word *low* exemplifies a prototypical metaphorical relationship. Its physical and emotional meanings belong to distinct domains yet remain systematically linked through the conceptual mapping between spatial position and emotional valence. In contrast, *blue* demonstrates a looser connection in that its color and emotional senses are distinct, but their association lacks systematic grounding. Although the metaphorical status of *blue* is debatable, it does not exhibit the systematicity typical of metaphors.

- (1) a. “The ceiling is low.” → HEIGHT  
b. “I am feeling low today.” → FEELING  
c. → POLY13: yes, POLY14: yes
- (2) a. “The sky is blue.” → COLOR  
b. “I feel blue.” → FEELING  
c. → POLY13: yes, POLY14: no

These dimensions derive from conceptual metaphor theory, and specifically the Metaphor Identification Protocol (Declercq and van Poppel, 2023; Group, 2007; Kovecses, 2010), to annotate words that could have emotional meaning through metaphorical extension (e.g., *low*). The annotation results could help reveal how emotional meanings emerge and relate to other meanings.

## 5. Dataset Analysis and Evaluation

### 5.1. Word characteristics

#### 5.1.1. Part-of-Speech Statistics

Summary word statistics are in Table 2. A total of 1,426 Korean words and 1,397 English words were included. Translation equivalents were identified for all of these, forming many-to-many mappings across Korean and English (Figure 1).

#### 5.1.2. Translational Equivalence

The majority of words had corresponding translations in the other language. For Korean words, 574 nouns (97.12%), 587 adjectives (96.86%), and 213 verbs (93.01%) had one or more translational equivalents in English. For English words, 495 nouns (97.44%), 630 adjectives (93.89%), and 213 verbs



	Nouns	Adjectives	Verbs	Total
<b>Korean words</b>	591	606	229	1,426
With translation(s) in English	574 (97.12%)	587 (96.86%)	213 (93.01%)	1,374 (96.35%)
Without translation(s) in English	17 (2.88%)	19 (3.14%)	16 (6.99%)	52 (3.65%)
<b>English words</b>	508	671	218	1,397
With translation(s) in Korean	495 (97.44%)	630 (93.89%)	213 (97.71%)	1,338 (95.80%)
Without translation(s) in Korean	13 (2.56%)	41 (6.11%)	5 (2.29%)	59 (4.22%)

Table 2: The statistics of Korean and English lexical entries with and without translational equivalents. Words without translational equivalents indicate lexical gaps. Polysemous words with either (1) multiple parts of speech, or (2) one of the meanings having translations in the other language but not in the other, were counted as separate lexical entries. See Appendix B for policies regarding counting polysemous word entries.

(97.71%) had one or more corresponding words in Korean.

The analysis of translational equivalents revealed a relatively similar degree of mapping complexity between the two languages. Korean words had an average of 1.61 translational equivalents in English ( $SD = 0.93$ ), and English words had a nearly identical average of 1.65 translational equivalents in Korean ( $SD = 1.15$ ). A total of 472 words in English (33.79%) and 542 words in Korean (38.01%) had two or more translational equivalents in the other languages. Some words demonstrated particularly rich translational mappings. The Korean adjective *우울하다* (*uulhata*) exemplified this complexity with six English equivalents: *depressed*, *dismal*, *gloomy*, *low*, *down*, and *blue*, capturing a range of physical and emotional states that Koreans express through a single word.

### 5.1.3. Lexical Gaps

Among Korean words, 17 nouns (2.88%), 19 adjectives (3.14%), and 16 verbs (6.99%) lacked translational equivalents in English. In English, 13 nouns (2.56%), 41 adjectives (6.11%), and 5 verbs (2.29%) had no corresponding translations in Korean.

These lexical gaps may indicate conceptual gaps, though this is not always the case. For example, some of those words indicate areas where each language expresses emotion concepts uniquely, like *답답하다* (*taptaphata*)<sup>9</sup> (Schmidt-Atzert and Park, 1999). No single English word is equivalent, nor is there a conventionalized phrasal way to express the concept it denotes. However, other lexical gaps do have phrasal equivalents, like *moody*, which has a multi-word equivalent to “having ups and downs”

<sup>9</sup>Feeling dissatisfied and uncomfortable when someone’s behavior or the situation makes it hard to meet one’s expectation (National Institute of Korean Language, 2025d; Schmidt-Atzert and Park, 1999).

(기분 변화가 심한 (*kipun byŏnhwaka simhan*)) in Korean. Systematically distinguishing these instances from each other is beyond the scope of this paper. However, because the identified lexical gaps include both cases, this dataset can be utilized to further explore empirical questions regarding the relationship between lexical gaps and conceptual gaps and the effects—if any—of the conventionality of multi-word descriptions for emotion concepts.

## 5.2. Annotation Results

A total of 819 Korean words (602 adjectives, 217 verbs) and 923 English words (678 adjectives, 245 verbs) were annotated.<sup>10</sup>

Annotation criteria differed by part of speech, which made averaging results across adjectives and verbs infeasible. We thus report results for adjectives, which constitute the largest portion of the dataset, and use all criteria (see Section 4.3 for details). Results of verb annotations are provided in Appendix C, and Figs. 5 to 7.

### 5.2.1. Agreement Scores

Agreement scores were calculated for the annotation criteria in Parts 1–3, as not all words had multiple meanings to be annotated. See Figures 3 and 4 in Appendix for the full pairwise annotation scores.

Overall, Korean annotations showed higher reliability (mean Cohen’s  $\kappa = 0.74$ ) than English annotations (mean  $\kappa = 0.60$ ) in average. This pattern may reflect greater consistency in the features of Korean emotion words or, alternatively, additional training adjustments introduced for the Korean word annotation after the annotation of English words.

<sup>10</sup>Discrepancies between the number of words annotated and those included in the translational mappings arose from post-hoc corrections to part-of-speech coding and translation mappings.

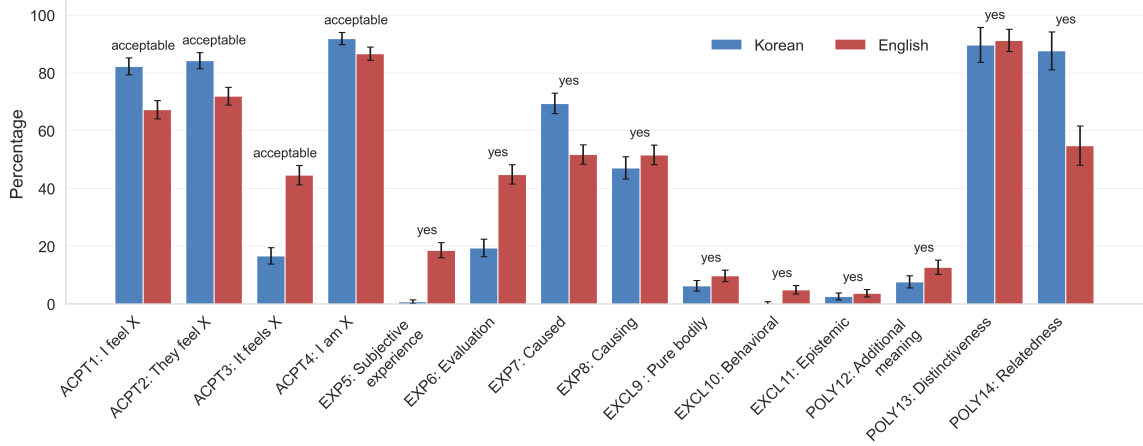


Figure 2: The percentages of annotation values for each annotation criterion for adjectives in Korean and English. The percentages for POLY12 indicate the ratio of words annotated as having more than one meaning in each language. The percentages for POLY13 and POLY14 were calculated among items annotated as having more than one meaning in POLY12. Error bars indicate 95% confidence intervals.

Agreement also varied by criterion and cross-linguistically. Exclusion criteria produced the highest agreement in both languages (mean EXCL9–11: Korean  $\kappa = 0.89$ , English  $\kappa = 0.82$ ). Causality criteria showed the lowest agreement across both languages (EXP7–8: mean Korean  $\kappa = 0.52$ , mean English  $\kappa = 0.45$ ), with the “causing” criterion (EXP8) being particularly challenging (Korean  $\kappa = 0.46$ , English  $\kappa = 0.45$ ). For the acceptability judgments (ACPT1–4), Korean annotators reached substantial agreement (mean  $\kappa = 0.73$ ), while English annotators reached moderate to substantial levels (mean  $\kappa = 0.58$ ). The sub-criteria for subjective experience and evaluation (EXP5–6) achieved higher agreement in Korean (mean  $\kappa = 0.76$ ) than in English (mean  $\kappa = 0.47$ ).

These inter-annotator agreement patterns highlight both the subjective nature of emotion concepts and the theoretical disagreements that surround them. In particular, the relatively low agreement on causality judgments likely reflects the abstract and complex reasoning required. Annotators likely relied on their subjective judgments that relied on personal experiences and their language use. As such, criteria like EXP7 (“Caused”) and EXP8 (“Causing”) should be applied cautiously in making inclusion or exclusion decisions when defining emotion words.

### 5.2.2. Annotation Trend

A summary of annotation trends for all annotated adjectives is presented in Figure 2. Annotations in the agreement set were adjudicated using majority vote to produce a single combined dataset for analysis.

Words in both languages exhibited higher acceptance in “I am” (ACPT4; Korean: 91.83%, English: 86.85%) than in “I feel” (ACPT1; Korean: 82.13%,

English: 67.23%) or in “They feel” (ACPT2; Korean: 84.28%, English: 71.91%) sentences. Most words were annotated as unacceptable in “It feels” (ACPT3) in Korean (83.36%), but not in English (55.14%).

Responses for causality criteria were more evenly distributed in both languages (EXP7 Korean: 69.18% yes, English: 51.56% yes; EXP8 Korean: 46.69% yes, English: 51.82% yes). Exclusion criteria (EXCL9–11) showed a highly consistent rejection rate (around 90%) in both languages.

More English adjectives had multiple meanings annotated (12.68%) than in Korean (7.48%). Cross-linguistic differences also appeared in relatedness judgments (POLY14; Korean 87.37%, English 55.00%), whereas distinctiveness judgments converged (POLY13; Korean 89.47%, English 92.22%). Annotation of additional meaning was correlated positively with bodily meanings in both languages, though more strongly in Korean ( $r = 0.46$ ) than English ( $r = 0.26$ ), suggesting a pattern of metaphorical extension from physical to emotional domains.

## 6. Future Applications

### 6.1. Theory-Driven Selection of Words

This dataset is theory-agnostic, as it does not commit to a single theoretical stance on how emotion words should be defined. Researchers can apply their preferred annotation criteria to select emotion words that align with their theoretical perspectives and research needs, enabling the curation of experimental stimuli for both emotion research and computational studies of emotion word semantics.

A conservative approach to selecting words from the dataset applies decision criteria from prior studies. Specifically, (1) a word should be acceptable in

“I feel”, “They feel”, and “I am” sentences, but not in “It feels” context (Part 1; *yes* for ACPT1, ACPT2, ACPT4; *no* for ACPT3); (2) if acceptable in “It feels”, it should express human evaluation rather than an inanimate subject’s experience (Part 2; *no* for EXP5 and *yes* for EXP6); and (3) it should not denote pure bodily sensations, behavioral expressions, or non-emotional epistemic states (Part 3; *no* for EXCL9–11).

Applying these criteria yields 425 single-word emotion adjectives in Korean and 317 in English. Examples of excluded words include *mean* (못되다 (*mottwaeta*) in Korean), which feels unnatural in “I feel X”, and *clueless* (어리둥절하다 (*ëritungchölhata*) in Korean), which denotes a non-emotional epistemic state.

## 6.2. Emotion Concepts and Metaphors

Moreover, this dataset supports both computational and behavioral research on emotion concepts. Emotion word embeddings could be utilized to conduct computational analyses of lexical gaps, revealing how multilingual language models represent language-specific words. Emotion word embeddings also enable the investigation of foundational questions about emotion structure, such as whether shared dimensions (e.g., valence and arousal) organize the emotion lexicon across languages (Niedenthal, 2008; Yik et al., 2023). High-quality translation mappings provided by the dataset, validated by bilingual speakers of both languages, offer a reliable gold standard for machine learning approaches.

Behavioral studies can further examine how conceptual gaps influence the communication or interpretation of those emotions (Rissman et al., 2023) using the emotion terms in this dataset. Finally, the annotations of multiple meanings in this dataset (POLY12–14) supports cross-cultural investigations of emotion-related metaphors. Comparing the identified metaphorical extensions, researchers can assess the universality versus specificity of emotion-related metaphors (Kövecses, 2003; Sauciuc, 2009).

## 7. Conclusion

The present dataset includes 1,426 Korean words and 1,397 English words, with many-to-many translational mappings documented between the two languages. Using a feature-based annotation approach, the dataset provides a theory-agnostic set of adjectives and verbs that can be selected as emotion words based on criteria drawn from prior studies. The dataset captures both within-language and cross-linguistic features of emotion words. Beyond documenting linguistic variation, the

dataset provides several principled bases for selecting emotion words and for comparing emotion-related metaphors across languages. It can serve as a resource for linguistic research, computational analysis of emotion concepts, and empirical studies of emotion.

## 8. Ethical Considerations and Limitations

The dataset does not involve serious ethical considerations, as it contains no personal or identifiable information. All annotators were assigned anonymized coder IDs.

Despite the dataset’s contributions, this study has certain limitations. Although the dataset was constructed to cover more than one language, there are many other languages to investigate for their emotion words. Regarding the annotation, the majority of the data was annotated by a single trained annotator, and the agreement set (about 10% of adjectives and verbs) has three annotators’ judgments adjudicated. This percentage was chosen to prioritize the annotators’ effort put into the large number of emotion words in the dataset with complex annotation criteria that required specific training with adequate domain knowledge. Future research could extend the approach to a broader range of languages and include a larger proportion of items annotated by multiple annotators.

## 9. Acknowledgements

We would like to thank the annotators for their dedication and hard work. We also thank anonymous reviewers for their valuable feedback and comments.

## 10. Bibliographical References

- David J Anderson and Ralph Adolphs. 2014. A framework for studying emotions across species. *Cell*, 157(1):187–200.
- Simon Baron-Cohen, Ofer Golan, Sally Wheelwright, Yael Granader, and Jacqueline Hill. 2010. Emotion word comprehension from 4 to 16 years old: A developmental survey. *Frontiers in evolutionary neuroscience*, 2:109.
- Renate Bartsch. 2002. Generating polysemy: Metaphor and metonymy. *Metaphor and metonymy in comparison and contrast*, 20:49–74.
- Luisa Bentivogli, Emanuele Pianta, and Fabio Pianesi. 2000. Coping with lexical gaps when building aligned multilingual wordnets. In *Proceed-*

- ings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece. European Language Resources Association.
- Katarzyna Bromberek-Dyzman, Rafał Jończyk, Monica Vasileanu, Anabella-Gloria Niculescu-Gorpin, and Halszka Bąk. 2021. Cross-linguistic differences affect emotion and emotion-laden word processing: Evidence from polish-english and romanian-english bilinguals. *International Journal of Bilingualism*, 25(5):1161–1182.
- Cambridge University Press. 2025. [Cambridge english dictionary](#).
- John Cunnison Catford. 1965. *A linguistic theory of translation*, volume 31. Oxford university press London.
- Injae Choi and Sang-Chin Choi. 2002. The effects of korean cultural psychological characteristics on coping styles, stress, and life satisfaction: Centering around cheong and weness. *Korean Journal Of Counseling And Psychotherapy*, 14(1):55–71.
- SJ Choi. 2008. The type and character of feeling verb. *Journal of the Society of Korean Language and Literature*, 58:127–159.
- Gerald L Clore, Andrew Ortony, and Mark A Foss. 1987. The psychological foundations of the affective lexicon. *Journal of personality and social psychology*, 53(4):751.
- Jana Declercq and Lotte van Poppel. 2023. Coding metaphors in interaction: A study protocol and reflection on validity and reliability challenges. *International Journal of Qualitative Methods*, 22:16094069231164608.
- David Dowty. 1991. Thematic proto-roles and argument selection. *language*, 67(3):547–619.
- Beverley Fehr and James A Russell. 1984. Concept of emotion viewed from a prototype perspective. *Journal of experimental psychology: General*, 113(3):464.
- Pragglejaz Group. 2007. Mip: A method for identifying metaphorically used words in discourse. *Metaphor and symbol*, 22(1):1–39.
- Janosch Haber and Massimo Poesio. 2024. Polysemy—evidence from linguistics, behavioral science, and contextualized language models. *Computational Linguistics*, 50(1):351–417.
- Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.
- Jong-sun Hong and Yeon-ju Jeong. 2009. Establishing the category of emotion verb and classifying emotion verbs. *Korean Linguistics*, 45:387–420.
- Vladimir Ivir. 1977. Lexical gaps: A contrastive view. *Studia Romanica et Anglica Zagradiensia: Revue publiée par les Sections romane, italienne et anglaise de la Faculté des Lettres de l'Université de Zagreb*, (43):167–176.
- Joshua Conrad Jackson, Joseph Watts, Teague R Henry, Johann-Mattis List, Robert Forkel, Peter J Mucha, Simon J Greenhill, Russell D Gray, and Kristen A Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. *Science*, 366(6472):1517–1522.
- Maarten Janssen. 2004. Multilingual lexical databases, lexical gaps, and sim u lida. *International Journal of Lexicography*, 17(2):137–154.
- Duyoung Jeon, Junho Lee, and Cheongtag Kim. 2022. User guide for kote: Korean online comments emotions dataset. *arXiv preprint arXiv:2205.05300*.
- Philip Nicholas Johnson-Laird and Keith Oatley. 1989. The language of emotions: An analysis of a semantic field. *Cognition and emotion*, 3(2):81–123.
- Philip Nicholas Johnson-Laird and Peter Cathcart Wason. 1977. *Thinking: Readings in cognitive science*. CUP Archive.
- Sandra So Hee Chi Kim. 2017. Korean" han" and the postcolonial afterlives of" the beauty of sorrow". *Korean Studies*, pages 253–279.
- Zoltán Kövecses. 2003. *Metaphor and emotion: Language, culture, and body in human feeling*. Cambridge University Press.
- Zoltan Kovecses. 2010. *Metaphor: A practical introduction*. Oxford university press.
- Zoltán Kövecses. 2012. *Emotion concepts*. Springer Science & Business Media.
- Peter Kuppens, Francis Tuerlinckx, James A Russell, and Lisa Feldman Barrett. 2013. The relation between valence and arousal in subjective experience. *Psychological bulletin*, 139(4):917.
- George Lakoff and Zoltán Kövecses. 1987. The cognitive model of anger inherent in american english. In Dorothy Holland and Naomi Quinn, editors, *Cultural Models in Language and Thought*, pages 195–221. Cambridge University Press.
- Won-Kyung Lee. 2006. The classification and properties of emotive verbs in korean. *Discourse and Cognition*, 13(1):133–161.



- Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2024. [Translation-based lexicalization generation and lexical gap detection: Application to kinship terms](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6891–6900, Bangkok, Thailand. Association for Computational Linguistics.
- Charalampos M Liapis, Aikaterini Karanikola, and Sotiris Kotsiantis. 2025. Enhancing sentiment analysis with distributional emotion embeddings. *Neurocomputing*, 634:129822.
- Johann-Mattis List, Simon J Greenhill, Cormac Anderson, Thomas Mayer, Tiago Tresoldi, and Robert Forkel. 2018. Clics2: An improved database of cross-linguistic colexifications assembling lexical data with the help of cross-linguistic data formats. *Linguistic Typology*, 22(2):277–306.
- Meichun Liu. 2016. Emotion in lexicon and grammar: lexical-constructional interface of mandarin emotional predicates. *Lingua Sinica*, 2(1):4.
- Tim Lomas. 2018. Experiential cartography and the significance of “untranslatable” words. *Theory & Psychology*, 28(4):476–495.
- Gary Lupyan. 2012. Chapter Seven - What Do Words Do? Toward a Theory of Language-Augmented Thought. *Psychology of Learning and Motivation*, 57:255–297.
- Rick L Morgan and David Heise. 1988. Structure of emotions. *Social psychology quarterly*, pages 19–31.
- National Institute of Korean Language. 2025a. [Korean-english learners' dictionary](#).
- National Institute of Korean Language. 2025b. [Korean-english learners' dictionary](#).
- National Institute of Korean Language. 2025c. [Korean-english learners' dictionary](#).
- National Institute of Korean Language. 2025d. [Korean-english learners' dictionary](#).
- National Institute of Korean Language. 2025e. [Standard korean language dictionary](#).
- NAVER. 2025. [Naver korean-english dictionary](#).
- Paula M Niedenthal. 2008. Emotion concepts. *Handbook of emotions*, 3:587–600.
- Oxford University Press. 2025. [Oxford english dictionary](#).
- In-Jo Park and Kyung-Hwan Min. 2005. Making a list of korean emotion terms and exploring dimensions underlying them. *Korean Journal of Social and Personality Psychology*, 19(1):109–129.
- Rebecca J Passonneau, Ansaf Salieb-Aouissi, Vikas Bhardwaj, and Nancy Ide. 2010. Word sense annotation of polysemous words by multiple annotators. In *LREC*.
- Elizabeth S Paul, Shlomi Sher, Marco Tamietto, Piotr Winkielman, and Michael T Mendl. 2020. Towards a comparative science of emotion: Affect and consciousness in humans and animals. *Neuroscience & Biobehavioral Reviews*, 108:749–770.
- Shahab Raji and Gerard De Melo. 2020. What sparks joy: The affectvec emotion database. In *Proceedings of the web conference 2020*, pages 2991–2997.
- Shin-Young Rhee and Il-Ju Ko. 2013. Measuring a valence and activation dimension of korean emotion terms using in social media. *Science of Emotion and Sensibility*, 16(2):167–176.
- Lilia Rissman, Qiawen Liu, and Gary Lupyan. 2023. Gaps in the lexicon restrict communication. *Open Mind*, 7:412–434.
- Anna Rumshisky and Olga Batiukova. 2008. Polysemy in verbs: systematic relations between senses and their effect on annotation. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 33–41.
- James A Russell. 1991. Culture and the categorization of emotions. *Psychological bulletin*, 110(3):426.
- Gabriela-Alina Sauciuc. 2009. The role of metaphor in the structuring of emotion concepts. *Cognitive Semiotics*, 5(1-2):244–267.
- Lothar Schmidt-Atzert and Hyun-Sook Park. 1999. The korean concepts dappdaphada and uulhada: A cross-cultural study of the meaning of emotions. *Journal of Cross-Cultural Psychology*, 30(5):646–654.
- Christine Storm and Tom Storm. 1987. A taxonomic study of the vocabulary of emotions. *Journal of personality and social psychology*, 53(4):805.
- Dong Tang, Yang Fu, Huili Wang, Bo Liu, Anqi Zang, and Tommi Kärkkäinen. 2023. The embodiment of emotion-label words and emotion-laden words: Evidence from late chinese–english bilinguals. *Frontiers in Psychology*, 14:1143064.



Bill Thompson, Seán G Roberts, and Gary Lupyan. 2020. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, 4(10):1029–1038.

Ebru Türker. 2013. A corpus-based approach to emotion metaphors in Korean: A case study of anger, happiness, and sadness. *Review of Cognitive Linguistics. Published under the auspices of the Spanish Cognitive Linguistics Association*, 11(1):73–144.

Anna Wierzbicka. 1999. Emotions across languages and cultures: Diversity and universals. *Cambridge UP*.

Jonathan Winawer, Nathan Witthoft, Michael C. Frank, Lisa Wu, Alex R. Wade, and Lera Boroditsky. 2007. [Russian blues reveal effects of language on color discrimination](#). *Proceedings of the National Academy of Sciences*, 104(19):7780–7785.

Chenggang Wu and Juan Zhang. 2025. Emotion word is an ambiguous concept: dividing, defining, and connecting emotion-label words and emotion-laden words. *Cognition and Emotion*, pages 1–15.

Michelle Yik, Chiel Mues, Irene NL Sze, Peter Kuppens, Francis Tuerlinckx, Kim De Roover, Felicity HC Kwok, Shalom H Schwartz, Maher Abu-Hilal, Damilola Fisayo Adebayo, et al. 2023. On the relationship between valence and arousal in samples across the globe. *Emotion*, 23(2):332.

## A. Modifications to Korean word annotations

For the Korean word annotations, the core instructions were provided in English, consistent with those used for the English word annotations. The sentence frames for ACPT1–4 and EXP5–8 were translated into Korean as follows (For EXP5–8, the sentences and subjects were translated identically to those used in ACPT1–4).

For acceptability judgments, the sentence contexts were translated as: “내가/나는 X고 느낀다.” (ACPT1), “그들이/은 X고/(이)라고 느낀다.” (ACPT2), “이것이(은)/저것이(은)/그것이(은) X고/(이)라고 느낀다.” (ACPT3), and “나는 X다.” (ACPT4). As Korean doesn’t have a single word equivalent to the English *it*, three different types of subjects were allowed (*이것* ‘this thing’, *저것* ‘that thing over there’, *그것* ‘that thing’) in ACPT3. If any one of these subject forms yielded an *acceptable* judgment, the item was coded as *acceptable*.

Additionally, annotators were allowed to use whichever word ending felt most natural in each sentence context, as Korean lexical items can appear

with different inflectional endings depending on the context. Additional columns were provided for annotators to indicate the most appropriate word form for each frame. For example, 슬프다 (*sŭlpŭta*) feels natural as 슬프다고 in “내가/나는 X고 느낀다.” (ACPT1) context, and therefore the annotator recorded 슬프다고 in a separate column as a natural word form for ACPT1.

## B. Policies on Counting Polysemous Word Entries

Polysemous words are counted as two separate entries under two conditions: (1) has meanings spanning multiple parts of speech (e.g., *worry* as a noun and verb were counted as two separate entries.), or (2) one of the meanings having translations in one language but not in the other (e.g., *needy* was counted as two entries, as one sense had a Korean equivalent and another did not). This approach was adopted to ensure that the total number remained interpretable; otherwise, the sum of the number of words across categories would not correspond to the total number of words in each language. Tables 3 to 5 summarizes the number of words overlapping across two conditions.

Table 3: Overlaps Across Parts of Speech in Korean

Translation		Noun	Verb	Adjective
With	Noun	–	0	0
	Verb	–	–	1
Without	Noun	–	0	–
	Verb	–	–	0

Table 4: Overlaps Across Parts of Speech in English

Translation		Noun	Verb	Adjective
With	Noun	–	35	14
	Verb	–	–	2
Without	Noun	–	1	–
	Verb	–	–	0

Table 5: Overlaps Across Translational Status Within Parts of Speech

Part of Speech	Korean	English
Noun	2	4
Verb	2	3
Adjective	3	11

## C. Verb Annotation Results

### C.1. Agreement Scores

The pairwise kappa scores, along with the aggregated scores across all pairs, of the verb agreement set are presented in Figs. 5 and 6. Overall, annotations of Korean verbs (mean Cohen's  $k = 0.82$ ; aggregated across all annotation criteria and coders) showed higher reliability than annotations of English verbs (mean Cohen's  $k = 0.69$ ; aggregated across all annotation criteria and coders), which mirrors the agreement pattern of adjective annotations.

Across individual criteria, questions in the exclusion criteria produced higher agreement (EXCL9–11; Korean  $k = 0.88$ , English  $k = 0.83$ ). In contrast, annotations on causality criteria exhibited lower and more variable agreement, particularly for Korean verbs across the questions (EXP7–8; Korean  $k = 0.72$ , English  $k = 0.48$ ). While the agreement on the “causing” question was at an “almost perfect” level (mean  $k = 0.92$ ), the “caused” question (EXP7) produced lower agreement (mean  $k = 0.52$ ) for Korean verbs. This pattern contrasts with that observed for adjectives, where “causing” criteria were challenging for both Korean and English annotators (Korean  $k = 0.46$ , English  $k = 0.45$ ). Overall, agreement on exclusion criteria was similarly high for both verbs and adjectives, whereas agreement on causality annotations was higher for verbs than for adjectives.

The agreement patterns of verbs also highlight the subjective nature of emotion concepts, particularly for causality judgments. At the same time, higher agreement on causality annotations for verbs indicates relatively less variation for verbs. The very high level of agreement for “causing” criterion suggests that the causative properties of Korean emotion verbs may be more consistently interpretable than those of English emotion verbs.

### C.2. Annotation Trend

A summary of annotation trends is in Figure 7. Annotations in the agreement set were adjudicated using majority vote to produce a single combined dataset for analysis.

In the causality criteria, the general trend of Korean adjectives showing higher acceptance for EXP7 but lower acceptance for EXP8 than English adjectives remains the same for verbs as well, but the difference between the two criteria was larger (EXP7 acceptance: Korean 68.51%, English 51.15%; EXP8 acceptance: Korean 7.23%, English 43.08%). Exclusion criteria (EXCL9–11) also showed highly consistent rejection in both languages, achieving around 90% rate of `no` response.

Unlike the discrepancy across languages for

adjectives, verbs had a similar rate of additional emotion-related meanings being annotated (POLY12; 7.76% in Korean, 7.83% in English). This rate is similar to that found for Korean adjectives, but remains lower than the rate for English adjectives. These findings may suggest a relatively higher degree of polysemy among English adjectives.

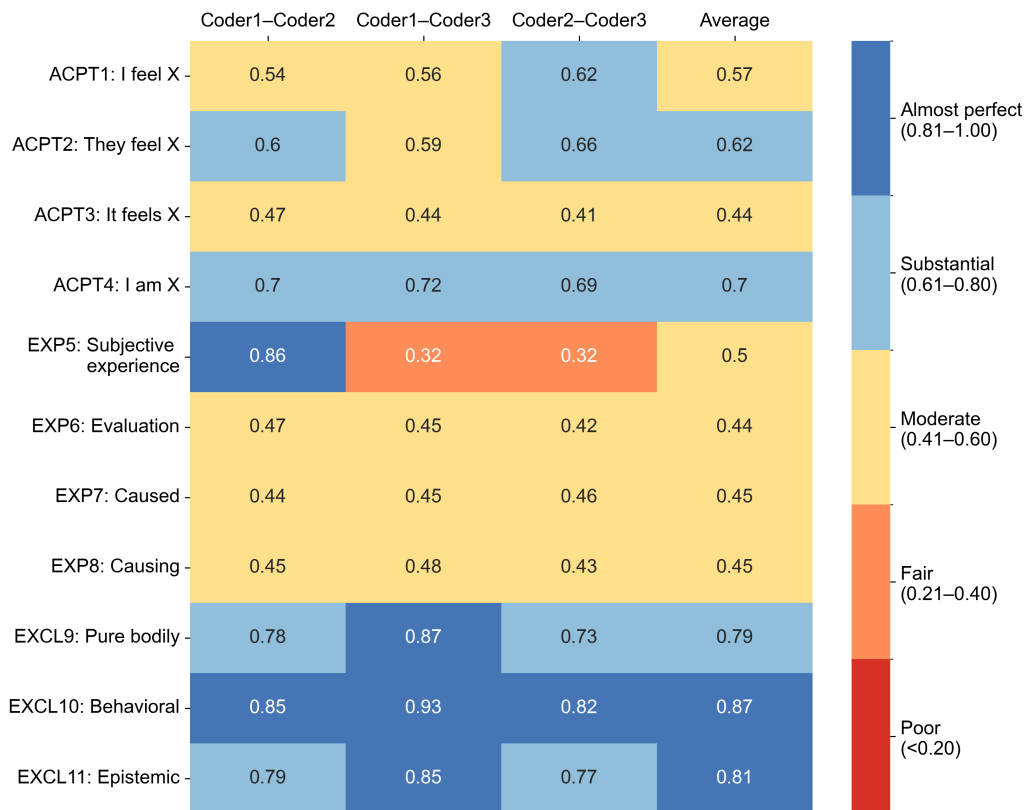


Figure 3: Agreement scores across annotation criteria for the English adjective agreement set.

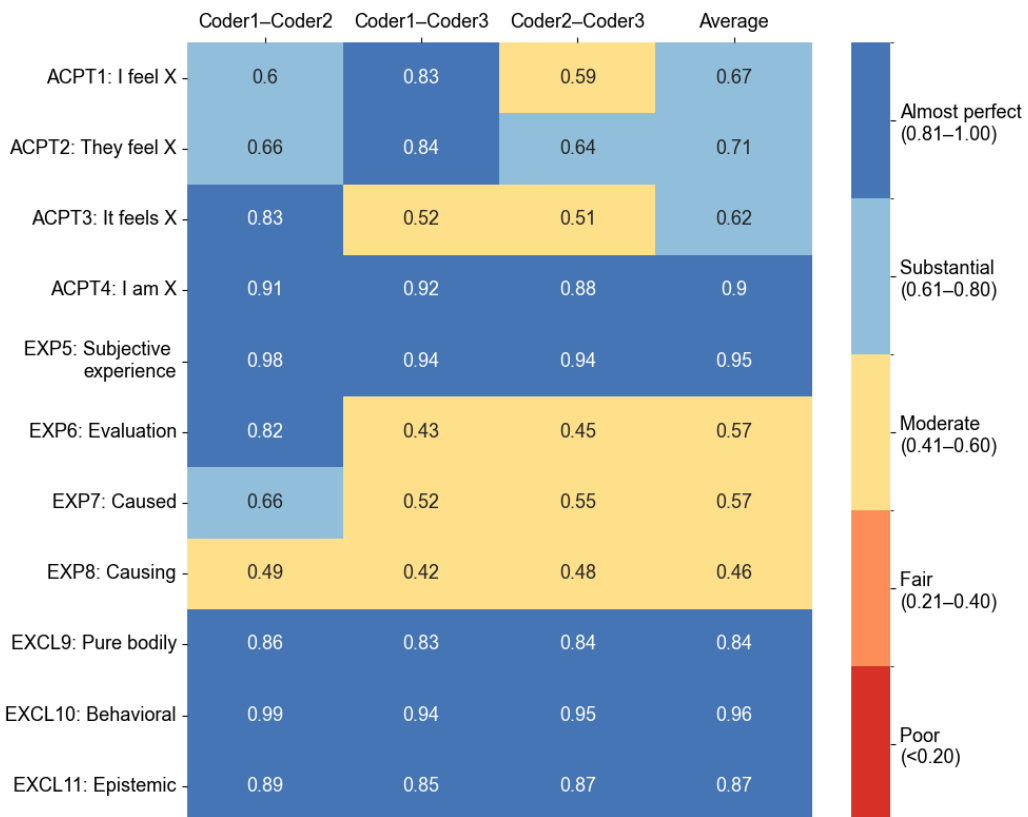


Figure 4: Agreement scores across annotation criteria for the Korean adjective agreement set.

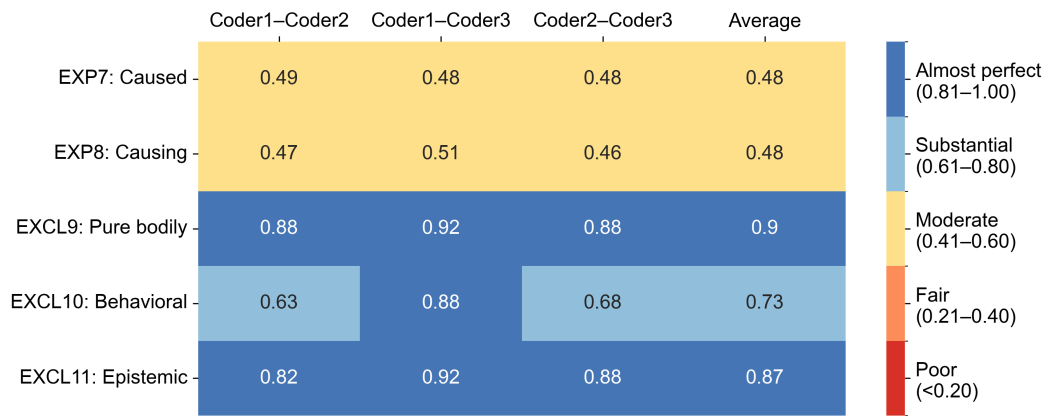


Figure 5: Agreement scores across annotation criteria for the English verb agreement set.

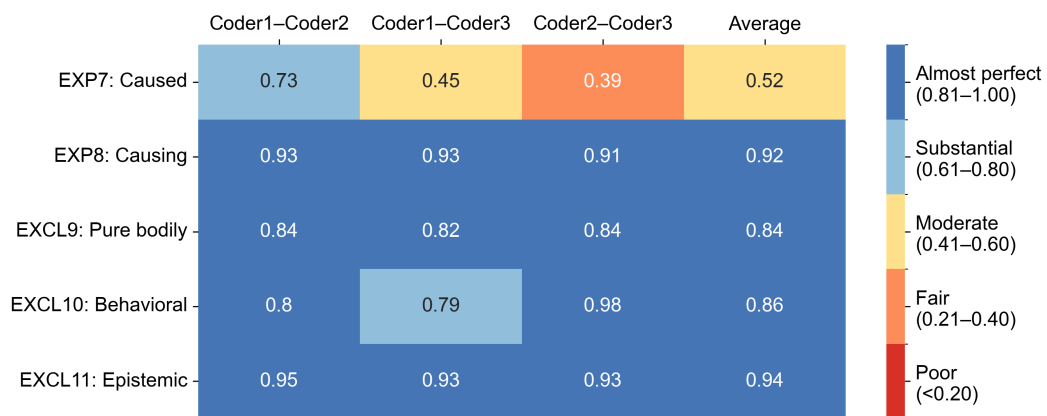


Figure 6: Agreement scores across annotation criteria for the Korean verb agreement set.

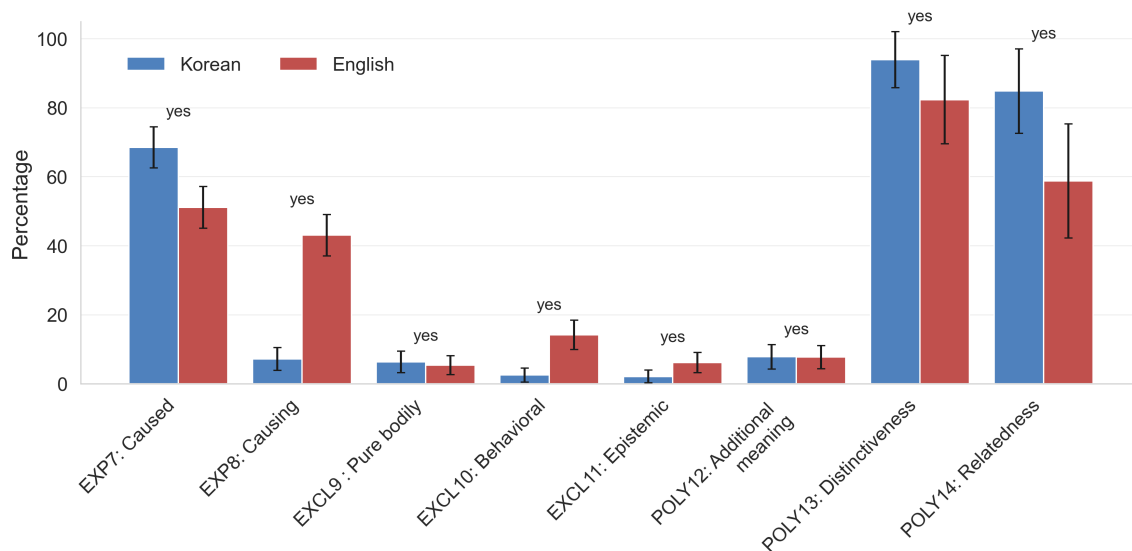


Figure 7: The percentages of annotation values for each criterion for verbs in Korean and English. The percentages for POLY12 indicate the ratio of words annotated as having more than one meaning in each language. The percentages for POLY13 and POLY14 were calculated among items annotated as having more than one meaning in POLY12. Error bars indicate 95% confidence intervals.

# Linguistic Distancing on Social Media: Indicators of Emotion Regulation Across Age Groups

Daniela Teodorescu<sup>α\*</sup>, Saif M. Mohammad<sup>†</sup>, Alona Fyshe<sup>α\*‡</sup>

<sup>α</sup>Alberta Machine Intelligence Institute, <sup>\*</sup>Department of Computing Science, University of Alberta

<sup>†</sup>National Research Council Canada,

<sup>‡</sup>Department of Psychology, University of Alberta

{dteodore,alona}@ualberta.ca, saif.mohammad@nrc-cnrc.gc.ca

## Abstract

Managing our emotional responses to events is key to emotional well-being, a process referred to as emotion regulation in psychology. Previous work has established that the degree to which we distance events is a type of emotion regulation. When we psychologically distance from events there can be markers in our language. These markers have been referred to as linguistic distancing. We build upon a previous metric to operationalize linguistic distancing, and explore how it changes across the lifespan. We explore this systematically by analyzing large amounts of social media text, a venue where people express their emotions. By investigating how distancing varies across age groups we can better understand how emotion regulation varies with age and provide initial benchmarks on social media data. We provide additional evidence further strengthening the hypothesis that linguistic distancing occurs in proportionally more instances with age. These findings align with past work in psychology which indicate improved well-being with older age. Better understanding how linguistic distancing changes with age is important because it functions as a marker of well-being and can inform effective health interventions. We provide a foundation for further exploring emotion regulation through linguistic distancing in text data.

**Keywords:** Linguistic Distancing, Age, Emotion Regulation, Social Media

## 1. Introduction

Our everyday emotional experiences are not static but rather dynamic and our emotions are constantly changing over time. The way in which our emotional experiences change over time creates an emotional trajectory, or what some refer to as an *emotion arc* (Mohammad, 2011; Reagan et al., 2016). Emotion regulation includes the processes by which “we influence which emotions we have, when we have them, and how we experience and express them” (Gross, 1998). More simply put, emotion regulation is the way in which we manage our emotions.

Emotion regulation is key to overall well-being; dysregulation of emotions is associated with psychopathologies such as major depressive disorder and social anxiety disorder (Aldao et al., 2010). There are various ways (or strategies) of regulating emotions and some of these are seen as *maladaptive*. While such strategies provide momentary relief from distress, they are not effective in the long term. Maladaptive strategies include rumination (rethinking about memories or negative experiences), avoidance (of a situation or emotional experience), and suppression (of emotions or thoughts), and each of these has been shown to be significantly associated with anxiety, eating disorders, and substance abuse (Aldao et al., 2010). On the other hand, some strategies are considered *adaptive* and are seen as healthy ways of managing emotions, e.g., reappraisal (changing how one thinks about

a situation to change emotional reactions), problem solving (taking actions to solve a problem), and acceptance (non-judgmental acceptance of emotions). Adaptive strategies are seen as aiding against psychopathology and use of such strategies are inversely associated with psychopathology (Aldao et al., 2010).

Given the importance of emotion regulation to well-being, understanding how it changes with age is important for providing appropriate support and interventions. The specific emotion regulation strategies used change throughout childhood, adolescence, young adulthood, and adulthood (Nook et al., 2020). For example, a shift has been found from emotion regulation strategies which deal with a situation behaviorally (e.g., escaping a situation), to more cognitive strategies such as seeking information (Brown et al., 1991; Altshuler et al., 1995). Although there are some conflicting findings on how exactly various strategies track with age, many studies point to older ages having “enhanced emotion regulation” (Urry and Gross, 2010), being better at regulating their emotions (Gross et al., 1997; Mikkelsen et al., 2024), and have better overall well-being than younger adults despite losses physically and cognitively (Urry and Gross, 2010; Stawski et al., 2008).

Another lens through which emotion regulation has been studied is psychological distancing. By creating distance from an event, one can better handle their emotions related to the event and view the event objectively (allowing for a reappraisal of the



event). Adopting a distanced perspective has been shown to down-regulate negative affect (Kross and Ayduk, 2017), and it is a common technique used in Cognitive-Behavioral Therapy (Beck and Beck, 2020) and Dialectical Behavior Therapy (Linehan, 1993). When psychologically distancing from an event, one is viewing the event from a third-person perspective, and often psychological distancing appears in language as markers. For example, when distancing, there are less first-person pronouns and more past and future tense verbs rather than present tense verbs. These markers are described via *linguistic distancing* (Nook et al., 2017). Linguistic distancing has been shown to be associated with successful emotion regulation and reduced negative affect (Nook et al., 2017). Further, in client-therapist transcripts, less linguistic distancing was associated with worse internalizing symptoms (inward focused behaviors often occurring with anxiety and depression) (Nook et al., 2022).

Through linguistic distancing we can study emotion regulation in text; although so far only client-therapist transcripts have been analyzed for select age groups. While linguistic distancing has been explored across childhood, adolescence, and young adults (Cohen et al., 2022; Nook et al., 2020), we do not know how it changes over adulthood which is what we address in our work. While one approach to measuring linguistic distancing across ages is through longitudinal research, that is very difficult as it would require text data from the same people across all the decades. Instead, we conduct cross-sectional research, which has its own benefits, but appropriate conclusions must be drawn that are different from what could be drawn from longitudinal research. Additionally, linguistic distancing could be influenced by the era in which people grew up; e.g., those in a certain age group may use more distancing due to world events, culture, social norms, etc. than other generations. Therefore, we explore how linguistic distancing changes in *contemporary* times using a platform which many of use to communicate: social media.

Social media is a platform where we frequently communicate and express our feelings. Platforms such as Reddit and X provide a space where we can freely express our thoughts at any time of day, connect with others, and possibly receive support. Given the important role social media plays in everyday communication, as researchers we make use of large amounts of data to analyze linguistic distancing in a systematic way across adulthood. We explore how linguistic distancing varies across age groups on social media. We make use of a social media dataset annotated with the age of the author at the time of posting to answer the following research questions:

- How does linguistic distancing vary across age

groups?

We hypothesize that linguistic distancing increases with age as people become better at regulating emotions based on literature in psychology.

- How do the various *dimensions* of linguistic distancing (i.e., social distancing, temporal distancing, passive voice and abstractness) vary across age groups? Do some dimensions change more than others across adulthood?

By answering these questions, we provide important findings on how linguistic distancing tracks across age groups on social media. Through linguistic distancing, our findings allow us to better understand emotion regulation on a larger scale across age, enabling further study of this phenomenon in psychology and the social sciences. We make our code publicly available.<sup>1</sup>

## 2. Related Work

Below we describe past work examining the relationship between emotion regulation, linguistic distancing and how they change across the lifespan. Afterwards, we describe markers of distancing in language.

### 2.1. Emotion Regulation, Linguistic Distancing & Age

Appropriately regulating emotions is key to mental health and well-being. Vast amounts of literature point to dysregulation of emotions being tied with psychopathology such as major depressive disorder and anxiety disorders (Gross and Muñoz, 1995; Sheppes et al., 2015; Aldao et al., 2010). The various ways in which we manage our emotions are related with health outcomes: adaptive strategies such as cognitive reappraisal are significantly associated with positive indicators of mental health, whereas maladaptive strategies such as suppression of emotions are significantly associated with negative indicators of mental health (Hu et al., 2014). Emotion regulation also changes throughout the lifespan. Infants have minimal to no control over their emotions, relying on their caregiver for support (Bowlby, 1969). Infants begin to develop approaches such as changing their eye gaze to focus on other objects as a way of signaling autonomous emotion regulation (Derryberry and Rothbart, 1988). As they become toddlers, they begin to understand language which opens a whole new world in terms of emotion regulation. Now, they can begin to understand instructions from their parents (Thompson, 1991) and they can

<sup>1</sup>[https://github.com/dteodore/Age--Linguistic\\_Distancing](https://github.com/dteodore/Age--Linguistic_Distancing)

begin to express how they feel and socialize with others, learning to differentiate between positive and negative emotions (Gross and Muñoz, 1995). In adolescence, more sophisticated strategies develop, such as reappraisal as well as maladaptive strategies (Gross and Muñoz, 1995; Hall et al., 1993; Félix-Ortiz et al., 1995). While overall findings point to older age being associated with better overall well-being, control of emotions, and enhanced emotion regulation (Urry and Gross, 2010; Gross et al., 1997), there are some discrepancies in exactly how emotion regulation changes with age, such as which strategies do we use more vs. less. For example, some work found *suppression* decreased across the life span (John and Gross, 2004; De France and Hollenstein, 2019), whereas others found *suppression* increased with age (Brummer et al., 2014). De France and Hollenstein (2019) found adolescents used *distraction* and *suppression* more than young adults, and young adults used more *rumination*, whereas Sütterlin et al. (2012) found no differences in *rumination* across ages. John and Gross (2004) found *reappraisal* increased across the lifespan, whereas De France and Hollenstein (2019) found no difference. Given the conflicting findings, we use social media data to examine how *reappraisal* changes across the lifespan.

We specifically study reappraisal through linguistic distancing. Linguistic distancing has been studied across children, adolescents, and young adults when performing a reappraisal task, however no differences across ages was found (Nook et al., 2020). Although there were changes in the exact strategies of cognitive reappraisal employed across ages: *changing circumstances* and *distancing* increased across age; *changing consequences* decreased across age; and adolescences used more *challenging reality* and less *problem-solving* compared to other ages. Linguistic distancing has also been studied specifically in adolescents: more linguistic distancing was associated with lower levels of hopelessness and higher levels of perceived agency, whereas greater use of linguistic distancing predicted fewer depressive symptoms in follow-ups (Cohen et al., 2022). While these studies provide findings for ages 10–23 (Nook et al., 2020) and 13–16 (Cohen et al., 2022), we are interested in changes across a vast range of ages i.e., 13–70's. Further, we investigate the natural use of linguistic distancing as it appears in everyday communication (e.g., social media), whereas past findings in psychology were based on self-reports and questionnaires (e.g., Emotion Regulation Questionnaire (ERQ) (Gross and John, 2003), Cognitive Emotion Regulation Questionnaire (CERQ) (Garnefski and Kraaij, 2007)) or reappraisal tasks. While past work has created a corpus for detecting coping

strategies in text (Troiano et al., 2024), which is very intertwined with emotion regulation (coping strategies take direct action to manage distress), we study naturally occurring text (rather than text from role-playing scenarios) and are interested in this phenomenon across ages.

## 2.2. Distancing in Language

Distancing can occur through various dimensions, such as in terms of time, space and socially (Construal Level Theory (Liberman and Trope, 2008; Trope and Liberman, 2010)). Therefore, past operationalizations of linguistic distancing include measures of *social distancing* and *temporal distancing*. Use of more first-person singular pronouns e.g., “I”, “me”, “mine” represents a more immersed perspective and less *social* distancing, which means that one is performing less psychological distancing. On the other hand, using more second- and third-person pronouns e.g., “she”, “they”, “them” represents using more *socially* distanced language, and therefore one is doing more psychological distancing. Using more present tense verbs represents less *temporal distancing* and less psychological distancing, whereas using more past- and future-tense signals more *temporal* distancing and therefore more psychological distancing (Nook et al., 2022). We use these definitions of *social* and *temporal* distancing in our work. The bidirectional relationship between emotion regulation and linguistic signatures of psychological distancing has been shown (Nook et al., 2017). Regulating emotions through psychological distancing has been associated with increased linguistic markers of social and temporal distance. Likewise, using distanced language also regulated emotions and reduced negative affect (Nook et al., 2017). Self-distancing (“I” vs. “you” pronouns) and social withdrawal (indicated by more “they” pronouns), has been explored on Reddit for healthcare workers during the COVID-19 pandemic, however we are interested in linguistic distancing on a broader scale (i.e., for those beyond the healthcare field) and in day-to-day contexts outside of the pandemic (Ireland et al., 2022).

When distancing from an event psychologically, this creates a more abstract mental representation of the event – leading to a high correlation between linguistic measures of psychological distancing and *abstractness* (Nook et al., 2025). Measures of linguistic distancing and *abstractness* have been found to be highly correlated; *abstractness* increased when cognitive reappraisal was used to regulate emotions; when people distanced their language, *abstractness* also increased; and lastly, increased *abstractness* when regulating emotions was correlated with regulation success (Nook et al., 2025). Given the established relationship between *abstractness* and linguistic distancing, we also in-

Age Group	#Posts	
	TUSC-City	Reddit
13–19	94,857	9,281,055
20–29	406,027	15,455,426
30–39	462,238	6,151,757
40–49	360,946	1,220,498
50–59	278,809	449,230
60–69	218,008	246,398
70–79	97,432	160,465

Table 1: The number of posts across the age groups in each subset of the *AgeCorpus*.

corporate a measure of *abstractness* in our metric of linguistic distancing.

Work in psychology has also pointed to the relationship between using passive voice and psychological distancing (Chan and Maglio, 2020a,b). *Passive voice* is where the object is placed before the verb such that the subject receives the action rather than performing it (e.g., passive voice: “The ball was thrown by the child” vs. active voice: “The child threw the ball”).<sup>2</sup> When reading passages in passive voice participants reported feeling temporal, hypothetical, and spatial distance from the events in the passage resulting in a more abstract mental construal (Chan and Maglio, 2020b; Trope and Liberman, 2003). Past work has also studied passive voice in the context of assault, where distancing removed the action from the actor and places blame on the victim (Bohner, 2001). When using passive voice, the agency of the subject is taken away, creating distance from the action.

Given the past literature demonstrating the relationship between linguistic distancing with *social* distancing, *temporal* distancing, *abstractness*, and *passive voice*, we construct a measure of linguistic distancing which includes all four dimensions. Analyzing how each of these dimensions changes across adulthood is informative for understanding behaviors and tendencies which is of interest not only to psychologists, linguists, and sociologists, but is also linked with mental health.

### 3. Dataset: AgeCorpus

We perform our experiments on *AgeCorpus*, a social media dataset containing posts annotated with the author’s age at the time of writing (Teodorescu et al., 2026). The dataset contains posts from both Reddit and X, making it a suitable dataset for exploring linguistic distance on social media platforms. Further, there are a large number of posts in the dataset, allowing us to aggregate results per age group. The dataset was collected by identifying in-

stances where authors self-disclosed their age e.g., “I am X years old” or “Me [20F] and my best friend [21M]”. Using various high quality pattern matching templates, the age of an author can be determined at a point in time based on the timestamp of the post (seed post). We show the pattern templates used in Appendix A. Using the age declared at the given timestamp in the seed post, the author’s age can be determined for any other post based on the difference of the timestamp to that of the seed post’s. The dataset consists of thousands of users each with hundreds of posts spanning from 2010–2022 for Reddit and 2020–2021 for X. We show the exact number of posts per age group in Table 1. The X component of the dataset consists of *TUSC-Country* and *TUSC-City* which differ in how they were collected. *TUSC-City* was collected using Twitter’s free API for American and Canadian cities during 2020–2021. Whereas *TUSC-Country* was collected using Twitter’s Academic API for tweets from the US and Canada during 2015–2021. Our results on the two subsets are very similar so we report those on *TUSC-City* since it is the larger subset and show the results on *TUSC-City* in the Appendix (Section C). We use this dataset for our experiments described in the next section.

## 4. Experiments

In the sections below we describe our methods for exploring how linguistic distancing varies across age groups.

### 4.1. Computing Linguistic Distancing in Text

We build on Nook et al. (2022)’s work on computing linguistic distancing in text. Specifically, the authors operationalize linguistic distancing as being composed of two components: *social* distancing and *temporal* distancing. *Social* distancing is measured through the proportions of pronouns present in the text. Fewer first-person singular pronouns are associated with proportionally more instances of distancing. Here, we analyze the proportions of pronouns present in an instance or a post. More specifically, *social* distancing is computed per post. Let  $f_{1p,sg}$ ,  $f_{1p,pl}$ ,  $f_{2p}$ ,  $f_{3p,sg}$ ,  $f_{3p,pl}$  denote the corresponding frequencies of first-person singular, first-person plural, second-person, third-person singular and third-person plural pronouns respectively. Then the social distancing (*social*) component of linguistic distancing (*LD*) per post or instance ( $t$ ) is computed as:

$$LD_{social}(t) = \frac{f_{2p} + f_{1p,pl} + f_{3p,sg} + f_{3p,pl}}{f_{2p} + f_{1p,pl} + f_{3p,sg} + f_{3p,pl} + f_{1p,sg}}$$

*Temporal* distancing is captured through the use of verb tense. More past and future tense verbs

<sup>2</sup><https://www.grammarly.com/blog/grammar/passive-voice/>



Dataset	Metric	df1	df2	F-statistic	P-value	Effect Size ( $est \omega^2$ )
Reddit	Linguistic	6	1203317.10	17703.63	$p < .001$	0.003
TUSC-City	Linguistic	6	539267.68	2830.26	$p < .001$	0.009

Table 2: The degrees of freedom (for the numerator and denominator), F-statistic, p-value, and effect size in the one-way Welch’s ANOVA test for differences in linguistic distancing between age groups.

are associated with proportionally more instances of distancing. We compute *temporal* distancing per post. Let  $f_{past}$ ,  $f_{present}$ ,  $f_{future}$  denote the corresponding frequencies of past-tense, present-tense and future-tense verbs. Then the temporal distancing (*temp*) component of linguistic distancing (*LD*) per post or instance ( $t$ ) is computed as:

$$LD_{temp}(t) = \frac{f_{past} + f_{future}}{f_{past} + f_{future} + f_{present}}$$

We further build on the metric for linguistic distancing by including a measure of *abstractness* and *passive voice*. We include *abstractness* in our measure of linguistic distancing as more abstract mental representations of an event are linked with more successful emotion regulation, and more abstract language is highly correlated with more distanced language (Nook et al., 2025). More details on this are described in Section 2. To quantify *abstractness* we use a word list with ratings of concreteness for forty-thousand words (Brysbaert et al., 2014).<sup>3</sup> Each word is rated on a 5-point scale from abstract to concrete which we invert to be consistent with other measures. We calculate the *abstractness* score ( $LD_{abs}$ ) per post ( $t$ ) as the average of the abstractness score of each word at index  $i$  ( $abs(t_i)$ ) of the instance with length  $n$ :

$$LD_{abs}(t) = \frac{1}{n} \sum_{i=1}^n abs(t_i)$$

Lastly, we determine whether a text contains *passive voice* or not using the PassivePy tool (Sepehri et al., 2023). PassivePy uses POS tagging, dependency parsing and rule-based matching for sentence structures/forms in order to determine whether passive voice is present in the text or not. We integrate this binary score of passive voice ( $pass(t)$ ) per instance ( $t$ ) into our metric for linguistic distancing.

$$LD_{pass}(t) = pass(t)$$

Each of these measures are on different scales. Therefore, we standardize all the scores for each dimension individually such that each dimension has scores with a mean of 0 and a standard deviation of 1.<sup>4</sup> Then we compute the linguistic distancing

(*LD*) score per post ( $t$ ) as the average of the four standardized ( $stand()$ ) dimensions:

$$LD(t) = \frac{stand(LD_{temp}(t)) + stand(LD_{social}(t)) + stand(LD_{abs}(t)) + stand(LD_{pass}(t))}{4}$$

**Linguistic Distancing Per Age Group:** The next step is to aggregate scores per age groups. For our analyses we consider the following age groups: 13–19 (teens), 20–29 (twenties), 30–39 (thirties), 40–49 (forties), 50–59 (fifties), 60–69 (sixties), 70–79 (seventies). We group instances based on the age of the author at the time of writing. Then we compute the linguistic distancing score per group as the average of all the scores of instances within the particular group.

## 4.2. Statistical Analyses

To compare whether there is a statistically significant difference in linguistic distancing measures across age groups, we perform a one-way ANOVA test. Before performing inferential tests, we ensure that our data met ANOVA’s prerequisites. Namely, we tested that the data are independent of each other, each independent variable is approximately normally distributed, and each group has roughly the same variance (homoscedasticity). Next, we use a qq-plot of the residuals to verify that each score is normally distributed for each group. In addition to the normality assumption, we can assume the estimate of the mean is normally distributed due to the large sample size (law of large numbers) according to the central limit theorem. To verify whether each group has approximately the same variance, we performed Levene’s test and the result was significant ( $p < 0.001$ ) for Reddit and TUSC-City, meaning that the age groups do not have equal variance in the Reddit and TUSC-City subsets of the dataset. We can assume each data point is independent of others since each post is generally independent of one another. While it is possible that while browsing social media, a post may influence others to some degree, this is largely not the case. Therefore, we will perform a one-way Welch’s ANOVA for TUSC-City and Reddit. Results for Levene’s tests can be found in Appendix B. After we determined our data met the assumptions, we performed the one-way ANOVA.

<sup>3</sup>We opt for this method of determining abstractness since Nook et al. (2025) use the LIWC Lexicon (Pennebaker et al.), which is not freely available.

<sup>4</sup>Standardizing is the same as computing the z-score.

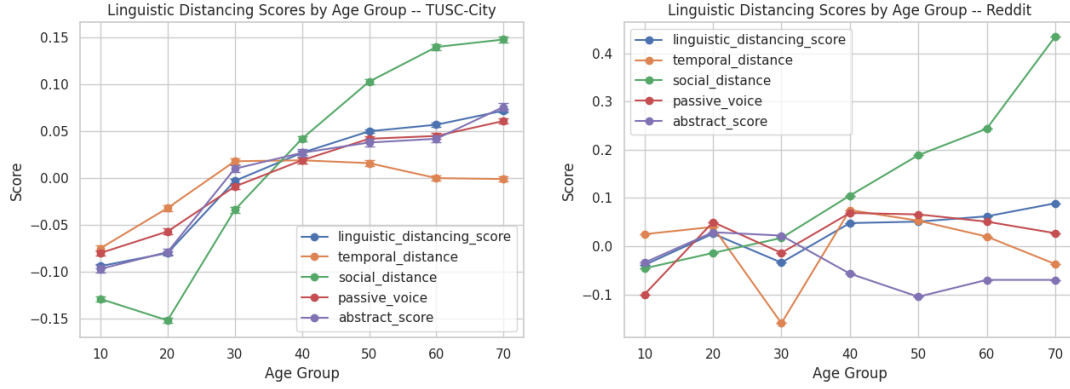


Figure 1: **Linguistic Distancing** scores for TUSC-City, and Reddit subset of the *AgeCorpus* (blue line). Individual components of linguistic distancing across age groups are shown: temporal distance (orange line), social distance (green line), passive voice (red line) and abstractness (purple line). Error bars represent the standard error of the mean.

## 5. Results

### 5.1. RQ1: How Does Linguistic Distancing Vary Across Age Groups?

In Figure 1, we show the trend of linguistic distancing across age groups. We see that regardless of the dataset (i.e., X or Reddit), linguistic distancing (blue line) occurs more *frequently* with age. This upwards trend is consistent across age groups (except with a dip in the 30’s on Reddit). This result is also reflected in the statistical analyses. We show in Table 2 the F-statistic and p-value for the ANOVA tests. For all datasets, there is a significant increase in linguistic distance with age (Reddit:  $F(6, 1203317.10) = 17703.63, p < 0.001, \eta^2 = 0.003$ ; TUSC-City:  $F(6, 539267.68) = 2830.26, p < 0.001, \eta^2 = 0.009$ ). The effect size on the X dataset is near 0.01 which is considered small, and the effect size on the Reddit dataset is even smaller.<sup>5</sup>

**Discussion:** Past literature has found that overall well-being, lower negative affect, higher positive affect and improved emotion regulation skills is associated with age (Urry and Gross, 2010; Charles et al., 2023; Stone et al., 2010; Isaacowitz, 2022). Our findings show that with age, there are proportionally more instances of distancing. Given the relationship between linguistic distancing and emotion regulation, our findings support more use of reappraisal with age and in turn better well-being in the older ages. The drop in the 30’s Reddit data is interesting as past work has found a dip in well-being during midlife (e.g., 35’s–50’s) (Blanchflower and Oswald, 2008; Blanchflower and Graham, 2021; Stone et al., 2010), however, there are

some theories opposing the U-shape of happiness (Galambos et al., 2020).

### 5.2. RQ2: How does Social Distancing, Temporal Distancing, Abstractness, and Passive Voice Change with Age?

In Figure 1, we see the trend for each of the dimensions of linguistic distancing across age groups. Every dimension (i.e., *temporal* distance, *social* distance, *abstractness*, *passive*) occurs more frequently with age, although some have a steeper slope. *Social* distancing has the steepest slope across datasets based on the graph. This means that the rate at which *social* distancing increases with age is higher than the rate for other dimensions. In addition, *social* distancing also starts off at a lower point on the y-axis than all the other dimensions, allowing more room to grow with age. However, on Reddit *abstractness*, *temporal*, and *passive* scores plateau/decrease after age 40 and onwards, and likewise with *temporal* distancing on X. This could be due to the nature of the platforms and the topics discussed on each e.g., often people describe situations on Reddit to obtain advice and therefore talk in a more present-tense, less abstract manner. Reddit also has a dip in linguistic distancing in the 30’s which is primarily driven by a decrease in temporal distance. This means that individuals are using more present tense verbs.

**Discussion:** While some of these dimensions have not been studied directly with age, they measure concepts that have been shown to change with age. For example, more *social* distancing is associated with fewer personal-pronouns used. Various literature point to the negative relationship between age and first-person singular pronouns (Schwartz et al., 2013; Pang et al., 2025; Pennebaker and Stone, 2003). The trends we see across X and

<sup>5</sup>Effect size of 0.01 is small; 0.06 is medium; and 0.14 is large. Effect sizes help us quantify whether differences are practically meaningful in the real-world.



Reddit support this pattern. In terms of *temporal* distancing, higher levels mean that fewer present tense verbs are used which is also supported by past work which points to more future-tense and fewer past-tense verbs with age (Pennebaker and Stone, 2003). While there are few works studying *passive voice* with age, older adults were found to produce fewer passive sentences than younger adults in Korean (Sung et al., 2024). Also, older adults (i.e., 70's) had more difficulty producing and understanding abstractness on tests (Albert et al., 1990). For abstractness, no statistical difference was found for the ability to define abstract words across participants aged 19–89, but the use of abstract vs. concrete words in everyday language was not examined (Pezzuti et al., 2021). While some work examines how each of these components vary across childhood or in terms of acquisition (e.g., linguistic distancing: Nook et al. (2020); Cohen et al. (2022); passive voice: Baldie (1976); Horgan (1978)), fewer work examines how they change across adulthood. Further, past work examining linguistic distancing does not consider the contributions of *temporal* vs. *social* distancing to linguistic distancing across age groups (Nook et al., 2017). Therefore, we provide new findings that lay the foundation for future exploration into what drives changes in linguistic distancing with age.

## 6. Conclusion

We study how linguistic distancing changes across adulthood to better understand emotion regulation with age. We use large social media datasets to systematically analyze how does linguistic distancing vary across teens to 70's on X and Reddit. We construct an interpretable measure of linguistic distancing based on four dimensions: *temporal* distancing, *social* distancing, *passive voice* and *abstractness*. We found that linguistic distancing occurs in proportionally more posts with age. When exploring how each of the dimensions changed across age groups, we found that *social* distancing has the highest rate of increase. We also saw that *temporal* distancing plateaus and decreases in later ages. Our work provides an initial foundation for understanding how linguistic distancing varies in text across age groups and supports further exploration into measuring linguistic distancing on social media. Future work could examine how such measures change across platforms discussing diverse topics (e.g., subreddits) and contrast results with LLM measures of linguistic distancing. Future work could also explore linguistic distancing in different languages and cultures, as psychological distancing and distancing in language may differ based on culture or contexts.

## Ethical Considerations

Our research interest is to study how linguistic distancing changes across adulthood at the aggregate/group level. This has applications in emotional development psychology and in public health (e.g., overall well-being and mental health). However, personal well-being is complex, private, and central to an individual's experience. Additionally, each individual expresses themselves differently through language, which results in large amounts of variation.

Our work on studying linguistic distancing should not be construed as detecting how people feel; rather, we draw inferences on the language used.

The inferences we draw in this paper are based on aggregate trends across large populations. We do not draw conclusions about specific individuals or momentary states of well-being.

## Limitations

The dataset used in this study relies on self-disclosure of age on social media. As with all scenarios involving self-disclosure, individuals may falsely report their age due to social pressures, or in order to relate to and engage with certain communities. Additionally, the templates used for pattern matching with age may not fully capture all possible ways to express age, so there may be age declarations not included in the dataset.

## Acknowledgments

This research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC), the Social Sciences and Humanities Research Council (SSHRC), the Digital Research Alliance of Canada (alliancecan.ca), and the Canadian Institute for Advanced Research (CIFAR). Compute resources were graciously supplied by Prairies DRI and the Digital Research Alliance of Canada. Alona Fyshe holds a Canada CIFAR AI Chair.

## 7. Bibliographical References

- MS Albert, J Wolfe, and G Lafleche. 1990. [Differences in abstraction ability with age](#). *Psychology and aging*, 5(1):94–100.
- Amelia Aldao, Susan Nolen-Hoeksema, and Susanne Schweizer. 2010. [Emotion-regulation strategies across psychopathology: A meta-analytic review](#). *Clinical Psychology Review*, 30(2):217–237.

- Jennifer L. Altshuler, Janice L. Genevaro, Diane N. Ruble, and Marc H. Bonstein. 1995. [Children's knowledge and use of coping strategies during hospitalization for elective surgery](#). *Journal of Applied Developmental Psychology*, 16(1):53–76.
- Brian J. Baldie. 1976. [The acquisition of the passive voice](#). *Journal of Child Language*, 3(3):331–348.
- J.S. Beck and A.T. Beck. 2020. *Cognitive Behavior Therapy, Third Edition: Basics and Beyond*. Guilford Publications.
- David G Blanchflower and Carol L Graham. 2021. [The u shape of happiness: A response](#). *Perspectives on psychological science : a journal of the Association for Psychological Science*, 16(6):1435–1446.
- David G. Blanchflower and Andrew J. Oswald. 2008. [Is well-being u-shaped over the life cycle?](#) *Social Science Medicine*, 66(8):1733–1749.
- Gerd Bohner. 2001. [Writing about rape: Use of the passive voice and other distancing text features as an expression of perceived responsibility of the victim](#). *British Journal of Social Psychology*, 40(4):515–529.
- John Bowlby. 1969. *Attachment and loss*. 79. New York: Basic Books.
- Kirk Brown, Katherine Covell, and Rona Abramovitch. 1991. [Time course and control of emotion: Age differences in understanding and recognition](#). *Merrill-Palmer Quarterly*, 37(2):273–287.
- Laura Brummer, Lusía Stopa, and Romola Bucks. 2014. [The influence of age on emotion regulation strategies and psychological distress](#). *Behavioural and Cognitive Psychotherapy*, 42(6):668–681.
- Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. [Concreteness ratings for 40 thousand generally known english word lemmas](#). *Behavior research methods*, 46(3):904–911.
- Eugene Y. Chan and Sam J. Maglio. 2020a. [The voice of cognition: Active and passive voice influence distance and construal](#). *Personality and Social Psychology Bulletin*, 46(4):547–558. PMID: 31390936.
- Eugene Y. Chan and Sam J. Maglio. 2020b. [The voice of cognition: Active and passive voice influence distance and construal](#). *Personality and Social Psychology Bulletin*, 46(4):547–558. PMID: 31390936.
- Susan T Charles, Jonathan Rush, Jennifer R Piazza, Eric S Cerino, Jaqueline Mogle, and David M Almeida. 2023. [Growing old and being old: Emotional well-being across adulthood](#). *Journal of personality and social psychology*, 125(2):455–469.
- Katherine A. Cohen, Akash Shroff, Erik C. Nook, and Jessica L. Schleider. 2022. [Linguistic distancing predicts response to a digital single-session intervention for adolescent depression](#). *Behaviour Research and Therapy*, 159:104220.
- Kalee De France and Tom Hollenstein. 2019. [Emotion regulation and relations to well-being across the lifespan](#). *Developmental psychology*, 55(8):1768–1774.
- D Derryberry and MK Rothbart. 1988. [Arousal, affect, and attention as components of temperament](#). *Journal of personality and social psychology*, 55(6):958–966.
- María Félix-Ortiz, Ricardo Muñoz, and Michael D. Newcomb. 1995. [The role of emotional distress in drug use among latino adolescents](#). *Journal of Child & Adolescent Substance Abuse*, 3(4):1–22.
- Nancy L Galambos, Harvey J Krahn, Matthew D Johnson, and Margie E Lachman. 2020. [The u shape of happiness across the life course: Expanding the discussion](#). *Perspectives on psychological science : a journal of the Association for Psychological Science*, 15(4):898–912.
- Nadia Garnefski and Vivian Kraaij. 2007. [The cognitive emotion regulation questionnaire](#). *European Journal of Psychological Assessment*, 23:141–149.
- James J. Gross. 1998. [The emerging field of emotion regulation: An integrative review](#). *Review of General Psychology*, 2(3):271–299.
- James J Gross and Oliver P John. 2003. [Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being](#). *Journal of personality and social psychology*, 85(2):348–362.
- James J. Gross and Ricardo F. Muñoz. 1995. [Emotion regulation and mental health](#). *Clinical Psychology: Science and Practice*, 2(2):151–164.
- JJ Gross, LL Carstensen, M Pasupathi, J Tsai, CG Skorpén, and AY Hsu. 1997. [Emotion and aging: experience, expression, and control](#). *Psychology and aging*, 12(4):590–599.
- Sharon M Hall, Ricardo F Muñoz, Victor I Reus, and Karen L Sees. 1993. Nicotine, negative affect, and depression. *Journal of consulting and clinical psychology*, 61(5):761.

- Dianne Horgan. 1978. [The development of the full passive](#). *Journal of Child Language*, 5(1):65–80.
- Tianqiang Hu, Dajun Zhang, Jinliang Wang, Ritesh Mistry, Guangming Ran, and Xinqiang Wang. 2014. [Relation between emotion regulation and mental health: a meta-analysis review](#). *Psychological reports*, 114(2):341–362.
- Molly Ireland, Kaitlin Adams, and Sean Farrell. 2022. [Tracking mental health risks and coping strategies in healthcare workers' online conversations across the COVID-19 pandemic](#). In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 76–88, Seattle, USA. Association for Computational Linguistics.
- Derek M Isaacowitz. 2022. [What do we know about aging and emotion regulation?](#) *Perspectives on psychological science : a journal of the Association for Psychological Science*, 17(6):1541–1555.
- Oliver P. John and James J. Gross. 2004. [Healthy and unhealthy emotion regulation: Personality processes, individual differences, and life span development](#). *Journal of Personality*, 72(6):1301–1334.
- E. Kross and O. Ayduk. 2017. [Chapter two - self-distancing: Theory, research, and current directions](#). volume 55 of *Advances in Experimental Social Psychology*, pages 81–136. Academic Press.
- Nira Liberman and Yaacov Trope. 2008. [The psychology of transcending the here and now](#). *Science (New York, N.Y.)*, 322(5905):1201–1205.
- Marsha M Linehan. 1993. *Skills training manual for treating borderline personality disorder*. Guilford press.
- Mai Bjørnskov Mikkelsen, Mia Skytte O'Toole, Emma Elkjaer, and Mimi Mehlsen. 2024. [The effect of age on emotion regulation patterns in daily life: Findings from an experience sampling study](#). *Scandinavian journal of psychology*, 65(2):231–239.
- Saif Mohammad. 2011. [From once upon a time to happily ever after: Tracking emotions in novels and fairy tales](#). In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 105–114, Portland, OR, USA. Association for Computational Linguistics.
- Erik C. Nook, Hayoung E. Ahn, Jessica L. Schleider, and Leah H. Somerville. 2025. [Emotion regulation is associated with increases in linguistic measures of both psychological distancing and abstractness](#). *Affective Science*, 6(1):63–76. Publisher Copyright: © The Author(s) 2024.
- Erik C. Nook, Thomas D. Hull, Matthew K. Nock, and Leah H. Somerville. 2022. [Linguistic measures of psychological distance track symptom levels and treatment outcomes in a large set of psychotherapy transcripts](#). *Proceedings of the National Academy of Sciences*, 119(13):e2114737119.
- Erik C Nook, Jessica L Schleider, and Leah H Somerville. 2017. [A linguistic signature of psychological distancing in emotion regulation](#). *Journal of experimental psychology. General*, 146(3):337–346.
- Erik C. Nook, Constanza M. Vidal Bustamante, Hyun Young Cho, and Leah H. Somerville. 2020. [Use of linguistic distancing and cognitive reappraisal strategies during emotion regulation in children, adolescents, and young adults](#). *Emotion*, 20(4):525–540. Publisher Copyright: © 2020 American Psychological Association.
- Dandan Pang, Sharath C. Guntuku, Garrick Sherman, Tingting Liu, Sunny Rai, Young-Min Cho, Rian You, Arnold B. Bakker, and Lyle H. Ungar. 2025. [Understanding gender and age differences in language use: cross-cultural insights from weibo and facebook](#). *Humanities Social Sciences Communications*, 12(1):1667. Copyright - © The Author(s) 2025. This work is published under <http://creativecommons.org/licenses/by-nc-nd/4.0/> (the “License”). Notwithstanding the ProQuest Terms and Conditions, you may use this content in accordance with the terms of the License; Last updated - 2025-11-07; SubjectsTermNotLitGenreText - United States–US; China.
- James W Pennebaker and Lori D Stone. 2003. [Words of wisdom: language use over the life span](#). *Journal of personality and social psychology*, 85(2):291–301.
- James W Pennebaker et al. Linguistic inquiry and word count: Liwc 2001.
- Lina Pezzuti, James Dawe, and Anna Maria Borghi. 2021. [Does mastering of abstract words decline with age?](#) *Educational Gerontology*, 47(12):527–542.
- Andrew J. Reagan, Lewis Mitchell, Dilan Kiley, Christopher M. Danforth, and Peter S. Dodds. 2016. The emotional arcs of stories are dominated by six basic shapes. *EPJ Data Science*, 5(1):31.



- H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. [Personality, gender, and age in the language of social media: The open-vocabulary approach](#). *PLOS ONE*, 8(9):1–16.
- Amir Sepehri, Mitra Sadat Mirshafiee, and David M. Markowitz. 2023. [Passivepy: A tool to automatically identify passive voice in big text data](#). *Journal of Consumer Psychology*, 33(4):714–727.
- Gal Sheppes, Gaurav Suri, and James J Gross. 2015. [Emotion regulation and psychopathology](#). *Annual review of clinical psychology*, 11:379–405.
- Robert S Stawski, Martin J Sliwinski, David M Almeida, and Joshua M Smyth. 2008. [Reported exposure and emotional reactivity to daily stressors: the roles of adult age and global perceived stress](#). *Psychology and aging*, 23(1):52–61.
- Arthur A. Stone, Joseph E. Schwartz, Joan E. Broderick, and Angus Deaton. 2010. [A snapshot of the age distribution of psychological well-being in the united states](#). *Proceedings of the National Academy of Sciences*, 107(22):9985–9990.
- Jee Eun Sung, Eunha Jo, Sujin Choi, and Jiyeon Lee. 2024. [Coordinating words and sentences: Detecting age-related changes in language production](#). *Journal of Speech, Language, and Hearing Research*, 67(1):211–220.
- Stefan Sütterlin, Muirne C. S. Paap, Stana Babic, Andrea Kübler, and Claus Vögele. 2012. [Rumination and age: Some things get better](#). *Journal of Aging Research*, 2012(1):267327.
- Daniela Teodorescu, Jan Wahle, and Saif M. Mohammad. 2026. Age and affect in language: How emotion expression on social media varies across adulthood.
- Ross A. Thompson. 1991. [Emotional regulation and emotional development](#). *Educational Psychology Review*, 3(4):269–307.
- Enrica Troiano, Sofie Labat, Marco Antonio Stranisci, Rossana Damiano, Viviana Patti, and Roman Klinger. 2024. [Dealing with controversy: An emotion and coping strategy corpus based on role playing](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1634–1658, Miami, Florida, USA. Association for Computational Linguistics.
- Yaacov Trope and Nira Liberman. 2003. [Temporal construal](#). *Psychological review*, 110(3):403–421.
- Yaacov Trope and Nira Liberman. 2010. [Construal-level theory of psychological distance](#). *Psychological review*, 117(2):440–463.
- Heather L. Urry and James J. Gross. 2010. [Emotion regulation in older age](#). *Current Directions in Psychological Science*, 19(6):352–357.

## A. Patterns to Match Age Declarations

In Table 3, we show the pattern templates used to match age declarations in seed posts in the *AgeCorpus*.

## B. Levene’s Test of Homogeneity of Variance

Levene’s test indicated that the assumption for homogeneity of variance was not violated for the effect of age group on linguistic distancing for TUSC-Country, but was violated for Reddit and TUSC-City subsets of the *AgeCorpus*. We show the results in Table 4.

## C. TUSC-Country Results

In the following Sections we show the number of posts in the TUSC-Country version of the dataset and the linguistic distancing results on this subset of the *AgeCorpus*.

### C.1. *AgeCorpus*: TUSC-Country

In Table 5, we show the number of posts in TUSC-Country, TUSC-City and Reddit subsections of the *AgeCorpus* for comparison.

### C.2. Linguistic Distancing Results

In Table 6, we show the results from the ANOVA tests for the difference in linguistic distancing across age groups. We include the rows for the Reddit and TUSC-City subsets of the *AgeCorpus* dataset as a comparison.

In Figure 2, we show the trends of linguistic distancing across age groups (blue line), as well as the trends for *temporal* distancing, *social* distancing, *passive voice* and *abstractness*. Largely the trends for TUSC-Country follow those of TUSC-City, with an increasing proportion of distancing occurring with age. Generally, all dimensions of distancing grow similarly across age groups, with social distancing having the steepest slope.

Regex	Example
[I] \bI(?:\s+am 'm)\s+(\d{1,2})\s+years?\s+old\b <i>I'm 30 year old</i>	[I] <i>I am 25 years old</i>
[I] \bI(?:\s+am 'm)\s+(\d{1,2}) (?=\s*(?:\$ [, .!?:;-] (?:and but so yet)\s)) <i>I'm 30, and ...</i>	[I] <i>I am 25.</i>
[I] \bI(?:\s+was \s+am 'm)\s+born\s+in\s+ (19\d{2} 20(?:0\d 1\d 2[0-4]))\s\b <i>I am born in 2005</i>	[I] <i>I was born in 1998</i>
[I] \bI(?:\s+was \s+am 'm)\s+born\s+in\s+'(\d{2})\s\b <i>I was born in '98</i> <i>I'm born in '05</i>	
[I] \bI\s+was\s+born\s+on\s+ (?:\d{1,2})(?:st nd rd th)?\s+? (?:January February March April May June  July August September October November December  Jan Feb Mar Apr May Jun Jul Aug Sep Sept Oct Nov Dec) \s+ (?:\d{1,2})(?:st nd rd th)?,?\s+? (19\d{2} 20(?:0\d 1\d 2[0-4]))\s\b <i>I was born on March 15th, 1998</i>	[I] <i>I was born on 15 March 1998</i>
[I] \bI\s+was\s+born\s+on\s+\d{1,2} [/-]\d{1,2}[/-](19\d{2} 20(?:0\d 1\d 2[0-4]))\s\b <i>I was born on 15-03-1998</i>	[I] <i>I was born on 03/15/1998</i>

Table 3: Regexes used to identify users from both the Reddit and X datasets.

Dataset	Metric	df1	df2	F-statistic	P-value
Reddit	Linguistic	6	32964822	10269.65	$p < .001$
TUSC-City	Linguistic	6	1918310	86.13	$p < .001$
TUSC-Country	Linguistic	6	11,940	1.12	$p = .348$

Table 4: The degrees of freedom, F-statistic, and p-value in Levene's test of Homogeneity of Variances for linguistic distancing and each dimension of linguistic distancing across the subsets of the *AgeCorpus*.

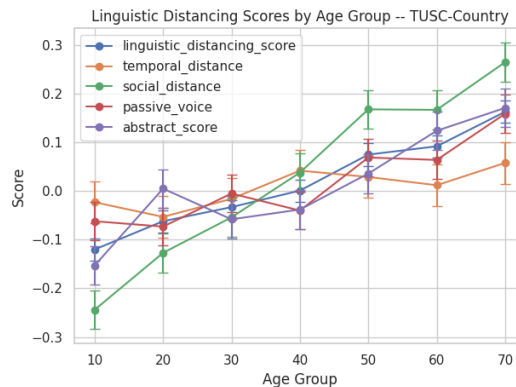


Figure 2: **Linguistic Distancing** scores for TUSC-Country subset of the *AgeCorpus* (blue line). Individual components of linguistic distancing across age groups are shown: temporal distance (orange line), social distance (green line), passive voice (red line) and abstractness (purple line). Error bars represent the standard error of the mean.



Age Group	#Posts		
	Country	City	Reddit
13–19	551	94,857	9,281,055
20–29	3,076	406,027	15,455,426
30–39	3,001	462,238	6,151,757
40–49	1,808	360,946	1,220,498
50–59	1,541	278,809	449,230
60–69	1,551	218,008	246,398
70–79	419	97,432	160,465

Table 5: The number of posts across the age groups in each subset of the *AgeCorpus*. “Country” refers to TUSC-Country and “City” refers to TUSC-City subsets of the dataset. We use this data in our experiments.

<b>Dataset</b>	<b>Metric</b>	<b>df1</b>	<b>df2</b>	<b>F-statistic</b>	<b>P-value</b>	<b>Effect Size (<i>est</i> <math>\omega^2</math>)</b>
Reddit	Linguistic	6	1203317.10	17703.63	$p < .001$	0.003
TUSC-Country	Linguistic	6	11940	29.66	$p < .001$	0.015
TUSC-City	Linguistic	6	539267.68	2830.26	$p < .001$	0.009

Table 6: The degrees of freedom (for the numerator and denominator), F-statistic, p-value, and effect size in the one-way ANOVA test for differences in linguistic distancing between age groups. Welch's ANOVA was performed for Reddit and TUSC-City.

# MOSAIC : a Corpus of Small-Group Interactions During a Collaborative Task

Amine Benamara<sup>1</sup>, Celine Clavel<sup>1</sup>, Brian Ravenet<sup>1</sup>, Nicolas Sabouret<sup>1</sup>,  
Mathilde Sassier-Roublin<sup>2</sup>, Julien Saunier<sup>2</sup>

<sup>1</sup>Université Paris-Saclay, CNRS, LISN, <sup>2</sup>Normandie Université, INSA Rouen, LITIS  
{benamara, clavel, ravenet, sabouret}@lisn.upsaclay.fr  
julien.saunier, mathilde.sassier-roublin@insa-rouen.fr

## Abstract

This paper presents MOSAIC (Multimodal Observations of Social Affect, Intimacy, and Cohesion), a multimodal interaction corpus of video and audio recordings of 17 groups of 4 participants (68 participants in total) playing a collaborative board game. The aim of this corpus collection is to support the design of socially interactive agents. We describe the experimental protocol of this corpus collection, the perceptive questionnaires completed by participants, the automatic annotation process of game specific elements and non-verbal behaviors, and the manual verbal annotations collected. We provide a preliminary characterization of this corpus with descriptive results for some of the perceptive scales used in this study. We illustrate the possibilities offered by this corpus on the question of interpersonal social relations and group dynamics.

**Keywords:** multimodal interaction, corpus, collaborative task

## 1. Introduction

Artificial agents are increasingly designed to collaborate with humans in small groups, for instance in education, training, or decision-making settings (Sebo et al., 2020). In such mixed teams, interaction quality depends not only on task performance but also on group cohesion, i.e., the extent to which members feel connected and committed to a shared goal (Mikalachki, 1969; Salas et al., 2015). Cohesion influences engagement, motivation, and collective performance (Beal et al., 2003), and therefore constitutes a central challenge for socially interactive agents operating in multiparty environments.

Recent work in social psychology indicates that cohesion partly emerges from situational intimacy i.e. short-term interpersonal processes involving positive affect, mutual understanding, and self-disclosure (Reis, 2017; Prager, 2000). For artificial agents to adapt their behavior in group interactions, computational models must thus capture how multimodal behaviors relate to perceived intimacy and cohesion over time.

However, this relationship cannot be directly studied using existing multiparty interaction corpora. Meeting corpora such as AMI focus on decision-making dynamics (Carletta et al., 2006), while datasets such as MULTISIMO target task-oriented collaboration (Koutsombogera and Vogel, 2018). Other resources, including GAME-ON, measure group cohesion but do not jointly capture interpersonal intimacy and multimodal behavioral signals (Maman et al., 2020). As a result, no current corpus aligns behavioral cues, perceived intimacy, and cohesion dynamics in short-term collaborative in-

teractions. While previous corpora measure either group-level outcomes or behavioral signals, none enable the investigation of how interpersonal relational processes unfold over time and differentially contribute to group-level cohesion in short-term collaborative settings.

To address this gap, we introduce MOSAIC (Multimodal Observations of Social Affect, Intimacy, and Cohesion), a multimodal corpus of four-person collaborative interactions collected during a cooperative word-association game. The dataset contains recordings from 68 participants organized into 17 groups, resulting in 136 interaction sessions, adding up to more than 6 hours of multimodal data. Each session captures the emergence of short-term group relationships.

The corpus combines multimodal behavioral recordings with aligned social perception measures, enabling the study of how interpersonal behaviors relate to perceived intimacy and group cohesion. It includes audio-video recordings, transcriptions, gaze targets, facial behavior signals, interaction phases derived from game logs, and post-interaction questionnaires measuring cohesion and intimacy (self- and other-report), while also controlling for prior familiarity between participants.

The MOSAIC corpus supports research in social signal processing, multimodal interaction modeling, and the design of socially adaptive artificial agents in group settings.

This paper provides:

- A multimodal multiparty interaction corpus designed to study situational intimacy and group cohesion
- A role-based interaction phase annotation

scheme automatically derived from game logs

- Multilevel social perception measures (self, interpersonal, and group) aligned with behavioral recordings
- Baseline analyses revealing differential temporal dynamics between intimacy and cohesion

Our preliminary analyses suggest that intimacy and cohesion follow partially distinct temporal dynamics, especially depending on prior familiarity between participants, highlighting the need for fine-grained multimodal corpora capturing both constructs simultaneously.

## 2. Related works

Understanding group interaction requires capturing interpersonal processes that are not directly observable but inferred from behavior. Psychological models describe intimacy as a relational process involving self-disclosure and partner responsiveness during interaction (Laurenceau et al., 1998; Reis, 2017; Prager, 2000). In computational and human-agent interaction research, this process has been operationalized through observable behaviors such as self-disclosure, positive affect, and mutual understanding (Potdevin, 2020). However, these approaches have primarily been investigated in dyadic settings, leaving open how such interpersonal mechanisms manifest in multiparty collaboration.

Psychological studies validate perceived intimacy without directly observing its multimodal behavioral realization, while HCI studies manipulate predefined behaviors assumed to express intimacy. Previous work has therefore not examined how intimacy is expressed through multimodal behaviors in spontaneous multiparty interaction.

In human-agent interaction, these mechanisms have been shown to influence interaction outcomes. Agents using self-disclosure or reciprocal behaviors increase trust, engagement, and social presence (Bickmore and Schulman, 2012; Lee and Choi, 2017; Potdevin, 2020). Similarly, robots expressing vulnerability improve human-human communication and cooperation within teams (Traeger et al., 2020). These findings highlight the functional importance of intimacy-related behaviors but do not provide observational data from natural multiparty interactions.

From a data perspective, social signal processing research relies on multimodal corpora to analyze group dynamics and train computational models (Bänziger et al., 2012; Soleymani et al., 2019; Kebe et al., 2024). Multiparty interactions are particularly challenging because participants simultaneously manage turn-taking, gaze, spatial positioning, and

multiple conversational threads (Gillet et al., 2022). As a result, existing datasets have typically been designed around specific interaction contexts and target processes.

Meeting corpora such as AMI (Carletta et al., 2006) capture decision-making and role distribution in professional discussions using audio, video, and manual behavioral annotations. Other datasets focus on collaborative task solving: MULTISIMO (Koutsombogera and Vogel, 2018) records small groups performing a quiz task and provides multimodal annotations for affect, engagement, and interaction strategies, while linking them to individual participant traits. Some corpora investigate emergent group states, for instance leadership (Sanchez-Cortes et al., 2011) or cohesion. The GAME-ON dataset (Maman et al., 2020) specifically targets group cohesion in a cooperative escape-game scenario and combines multimodal behavioral recordings with group perception questionnaires.

Although these resources provide rich behavioral recordings and annotations, they typically capture either interaction behavior (e.g., coordination, engagement, roles) or collective perception (e.g., cohesion, leadership). They do not explicitly model interpersonal mechanisms underlying these group states, and in particular they lack aligned measures of perceived intimacy between participants during the interaction. Consequently, existing corpora do not allow studying how multimodal interaction patterns between individuals contribute to the temporal emergence of group cohesion in short-term collaborative settings.

## 3. Corpus Description and Data Acquisition

### 3.1. Data availability

Due to ethical constraints and participant consent limitations, video recordings cannot be publicly shared. Access to anonymized behavioral data is available upon request to the authors for research purposes, subject to a confidentiality agreement. This restriction is designed to protect participant anonymity while still providing access to the essential data needed for result reproducibility.

### 3.2. Participants

We recruited 68 participants, split into 17 groups of 4 (50 female, 18 male) aged between 18 and 19 ( $m=18.2$  ;  $std=0.4$ ). The participants were first and second year students from the University Bachelor of Technology. Participants received a cookie and 0.5 point bonus on their psychology course grade (small enough not to constitute a form of coercion

or pressure, while still serving as an incentive), with an alternative task offered for non-participants, ensuring equitable and ethical compensation. We collected a total of 136 videos (one video for each participant and for each session) with a total duration of 361.5 minutes (6h 1min 30s). Each game session lasted between 8min 43s and 14min 17s ( $m=10\text{min}38\text{s}$ ,  $\text{std}=1\text{min}2\text{s}$ ).

### 3.3. Data Collection and Recording Setup

The MOSAIC corpus was collected during structured collaborative interactions involving four participants engaged in a cooperative word-association task. The data collection protocol combined a controlled game environment with synchronized multimodal recording.

#### 3.3.1. Collaborative task

Participants played two rounds of a computerized version of *Cross Clues*, a cooperative board game designed by *Blue Orange Games*. In this game, players have to make associations between words to score a maximum number of points during a restricted time.

The game board is composed of a 4x4 grid in which each column and each row is associated with a unique word. Each cell is thus associated with a pair of two words ( $w_c, w_r$ ). At the beginning of the game, each player draws a card from the deck of 16 cards, each one designating the coordinates of a cell (e.g. "B3"), and keeps its content secret. Each player starts to think about a clue word linked to both words associated with the coordinates on their card. For instance, if column B is associated to the word "banana" and line 3 with the word "tissue", the player must find a single clue word that would suggest to the others these two words, so that they can guess the correct coordinate B3. At any time, any player can take the floor to suggest a clue word, taking the role of *suggester* during this turn. The 3 other players take the role of *guessers* during this turn, and have to guess the coordinates on the *suggester*'s card. The *suggester* can only announce their clue word to the other players and cannot provide any other information. The *guessers* discuss to agree on a proposition of coordinates, then select the cell corresponding to their choice on the touchscreen. The *suggester* can then validate or invalidate their guess. If the proposition is correct, the card is placed on the grid. If it is incorrect, the card is discarded without revealing it. A new turn then begins. The roles can either change, if a different player from the previous turn decides to suggest a clue word, or stay the same if the same player decides to suggest another clue word. The session ends when all the 16 coordinates cards have been played (placed on the grid or discarded), or when

the time has elapsed. The task structure induces repeated cycles of clue production, collaborative reasoning, agreement negotiation, and feedback, creating rich multiparty interaction data.

The task was selected to elicit spontaneous yet structured multiparty interaction involving turn-taking, joint decision-making, and role alternation. The game's mechanics (time pressure, iterative turns, and role-based interactions) encourage participants to adapt their communication, offer support, and resolve ambiguities collaboratively. Success depends on effective communication, mutual understanding, and coordinated effort, as one participant (the *suggester*) provides clue words while the others (the *guessers*) collectively interpret and agree on the correct coordinates. This dynamic also requires active listening, perspective-taking, and consensus-building, as participants must align their interpretations and justify their reasoning to reach a shared decision. These interactions can also require trust, cooperation, and emotional engagement between participants, reflected by reactions to incorrect proposition, celebration of success, and role adjustments. Additionally, the cooperative structure fosters shared emotional experiences, like tension, humor, or collective achievement, which strengthen interpersonal bonds.

#### 3.3.2. Recording Environment

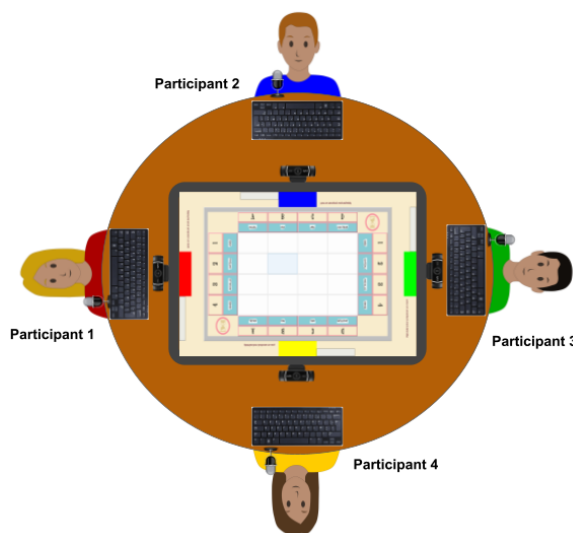


Figure 1: Illustration of the experimental setting.

An illustration of the setting is displayed in Figure 1. Four participants are seated around a round table, with a tactile screen in the center of the table. Each participant is assigned a color (red, green, blue or yellow), corresponding to a specific seat around the table. They are given a badge holder containing a card of the color assigned to them, which they must wear throughout the experiment.



A microphone (Lavalier II) is then attached to the necklace of the badge holder of each participant to capture their speech (AAC-LC 48 kHz - 162 kb/s). Four identical cameras (Logitech C922 Pro) are placed around the screen to capture each participant's image during the game from the opposite side (H.264 720p 60fps - 2500kb/s). Four keyboards are also placed near each participant. A video of the content of the screen is recorded using OBS Studio.

### 3.3.3. Experimental procedure

This study received an approval of the ethics committee of the Paris Saclay University (CER-Paris-Saclay-2024-100). Participants were recruited in groups of four and completed the experiment in a single session.

After providing informed consent, participants received instructions about the game rules and interface. A 5-minute familiarization round was conducted to ensure understanding of the mechanics.

Each group then played two consecutive game sessions of approximately 10 minutes each. After each session, participants completed a set of questionnaires measuring their perception of the interaction

## 3.4. Multimodal Data Streams

The corpus contains synchronized audio, video, game logs, gaze annotations, facial action unit activations, and speech transcriptions.

### 3.4.1. Game logs and annotation scheme

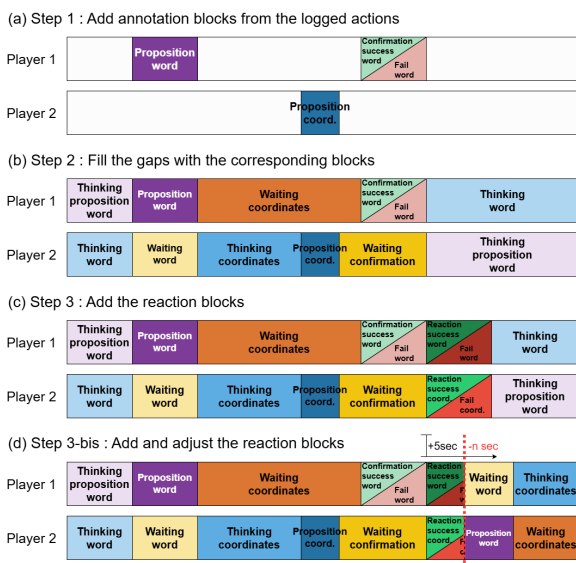


Figure 2: Annotation process of game phases from game logs.

We defined a role-based annotation scheme to automatically derive interaction phases from game logs. Game interface actions were logged with timestamps, including: clue word submission, coordinate proposition, confirmation of success or failure. Based on these logged events, we generated annotation blocks defined by a start time, end time, and categorical label. We distinguish actions made by the *suggester* and the ones made by the *guessers*. For the *suggester* role, five interaction phases are defined : thinking about a word, suggesting a word, waiting the coordinates proposition of the *guessers*, confirming the proposition (success or fail) and the reaction to the result (success or fail). For the *guessers*, six interaction phases are defined : thinking about a word, waiting for the *suggester* to give the word, thinking about possible coordinates, proposition of coordinates, waiting the confirmation from the *suggester* and finally the reaction to the result (success or fail). We also have one category associated to the end of the session and one to a participant canceling a proposition. The annotation scheme is thus composed of a total of 16 annotation categories, with 7 categories for the *suggester* role, 7 categories for the *guesser* role and 2 unrelated to the role (End session and Cancel proposition). For higher-level analyses, categories can be grouped into five broader process types (Collaboration, Observation, Thinking, Reaction, and Game-related actions). Their distribution across the corpus is presented in Section 4. The annotation process follows three steps: (1) Creation of base blocks from explicit logged actions (2) Temporal inference of intermediate phases (e.g., thinking, waiting) (3) Addition of reaction phases with a maximum duration of 5 seconds (empirically determined).

The process is illustrated on Figure 2, with a simplified example of two players' timelines. Using this annotation process, we annotated a total of 9168 blocks.

### 3.4.2. Performance scores

Using the logs of the game, we provide performance measures for each game session and for the combined session, from both group-level (number of successful and failed attempts) and individual-level, such as number of total propositions (fail or success) as *guesser* (coordinates) and *suggester* (clue-word).

### 3.4.3. Facial Action Unit Extraction

Facial behavior signals were automatically extracted using OpenFace (Baltrusaitis et al., 2018), a widely used toolkit for facial behavior analysis. OpenFace can, among others, detect and extract the intensity (value between 0.0 and 5.0) of 17 ac-

tion units (AU) on a human face image or video. We used this tool to work on 13 facial behaviors. Based on our expertise with OpenFace data, we provide, in addition to raw data, filtered and preprocessed data suitable for analysis. We excluded the AU 45 (blink), as we do not focus on blinks in our analysis, the AUs 25 and 26 (lips part and jaw drop), because it is often confused with speaking, and AU 14, because it has been shown that its detection is unreliable (Namba et al., 2021). Additionally, we used the AUs 6 and 12 together, to detect smiles, and the AUs 1 and 2, both separated, to distinguish inner brows raising from outer brows raising, and together, to detect brows raised from both sides.

OpenFace also provides a confidence value, that indicates how much confident (value between 0 and 1) the system is about the detection of the face. To minimize false detections and remove noise, we ignore all detections with a confidence lower than 0.8, .

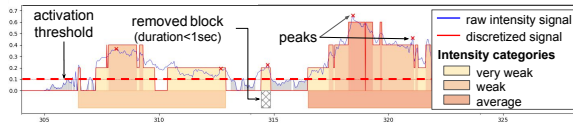


Figure 3: Illustration of the extraction process of Action Units (AUs). The raw signal from Openface (plain blue line) is first filtered using an activation threshold (dashed red line), to remove noise (grey areas). The peaks (red crosses) and their associated time boundaries (not represented to avoid cluttering the figure) are identified. The signal is then discretized (plain red line), and associated with intensity categories. Successive blocks are then merged in a single block with the maximum intensity of the merged blocks. Finally, blocks shorter than 1 second are removed (grey hatched blocks).

The processing pipeline consisted of: : (1) Normalization of raw AU intensity signals, (2) Peak detection using SciPy (minimum height = 0.05; minimum inter-peak distance = 80 frames), (3) Discretization of intensity values into five categories (Very weak [0.1-0.3[, Weak [0.3-0.5[, Average [0.5-0.7[, Strong [0.7-0.9[, Very strong[0.9-1]), (4) Conversion into temporal annotation blocks defined by sustained activation at the same intensity level, (5) Merging of successive blocks separated by gaps below 1 second, (6) Removal of blocks shorter than 1 second.

We illustrate this process on Figure 3, with a 20 seconds signal of a single AU, extracted from our corpus.

### 3.4.4. Gaze targets

Gaze vector estimation refers to the task of extracting gaze direction vectors from a video or an image

of a human face. SpaTaTrack<sup>1</sup> (Spatial gaze Target Tracking) is a tool that uses the output from a camera to extract gaze direction vectors and identify the gaze targets of a human in a video. It allows to define the scene configuration by placing the different objects using a global coordinates system, and project the gaze vectors in this scene to identify the target of the gaze. A visual interface allows manual inspection and validation before extracting the targets in batch.

We used the SpaTaTrack tool configured with our experimental setup, and extracted for each participant and for each frame the gaze targets, which can take one of the following values :

- screen : when the target is either the table or the screen,
- red, blue, green or yellow : when the target is one of the participants,
- other : when the gaze vector does not intersect with any of the defined objects or zones.

We extracted a total of 35567 gaze target blocks, with 41.4% screen, 37.2% participant (red, blue, green or yellow) and 21.4% other.

### 3.4.5. Verbal transcriptions

Due to the close proximity of participants, individual audio tracks were not acoustically isolated, which compromised the performance of standard automatic transcription, specifically during crosstalk. To address this challenge, we used the HappyScribe service<sup>2</sup>, which provided time-stamped annotations and diarization performed by professional linguists experienced in transcription tasks. The annotation process was conducted offline, allowing annotators to carefully review the audio. The service includes quality-control procedures to ensure the accuracy and reliability of transcriptions. While the task was generally straightforward, it became more challenging during segments where more than two participants spoke simultaneously. Each speech segment was divided into block units, defined as continuous utterances separated by pauses of at least 300 milliseconds. Each block is associated with a start time, end time, speaker identifier, and textual content. All transcripts are temporally aligned with the corresponding audio, video, and interaction phase annotations through shared timestamps.

### 3.4.6. Social perception measures

In addition to multimodal behavioral recordings, the corpus includes session level subjective measures

<sup>1</sup><https://github.com/mlamine21/SpaTaTrack>

<sup>2</sup>More information is available at: <https://www.happyscribe.com>

collected after each game round. The following questionnaires were administrated :

**Group Cohesion Questionnaire (GCQ)** (Sassier-Roubin et al., 2026) : Measure of individual perception of group cohesion according to three dimensions : Social (13 items), Task (3 items) and Communication (3 item). Responses are provided on a 7-point Likert scale and reflect group-level perception.

**Social Belonging Scale** (Richer and Vallerand, 1996) : Measures perceived belongingness and acceptance within the group. Responses are provided on a 7-point Likert scale along 5 items for the Acceptation dimension, and 5 items for the Intimacy dimension. Scores reflect individual-level perception.

**Virtual Intimacy Scale (VIS)** (Potdevin, 2020). Measures perceived intimacy in social interaction. Two versions were administered: (1) a self-report (Self-VIS), in which participants rated their own intimacy, (2) and an other-report (Front-VIS), in which they evaluated the intimacy of the person facing them during the game. The use of these two complementary measures was intended to capture both their mutual influence and potential incongruence between self and partner perceptions. Responses are provided on a continuous 0–100 scale across three dimensions : honesty and genuineness (5 items), positivity (4 items) and mutual comprehension (6 items).

**Pre-existing relationship** : Participants rated how well they knew each other participant on a 5-point scale (1 = Not at all, 5 = Very well).

## 4. Results

### 4.1. Dataset characteristics

#### 4.1.1. Annotation categories distribution

Interaction phase annotations were grouped into five higher-level process types:

**Collaboration** (53.2% of the blocks): This group concerns both the *suggester* ("Waiting coordinates") who can give clues using their non-verbal behavior to give indirect feedback to *guessers* during their thinking process, and the *guessers* ("Thinking coordinates"), who share in a more direct way their thinking process and their opinion to the other players, based on cues such as interruptions, overlapping speech, attention allocation, agreement, disagreement or doubt expression.

**Observation** (16.9%): moments when *guessers* observe the *suggester* writing a word ("Waiting word") and gather information about how the *suggester* feel about their word: "do they look confident? are they happy about their proposition?" or observe both the *suggester* and the other *guessers*

reacting about their coordinates proposition ("Waiting confirmation").

**Thinking** (15.3%): as a mental processes before a word is suggested, indicating a reflexion process: "how can I link these words? Are there other possible coordinates that could confuse them?". Participants can either emphasize that they are about to suggest a word ("Thinking proposition word" or "Thinking word"), or that they probably won't suggest a word because they can't find an easy link between the words associated to their coordinates ("Thinking word").

**Reactions** (8.8%): emotional expression of the *suggester* right before they confirm coordinates ("Confirmation success/fail") and both *suggester* and *guessers* the moment after the answer is revealed ("Reaction success/fail coordinates/word").

**Game-related** (5.4%): mostly unrelated with group or interpersonal processes, refers to moments where the *suggester* is writing their word on the keyboard ("Proposition word"), when the *guessers* are clicking on the coordinates they chose ("Proposition coordinates"), the "End session" and the "Cancel proposition" categories.

Their distribution across the corpus (percentage of annotation blocks) and the categories they contain is illustrated in Figure 4.

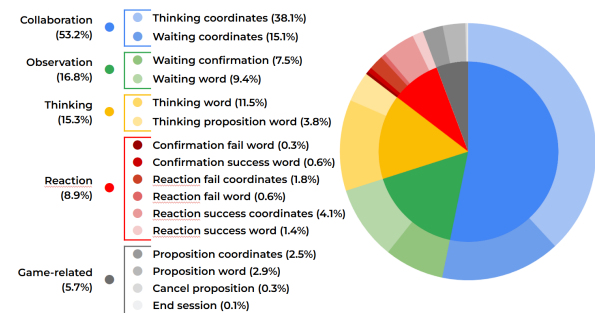


Figure 4: Distribution of interaction phase categories across the corpus.

#### 4.1.2. Facial action unit distribution

A total of 27,648 facial activation blocks were extracted across the corpus. The facial behaviors we used can be found on the pie charts (Figure 5), with percentages indicating their respective proportions.

#### 4.1.3. Gaze target distribution

A total of 35567 gaze target blocks were extracted across the corpus. Gaze was distributed toward the screen in 41.4% of annotated blocks, toward another participant (red, blue, green or yellow) in 37.2% of the blocks and categorized as other in

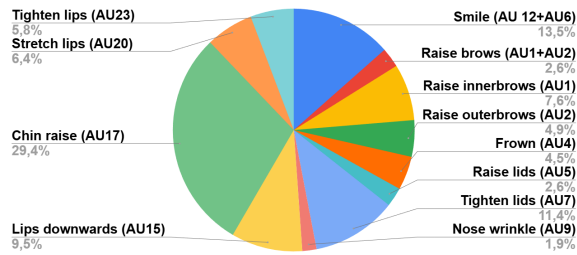


Figure 5: Distribution of extracted facial action units across the corpus.

21.4% of the cases. This distribution reflects the relative proportion of attention allocated to the shared task interface, co-participants, and non-defined targets.

#### 4.1.4. Speech statistics

The corpus is composed of 11963 blocks, with a total speech duration of 3 hours, 33 minutes, 26.74 seconds (12806.74 seconds), and a word count of 44565. The average block duration is 1.07 second, with an average of 3.73 words per block.

#### 4.1.5. Social perception measures

**Cohesion:** The overall cohesion score (Total dimension) indicates a high level of perceived cohesion across all participants ( $M = 5.97$ ,  $SD = 0.84$ ). Cohesion was particularly high along the Task dimension ( $M = 6.50$ ,  $SD = 0.70$ ) and the Communication dimension ( $M = 6.22$ ,  $SD = 0.76$ ) with lower variability compared to the Social dimension ( $M = 5.74$ ,  $SD = 1.10$ ).

**Intimacy:** Participants reported a high overall level (Total dimension) of intimacy for both the self-report ( $M = 78.79$ ,  $SD = 13.04$ ) and the front-report ( $M = 78.80$ ,  $SD = 13.75$ ). The Positivity dimension received the highest scores (Self :  $M = 83.93$ ,  $SD = 13.02$ , Front :  $M = 84.43$ ,  $S = 12.62$ ), followed by the Mutual Comprehension dimension (Self :  $M = 79.02$ ,  $SD = 15.95$ , Front :  $M = 78.50$ ,  $SD = 17.27$ ), while the Honesty dimension was rated comparatively low, yet still high (Self :  $M = 73.34$ ,  $SD = 15.61$ , Front :  $M = 73.60$ ,  $SD = 16.04$ ).

**Correlations:** We computed correlation coefficients (Spearman when the data did not meet the assumption of normality, Pearson otherwise) to investigate how the components of intimacy and cohesion interact. Concerning the self-reported intimacy, the analysis revealed significant associations between all dimensions of both cohesion and intimacy. However, we can note substantial differences in the strength of these associations, with strong correlations with the social dimension of the cohesion (Honesty(H) :  $r = 0.520$ , Positivity(P) :  $r = 0.620$ , Comprehension(C) :  $r = 0.607$ ;  $p < 0.001$  for

all dimensions), followed by the communication dimension with moderate associations (H :  $r = 0.322$ , P :  $r = 0.414$ , C :  $r = 0.442$ ;  $p < 0.001$  for all dimensions) and finally weak associations for the Task dimension (H :  $r = 0.235$ , P :  $r = 0.289$ , C :  $r = 0.273$ ;  $p < 0.001$  for all dimensions). We observed a similar pattern with the front-reported intimacy, but with weaker associations for the Social dimension (H :  $r = 0.357$ , P :  $r = 0.474$ , C :  $r = 0.509$ ;  $p < 0.001$  for all dimensions), the Communication dimension (H :  $r = 0.190$   $p < 0.05$ , P :  $r = 0.314$   $p < 0.001$ , C :  $r = 0.392$   $p < 0.001$ ) and even non significant for the Task dimension (H :  $r = 0.091$   $p < 0.15$ , P :  $r = 0.208$   $p < 0.01$ , C :  $r = 0.300$   $p < 0.001$ ).

## 4.2. Evolution of subjective measures across sessions

### 4.2.1. Statistical procedure

We analyzed the evolution of the scores across the 2 game sessions to investigate the dynamics created throughout the experiment, using one-sided Wilcoxon signed-rank tests when the data distributions did not meet the assumption of normality, paired t-test otherwise, with the 'less' alternate hypothesis (Session 1 < Session 2), indicating a hypothesized increase after Session 2.

As the concepts studied (cohesion and intimacy) are dependent on existing links between the participants, we used the Links scale (see Section 3.4.6) to calculate a mean score (scale from 1 to 5) for each participant based on the strength of their existing relationship with the other three players. Participants were then divided into two groups: the Unfamiliar group, who reported not knowing the other participants well (mean score < 3,  $n=21$ ,  $M = 2.16$   $SD = 0.54$ ) and the Familiar group, who reported knowing them well (mean score  $\geq 3$ ,  $n=47$ ,  $M = 3.77$   $SD = 0.58$ ).

### 4.2.2. Evolution across sessions

**Cohesion:** Across all participants, significant increase was observed for total cohesion ( $p = 3.50e-02$ ) score, reflecting a global increase in the perception of the cohesion after the second session. Social cohesion also exhibited a significant increase ( $p = 3.81e-02$ ), but for the Task ( $p = 5.56e-02$ ) and Communication ( $p = 5.53e-01$ ) dimensions, the differences were not statistically significant, while still showing a tendency to increase for the Task dimension.

**Intimacy:** Perceived intimacy increased significantly across sessions for both self-report and front-report when considering all participants ( $p < 0.001$  for all dimensions).



### 4.2.3. Effect of Links

**Cohesion:** The participants in the Unfamiliar group registered a significant increase in the cohesion score, for the Social dimension ( $p < 0.001$ ), the Task dimension ( $p = 4.51e-02$ ) and Total cohesion ( $p < 0.01$ ), while the Communication dimension did not show a significant increase ( $p = 2.85e-01$ ).

For the Familiar group, the perceived cohesion did not show a significant increase for any of the dimensions across the sessions (Social :  $p = 5.50e-01$ , Task :  $p = 2.80e-01$ , Communication :  $p = 7.02e-01$ , Total :  $p = 5.48e-01$ ).

**Intimacy:** For the Unfamiliar group, overall intimacy showed a significant increase (self and front :  $p < 0.01$ ) as well as the Honesty dimension (self :  $p < 0.01$ , front :  $p < 0.001$ ) and the Positivity dimension (self :  $p < 0.01$ , front :  $p < 0.05$ ). No significant increase was observed for the self-reported Comprehension dimension ( $p = 1.52e-01$ ), whereas significant increase emerged in the front-report for the same dimension ( $p = 3.98e-02$ ).

For the familiar group, overall intimacy and all dimensions showed a significant increase ( $p < 0.001$ ) for both the self and front report.

## 5. Discussion

**Cohesion:** The participants perceived high levels of cohesion along the Task and Communication dimensions. No significant evolution across sessions were observed for these dimensions. The social cohesion scores, while still above average, were relatively lower than the other dimensions, but showed a significant increase across sessions. According to the literature, cohesion can build around the shared objective of the group, later evolving and deepening through the social dimension, reinforcing bonds within the group, encouraging vulnerability exposure and risk taking to achieve better results (Forsyth and Burnette, 2010; Marmarosh and Sproul, 2021). These results could suggest an already established cohesion on the Task and Communication level, leading participants to reinforce and focus on social cohesion. This initial high level of task and communication cohesion, specifically in Familiar groups, could be due to participants having already interacted together in a collaborative context (e.g. group works in university setting) and possibly already sharing effective task and communication-level collaboration strategies.

The results also showed that Social and Task cohesion increased for Unfamiliar groups but not for Familiar groups. The Communication dimension seems to remain stable across the sessions, even if we can observe a slightly higher increase in the Unfamiliar group. This reveals an interesting impact of familiarity on cohesion, suggesting that building

cohesion plays a more critical role among unfamiliar participants. As stated by (Ogrodniczuk et al., 2021), cohesion can move from initial uncertainty, starting from members getting to know one another and build trust, to deeper collaboration at a later stage, where cohesion gradually develops. This indicates that familiarity may impact the need for social bonding to enhance cohesion, as established relationships already reached a baseline level of cohesion, highlighting the importance of social skills in emerging groups. Moreover, these results indicate that the experimental setting allowed the observation of meaningful changes in perceived cohesion across sessions, providing a relevant framework for studying cohesion, specifically through the social dimension, and its dynamics in collaborative interactions. However, the high overall scores is a limitation of this dataset, as it makes it difficult to identify possible indicators associated to low cohesion scores.

**Intimacy:** The intimacy scores from both self and front report revealed high level of perceived intimacy. According to (Kozlowski and Chao, 2012), first impressions guide initial group interactions, often relying on visible traits such as race, age, and gender. They also suggest that groups with similar surface-level characteristics may experience stronger initial attraction among members. The high initial level of overall intimacy could then be partly explained by the shared characteristics (age, social class, education level, high semantic alignment and shared reference) of the participants, specifically for the Mutual Comprehension dimension scores. The collaborative game context, with low (for competitive participants who seek performance) to non-existent stakes, probably helped in creating a positive and relaxed atmosphere, and could explain the Positivity dimension displaying the highest scores among the dimensions. The results should then be taken carefully by considering the specificities of our sample.

The perception of intimacy increased significantly across sessions, and familiarity did not seem to have an impact on intimacy evolution, which displayed a significant increase for all dimensions except for the self-reported intimacy for the Mutual comprehension dimension in the unfamiliar group, which remained stable. This suggests that while participants did not perceive a change in their own attempt to enhance Mutual understanding, they perceived more easily this change in their interaction partner when social cohesion increased. Additionally, we can conclude that while cohesion seems to reach a peak in cohesion level, intimacy keeps increasing, potentially helping reach a higher peak of cohesion after longer period of collaborative interactions.

**Correlations:** The first conclusion suggested by



the results is that intimacy has few impact on the perception of cohesion along the Task dimension. For the Communication and Social dimensions of cohesion, results seem to indicate a strong association with both Positivity and Comprehension dimensions of intimacy. The results for the Honesty dimension appear more mitigated, and do not allow to draw conclusions on its link with the perception of cohesion.

## 6. Limitations

While the MOSAIC corpus provides valuable insights into multiparty interactions, several limitations must be acknowledged. First, the small sample size (68 participants) and demographically narrow population (18–19-year-old university students) restrict the generalizability of the findings to broader age groups or diverse cultural and social contexts. Additionally, the high and homogeneous scores for intimacy and cohesion across participants limit the dataset's ability to capture contrastive dynamics in terms of intimacy and cohesion. Finally, the automated extraction of gaze targets and facial expressions lacks human validation.

## 7. Perspectives and use cases

We introduced the MOSAIC corpus, an annotated video dataset aimed at exploring the links between cohesion, intimacy, and synchronized multimodal behaviors (action units, gaze targets and speech content) associated to these concepts. We described the segmentation used on the corpus based on the unfolding phases of the collaborative task, and suggested a grouping based on high-level processes, allowing the interaction to be studied from different angles. The collaboration interaction phases can be used to study group dynamics while focusing on decision-making or negotiation, and the observation and thinking phases can help gain insights on mental processes and their expression, and the reaction phases could be useful to explore emotional expressions in a collaborative context.

In the first analysis of questionnaires we provided, we focused on individual perceptions of cohesion and intimacy. This work could be extended to group-level analysis to explore for example disparity and homogeneity in the group's perception, or to examine how these constructs could be related to the group's performance in terms of success and communication quality. Additionally, we could further explore the interactions between the different measures by taking into account familiarity in the correlation analysis, and including the results for the social belonging scale.

This corpus could extend the contributions on the links between multimodal behaviors and high

levels of intimacy (Shaver and Reis, 1988; Soleymani et al., 2019; Potdevin, 2020; Kang et al., 2012) and on the identification of markers of high levels of cohesion (Maman, 2022). The observation of an evolution of both cohesion and intimacy could additionally inform on the specific changes related to language and non-verbal behaviors that led to this evolution. Several additional annotations could also be introduced to extend this corpus, such as verbal features like interruptions, laughter, sentiments and intentions associated to utterances, or cross-modal features, like emotional states or communication quality.

Finally, this corpus can be used to inform the design of socially interactive agents displaying intimacy behaviors (verbal and non-verbal) during the same collaborative task. The annotation categories described in Section 3.4.1 and used in this work as an analysis framework also represent the different possible states of an agent interacting in this task. The contextual analysis of the participants behaviors during each of these states will help to identify and then reproduce behavior patterns associated with different intimacy levels on socially interactive agents. These generated behaviors will then be evaluated in interaction with humans to explore their relevance and realism, as well as the impact of different embodiments (humans vs robots or virtual agents) on the perception of cohesion and perceived intimacy.

## 8. Acknowledgments

This work was supported by a French government grant managed by the Agence Nationale de la Recherche as part of the France 2030 program, reference ANR-22-EXEN-0004 (PEPR eSEMBLE / PC3 MATCHING).

## 9. References

- Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis Philippe Morency. 2018. [OpenFace 2.0: Facial behavior analysis toolkit](#). In *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, pages 59–66.
- Daniel J. Beal, Robin R. Cohen, Michael J. Burke, and Christy L. McLendon. 2003. [Cohesion and Performance in Groups: A Meta-Analytic Clarification of Construct Relations](#). *Journal of Applied Psychology*, 88(6):989–1004.
- Timothy Bickmore and Daniel Schulman. 2012. [Empirical Validation of an Accommodation Theory-Based Model of User-Agent Relationship](#). In

- David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, Gerhard Weikum, Yukiko Nakano, Michael Neff, Ana Paiva, and Marilyn Walker, editors, *Intelligent Virtual Agents*, volume 7502, pages 390–403. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.
- Tanja Bänziger, Marcello Mortillaro, and Klaus R. Scherer. 2012. [Introducing the Geneva Multimodal expression corpus for experimental research on emotion perception](#). *Emotion*, 12(5):1161–1179.
- Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. [The AMI Meeting Corpus: A Pre-announcement](#). In Steve Renals and Samy Bengio, editors, *Machine Learning for Multimodal Interaction*, volume 3869, pages 28–39. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.
- Donelson R. Forsyth and Jeni Burnette. 2010. Group Processes. In *Advanced Social Psychology: The State of the Science*, pages 495–534. Oxford : Oxford University Press.
- Sarah Gillet, Marynel Vázquez, Christopher Peters, Fangkai Yang, and Iolanda Leite. 2022. [Multi-party Interaction Between Humans and Socially Interactive Agents](#). In *The Handbook on Socially Interactive Agents: 20 years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics Volume 2: Interactivity, Platforms, Application*, 1 edition, volume 48, pages 113–154. Association for Computing Machinery, New York, NY, USA.
- Sin-Hwa Kang, Jonathan Gratch, Candy Sidner, Ron Artstein, Lixing Huang, and Louis-Philippe Morency. 2012. [Towards building a virtual counselor: modeling nonverbal behavior during intimate self-disclosure](#). In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 63–70.
- Gaoussou Youssouf Kebe, Mehmet Deniz Birlıkci, Auriane Boudin, Ryo Ishii, Jeffrey M. Girard, and Louis-Philippe Morency. 2024. [GeSTICS: A Multimodal Corpus for Studying Gesture Synthesis in Two-party Interactions with Contextualized Speech](#). In *Proceedings of the ACM International Conference on Intelligent Virtual Agents*, pages 1–10, GLASGOW United Kingdom. ACM.
- Maria Koutsombogera and Carl Vogel. 2018. [Modeling collaborative multimodal behavior in group dialogues: The MULTISIMO corpus](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Steve W. J. Kozlowski and Georgia T. Chao. 2012. [The Dynamics of Emergence: Cognition and Cohesion in Work Teams](#). *Managerial and Decision Economics*, 33(5-6):335–354. \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1002/mde.2552>.
- Jean-Philippe Laurenceau, Lisa Barrett, and Paula Pietromonaco. 1998. [Intimacy as an interpersonal process: the importance of self-disclosure, partner disclosure, and perceived partner responsiveness in interpersonal exchanges](#). *Journal of Personality and Social Psychology*, 74:1238–51.
- SeoYoung Lee and Junho Choi. 2017. [Enhancing user experience with conversational agent for movie recommendation: Effects of self-disclosure and reciprocity](#). *International Journal of Human-Computer Studies*, 103:95–105.
- Lucien Maman. 2022. *Automated analysis of cohesion in small groups interactions*. Ph.D. thesis, Institut Polytechnique de Paris, Paris.
- Lucien Maman, Eleonora Ceccaldi, Nale Lehmann-Willenbrock, Laurence Likforman-Sulem, Mohamed Chetouani, Gualtiero Volpe, and Giovanna Varni. 2020. [GAME-ON: a multimodal dataset for cohesion and group analysis](#). *IEEE Access*, 8:124185–124203.
- Cheri L. Marmarosh and Amy Sproul. 2021. [Group cohesion: Empirical evidence from group psychotherapy for those studying other areas of group work](#). In *The psychology of groups: The intersection of social psychology and psychotherapy research*, pages 169–189. American Psychological Association, Washington, DC, US.
- Alexander Mikalachki. 1969. *Group cohesion reconsidered : a study of blue collar work groups*. University of Western Ontario School of Business Administration.
- Shushi Namba, Wataru Sato, Masaki Osumi, and Koh Shimokawa. 2021. [Assessing automated facial action unit detection systems for analyzing cross-domain facial expression databases](#). *Sensors*, 21(12):4222.
- John S. Ogradniczuk, Joanna Cheek, and David Kealy. 2021. [Group therapy development: Implications for nontherapy groups](#). In *The psychology*

- of groups: *The intersection of social psychology and psychotherapy research*, pages 231–248. American Psychological Association, Washington, DC, US.
- Delphine Potdevin. 2020. *Vers des agents conversationnels animés sociaux: Quelle influence de l'intimité virtuelle sur l'expérience utilisateur et la relation-client?* PhD Thesis, Université Paris-Saclay.
- Karen J. Prager. 2000. *Intimacy in personal relationships*. In *Close relationships: A sourcebook*, pages 228–243. SAGE Publications, Inc.
- Harry T Reis. 2017. The interpersonal process model of intimacy: Maintaining intimacy through self-disclosure and responsiveness. In J. Fitzgerald (Ed.), *Foundations for couples' therapy: Research for the real world*, pages 216—225.
- Sylvie F. Richer and Robert J. Vallerand. 1996. *l'Echelle du sentiment d'appartenance sociale (ESAS)*. *Manuscrit inédit, Laboratoire de recherche sur le comportement social, Université du Québec à Montréal*.
- Eduardo Salas, Rebecca Grossman, Ashley M. Hughes, and Chris W. Coultas. 2015. *Measuring Team Cohesion: Observations from the Science*. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 57(3):365–374.
- Dairazalia Sanchez-Cortes, Oya Aran, and Daniel Gatica-Perez. 2011. *An audio visual corpus for emergent leader analysis*. In *Workshop on multi-modal corpora for machine learning: taking stock and road mapping the future, ICMI-MLMI*.
- Mathilde Sassier-Roublin, Amine Benamara, Céline Clavel, Julien Saunier, and Alexandre Pauchet. 2026. *Measuring Group Cohesion: Development and Validation of the Group Cohesion Questionnaire (GCQ)*. Working paper or preprint.
- Sarah Sebo, Brett Stoll, Brian Scassellati, and Malte F. Jung. 2020. *Robots in groups and teams: A literature review*. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–36.
- R. d Shaver and P. Reis. 1988. *Intimacy as an interpersonal process*. USA.
- Mohammad Soleymani, Kalin Stefanov, Sin-Hwa Kang, Jan Ondras, and Jonathan Gratch. 2019. *Multimodal Analysis and Estimation of Intimate Self-Disclosure*. In *2019 International Conference on Multimodal Interaction*, pages 59–68, Suzhou China. ACM.
- Margaret L. Traeger, Sarah Strohkorb Sebo, Malte Jung, Brian Scassellati, and Nicholas A.
- Christakis. 2020. *Vulnerable robots positively shape human conversational dynamics in a human–robot team*. *Proceedings of the National Academy of Sciences*, 117(12):6370–6375.

# Age and Affect in Language: How Emotion Expression on Social Media Varies Across Adulthood

Daniela Teodorescu\*, Jan Philip Wahle†, Saif M. Mohammad‡

\*University of Alberta, †University of Göttingen, ‡National Research Council Canada  
\*dteodore@ualberta.ca, †wahle@uni-goettingen.de, ‡saif.mohammad@nrc-cnrc.gc.ca

## Abstract

As we age, the way we experience and express emotions changes. This is because of a number of factors, including: changes in our body, differing types of experiences at different ages, improved emotion regulation strategies, and increasing experience of dealing with affective situations. However, work in psychology points to differing findings on how emotions, typically valence and happiness, changes with age. Psychologists measure happiness and well-being through questionnaires, which can have biases and result in limited data. Thus corpus analyses can provide useful complementary insights. We compile and release a large dataset of social media posts annotated with the age of the author at the time of posting. We refer to it as *AgeCorpus*. Using this dataset, we apply simple and interpretable methods to explore research questions pertaining to how social media posts, especially emotion expression through these posts, varies by age groups. Analyzing the emotions expressed in the posts, we find that the average valence increases until the middle ages, and then decreases; arousal decreases (Reddit)/plateaus with age (Twitter); and dominance follows the inverted U-shape (Reddit)/increases with age (Twitter). For categorical emotions, we find they follow the inverted U-shape on Reddit and increase in intensity with age on Twitter. We hope our dataset enables further research into age related phenomenon, such as well-being and language use.

**Keywords:** Age, Affect, Emotion, Valence, Arousal, Dominance, Categorical Emotions, Corpus creation, Social Media

## 1. Introduction

The way in which we perceive and express emotions is dependent on many aspects (e.g., environment, life events, personality dispositions, etc.), and change with age. With age and increased exposure to emotional situations, people's ability to regulate emotions improves and therefore older adults report higher levels of well-being (Urry and Gross, 2010; Gross et al., 1997). The way in which we perceive emotions is important as it directly impacts how we process that emotion information, respond, and act on that decision (Appraisal Theory of Emotion (Arnold, 1960)).

Literature in psychology has found that there are general trends in emotion expression with age. However, these findings do not align and point to differing trends which can be grouped into 3 broad categories: "U-shape" (Blanchflower and Oswald, 2008), "Inverted U-shape", and linear or there is no change with age. The U-shape theory says that generally happiness is highest for those in their early twenties to thirty-five, dips during midlife (e.g., 35-50 years old), and rises again during fifties to seventies, following a U-shaped curve. On the other hand, in the inverted U-shape theories happiness peaks during midlife, and is lowest in 20's to early 30's, and likewise again in older age. Whereas other work points to there being no relationship between emotions and age, with positive and negative emotions being evenly distributed across the life-course.

Understanding how emotions change across the

lifespan is critical for answering questions related to what changes (e.g., experience, regulation, expression) and why changes occur across age (e.g., cognition, biology, culture). Answering such questions is important to determine whether such changes are adaptive or problematic. If problematic, are there appropriate interventions that can be taken to help improve health outcomes and well-being and when should they be taken? Such findings can support clinicians provide more tailored care and appropriate interventions when needed. A better understanding the relationship between age and emotion is of interest to numerous fields: psychology (e.g., how does emotion experience, expression and regulation change across the lifespan), mental health (e.g., what does healthy changes look like vs. those associated with a pathology), education (e.g., emotional experiences are key to learning, how can classrooms best support age-related emotional processing), behavioral economics (e.g., how does emotions changes with age impact decisions and consumer trends) and more.

In much of the psychology work on age and emotions, the data studied was obtained from surveys or census asking how happy people feel, or how satisfied they feel. Surveys can have biases (e.g., social pressures to appear happy, population representation biases, etc.), and are timely to collect for large populations. Further, the amount of data that can be collected from surveys is rather limited. In this work, we analyze large amounts of data systematically and efficiently by looking at the emotions expressed in text, specifically social



media. While people are not asked to describe how they feel when posting on social media, inadvertently through the words used, they express how they are feeling. Therefore, we can often infer how people feel through the emotions expressed in their posts. Further, the text we write on social media also impacts others – providing another angle on this question. We create a large corpus of social media posts and determine whether the expression of emotion in text varies across age groups, and if so, how does it vary. Further, most if not all of the work in psychology measures one dimension of emotion – valence (the pleasure–displeasure, happiness–unhappiness, or positive–negative dimension). Affect encompasses valence, arousal, and dominance and are thought to be the primary dimensions of emotion (Osgood et al., 1957; Russell, 2003; Russell and Mehrabian, 1977). Whereas categorical emotions include joy, sadness, fear, etc. such as Ekman’s 6 emotions (anger, fear, joy, sadness, disgust, surprise) (Ekman, 1992) and Plutchik’s Wheel of Emotions (additionally including anticipation and trust) (Plutchik, 1980). Emotions are multifaceted and it is important to understand how do all types of emotions which we experience every day (e.g., fear, anger) and affect vary across age groups? Each and every one of these dimensions has an impact on overall well-being.

In this paper we introduce *AgeCorpus* – a dataset of social media posts annotated with the age of the author at the time of writing. We use *AgeCorpus* to explore research questions on how emotion varies across age groups. We provide novel insights from two social media platforms, X (formerly known as Twitter) and Reddit, and contribute a large dataset for affective science research. However, there are several ways one could go about studying emotion change with age. One approach is having a group of individuals and following how their emotion expression in text changes as they age through their 20’s, 30’s, and onwards, allowing us to study *how emotion expressed changes across age* with a consistent cohort. Although such data sources are limited, so this would be a challenge. Therefore, we ask a more specific question - how does emotion expression vary across age groups *during specific periods of time*. This date range, or period of time, is a parameter and can be set to any value e.g., from a few years to a decade. We create a corpus of *contemporary* age utterances, as our dataset covers 2010–2022 for Reddit and 2020–2021 for Twitter. We refer to this dataset as *AgeCorpus*. We add this nuance to our research question as being in your 20’s during the 1940’s is quite different from being in your 20’s during the early 2000’s. For simplicity, from here onwards we will refer to our research question as “how does emotion vary across age groups”.

Using *AgeCorpus* dataset we investigate the following research questions:

1. What patterns do we see in posting characteristics with age? This can help inform research on well-being such as posting frequency may be related to social withdrawal, or how use of pronouns can be indicative of mental health.
2. Does emotion expression vary by age? How?

We will make our code and dataset publicly available.<sup>1</sup> Our data is also a part of the ABCDE (Affect, Body, Cognition, Demographics, and Emotion) corpus (Wahle et al., 2026) where there are various additional annotations included beyond age, allowing for a wide range of research questions on the relationship between age and other features to be explored.

## 2. Related Work

### 2.1. Psychology, Age, and Emotions

Many theories in psychology have explored how well-being and happiness change with age. One branch of theories describe happiness levels as following a “U-shape” with age, whereas others say it follows an “Inverted U-Shape”. More so, others say there is simply no relationship with age. Blanchflower and Oswald (2008) introduced the U-shape theory, demonstrating that psychological well-being is at its lowest and mental distress at its highest during midlife. Potential confounding factors such as income, education, marital status, and region have been considered and the U-shape still holds. Similar trends were found regardless of employment status (Clark and Oswald, 1994). Likewise, Deaton (2008) found that mean life satisfaction is U-shaped with age for high income countries, whereas declines with age for middle-income countries but to a lesser extent for low-income countries. Similarly, Kim et al. (2023) noted the U-shape relationship between well-being and age, however it was moderated by personality. On the other hand, some work in psychology points to an inverted U-shape between age and well-being with happiness increasing until 50’s/60’s and decreasing afterwards (East-erlin, 2006), even after considering extraversion level (Mroczek and Spiro, 2005). Whereas Myers and Diener (1995) claimed that regardless of gender, race, and socioeconomic status, satisfaction with life remained constant with age.

Work in psychology has also studied the relationship between patterns in language use and well-being. For example, the proportion of pronouns can be an indicator of depression (consistently found in a meta-survey (Edwards and Holtzman, 2017)), and the pattern with which emotions change over

<sup>1</sup>[https://github.com/dteodore/Age--Affect\\_AgeCorpus](https://github.com/dteodore/Age--Affect_AgeCorpus)



time (i.e., referred to as *emotion dynamics*) is associated with mental health diagnoses (e.g., more negative affect over time and higher fluctuations of emotional states are associated with anxiety, depression (Houben et al., 2015; Heller et al., 2018)).

## 2.2. NLP, Age, and Emotions

Past NLP work created age corpora for supporting the development of age-detection models (Chew et al., 2021; Rosenthal and McKeown, 2011; Peersman et al., 2011) and personalized models (e.g., search and recommendations (Tigunova et al., 2020)). Other work, created age corpora to enable research in the social sciences (e.g., Reddit data with demographic and personality labels (Gjurković et al., 2021)) and to mitigate bias in language technologies (e.g., demographic-aware word embeddings (Welch et al., 2020)).

Few works, studied how emotions varies across age groups. Hipson and Mohammad (2020) explored the emotion expressed in poems for grade 1–12 students and found decreased valence during mid-adolescence, and arousal and dominance peaked during adolescence. We focus on how emotion varies across adulthood in the social media domain. Schwartz et al. (2013) found that in Facebook posts there was a positive relationship between positive emotion and age, whereas a negative relationship for negative emotion with age. Lastly, the most related is by Li (2023), who explored whether older people expressed more positive emotions than younger people for 8 discrete emotions with the spoken portion of the BNC (fiction, magazines, newspapers, and academic). Overall, their findings aligned with the U-shaped theory in psychology – older people expressed less negative emotions (except for sadness), but also there were fluctuations in how positive emotions increased with age. In contrast, we create and explore an utterance–age resource for Reddit and Twitter.

## 3. AgeCorpus Dataset

We create a dataset of textual utterances/posts from social media annotated with the speaker’s age at the time of posting. Such a dataset enables a wide array of research such as on how language characteristics vary by age (e.g., post length, metaphor usage, etc.), or for studying aspects of well-being (e.g., loneliness, mental health) across age groups. We focus on social media as a source of text data as it allows us to study large amounts of data and is more accessible compared to other text mediums (e.g., essays, discussion forums). We chose Reddit and Twitter as our platforms as these are two representative social media sources which are rich in text data. Rather than using APIs to

scrape these social media platforms, we use existing dumps of social media. Future research could extend this work to multimodal platforms such as Instagram or TikTok. We make our dataset publicly available.<sup>2</sup>

Social media data instances (e.g., Tweets, Reddit posts) often have various metadata information associated with them such as the time of posting, user id, number of likes, reposts, retweets, location, etc. However, the age of the author is not included in the metadata. Some people choose to include their age in their profile or in their social media bio, however this is often noisy data and may be outdated. In our work, we make use of self-declarations of age in posts: e.g., *“I’m so happy to be turning 25, happy birthday to me!”* or *“I’m 30 and relate to this...”*. Using the age mentioned in the self-declarations and the time stamp of that post, we can determine the age of the person at the time of the posting. Then we can determine their age at the time they made any other (past or future) target post, by simply determining the difference in time between the target post and the time of the self-declaration post.

People can self-declare their age in many ways, but there are certain templates that are common: e.g., *“I’m turning 25 years old today!”*. We make use of these templates to automatically identify self-declarations.<sup>3</sup> In Appendix A, we show the regexes used to capture these templates. We note that age declarations can appear in two forms: people declare their age at the time of posting; and those that declare their date of birth (e.g., birth year). We make use of high-precision templates since these templates can sometimes capture false positives (e.g., *“When I was 26 years old ...”*). Further, an author can have multiple age declarations, so we increase the precision of the self-declarations using the approach described in the section below.

### 3.1. Resolving Multiple Age Declarations

We determine the birth year from the self-declaration of age and time of posting. For any user that is below 13 years old, we remove their posts (minimum age to use these social media platforms). For users with several birth year declarations, birth years that are similar (within 2 years) are clustered together. Each cluster is assigned a weight, and the cluster with the highest weight is selected. The final birth year is calculated as the weighted average of that cluster. We remove posts from users with ages determined to be outside of 13–100 (as users older than 100 are unlikely). This methodol-

<sup>2</sup>[https://github.com/dteodore/Age--Affect\\_AgeCorpus](https://github.com/dteodore/Age--Affect_AgeCorpus)

<sup>3</sup>We do not aim to capture all possible ways; just commonly used ways.

Dataset	Time Period	# Users	Total # Utterances	User Avg. # Utterances (Std dev.)
Reddit	2010–2022	1,047,553	33,495,292	31.97 (841.89)
TUSC-City	2020–2021	48,712	7,486,672	153.69 (230.715)

Table 1: The number of users and number of utterances with age annotations obtained from the Reddit and Twitter datasets.

ogy ensures that if multiple ages are declared, the most probable one is selected.

### 3.2. Reddit Dataset

Reddit data is not easily accessible due to API changes so we use past Reddit Pushshift dumps from 2010 up to and including 2022.<sup>4</sup> We search for posts matching the age declaration templates and identify potential users. Next, we collect all of the users' posts.

### 3.3. Twitter Dataset

We use an existing Twitter corpus, TUSC-City (Vishnubhotla and Mohammad, 2022), which contains geo-located English Tweets from US and Canada. It consists of posts collected via the free Twitter API by using geo-location and random sampling from 46 American and Canadian cities.

We likewise search for users from TUSC-City using the above mentioned matching templates, and obtain the posts for these users.

The number of users, utterances, and other details of the dataset are in Table 1, and the number of users in each age group is in Appendix C. Most age groups have about 2,000 users from Twitter and about 20,000 users from Reddit, and the smallest age group had 360 users, ensuring a strong enough signal when performing the below experiments.

## 4. How do Social Media Post Characteristics Change with Age?

Using *AgeCorpus* we explore a series of questions on how social media posting habits and social media posts vary across age groups. In the following section, we explore how emotion word usage varies with age. We remind readers (as mentioned in the Introduction), we will use for short how emotion expression “varies across age groups” although we are referring to more specifically “varies across age groups during the time periods of the dataset” (e.g., 2010–2020 for Reddit and 2020–2021 for Twitter).

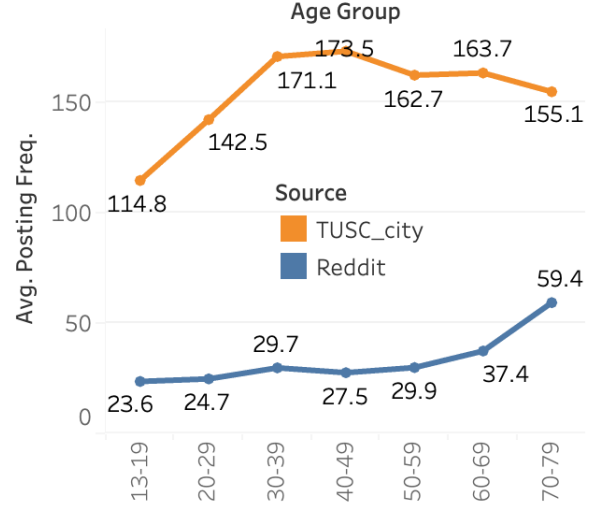


Figure 1: Posting frequency on Reddit and Twitter.

### 4.1. How does Posting Frequency Change with Age?

We ask how posting habits, such as the frequency of posting, varies across age groups. This is an important question as it is informative for research on well-being (e.g., loneliness) which uses the amount of online activity as a signal. It has been shown that individuals who experience depression, may withdraw from social media use (De Choudhury et al., 2021), and there were significant differences in Facebook posting quantities for individuals diagnosed with depression (Smith et al., 2017).

*Experiment Setup:* We find the number of posts for each author and then take the average across authors in that respective age group.

*Results:* We show the results in Figure 1. On Twitter, there is an increase in posting frequency until the 40's, with a decrease afterwards. On Reddit, we see an increasing trend in the frequency of posting across age groups up until the 40's, with a decrease afterwards until the 60's to 70's.

*Discussion:* These findings contrast those by Chang et al. (2015) who found that age was negatively associated with self-posting on Facebook, and found there is less activity in general for older adults on Facebook (McAndrew and Jeong, 2012).

<sup>4</sup>We do not consider more recent Reddit data as it contains numerous LLM- and bot-generated posts.

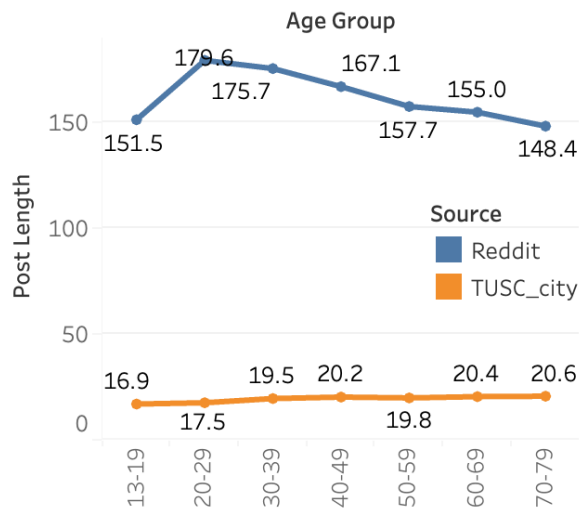


Figure 2: Post length (tokens) by age group.

#### 4.2. How do Post Lengths Change with Age on Reddit and Twitter?

Is it that as we age, we tend to share longer messages online, or is it the opposite? Or, is there no clear trend? These findings provide insights into the writing style of each age group.

*Experiment Setup:* The experiment is simple – we compute the average length of each post in terms of number of tokens, average the length of posts per author, and compute the average post length per age group. This allows us to see how post length varies across groups and if there are consistent trends across the two platforms.

*Results:* The results are shown in Figure 2. In the Reddit data we see a marked difference between the teens and 20's, compared to the Twitter data. Generally, there is an increasing trend in post length on TUSC, whereas the Reddit data shows the opposite trend with decreasing post length from the 20's to 70's. Overall, we see more variation in post length on Reddit than Twitter, which could be due to the character limit on Twitter (280 characters).

*Discussion:* Previous literature found that post length in blogs was longer for older adults (Rosen-thal and McKeown, 2011), post length increased on a health forum with age (Shrestha et al., 2016), and the total number of content words in Facebook messages had a positive relationship with age (Schwartz et al., 2013). These findings align with the trends found on Twitter. Whereas other literature found no relationship between social network post length and age (Peersman et al., 2011). The opposing trends between Twitter and Reddit could be explained by how these platforms are used for different purposes.

Age Group	Most popular Subreddits
13–19	teenagers, quzangle, randonaut_reports, AskReddit, buildapc
20–29	AskReddit, NoStupidQuestions, personalfinance, buildapc, NoFap
30–39	funwithwords, Vocabulario, EsperantoFeed, KazakhFeed, CatalanFeed
40–49	stopdrinking, AskReddit, NoStupidQuestions, personalfinance, keto
50–59	SandersAlerts, nba, stopdrinking, MillennialBets, personalfinance
60–69	gw2devtrackalt, Wallstreetsilver, personalfinance, Jokes, teenagers
70–79	WorldNewsinPictures, CMKMArchive, u_WorldNewsinPictures, teenagers, pathofexile

Table 2: The top 5 most popular subreddits per age group in descending order.

#### 4.3. How does Most Popular Subreddits Change with Age?

We are interested in the top subreddits each age group posts in and what changes, if any, we see across age groups. This informs us of the interests and communities that each group engages with.

*Experiment Setup:* We find the subreddit counts of posts for each age group, and consider the top 5 most popular. We also performed these experiments with up until the top 20 subreddits for each age group, and summarize the results below. We show the top 20 most popular subreddits in Appendix D. This research question is only applicable to the Reddit subset of AgeCorpus.

*Results:* The results are shown in Table 2. We see that in the teens and 20's there is interest in r/teenagers, asking questions online, and hobbies such as building personal computers and an interest in finance. In the 30's people engage in interests with languages, vocabulary and words (e.g., r/funwithwords, r/Vocabulario, r/Word\_of\_The\_Hour). In the 40's and 50's health-conscious interests appear such as reducing drinking and keto diets, and home improvement. In the 60's, 70's other interests such as finances, games, and news appear which could be as people retire they are passing time looking at entertainment and news on their computers. Overall, we see little overlap between groups except when looking at the top 20 subreddits we see sustained interest in r/personalfinance for 20's to 60's, as well asking questions online (e.g., r/AskReddit and r/NoStupidQuestions across all age groups). Looking at the top 20 subreddits we see overall that the teens to 20's is a different time than 30's, as people are in transitory periods of their life seeking advice from r/relation-

ship\_advice and r/offmychest, and have interests in gaming (r/gaming, r/leagueoflegends). The teens and 20's is also a time when mental health becomes a prevalent topic (r/depression, r/Anxiety), and it reappears in the 40's (e.g., r/depression, r/ADHD). We also see topics such as r/Menopause appear in the 50's.

*Discussion:* There is no vast literature exploring which subreddits are popular among each age group, but Chew et al. (2021) found that teenagers commonly post in r/teenagers (age 13–20), whereas older adults (age 21–54) posted more frequently in r/news, which aligns with our findings. Monti et al. (2023) studied topics discussed in r/news and found that generally younger people (teenagers) tended to discuss more about arts, technology, and entertainment, while older adults (30's and 40's) discussed business, politics, and green/environmental topics.

#### 4.4. How does Pronoun Usage Change with Age?

We explore how the pronoun usage in posts on social media varies with age. This informs us of people's focus e.g., themselves, those in direct contact with them, or others. We analyze subject and object pronouns listed in Appendix B.

*Experiment Setup:* We use a binary flag to indicate whether a post contains a given pronoun or not. Then, we calculate the percentage of posts that have each pronoun across age groups.

*Results:* In Figure 3, we show the percentage of pronoun usage for subject and object pronouns across the platforms. Across groups, the most common pronoun is *I*, followed by *you*, *me*, *they*, *we* and *them*. Overall, first-person pronoun usage (e.g., *I*, *me*) increases from teens to 20's, and decreases from 30's to 70's. Whereas second-person (e.g., *you*) tends to have an inverse pattern, increasing until 40's and then plateauing. For third-person pronouns (e.g., *we*, *they*) we see the overall trend increasing with age. It appears that for *them*, proportions are relatively constant with age, and just slightly increase for TUSC-City.

*Discussion:* There is little work comparing how each age group uses first- vs. second- vs. third-person pronouns, however Hendriks et al. (2008) found that elderly adults produced pronouns significantly more often than young adults when describing picture stories and referring to past topics, and more potentially ambiguous pronouns (Hendriks et al., 2014). Recently, the percentage of tweets containing various pronouns has been explored, finding that about 60% of tweets had “I” and 30% had “you” ((Mohammad, 2026), Figure 3). More recent work in the area of pronoun usage has focused on contrasting self-disclosed use of *he*, *she*,

and *them* in profile descriptions on social media (Jiang et al., 2023; Lauscher et al., 2022), attitudes towards pronouns (Sendén et al., 2021), pronoun resolution (Andy et al., 2020; Seminck and Amsili, 2017; Yu et al., 2019; Chada, 2019), or errors in translating pronouns (Lauscher et al., 2023; Luong and Popescu-Belis, 2016), rather than looking at overall trends in pronoun usage across age groups.

## 5. How do Emotions Expressed in Social Media Vary with Age?

We explore how does emotion expression on social media vary across age groups. These findings are important for diverse research questions such as how do characteristics of language use change with age, and in interdisciplinary work such as examining the relationship between emotion in text and mental health. For example, more negative emotion expressed in text has been found to be associated with mental health diagnoses such as depression, ADHD, etc. (Seabrook et al., 2018; De Choudhury et al., 2013, 2021). Further, the way in which emotion expression in text changes over time (e.g., emotional variance) has been linked to mental health diagnoses (Teodorescu et al., 2023a).

We explore two questions about how emotions vary across age groups in the two subsections below. We answer these questions by looking at the three primary dimensions of affect (valence, arousal and dominance), and four important categorical emotions (anger, fear, joy and sadness). We apply common preprocessing to the text by first tokenizing, lowercasing and obtaining the emotion scores. We use the NRC Emotion Lexicon (Mohammad and Turney, 2012) for anger, fear, joy, and sadness, where scores range from 0 (lowest amount of the emotion) – 1 (highest amount of the emotion). We use the NRC Valence Arousal Dominance Lexicon (Mohammad, 2018) for valence, arousal and dominance, where scores range from -1 (lowest V/A/D) to 1 (highest V/A/D).<sup>5</sup> Past work has shown that lexical approaches produce emotion arcs that are highly correlated with the true (gold) arcs (correlations > 0.9) (Teodorescu and Mohammad, 2023).

### 5.1. How does Average Emotion Change with Age?

*Experiment Setup:* We calculate the emotion score per post by taking the average of the emotion

<sup>5</sup><http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm>  
<https://saifmohammad.com/WebPages/nrc-vad.html>



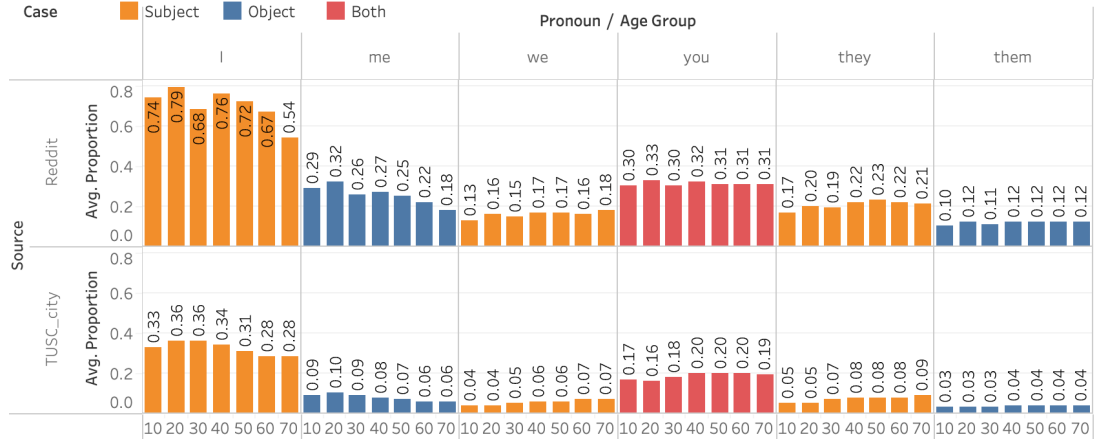


Figure 3: Pronoun usage by age group.

scores for each word in the post. Then, we compute the average emotion score per user by taking the average across their posts. To compare how emotional states vary across age groups, we average the emotion scores across all users per group to analyze the results at the aggregate level.

**Results:** We plot the results for average valence, arousal, dominance in Figure 4 for TUSC and Reddit, and the average anger, fear, joy, and sadness in Figure 5 for TUSC and Reddit.

**Valence, Arousal, Dominance:** In both TUSC and Reddit, we see an increase in average valence from teens up until the 30's, and a steady decrease from the 30's till the 70's. We see the reverse pattern for arousal with decreasing scores on Reddit from teen's up until the 70's. For TUSC, we see an increase from teens to 20's, and then arousal intensities plateau until the 70's. This means that younger and older groups are expressing similar levels of arousal in TUSC, whereas there is a consistent decrease in arousal with age on Reddit. For dominance, there are increasing scores with age from teen's to the 70's in TUSC, meaning that language which is more associated with control is used with age. However, there is an inverted U-shape on Reddit, where dominance reaches a peak in the 20's and then decreases with age. Generally, it appears that valence follows the inverted U-shape, arousal decreases (Reddit)/plateaus (TUSC) with age, and dominance follows the inverted U-shape (Reddit)/increases (TUSC) with age. Overall, we see consistent trends between platforms for valence but differing patterns for arousal and dominance.

**Discussion:** Work in psychology has found various trends in terms of happiness across ages. The majority of the literature points to the U-shape of happiness, and some to the inverted-U or bimodal trends. Our findings support the inverted U-shape, meaning that younger and older adults experience less valence/happiness compared to those in their middle ages. Past work has found

that when children wrote poems, there was decreasing valence and increasing arousal and dominance across grades 1–12 (Teodorescu et al., 2023b). No work examines arousal or dominance expression in text across adulthood, so we establish baseline findings for these dimensions.

**Anger, Fear, Joy, Sadness:** We see inverted U-shape patterns appearing for anger, fear, joy and sadness on Reddit, whereas in TUSC we found increasing intensities for all categorical emotions with age. However, we note that the inverted U-shape on Reddit is not symmetrical e.g., anger and joy is highest for teens than for those in their 70's. In terms of categorical emotions, we find that the negative emotions (i.e., anger, fear, sadness) show similar trends of positive emotions such as valence. Valence was highest for those in midlife, while anger, fear, joy and sadness were highest on Reddit. We expected that valence and joy would have similar trends as both are positive emotions/dimensions and overall they are both have an inverted U-shape (with joy increasing in TUSC).

**Discussion:** Past work has found more anger and sadness for older adults, however it was noted that sadness only increased for those with decreased perceived control (Blanchard-Fields and Coats, 2008; Wrosch et al., 2018). Haase et al. (2012) point to the adaptive function of anger and sadness at certain stages of life, as higher levels of self-reported sadness in response to a neutral film were associated with higher well-being for older adults. Higher self-reported anger was associated with higher well-being for middle-aged adults (rather than for younger and older adults). Anger is said to be more adaptive to younger adults when they have many resources and are trying to determine their niche in society, and reach goals, whereas sadness is adaptive to older adults when dealing with unachievable goals, diminishing resources, loss, and to elicit support (Haase et al., 2012; Kunzmann and Wrosch, 2024). Other work has found that the intensity of anger, fear, joy and sadness expressed in



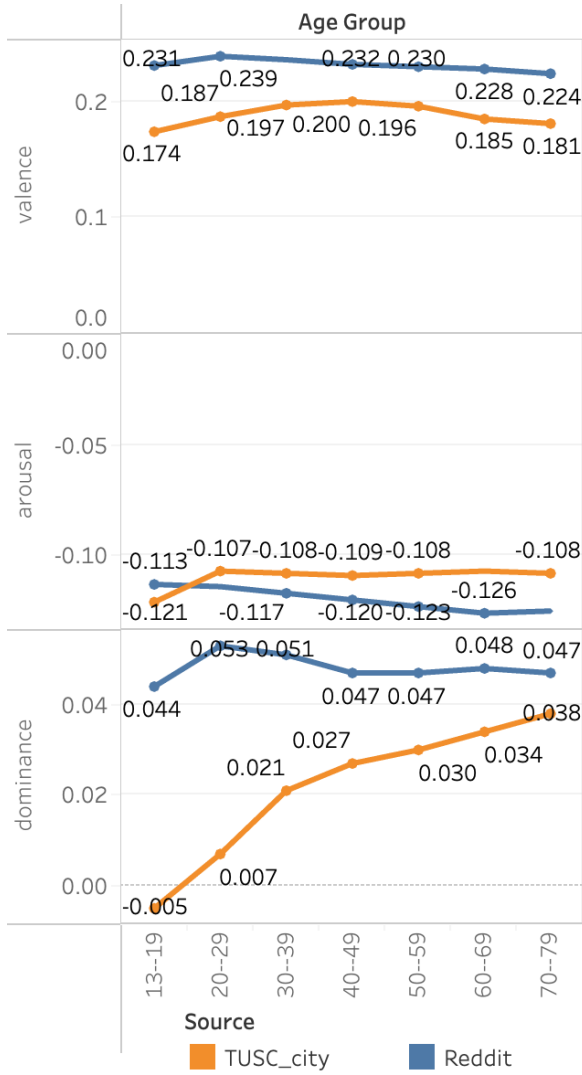


Figure 4: **Average valence, arousal, dominance** across age groups for TUSC-City and Reddit. Note that the lexicon scores for valence, arousal, dominance range from -1 to 1.

poems written by children increases across grades 1–12 (Teodorescu et al., 2023b).

## 5.2. How does Emotional Variance Change with Age?

*Experiment Setup:* Using the emotion score per post, we calculate the standard deviation across the posts per user. Then we average the standard deviations per age group to compare how emotional variance varies with age.

*Results:* We show the results in Figure 6 for valence, arousal, and dominance variation on TUSC and Reddit.

*Valence, Arousal, Dominance:* In Twitter, we see an overall U-shape for valence, arousal and dominance variation, and then a slight drop in the 60's and 70's. On Reddit, we also see a U-shape for valence, arousal and dominance variation across

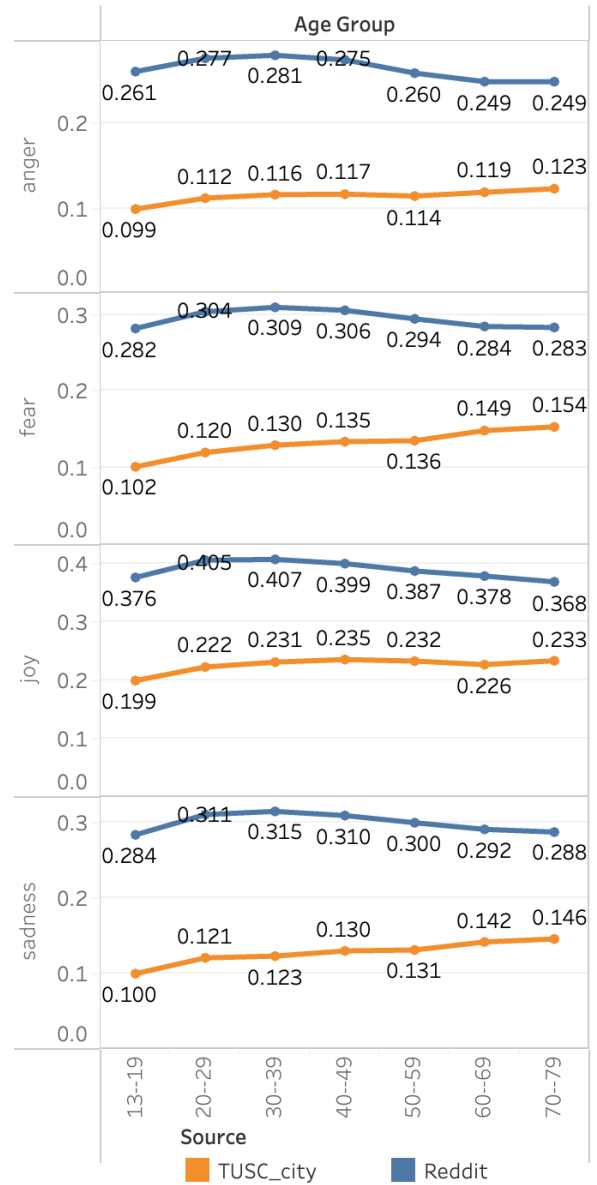


Figure 5: **Average anger, fear, joy, sadness** across age groups for TUSC-City and Reddit.

age groups, although teens have markedly higher variation.

*Discussion:* This is the first work that studies how the variation of emotions changes across age groups, contributing novel findings to affective developmental psychology. Work in psychology has pointed to the relationship between emotional variance and mental health diagnoses (e.g., ADHD, anxiety, depression (Houben et al., 2015; Heller et al., 2018; Seabrook et al., 2018; Teodorescu et al., 2023a)), however little is known about the pattern of emotion variation with age. While we see similar U-shape trends between Reddit and Twitter, teens had high variation on Reddit. This could be due to the nature of the platforms. Reddit allows the space for longer posts, such as stories, where we may expect younger crowds to have more vary-

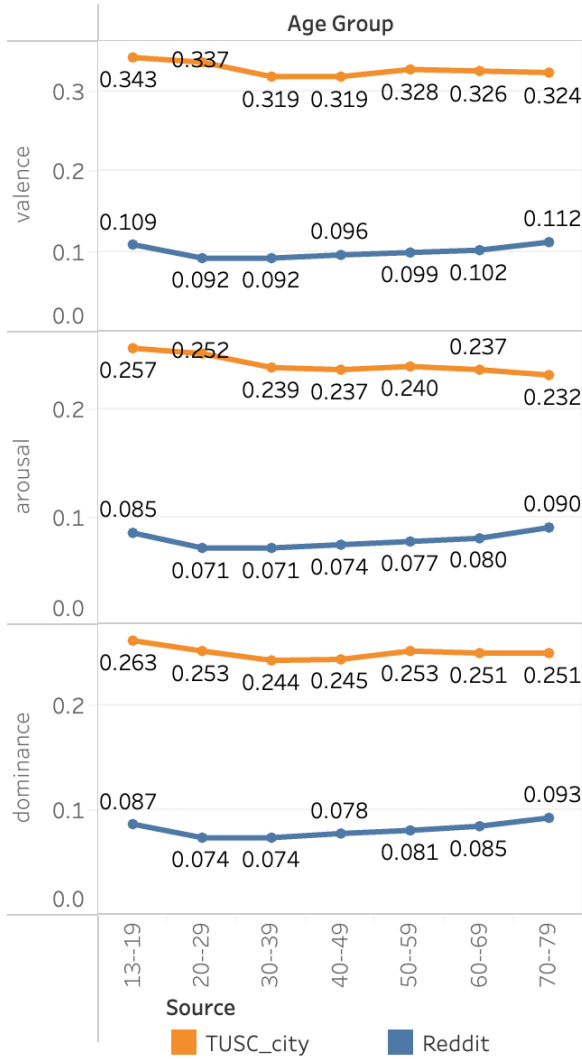


Figure 6: **Valence, arousal, and dominance variation** across age groups for TUSC-City and Reddit. Note that the lexicon scores for valence, arousal, dominance range from -1 to 1.

ing emotional experiences during transitory periods of life. Whereas Twitter is more suitable for shorter updates/posts and generally has an older audience. Future work should explore emotional variation on other platforms and through surveys across age groups to further findings across disciplines.

## 6. Conclusion

We created a dataset of social media utterances annotated with the author’s age at the time of posting. This unique dataset has social media posts from Reddit during 2010–2022 and Twitter from 2020–2021. Such a dataset enables diverse research in the social sciences, psychology, and NLP. With the dataset we explored how emotion expression in text varied across age groups. We found valence followed the inverted-U shape, arousal decreased with age (Reddit)/plateaus (TUSC), and dominance

followed the inverted-U shape (Reddit)/increased with age (TUSC). For the categorical emotions, we found all emotions followed an inverted U-shape on Reddit and increased with age in TUSC. We found generally similar trends of U-shaped emotional variance on Reddit and TUSC. Such insight into how online behavior changes with age is informative for research in mental health and well-being. Further, this dataset can be used to explore the relationship between age and many other aspects of human behavior and language, such as emotion regulation, interoception, anxiety, etc. For example, body part mentions in tweets has been linked with poorer health outcomes (Wu et al., 2025) and this relationship could be investigated across ages.

## Limitations

This dataset relies on self-disclosure of age on social media. As with all scenarios involving self-disclosure, individuals may falsely report their age due to social pressures, or in order to relate to and engage with certain communities. Additionally, our templates may not fully capture all possible ways to express age, so there may be age declarations not included in our dataset.

## Ethical Considerations

There are ethical considerations around creating such an age corpus. The Reddit User Agreement mentions that users agree not to disclose sensitive information of other users and they consent that their posts are publicly available (e.g., through API to other services).

Our research interest is to study emotional development through adulthood at the aggregate/group level. This has applications in emotional development psychology and in public health (e.g., overall well-being and mental health). However, emotions are complex, private, and central to an individual’s experience. Additionally, each individual expresses emotion differently through language, which results in large amounts of variation. Therefore, several ethical considerations should be accounted for when performing any textual analysis of emotions (Mohammad, 2022, 2023). The ones we would particularly like to highlight are listed below:

- Our work on studying emotion word usage should not be construed as detecting how people feel; rather, we draw inferences on the emotions that are conveyed by users via the language that they use.
- The language used in an utterance may convey information about the emotional state (or perceived emotional state) of the speaker, listener, or someone mentioned in the utterance.

However, it is not sufficient for accurately determining any of their momentary emotional states. Deciphering the true momentary emotional state of an individual requires extralinguistic context and world knowledge. Even then, one can be easily mistaken.

- The inferences we draw in this paper are based on aggregate trends across large populations. We do not draw conclusions about specific individuals or momentary emotional states.

## Bibliographical References

- Anietie Andy, Chris Callison-Burch, and Derry Tanti Wijaya. 2020. [Resolving pronouns in Twitter streams: Context can help!](#) In *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 133–138, Barcelona, Spain (online). Association for Computational Linguistics.
- Magda B Arnold. 1960. Emotion and personality. vol. i. psychological aspects.
- Fredda Blanchard-Fields and Abby Heckman Coats. 2008. [The experience of anger and sadness in everyday problems impacts age differences in emotion regulation.](#) *Developmental psychology*, 44(6):1547—1556.
- David G. Blanchflower and Andrew J. Oswald. 2008. [Is well-being u-shaped over the life cycle?](#) *Social Science & Medicine*, 66(8):1733–1749.
- Rakesh Chada. 2019. [Gendered pronoun resolution using BERT and an extractive question answering formulation.](#) In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 126–133, Florence, Italy. Association for Computational Linguistics.
- Pamara F. Chang, Yoon Hyung Choi, Natalya N. Bazarova, and Corinna E. Löckenhoff. 2015. [Age differences in online social networking: Extending socioemotional selectivity theory to social network sites.](#) *Journal of Broadcasting & Electronic Media*, 59(2):221–239. PMID: 31223198.
- Robert Chew, Caroline Kery, Laura Baum, Thomas Bukowski, Annice Kim, and Mario Navarro. 2021. [Predicting age groups of reddit users based on posting behavior and metadata: Classification model development and validation.](#) *JMIR Public Health Surveill*, 7(3):e25807.
- Andrew E. Clark and Andrew J. Oswald. 1994. [Unhappiness and unemployment.](#) *The Economic Journal*, 104(424):648–659.
- Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. [Social media as a measurement tool of depression in populations.](#) In *Proceedings of the 5th Annual ACM Web Science Conference, WebSci '13*, page 47–56, New York, NY, USA. Association for Computing Machinery.
- Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2021. [Predicting depression via social media.](#) *Proceedings of the International AAAI Conference on Web and Social Media*, 7(1):128–137.
- Angus Deaton. 2008. [Income, health, and well-being around the world: Evidence from the gallup world poll.](#) *Journal of Economic perspectives*, 22(2):53–72.
- Richard A. Easterlin. 2006. [Life cycle happiness and its sources: Intersections of psychology, economics, and demography.](#) *Journal of Economic Psychology*, 27(4):463–482.
- To'Meisha Edwards and Nicholas S. Holtzman. 2017. [A meta-analysis of correlations between depression and first person singular pronoun use.](#) *Journal of Research in Personality*, 68:63–68.
- Paul Ekman. 1992. [An argument for basic emotions.](#) *Cognition and Emotion*, 6(3-4):169–200.
- Matej Gjurković, Vanja Mladen Karan, Iva Vukojević, Mihaela Bošnjak, and Jan Snajder. 2021. [PANDORA talks: Personality and demographics on Reddit.](#) In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 138–152, Online. Association for Computational Linguistics.
- JJ Gross, LL Carstensen, M Pasupathi, J Tsai, CG Skorpén, and AY Hsu. 1997. [Emotion and aging: experience, expression, and control.](#) *Psychology and aging*, 12(4):590—599.
- Claudia M Haase, Benjamin H Seider, Michelle N Shiota, and Robert W Levenson. 2012. [Anger and sadness in response to an emotionally neutral film: evidence for age-specific associations with well-being.](#) *Psychology and aging*, 27(2):305—317.
- Aaron S Heller, Andrew S Fox, and Richard J Davidson. 2018. Parsing affective dynamics to identify risk for mood and anxiety disorders. *Emotion*, 19(2):283–291.
- Petra Hendriks, Christina Englert, Ellis Wubs, and John Hoeks. 2008. [Age differences in adults' use of referring expressions.](#) *Journal of Logic, Language and Information*, 17(4):443–466.

- Petra Hendriks, Charlotte Koster, and John C.J. Hoeks. 2014. [Referential choice across the lifespan: why children and elderly adults produce ambiguous pronouns](#). *Language, Cognition and Neuroscience*, 29(4):391–407. PMID: 24771955.
- Will Hipson and Saif M. Mohammad. 2020. [PoKi: A large dataset of poems by children](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1578–1589, Marseille, France. European Language Resources Association.
- Marlies Houben, Wim Van Den Noortgate, and Peter Kuppens. 2015. [The relation between short-term emotion dynamics and psychological well-being: A meta-analysis](#). *Psychological Bulletin*, 141(4):901–930.
- Julie Jiang, Emily Chen, Luca Luceri, Goran Muric, Francesco Pierri, Ho-Chun Herbert Chang, and Emilio Ferrara. 2023. [What are your pronouns? examining gender pronoun usage on twitter](#).
- Joo Hyun Kim, Eunsoo Choi, Namhee Kim, and Incheol Choi. 2023. [Older people are not always happier than younger people: the moderating role of personality](#). *Applied Psychology: Health and Well-Being*, 15(1):275–292.
- Ute Kunzmann and Carsten Wrosch. 2024. [Not all negative emotions are equal - sadness and anger develop differently and their adaptivity is age-graded](#). *Current Opinion in Psychology*, 55:101766.
- Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. [Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1221–1232, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Anne Lauscher, Debora Nozza, Ehm Miltersen, Archie Crowley, and Dirk Hovy. 2023. [What about “em”? how commercial machine translation fails to handle \(neo-\)pronouns](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 377–392, Toronto, Canada. Association for Computational Linguistics.
- Jialei Li. 2023. [Emotion and ageing in discourse: Do older people express more positive emotions? Corpus-based Studies across Humanities](#), 1(1):23–50.
- Ngoc Quang Luong and Andrei Popescu-Belis. 2016. [Pronoun language model and grammatical heuristics for aiding pronoun prediction](#). In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 589–595, Berlin, Germany. Association for Computational Linguistics.
- Francis T. McAndrew and Hye Sun Jeong. 2012. [Who does what on facebook? age, sex, and relationship status as predictors of facebook use](#). *Computers in Human Behavior*, 28(6):2359–2365.
- Saif Mohammad. 2018. [Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.
- Saif Mohammad. 2023. [Best practices in the creation and use of emotion lexicons](#). In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1825–1836, Dubrovnik, Croatia. Association for Computational Linguistics.
- Saif M. Mohammad. 2022. [Ethics sheet for automatic emotion recognition and sentiment analysis](#). *Computational Linguistics*, 48(2):239–278.
- Saif M. Mohammad. 2026. [When are we worried? temporal trends of anxiety and what they reveal about us](#).
- Saif M. Mohammad and Peter D. Turney. 2012. [Crowdsourcing a word–emotion association lexicon](#). *Computational Intelligence*, 29(3):436–465.
- Corrado Monti, Jacopo D’Ignazi, Michele Starnini, and Gianmarco De Francisci Morales. 2023. [Evidence of demographic rather than ideological segregation in news discussion on reddit](#). In *Proceedings of the ACM Web Conference 2023*, WWW ’23, page 2777–2786, New York, NY, USA. Association for Computing Machinery.
- Daniel K. Mroczek and Avron Spiro. 2005. [Change in life satisfaction during adulthood: Findings from the veterans affairs normative aging study](#). *Journal of personality and social psychology*, 88(1):189–202.
- David G. Myers and Ed Diener. 1995. [Who is happy?](#) *Psychological Science*, 6(1):10–19.
- Charles E Osgood, George J Suci, and Percy Tanenbaum. 1957. *The measurement of meaning*. University of Illinois Press.
- Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. [Predicting age and gender in online social networks](#). In *Proceedings of the 3rd International Workshop on Search*



- and Mining User-Generated Contents, SMUC '11, page 37–44, New York, NY, USA. Association for Computing Machinery.
- Robert Plutchik. 1980. [Chapter 1 - a general psychoevolutionary theory of emotion](#). In Robert Plutchik and Henry Kellerman, editors, *Theories of Emotion*, pages 3–33. Academic Press.
- Sara Rosenthal and Kathleen McKeown. 2011. [Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations](#). In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 763–772, Portland, Oregon, USA. Association for Computational Linguistics.
- James A Russell. 2003. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145.
- James A Russell and Albert Mehrabian. 1977. [Evidence for a three-factor theory of emotions](#). *Journal of Research in Personality*, 11(3):273–294.
- H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. [Personality, gender, and age in the language of social media: The open-vocabulary approach](#). *PLOS ONE*, 8(9):1–16.
- Elizabeth M Seabrook, Margaret L Kern, Ben D Fulcher, and Nikki S Rickard. 2018. [Predicting depression from language-based emotion dynamics: Longitudinal analysis of facebook and twitter status updates](#). *J Med Internet Res*, 20(5):e168.
- Olga Seminck and Pascal Amsili. 2017. [A computational model of human preferences for pronoun resolution](#). In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 53–63, Valencia, Spain. Association for Computational Linguistics.
- Marie Gustafsson Sendén, Emma Renström, and Anna Lindqvist. 2021. [Pronouns beyond the binary: The change of attitudes and use over time](#). *Gender & Society*, 35(4):588–615.
- Prasha Shrestha, Nicolas Rey-Villamizar, Farig Sadeque, Ted Pedersen, Steven Bethard, and Tamar Solorio. 2016. [Age and gender prediction on health forum data](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3394–3401, Portorož, Slovenia. European Language Resources Association (ELRA).
- Robert J Smith, Patrick Crutchley, H Andrew Schwartz, Lyle Ungar, Frances Shofer, Kevin A Padrez, and Raina M Merchant. 2017. [Variations in facebook posting patterns across validated patient health conditions: A prospective cohort study](#). *J Med Internet Res*, 19(1):e7.
- Daniela Teodorescu, Tiffany Cheng, Alona Fyshe, and Saif Mohammad. 2023a. [Language and mental health: Measures of emotion dynamics from text as linguistic biosocial markers](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3117–3133, Singapore. Association for Computational Linguistics.
- Daniela Teodorescu, Alona Fyshe, and Saif Mohammad. 2023b. [Utterance emotion dynamics in children's poems: Emotional changes across age](#). In *Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 401–415, Toronto, Canada. Association for Computational Linguistics.
- Daniela Teodorescu and Saif Mohammad. 2023. [Evaluating emotion arcs across languages: Bridging the global divide in sentiment analysis](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4124–4137, Singapore. Association for Computational Linguistics.
- Anna Tigunova, Paramita Mirza, Andrew Yates, and Gerhard Weikum. 2020. [RedDust: a large reusable dataset of Reddit user traits](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6118–6126, Marseille, France. European Language Resources Association.
- Heather L. Urry and James J. Gross. 2010. [Emotion regulation in older age](#). *Current Directions in Psychological Science*, 19(6):352–357.
- Krishnapriya Vishnubhotla and Saif M. Mohammad. 2022. [Tweet Emotion Dynamics: Emotion word usage in tweets from US and Canada](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4162–4176, Marseille, France. European Language Resources Association.
- Jan Philip Wahle, Gipp Bela Vishnubhotla, Krishnapriya, and Saif M. Mohammad. 2026. Affect, body, cognition, demographics, and emotion: The abcde of text features for computational affective science. In *Proceedings of the 1st Workshop on Computational Affective Science (CAS 2026)*, Palma de Mallorca, Spain. European Language Resources Association (ELRA).



Charles Welch, Jonathan K. Kummerfeld, Verónica Pérez-Rosas, and Rada Mihalcea. 2020. [Compositional demographic word embeddings](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4076–4089, Online. Association for Computational Linguistics.

Carsten Wrosch, Meaghan A Barlow, and Ute Kunzmann. 2018. [Age-related changes in older adults' anger and sadness: The role of perceived control](#). *Psychology and aging*, 33(2):350—360.

Sophie Wu, Jan Philip Wahle, and Saif M. Mohammad. 2025. [The language of interoception: Examining embodiment and emotion through a corpus of body part mentions](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23375–23399, Suzhou, China. Association for Computational Linguistics.

Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changshui Zhang. 2019. [What you see is what you get: Visual pronoun coreference resolution in dialogues](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5123–5132, Hong Kong, China. Association for Computational Linguistics.

## **A. Patterns to Match Age Declarations**

In Table 3, we show the patterns used to find age declarations in the seed posts.

## **B. Subject and Object Pronouns**

We study the occurrences of the following pronouns shown in Table 4 to determine if there are patterns across age categories.

## **C. Number of Users per Age Group**

In Table 5, we show the number of users in each age group in the *AgeCorpus*. This data was used to perform the experiments.

## **D. Top 20 Most Popular Subreddits per Age Group**

In Table 6, we show the 20 most popular subreddits by age group. We describe the findings in Section 4.3.

Regex	Example
<code>\bI(?:\s+am 'm)\s+(\d{1,2})\s+years?\s+old\b</code>	<i>I am 25 years old</i> <i>I'm 30 year old</i>
<code>\bI(?:\s+am 'm)\s+(\d{1,2})</code> <code>(?=\s*(?:\$ [,.!?;:\-] (?:and but so yet)\s))</code>	<i>I am 25.</i> <i>I'm 30, and ...</i>
<code>\bI(?:\s+was \s+am 'm)\s+born\s+in\s+</code> <code>(19\d{2} 20(?:0\d 1\d 2[0-4]))\b</code>	<i>I was born in 1998</i> <i>I am born in 2005</i>
<code>\bI(?:\s+was \s+am 'm)\s+born\s+in\s+'(\d{2})\b</code>	<i>I was born in '98</i> <i>I'm born in '05</i>
<code>\bI\s+was\s+born\s+on\s+</code> <code>(?:\d{1,2}(?:st nd rd th)?\s+)?</code> <code>(?:January February March April May June </code> <code>July August September October November December </code> <code>Jan Feb Mar Apr May Jun Jul Aug Sep Sept Oct Nov Dec)</code> <code>\s+(?:\d{1,2}(?:st nd rd th)?\s+)?</code> <code>(19\d{2} 20(?:0\d 1\d 2[0-4]))\b</code>	<i>I was born on 15 March 1998</i> <i>I was born on March 15th, 1998</i>
<code>\bI\s+was\s+born\s+on\s+\d{1,2}</code> <code>[/\-]\d{1,2}[/\-](19\d{2} 20(?:0\d 1\d 2[0-4]))\b</code>	<i>I was born on 03/15/1998</i> <i>I was born on 15-03-1998</i>

Table 3: Regexes used to identify users from both the Reddit and Twitter datasets.

Subject	Object
I	me
you	you
he	him
she	her
it	it
we	us
they	them

Table 4: Subject and object pronouns we studied occurrences of in the dataset.

Age Group	#Posts	
	TUSC-City	Reddit
13–19	94,857	9,281,055
20–29	406,027	15,455,426
30–39	462,238	6,151,757
40–49	360,946	1,220,498
50–59	278,809	449,230
60–69	218,008	246,398
70–79	97,432	160,465

Table 5: The number of posts across the age groups in each subset of the *AgeCorpus*. “City” refers to TUSC-City subset of the dataset. We use this data in our experiments.

Age Group	Most popular Subreddits
13–19	teenagers, quzangle, randonaut_reports, AskReddit, buildapc, NoFap, NoStupidQuestions, Advice, GlobalOffensiveTrade, leagueoflegends, pcmasterrace, teenagersnew, depression, techsupport, trees, alt_source_bot_log, tipofmytongue, offmychest, relationship_advice, gaming
20–29	AskReddit, NoStupidQuestions, personalfinance, buildapc, NoFap, depression, Advice, techsupport, leagueoflegends, offmychest, relationship_advice, tipofmytongue, raisedbynarcissists, AskDocs, legaladvice, trees, Fireteams, r4r, gaming, Anxiety
30–39	funwithwords, Vocabulario, EsperantoFeed, KazakhFeed, CatalanFeed, personalfinance, GalicianFeed, Lessico, SwedishFeed, Traductions, Word_of_The_Hour, HindiFeed, AskReddit, PortugueseFeed, Sprache, NoStupidQuestions, BCSFeed, WelshFeed, DestinyTheGame, BengaliFeed
40–49	stopdrinking, AskReddit, NoStupidQuestions, personalfinance, keto, HomeImprovement, raisedbynarcissists, Jokes, legaladvice, techsupport, Isletforum, DestinyTheGame, buildapc, NoFap, depression, getdisciplined, tipofmytongue, AskDocs, ADHD, sysadmin
50–59	SandersAlerts, nba, stopdrinking, MillennialBets, personalfinance, exmormon, Jokes, AskReddit, NoStupidQuestions, keto, volunteer, legaladvice, open_bots_test, techsupport, Menopause, leagueoflegends, The_Donald, buildapc, raisedbynarcissists, HomeImprovement
60–69	gw2devtrackalt, Wallstreetsilver, personalfinance, Jokes, teenagers, TargetedEnergyWeapons, anychat, stopdrinking, NoStupidQuestions, zero hedge, ketoscience, amcstock, AskReddit, wow, exmormon, techsupport, LibraryofBabel, playcodformoney, leagueoflegends, buildapc
70–79	WorldNewsinPictures, CMKMArchive, u_WorldNewsinPictures, teenagers, pathofexile, makingascenemag, bestchange, CattleExchange, nba, PollsAndSurveys, BollyBlindsNGossip, Jokes, u_classeAS2002, u_From_Him_For_Him, NoStupidQuestions, guns, AskReddit, Gunsforsale, exjw, FashionWinter

Table 6: The top 20 most popular subreddits per age group in descending order.

# Multimodal Affective Modelling in an Intelligent Tutoring System for Foreign Language Learning

Dionysios Koulouris<sup>1</sup>, Athanasios Kallipolitis<sup>1</sup>, Melina Tziomaka<sup>1</sup>,  
Argyrios Zafeiriou<sup>1</sup>, Stamatios Orfanos<sup>1</sup>, Andreas Menychtas<sup>1</sup>, Ilias Maglogiannis<sup>1</sup>,  
Stamatia Michalopoulou<sup>2</sup>, George Tsoulouhas<sup>2</sup>, Athina Sioupi<sup>2</sup>, Voula Giouli<sup>2</sup>

<sup>1</sup>University of Pireus, <sup>2</sup>Aristotle University of Thessaloniki,

Greece

{nkoulouris, nasskall, tziomakamel, zafeiriou, sorfanos, amenychtas, imaglo}@unipi.gr  
george.tsoulouhas@athenarc.gr, {smichalo, sioupi, pgiouli}@del.auth.gr

## Abstract

Foreign language learning is a cognitively and affectively demanding process, in which fluctuations in attention and motivation can negatively impact learner engagement. Emotions play a central role in this process, yet they are rarely modelled in a systematic, data-driven manner in authentic learning environments. This paper presents a prototype affective computing architecture that incorporates various modalities (audio, video, biosignals) to facilitate real-time or near-real-time emotion recognition in an educational scenario; the architecture is integrated within an emotion-aware and adaptive Language Learning application that harnesses Large Language Models in view of providing appropriate educational scenarios to learners. The system comprises modules for acquiring data for each modality and a processing pipeline for synchronizing and analyzing heterogeneous affective signals. We demonstrate both the feasibility and applicability of the approach through a proof-of-concept implementation and discuss its relevance for studying learner affect and supporting affect-aware educational scenarios. The results highlight both the applicability of multimodal affective data in educational settings and the need for further research on their pedagogical interpretation and use.

**Keywords:** affective computing, emotion recognition, educational application, affect-aware language learning

## 1. Introduction

Foreign language learning is a demanding process for both learners and educators, requiring sustained attention and concentration. Empirical studies show that students often experience negative emotions such as stress, anxiety, fatigue, and short lapses during instruction, which lead to reduced engagement and distraction (Hlas et al., 2019). At the same time, research has shown that positive emotions - such as enjoyment and satisfaction - play a significant role in shaping learners' motivation and sustained engagement (Dewaele and MacIntyre, 2014; Dewaele and Macintyre, 2016; Kantaridou and Psaltou-Joycey, 2023).

The paper presents a multimodal affective computing module integrated into EmoBot (Kallipolitis et al., 2026), an agentive language learning system that employs Large Language Models (LLMs) to deliver CEFR<sup>1</sup>-aligned practice and formative assessment activities within an interactive educational environment. While the system includes safeguards for detecting and pedagogically addressing toxic or inappropriate language, its primary contribution lies in the integration of a multimodal affective computing module. This module enables the investigation of learners' affective states during language

learning interactions, positioning it not merely as an adaptive tutor but as a platform for computational affective science in education, capable of shedding light on affective engagement, emotional regulation, and learner behavior over time.

The rest of the paper is structured as follows: Section 2 outlines the aim and scope of the research; Section 3 presents the related work, focusing on multimodal emotion analysis and affective computing in foreign language learning, and identifies the literature gap that highlights the need for a biosignal-enabled multi-modal dual approach architecture for efficient emotion analysis in foreign language learning contexts. The system design, the architecture of the incorporated modules, emphasizing biosignals, and the technical implementation of the proposed system along with the adaptive affect-aware mechanisms, are detailed in Section 4. Section 5 demonstrates the system in practice, providing valuable insights into the proof-of-concept application and into a theory-driven qualitative evaluation. The challenges of the approach are discussed in Section 6 followed by conclusions, lessons learned, and directions for further research, in Section 7.

<sup>1</sup>Common European Framework of Reference for Languages (Council of Europe, 2001)

## 2. Aim and scope of the research

Recent advances in Natural Language Processing (NLP) have enabled increasingly accurate sentiment and emotion detection from textual and audio data. Neural architectures and LLMs have improved the modelling of affective cues in written and spoken language, supporting a range of applications in dialogue systems that enable human-computer interaction.

Parallel progress in affective computing has expanded both the methodological scope and the range of modalities through which emotional states can be inferred, especially in the field of health (Wang et al., 2024a). Computational emotion analysis increasingly relies on multimodal approaches, incorporating textual information, visual data, audio signals, and, more recently, physiological measurements recorded in real time or near-real time (NRT). The integration of heterogeneous yet complementary modalities has significantly enhanced the robustness of affective inference systems. Within this context, advanced human-computer interaction methods are aimed at identifying emotion through various sources. By utilizing intelligent pattern recognition techniques, like Neural Networks (NNs) and Support Vector Machines (SVMs), the latest emotion recognition trends include image analysis (Kundu and Saravanan, 2017) or textual analysis (Alslaity and Orji, 2024) with encouraging results.

At the same time, progress in mobile and wearable technologies has enabled the recognition of users' physiological signals through commodity devices (Dias and Cunha, 2018) or their incorporation into self-assessment procedures (Saganowski et al., 2023). Such wearable biosensor technologies are increasingly explored in learning applications, significantly impacting the field and providing insights regarding learners' performance (Hernández-Mustieles et al., 2024). More broadly, when multiple heterogeneous yet informative data sources are available, multimodal frameworks are widely adopted to support more comprehensive modelling of affective states (Lyons, 2015).

Within educational settings, various modalities have been used to monitor learner interaction and behavior. In most cases, systems, however, rely on a single interaction interface (e.g., text and audio capture during online instruction), typically performing modality-specific analysis followed by feature-level fusion (Miao, 2025). Nevertheless, the integration of multimodal affective inference into pedagogically grounded tutoring systems remains comparatively under-explored.

Addressing this gap, the EmoBot project (Kallipolitis et al., 2026) proposes a multimodal affective module integrated into an LLM-based, CEFR-aligned intelligent tutoring system. The sys-

tem is currently tailored to learners of German as a foreign language. Rather than presenting the tutoring system as a whole, this work focuses specifically on the design and integration of the affective module within an agentive, NLP-driven architecture. The study therefore concentrates on the operationalization, synchronization, and fusion of multimodal affective signals, as well as on how these signals can inform adaptive interaction in foreign language learning scenarios.

A novelty of this work lies in the inclusion of various data sources through different user interaction points and the application of a dual data analysis architecture. The affective module is enhanced with the modalities of text, audio, and visual inputs from the learner's computer and biosignals, from the learner's already-owned smartwatch, offering a complete example for an end-to-end emotion-analysis-enabled language learning system (Koulouris et al., 2025). In this work, we highlight the importance of applying intelligent multi-resource and multimodal methods by implementing the proposed novel EmoBot architecture into an emotion recognition paradigm fed by textual, vocal, visual, and biosignal data. The proposed system uses the dual-architecture approach to collect, isolate, sync, and manipulate the data types of each modality and provide user feedback. This feedback is vital for learning, as evidenced by this proof-of-concept technical implementation.

One step further, a key contribution of the work is the transfer of intelligent mHealth-inspired methodologies to the domain of education, combining distributed systems, commodity wearable devices, AI, advanced data manipulation, and data fusion techniques to support affect-aware interaction in foreign language learning contexts.

## 3. State of the art

This section reviews the existing literature on emotion analysis in educational contexts. It focuses on multimodal approaches and their application to foreign language learning (FLL) and highlights current methodologies and technological trends. Finally, it identifies existing limitations and, thereby, establishes the research gap addressed by the present study.

### 3.1. Multimodal emotion analysis

Multimodal sentiment analysis is currently a central topic of research, enabling advanced human-machine interaction (Zhu et al., 2024) and providing a robust source of information (Ezzameli and Mahersia, 2023). The most common approaches include the use of visual analysis combined with NLP methods, either at neutral time (Hu and Flax-



man, 2018) or in real time. Real time examples include the utilization of speech (Kim et al., 2007), images or health-oriented data (Huang et al., 2016) such as electroencephalograms, respiratory belts, and electrodermal activity monitors (Nandi et al., 2021).

While these solutions provide classification for valence and arousal, they still require sophisticated, and particularly expensive, hardware and cannot run on commodity mobile devices. Other mHealth-oriented solutions include the utilization of feature extraction from textual and visual modalities, by utilizing Convolutional Neural networks (CNNs), and relying only on the commodity device hardware (Poria et al., 2016). Some approaches utilize text sentiment analysis, usually using SVM text classification models (Song, 2021).

Nevertheless, the integration of such multimodal architectures into pedagogically grounded educational systems remains comparatively limited.

### 3.2. Affective Computing in Language Learning

Intelligent Tutoring Systems (ITSs) have been increasingly adopted in instructional settings, including language education, as tools for supporting interactive and personalized learning (Belda-Medina and Calvo-Ferrer, 2022; Barrot, 2023; Antoniou-Kritikou et al., 2024). Empirical research indicates that their integration into educational contexts can enhance learner performance, foster critical thinking skills, and provide targeted and individualized instructional support (Davar et al., 2025), while also contributing to positive learner experiences and increased satisfaction (Kuhail et al., 2022). Despite acknowledged limitations, several studies report that the pedagogical benefits of conversational agents generally outweigh their drawbacks for both learners and educators (Huang et al., 2022).

Recent advances in LLMs have further enhanced the capabilities of ITSs by enabling adaptive interaction, flexible guidance, automated educational content generation - especially in low-resource settings (Samuel et al., 2024), and instruction tailored to individual learner needs. In the context of FLL, conversational agents have been shown to offer particular advantages for learners who are reluctant to engage in classroom interaction due to stress, anxiety, or lack of confidence, suppressing negative emotions and providing a low-pressure environment for language practice and experimentation (Alwazzan, 2024).

In the field of education, various emotion analysis methods have been proposed. Emotion recognition was used for sentiment analysis in learning, focusing on identifying emotions and having learning as the centre of interest (Barrón Estrada et al.,

2019). This method is also implemented in distant learning environments, where it monitors learners' perception, learning processes, and communication methods and highlights the importance of knowledge acquisition and decision-making (Duraes et al., 2021). Within the FLL context, there are systems that monitor the emotions experienced by learners (MacIntyre and Gregersen, 2012), the dynamics of which affect the learning outcome and proficiency (Wang et al., 2024b). Anxiety is a metric that is also measured in foreign language e-learning systems (Ismail and Hastings, 2019).

The utilization of the biosignal modality for emotion recognition is applied in web-based e-learning scenarios and provides the tutor with insights tailored to providing appropriate learning assistance or modifying guidance (Chen and Lee, 2011). Finally, camera-based systems with capabilities to recognize facial expressions using CNNs are studied in classrooms for assessing students' engagement, providing promising results (Pabba and Kumar, 2021).

Although the literature reports multimodal approaches in educational settings, there is limited work addressing languages other than English—such as German—particularly in systems that integrate biosignal-enabled emotion and sentiment analysis within affective computing frameworks. Existing approaches tend to rely either on single-modality inference or on multimodal fusion techniques that require sophisticated and often costly hardware. Our work addresses this gap by proposing a multimodal emotion analysis system that incorporates intelligent mHealth architectures and methods, enabling real-time or near-real-time estimation of learners' emotional states using commodity devices.

## 4. Methodology

The methodological approach adopted in this work aims to use affective signals in an authentic FLL setting in view of monitoring how learners' affective states emerge and evolve during interaction with an agentive educational system. Rather than treating emotion recognition as an end in itself, affective inference is used as a means for studying learner affective states and regulation within CEFR-aligned learning activities that support written production, reading comprehension, and use of German, oral, and listening comprehension. To this end, a multimodal affective analysis framework is embedded within EmoBot, enabling the continuous observation of affective signals during learning sessions.

The proposed system methodology is structured into three subsections: system design specifications, architectural details, and the implementation of the prototype.

## 4.1. System Design

The system is designed to facilitate efficient emotion analysis in the educational scenario of language learning through simultaneous utilization of biosignal, visual, and audio modalities. The pilot implementation is tailored to the needs of learners of German as a foreign language, as a proof-of-concept. The data inputs are recorded through handheld commodity devices, such as smartwatches and personal computers. A Cloud Processing Platform (CPP) retrieves the input signals, performs data processing and per-modality emotion classification, before fusing the data to extract an overall emotional status. In addition to the ML methods that are used, the CPP integrates an alternative LLM approach for analysing emotion, offering a dual architecture approach that recognizes emotion in real-time or NRT. The system design is shown in Figure 1.

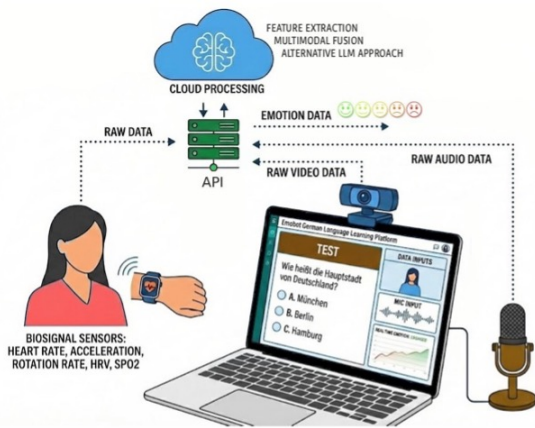


Figure 1: Overview of EmoBot. Data from different modalities (biosignals, audio, and visual) are transmitted to a cloud processing platform. Methods like feature extraction, data fusion, and LLMs are utilized to infer emotion in real-time or near-real-time.

## 4.2. System Architecture

The design of EmoBot follows a high-level interaction pipeline in which learner input is mediated through human–computer interaction, analyzed for affective cues, and used to inform adaptive pedagogical responses. Language understanding and generation are implemented through LLM-based NLP modules that support CEFR-aligned activity generation as well as feedback and explanation via retrieval-augmented generation. Within this pipeline, the emotion recognition module is aimed at providing affective signals that guide interaction flow and adaptation, enabling the system to adjust feedback, pacing, and task selection in a pedagogically appropriate manner.

The architecture of the proposed system fulfils the requirements of monitoring biosignals, along with recording the user expressions, importing the data in real time through a streaming bus service, synchronizing, extracting emotion from each modality and fusing. Hence, the system consists of a web e-Learning platform, which includes the learning interface and handles the visual and audio data input. Moreover, a smartwatch app is developed for biosignals recoding and a cloud environment incorporates a data input stream, an Affective Processing Unit (APU), an LLM unit and a Late Fusion Engine. An overview of the proposed architecture is shown in Figure 2.

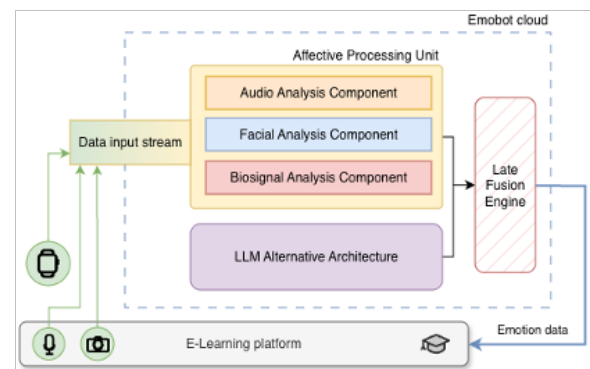


Figure 2: The EmoBot emotion analysis system architecture. Data relevant to each modality are retrieved from the educational platform (audio, visual data) and from a smartwatch application. These data are then fed into a unified input stream and transmitted to a cloud-based service for processing. Therein, the Affective Processing Unit (APU) handles the synchronization, feature extraction, and prediction generation. Data fusion is, then, performed through a Late Fusion Engine, enhanced and evaluated alternatively by LLMs, before returning a single emotional status to the educational platform.

The data input stream retrieves data from the three modalities in real time and transfers them to the APU. Therein, the Audio Analysis Component is responsible for extracting audio waveform data and performing emotion-based feature extraction. The Facial Analysis Component handles camera frames and the Biosignal Analysis Component analyses vital signs and IMU <sup>2</sup> data, performing also feature extraction. The results are synced and fused, incorporated alternatively by LLMs, contributing to a dual architecture approach.

<sup>2</sup>Intertrial Measurement Unit – A sensor which includes accelerometer and gyroscope and calculates accelerations and rotation rates.

### 4.3. Implementation

The system implementation is divided into three platforms: web, smartwatch and cloud.

#### 4.3.1. Web

The web system initiates once a learning session begins and utilizes WebRTC to send video and audio to the cloud platform. It is implemented in the NodeJS environment. Moreover, through a data streaming service implemented using Firebase Database, the web platform awaits emotion data from the cloud in NRT.

#### 4.3.2. Smartwatch

The smartwatch application is developed in watches that allow the custom application development, such as iOS WearOS and Android WatchOS. The recording session is initiated by the user by entering a 5-digit code generated over the web and retrieving the current session data. Recorded measurements include heart rate, SpO2, HRV calculated on device, accelerations, and rotation rates.

#### 4.3.3. Cloud

The cloud service exposes the data input stream through an API and an event-based data streaming functionality. The input interface retrieves the data from the aforementioned modalities and initiates the appropriate analysis component. The Audio Analysis Component utilizes a CNN-LSTM architecture with MFCC features and bidirectional temporal modelling to extract emotion from speech. The Facial Analysis Component utilizes methods for real-time facial expression recognition, with planned extensions using MediaPipe for landmark detection and 3D-CNNs for temporal expression analysis. The Biosignal Analysis Component utilizes datasets like the WESAD dataset (Schmidt et al., 2018) or SVM approaches (Hakim et al., 2018) to efficiently calculate emotions of stress, happiness, anger, or sadness.

Finally, an alternative LLM-based architecture is designed to extract fused affective information from audio and visual modalities. It is currently under development, and no fully operational prototype has been implemented to date. However, preliminary integration experiments have been conducted to assess its feasibility.

### 4.4. Adaptive Tutoring and human-computer interaction

The adaptive tutoring layer of EmoBot leverages affective inference by translating multimodal emotion

predictions into pedagogically meaningful interaction strategies. Crucially, this affect-driven adaptation is tightly integrated with the system's ability to generate CEFR-aligned instructional content through an LLM-based educational module. Thus, adaptation occurs not only at the level of feedback style and tone but also at the level of task design, linguistic complexity, and progression across proficiency levels.

This CEFR-aware generation mechanism provides the structural backbone upon which affective adaptation operates. Rather than dynamically generating arbitrary content, the system selects or generates tasks within a constrained proficiency band and then modulates their complexity and scaffolding intensity in response to affective signals.

## 5. Evaluation

This section highlights the results of the proof-of-concept prototype, including the system in practice as a paradigm validation and a qualitative evaluation.

### 5.1. System in practice

For the proof-of-concept smartwatch prototype, a WearOS application is developed that can measure heart rate and SpO2 in real time, as well as HRV in NRT. Additionally, the data from the onboard IMU sensor are included in a batch that is sent to the cloud every 30 seconds. The authentication is performed by entering a 5-digit code bound to the current user session. Snapshots of the system in the authentication and measurement screens are shown in Figure 3.

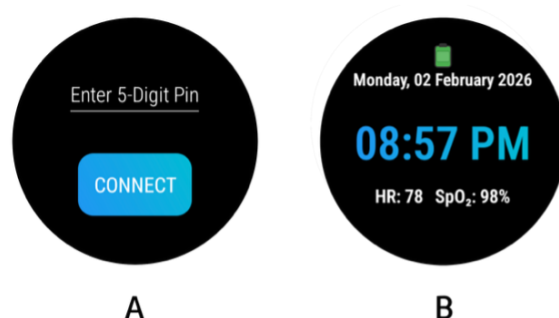


Figure 3: The mobile interface, including the authentication screen (A) and the live measurement screen (B), indicating the current heart rate and SpO2.

For the web interface, an HTML and JS-based iframe was created, including an interface for the camera and audio feed, and was included in the E-Learning platform. The platform is also responsible for showing the 5-digit code and rendering

the emotion result from the streaming output of the cloud service. A snapshot of the web interface is shown in Figure 4

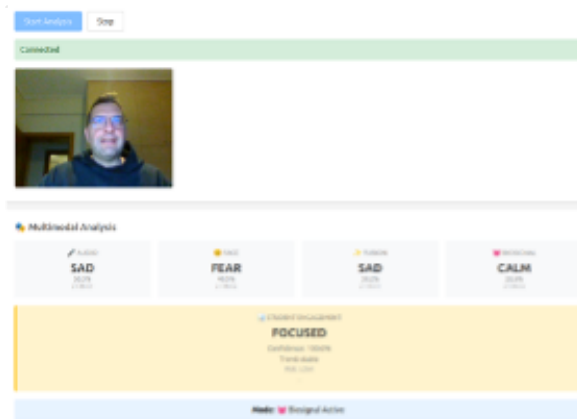


Figure 4: The web interface, including a camera preview utilizing WebRTC for audio and video input streams and multimodal analysis results in real time.

The components of the cloud service are built using a decoupled approach, in Python, dividing the data handling and the model inference. An LLM is utilized in the Late Fusion Engine to fuse the prediction of each modality into a general output, before notifying the E-Learning platform.

## 5.2. Paradigm evaluation

This proof-of-concept prototype acts as an example of the proposed architecture’s implementation. The three modalities are validated using state-of-the-art experimental protocols. For the audio modality, a comprehensive evaluation across eight datasets (English and German) identified the CNN-LSTM architecture as the superior model, achieving up to 81.3% accuracy on German speech with high stability and cross-lingual generalization, significantly outperforming complex alternatives like Efficient-Net.

For the visual modality, the DeepFace library (Serengil and Ozpinar, 2021) was employed for facial emotion recognition. DeepFace utilizes the Mini-Xception architecture (Arriaga et al., 2017), a lightweight CNN based on depthwise separable convolutions, trained on the FER-2013 dataset. The model achieves approximately 66% accuracy on the FER-2013 benchmark, which is comparable to human-level performance on the same dataset ( $65 \pm 5\%$ ). With only around 60,000 parameters, Mini-Xception is highly efficient and well-suited for real-time or near-real-time deployment on commodity hardware, making it a practical choice for the visual emotion recognition component of the proposed system. These independent validations

prove that the system uses the best possible methods and mitigates individual modality risks.

The alternative LLM-based architecture is under development and has not yet been fully deployed. It is designed to perform valence and arousal estimation from both audio and visual modalities, to produce a multimodal fused output. Preliminary integration experiments suggest that BERT-based models, particularly DistilBERT (Sanh et al., 2019), show promise in meeting the defined requirements. However, its evaluation falls outside the scope of this study, which focuses on biosignal-based affective computing in distributed systems using commodity devices.

## 5.3. Qualitative evaluation

In addition to component-level quantitative validation, and in the absence of a user study, a structured qualitative evaluation was conducted to assess the theoretical coherence, pedagogical interpretability, and instructional consistency of the affect-aware tutoring system within a CEFR-aligned FLL context.

**Theoretical coherence: representation of affect and FLL theory.** Our system models affect along the valence-arousal dimensions following Russell (1980) and Bradley and Lang (1994). This dimensional representation is consistent with the data captured via the various modalities including biosignals, while it also aligns with SLA research distinguishing positive (enjoyment, satisfaction) and negative (anxiety, fear of evaluation) affect (Dewaele and MacIntyre, 2014) in language learning contexts. Moreover, observed affect trajectories during task escalation (e.g., increasing arousal during time-constrained production tasks) correspond to theoretical expectations regarding foreign language anxiety (FLCA). Similarly, stabilized valence during scaffolded support aligns with positive psychology perspectives in SLA (MacIntyre and Gregersen, 2012). In this regard, the dimensional model produces affect patterns that are theoretically consistent with established emotion research in FLL contexts.

**Pedagogical interpretability.** The recognition and modelling of learners’ emotional states provide an insight into variations in emotional states throughout the learning process. SLA literature identifies anxiety, fear of negative evaluation, and reduced motivation as key adult learner constraints (Schumann, 1975; Gao and Liu, 2022). System-level affect trajectories plausibly correspond to elevated arousal in difficult tasks, stabilized valence following immediate formative feedback, and reduced activation during highly scaffolded comprehension tasks. These patterns align with both anxiety research and positive psychology perspectives in SLA (MacIntyre and Gregersen, 2012).



**Affect-aware adaptation and feedback.** Systematic and timely feedback has been found to enhance performance and sustain learner engagement and system retention (Katinskaia et al., 2018; de Haas et al., 2020), while encouragement contributes to progression across CEFR proficiency levels (Council of Europe, 2001). Formative feedback is widely regarded as a major determinant of learning outcomes (Hattie and Timperley, 2007) and is particularly critical for the development of written communicative language competence. Feedback should be immediate, transparent, and explicitly linked to CEFR can-do descriptors, guiding learners toward greater alignment with targeted proficiency levels and enabling iterative performance regulation in relation to descriptor-based performance criteria (Hattie and Timperley, 2007).

In the proposed architecture, affect-aware adaptations are embedded implicitly within the pedagogical interaction. The system does not explicitly label learners' emotional states (e.g., "You seem anxious"); instead, it modulates tone, pacing, task difficulty, and linguistic complexity. This approach implements affect-informed scaffolding by adjusting instructional support according to inferred valence-arousal patterns and gradually reducing assistance as learners stabilize, consistent with theoretical principles (Vygotsky, 1978).

These adaptations remain aligned with CEFR descriptors to ensure structured proficiency progression. This approach is grounded in SLA literature, which highlights the importance of fostering psychological safety and learner autonomy within a non-threatening interactional environment (Pachler et al., 2009), thereby reducing anxiety and promoting active participation in communicative language activities (Council of Europe, 2001).

Moreover, emotional variables, particularly learner self-confidence in communicative tasks, influence task completion and sustained engagement (Council of Europe, 2001), especially in evaluative contexts marked by perceived inadequacy or anxiety. Instructional design in our system supports learner agency and strategic competence through interactive scenarios and systematic positive reinforcement. Such affective engagement underpins persistence and successful progression across learning trajectories.

One step further, beyond real-time inference, the system maintains session-level and cross-session affective profiles indexed to task categories and CEFR descriptors. This longitudinal modelling enables detection of recurring task-affect couplings, such as elevated negative valence during reading comprehension at a specific proficiency level or reduced activation in written production; identified patterns guide medium-term adaptation through targeted reinforcement and controlled task

re-sequencing. This approach is consistent with SLA evidence linking learning efficiency to sustained attention, structured repetition, and strategic competence deployment (Council of Europe, 2001).

Overall, this systematic account of emotions is essential for identifying their contributing factors and for guiding research methodologies designed to mitigate their effects. From a computational perspective, the system can be interpreted as a multi-modal affect-adaptive control architecture in which valence and arousal estimates function as state variables governing pedagogical adaptation.

## 6. Discussion

In this section, we discuss challenges that arise from this work. Although many affective computing applications are grounded in basic emotion typologies (Ekman, 1992; Plutchik, 1980), the affective constructs that are pedagogically salient in educational settings only partially overlap with these predefined emotion categories. Basic emotion frameworks were primarily developed to describe evolutionarily grounded, universally recognizable affective expressions (e.g., anger, fear, happiness), often associated with discrete and prototypical facial configurations. In contrast, educational contexts frequently involve more subtle, cognitively intertwined affective states such as confusion, cognitive overload, sustained engagement, frustration, boredom, or curiosity. These states cannot always be mapped onto single discrete categories and may reflect blends of arousal, valence, and task-related appraisal.

This conceptual gap is also attested in the structure of widely used benchmark datasets, which predominantly annotate prototypical facial expressions corresponding to basic emotions under controlled conditions. In the present work, emotion categories were harmonized into a unified eight-class scheme, namely, Anger, Disgust, Fear, Happiness, Neutral, Sadness, Contempt, and Surprise (Ekman and Friesen, 1986), to reconcile discrepancies across source datasets (see Section 5.2). While this harmonization improves cross-dataset consistency for model training and evaluation, it does not fully address the broader issue that these categories only partially capture educationally meaningful affective states.

Two methodological challenges arise from this discrepancy. First, a label mismatch problem emerges: models trained on basic or general emotion datasets must rely on indirect or heuristic mappings when applied to pedagogical scenarios. Second, benchmark datasets often contain expressions recorded in controlled environments, whereas classroom or self study interactions involve low intensity, context dependent affective fluctuation.



tuations embedded within complex cognitive processes.

These limitations highlight the need for domain-specific datasets that go beyond discrete basic emotion labels and incorporate pedagogically grounded taxonomies. Such datasets should ideally include temporally aligned signals, task metadata, and dimensional affect annotations (e.g., valence–arousal trajectories), enabling the study of affect as a dynamic and context-sensitive process. Developing and validating such resources remains an important direction for advancing affect-aware educational systems.

## 7. Conclusion

To conclude, we have presented a multimodal affective computing architecture integrated into an intelligent language learning system delivering CEFR-aligned activities within an interactive environment. The focus is on the design and system-level validation of a distributed framework that combines audio, visual, and physiological signals acquired through commodity devices to support real-time or near-real-time affect inference.

The proposed architecture and the empirical results demonstrated by the working prototype underscore opportunities for further research. As a future direction, different methods can be used for modality-based predictions. Moreover, additional hardware such as IMU sensors or cameras can be tested. A larger-scale deployment of the system is underway, and, hence, results from applying the system to real students will be carried out. The state-of-the-art affective emotion recognition architecture, along with the developed paradigm, validates its applicability within the context of computational emotion recognition. The contribution directs towards continued research in the field, particularly in relation to its pedagogical and cognitive implications for Foreign Language Learning, with specific relevance to the acquisition of languages other than English.

Finally, future work is also underway to empirically validate the system with end-users and to develop a pedagogically grounded multimodal dataset designed to support reproducible research in affect-aware educational applications. All data will be made publicly available for reproducibility purposes, subject to ethical and privacy constraints.

## 8. Limitations and ethical considerations

At the time of submission, the proposed framework has undergone system-level validation but has not yet been deployed in a large-scale learner study.

No learner self-reports, expert annotations, or controlled comparative experiments were conducted at this stage. Consequently, affect inference accuracy and measurable learning gains remain to be empirically validated. A controlled evaluation with learners of German as a foreign language is currently in progress.

Furthermore, biosignals are probabilistic indicators of affective states and may be sensitive to individual variability, contextual noise, and model bias. Although temporal smoothing and bounded adaptation reduce instability, misclassification of affect may lead to suboptimal instructional modulation.

From an ethical perspective, affective computing in educational settings raises concerns related to privacy, transparency, and autonomy. The system processes multimodal signals that may be considered sensitive data; therefore, data minimization, secure storage, and explicit informed consent are required.

## 9. Acknowledgements

This research was supported by the National Recovery and Resilience Plan (NRRP) “Greece 2.0” under the “Clusters of Research Excellence” (CREs) program, SUB1.1, with project code OΠΣ ΤΑ 5180519 and title “Interactive Agent with Emotional Intelligence for Second/Foreign Language Learning”, Acronym: “EmoBot”.

## 10. Bibliographical References

### References

- Alaa Alsiaity and Rita Orji. 2024. [Machine learning techniques for emotion detection and sentiment analysis: current state, challenges, and future directions](#). *Behaviour & Information Technology*, 43(1):139–164.
- Mona Saleh Alwazzan. 2024. [Investigating the Effectiveness of Artificial Intelligence Chatbots in Enhancing Digital Dialogue Skills for Students](#). *Başlık*, 13(2):573–584.
- Ioanna Antoniou-Kritikou, Voula Giouli, George Tsoulouhas, and Constandina Economou. 2024. [Using LLMs in a language teaching and learning application](#). *ECRIM*, 136. Special Theme: Large Language Models; Guest Editors: Diego Collarana Vargas and Nassos Katsamanis.
- Octavio Arriaga, Matias Valdenegro-Toro, and Paul Plöger. 2017. Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.07557*.

- Jessie S. Barrot. 2023. [Using Chatgpt for second language writing: Pitfalls and potentials](#). *Assessing Writing*, 57:100745.
- Maria Barrón Estrada, Ramón Zatarain Cabada, and Raul Oramas. 2019. [Emotion Recognition for Education using Sentiment Analysis](#). *Research in Computing Science*, 148(5):71–80.
- Jose Belda-Medina and José Ramón Calvo-Ferrer. 2022. [Using Chatbots as AI Conversational Partners in Language Learning](#). *Applied Sciences*, 12(17):8427.
- Margaret M. Bradley and Peter J. Lang. 1994. Measuring emotion: The Self-Assessment Manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59.
- Chih-Ming Chen and T.-H Lee. 2011. [Emotion recognition and communication for reducing second-language speaking anxiety in a web-based one-to-one synchronous learning environment](#). *British Journal of Educational Technology*, 42:417–440.
- Council of Europe. 2001. [Common European Framework of Reference for Languages: Learning, Teaching, Assessment](#). Council of Europe, Strasbourg.
- Narius Farhad Davar, M. Ali Akber Dewan, and Xiaokun Zhang. 2025. [AI Chatbots in Education: Challenges and Opportunities](#). *Information*, 16(3).
- Mirjam de Haas, Paul Vogt, and Emiel Krahmer. 2020. The effects of feedback on children's engagement and learning outcomes in robot-assisted second language learning. *Frontiers in Robotics and AI*, 7:101.
- Jean-Marc Dewaele and Peter Macintyre. 2016. [Foreign Language Enjoyment and Foreign Language Classroom Anxiety. The right and left feet of FL learning?](#), pages 215–236. Bristol: Multilingual Matters.
- Jean-Marc Dewaele and Peter D. MacIntyre. 2014. [The two faces of Janus? Anxiety and enjoyment in the foreign language classroom](#). *Studies in Second Language Learning and Teaching*, 4(2):237–274.
- Duarte Dias and João Paulo Silva Cunha. 2018. [Wearable Health Devices—Vital Sign Monitoring, Systems and Technologies](#). *Sensors (Basel, Switzerland)*, 18.
- Dalila Duraes, Ramon Toala, and Paulo Novais. 2021. [Emotion Analysis in Distance Learning](#), pages 629–639. Springer International Publishing.
- Paul Ekman. 1992. [An argument for basic emotions](#). *Cognition and Emotion*, 6(3-4):169–200.
- Paul Ekman and Wallace V. Friesen. 1986. [A new pan-cultural facial expression of emotion](#). *Motivation and Emotion*, 10(2):159–168.
- Kaouther Ezzameli and Hela Mahersia. 2023. [Emotion recognition from unimodal to multi-modal analysis: A review](#). *Information Fusion*, 99:101847.
- Lixiang Gao and Honggang Liu. 2022. [Revisiting students' foreign language learning demotivation: From concepts to themes](#). *Frontiers in Psychology*, 13.
- Lutfi Hakim, Adhi Dharma Wibawa, Evi Septiana Pane, and Mauridhi Hery Purnomo. 2018. [Emotion Recognition in Elderly Based on Spo2 and Pulse Rate Signals Using Support Vector Machine](#). In *2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS)*, pages 474–479.
- John Hattie and Helen Timperley. 2007. [The Power of Feedback](#). *Review of Educational Research*, 77(1):81–112.
- María A. Hernández-Mustieles, Yoshua E. Lima-Carmona, Maxine A. Pacheco-Ramírez, Axel A. Mendoza-Armenta, José Esteban Romero-Gómez, César F. Cruz-Gómez, Diana C. Rodríguez-Alvarado, Alejandro Arceo, Jesús G. Cruz-Garza, Mauricio A. Ramírez-Moreno, and Jorge de J. Lozoya-Santos. 2024. [Wearable biosensor technology in education: A systematic review](#). *Sensors*, 24(8).
- Anne Cummings Hlas, Krista Neyers, and Sarah Molitor. 2019. [Measuring student attention in the second language classroom](#). *Language Teaching Research*, 23:107 – 125.
- Anthony Hu and Seth Flaxman. 2018. [Multimodal sentiment analysis to explore the structure of emotions](#). In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD '18*, page 350–358, New York, NY, USA. Association for Computing Machinery.
- Weijiao Huang, Khe Foon Hew, and Luke K. Fryer. 2022. [Chatbots for language learning—are they really useful? a systematic review of chatbot-supported language learning](#). *Journal of Computer Assisted Learning*, 38(1):237–257.

- Xiaohua Huang, Jukka Kortelainen, Guoying Zhao, Xiaobai Li, Antti Moilanen, Tapio Seppänen, and Matti Pietikäinen. 2016. [Multi-modal emotion analysis from facial expressions and electroencephalogram](#). *Comput. Vis. Image Underst.*, 147(C):114–124.
- Daneih Ismail and Peter Hastings. 2019. [Identifying foreign language anxiety when using an e-learning system](#). pages 131–140.
- Athanasios Kallipolitis, Dionysios Koulouris, Melina Tziokama, Kosmas Pinitas, Argyrios Zafeiriou, Andreas Menychtas, Ilias Maglogiannis, Voula Giouli, Athina Sioupi, Stamatia Michalopoulou, George Tsoulouhas, Michail Katras, Panagiotis Charalampopoulos, and Aristotelis Stamopoulos. 2026. Conversational agent with emotional intelligence for foreign language learning. In *Proceedings of the 14th International Conference on Information and Education Technology (IEEE-ICIET 2026)*, Koriyama, Japan. IEEE.
- Zoe Kantaridou and Angeliki Psaltou-Joycey. 2023. [Positive emotions and self-regulation strategies in EFL classroom situations](#). In *Selected papers on theoretical and applied linguistics*, pages 401–419.
- Anisia Katinskaia, Javad Nouri, and Roman Yangarber. 2018. Revita: a language-learning platform at the intersection of its and call. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).
- Samuel Kim, Panayiotis G. Georgiou, Sungbok Lee, and Shrikanth S. Narayanan. 2007. [Real-time emotion detection system using speech: Multi-modal fusion of different timescale features](#). *2007 IEEE 9th Workshop on Multimedia Signal Processing*, pages 48–51.
- Dionysis Koulouris, Melina Tziomaka, Argyrios Zafeiriou, Kosmas Pinitas, Athanasios Kallipolitis, Andreas Menychtas, and Ilias G. Maglogiannis. 2025. [Design of a multimodal affective module for emotion recognition in adaptive language learning systems](#). *2025 10th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*, pages 1–6.
- Mohammad Amin Kuhail, Nazik Alturki, Salwa Alramlawi, and Khlood Alhejori. 2022. [Interacting with educational chatbots: A systematic review](#). *Education and Information Technologies*, 28(1):973–1018.
- Tuhin Kundu and Chandran Saravanan. 2017. [Advancements and recent trends in emotion recognition using facial image analysis and machine learning models](#). *2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEEC-COT)*, pages 1–6.
- Agnieszka Lyons. 2015. *Multimodality*, chapter 18. John Wiley & Sons, Ltd.
- Peter MacIntyre and Tammy Gregersen. 2012. [Emotions that facilitate language learning: The positive-broadening power of the imagination](#). *Studies in Second Language Learning and Teaching*, 2(2):193–213.
- Guixian Miao. 2025. [A text-audio multimodal weighted network for emotion recognition to enhance interactivity in online education](#). *Journal of Computational Methods in Sciences and Engineering*, 25:5764 – 5777.
- Arijit Nandi, Fatos Xhafa, Laia Subirats, and Santi Fort. 2021. Real-time multimodal emotion classification system in e-learning context. In *Proceedings of the 22nd Engineering Applications of Neural Networks Conference*, pages 423–435, Cham. Springer International Publishing.
- Chakradhar Pabba and Praveen Kumar. 2021. [An intelligent system for monitoring students' engagement in large classroom teaching through facial expression recognition](#). *Expert Systems*, 39.
- Norbert Pachler, Harvey Mellar, Caroline Daly, Yishay Mor, and Dylan Wiliam. 2009. [Scoping a vision for formative e-assessment: a project report for JISC](#). WLE Centre and JISC, London.
- Robert Plutchik. 1980. [A general psychoevolutionary theory of emotion](#). *Emotion: Theory, research, and experience*, 1:3–33.
- Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. [Convolutional mkl based multimodal emotion recognition and sentiment analysis](#). In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 439–448.
- James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.
- Stanisław Saganowski, Bartosz Perz, Adam G. Polak, and Przemysław Kazienko. 2023. [Emotion recognition for everyday life using physiological signals from wearables: A systematic literature review](#). *IEEE Transactions on Affective Computing*, 14(3):1876–1897.

- Vinay Samuel, Houda Aynaou, Arijit Chowdhury, Karthik Venkat Ramanan, and Aman Chadha. 2024. [Can LLMs augment low-resource reading comprehension datasets? opportunities and challenges](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 307–317, Bangkok, Thailand. Association for Computational Linguistics.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.
- Philip Schmidt, Attila Reiss, Robert Duerichen, Claus Marberger, and Kristof Van Laerhoven. 2018. [Introducing wesad, a multimodal dataset for wearable stress and affect detection](#). In *Proceedings of the 20th ACM International Conference on Multimodal Interaction, ICMI '18*, page 400–408, New York, NY, USA. Association for Computing Machinery.
- John H. Schumann. 1975. [Affective factors and the problem of age in second language acquisition](#). *Language Learning*, 25(2):209–235.
- Sefik Ilkin Serengil and Alper Ozpinar. 2021. [Hyperextended lightface: A facial attribute analysis framework](#). *2021 International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4.
- Gang Song. 2021. [Sentiment analysis of japanese text and vocabulary learning based on natural language processing and svm](#). *Journal of Ambient Intelligence and Humanized Computing*.
- Lev S. Vygotsky. 1978. *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.
- Ke Wang, Lin Wei, Yuanyuan Liu, Jingying Chen, Yibing Zhan, Hua Jin, Chongchong Qi, and Zhe Chen. 2024a. Affective computing for health-care: Recent trends, applications, challenges, and beyond. In *CSIG Conference on Emotional Intelligence*, pages 3–19. Springer.
- Peng Wang, Lesya Y. Ganushchak, Camille Welie, and Roel van Steensel. 2024b. [The dynamic nature of emotions in language learning context: Theory, method, and analysis](#). *Educational Psychology Review*, 36.
- Xianxun Zhu, Chaopeng Guo, Heyang Feng, Yao Huang, Yichen Feng, Xiangyang Wang, and Rui Wang. 2024. [A Review of Key Technologies for Emotion Analysis Using Multimodal Information](#). *Cognitive Computation*, 16:1504 – 1530.



# PoETIC: A Re-framing of Context Dependent Emotion Detection

Nirmal Surange, Manish Shrivastava

Language Technologies Research Center  
Kohli Center for Intelligent Systems, IIIT Hyderabad, India  
nirmal.surange@research.iiit.ac.in, m.shrivastava@iiit.ac.in

## Abstract

Emotion classification has been extensively studied, with numerous datasets enabling progress in both textual and multimodal settings. However, most existing text-based resources treat emotion as an utterance-level property, assuming that the emotional content is fully encoded in the sentence itself. This assumption is problematic: in the absence of paralinguistic cues such as prosody, facial expressions, or emojis, textual emotions are often highly context-dependent. Many utterances lack explicit emotion markers, and even when present, such cues may be overridden by broader situational context. Sentence-level emotion annotation, thus, is driven by the annotator's ability to imagine the context in which the given utterance would elicit a given emotion. An utterance may be able to express an emotion completely (Emotion Obvious), or it can express an emotion when imagined in a certain context (Emotion Plausible). Also, for an utterance, certain emotions might be implausible to express given the specific wording of a sentence (Emotion-Implausible). To address these issues, we create a new paradigm for emotion classification by categorizing utterance and emotion pairs into context-dependency classes. We present the PoETIC benchmark dataset, where sentences in the GoEmotions dataset are human-annotated for the three aforementioned classes across seven emotions (Fear, Anger, Sadness, Joy, Disgust, Surprise, and Neutral). We observe that gold-tagged emotions in GoEmotions do not have a clear correlation with human judgment with respect to the ability to express other emotions, given different contexts. Human annotators identify significantly more plausible emotions for a given utterance if asked to imagine a plausible context per utterance-emotion pair. We also present baselines using three popular large language models and two "small" language models in zero-shot and few-shot settings on the benchmark dataset.

**Keywords:** Emotion Recognition, Benchmark, Language resource

## 1. Introduction

Emotion classification is a well-established domain within natural language processing (NLP), with significant applications in sentiment analysis, dialogue systems, and affective computing.

Over recent decades, a variety of benchmark datasets have been developed to facilitate the study of emotion detection in textual data. For instance, the ISEAR dataset compiled short statements categorized by fundamental emotions (Scherer and Wallbott, 1994), while EmoBank introduced valence-arousal-dominance (VAD) scores applicable to sentences (Büchel and Hahn, 2017). Furthermore, GoEmotions (Demszky et al., 2020) offered detailed annotations for 27 emotions across 58,000 Reddit comments. More recently, the BRIGHTER dataset expanded this research into multiple languages, featuring human-annotated labels for basic emotions (Muhammad et al., 2025). Concurrently with text-exclusive resources, multimodal datasets such as IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), and SEMAINE (McKeown et al., 2011) underscore the critical role of audio-visual cues in emotion recognition. Collectively, these datasets have considerably advanced research in both textual and multimodal emotion classification.

Current textual emotion datasets predominantly employ an utterance-based classification paradigm, where a sentence or short text snippet is directly

labeled with one or more emotions, often assuming that the emotional content resides entirely within the utterance itself. This simplification, although practical for supervised learning, neglects the inherently context-dependent nature of emotion interpretation.

Most utterances do not contain strong lexical emotion cues. Even when emotion-laden words are present, their interpretation may be overridden or reshaped by the surrounding context. In the absence of paralinguistic signals such as prosody, emojis, or facial expressions, textual emotion becomes highly context-dependent. The same sentence can plausibly convey different emotions when presented in different contexts. For example, while *"I can't believe this happened"* conveys 'surprise', it could signal joy at a positive surprise, anger at an injustice, or sadness at a sudden, unexpected loss. Without context, the interpretation remains ambiguous.

Therefore, we posit that text-based emotion detection faces the following principal challenges:

- Contextual dependence in text-only environments (in the absence of paralinguistic signals such as prosody or emojis).
- Absence of apparent emotional cues in text can be reinterpreted or overridden by the context.

These observations motivate a shift from



utterance-level emotion labeling to modeling contextual dependence. We introduce a task that determines whether an utterance: 1. *“obviously”* conveys the given emotion, 2. *can* convey the given emotion in certain contexts, or 3. *cannot* convey the given emotion in any conceivable context. To allow for context-based emotion interpretation, we formulate the task at the utterance–emotion pair level, and classify each pair into one of the three context-dependence categories (Section 4.1).

We present a new language resource: **“Plausibility of Emotion Tags for Imagined Contexts (PoETIC)”**<sup>1</sup>, a human-annotated benchmark of 7k utterance-emotion pairs annotated for contextual dependence. The utterances are taken from the test set of the GoEmotions (Demszky et al., 2020) dataset.

Given the strong performance of large language models (LLMs) on emotion classification, it is natural to investigate whether they can achieve human-level performance on this task. Therefore, we also report and discuss baseline results using both proprietary LLMs and self-hosted small language models.

## 2. Related Work

Most of the recent work around emotion classification is on the six basic emotions identified in Ekman (1992): happiness, sadness, fear, anger, disgust, and surprise. These emotions are considered fundamental and universally recognized across cultures, often characterized by distinct physiological and facial expressions.

Several benchmark datasets have addressed emotion classification in a variety of ways and have established emotion classification as a core NLP task. The ISEAR dataset (Scherer and Wallbott, 1994) collected short self-reports of emotional experiences, each labelled with one of seven basic emotions. EmoBank (Büchel and Hahn, 2017) annotated sentences with dimensional valence–arousal–dominance (VAD) scores, providing a graded view of affect. GoEmotions (Demszky et al., 2020) introduced a large-scale dataset of 58k Reddit comments annotated with 27 fine-grained emotions, demonstrating the feasibility of fine-grained classification. More recently, the BRIGHTER dataset (Muhammad et al., 2025) expanded emotion resources across multiple languages, providing high-quality human annotations for the seven basic emotions.

In contrast to text-only resources, multimodal datasets highlight the importance of audio-visual context in emotion recognition. IEMOCAP (Busso et al., 2008) provided dyadic conversations with

aligned speech, motion-capture, and text. MELD (Poria et al., 2019) introduced a multimodal corpus of TV dialogues with aligned audio, video, and transcripts for multi-party conversation emotion recognition. Similarly, the SEMAINE database (McKeown et al., 2011) captured emotionally rich human–agent interactions with multimodal annotation.

These datasets have fueled progress in supervised learning for emotion detection but remain utterance-based, providing labels without contextual grounding. However, understanding emotions in the broader context of discourse remains a distant goal. Some resources are available for context-based emotion tagging in dialogues, such as the IEST task at WASSA2018 (Klinger et al., 2018) and EmoContext task at SemEval2019 (Chatterjee et al., 2019). These datasets highlight how paralinguistic and situational cues disambiguate emotions that text alone cannot resolve. However, they are costly to build and domain-limited, leaving the challenge of context-aware modeling in purely textual settings open. Notably, Yang et al. (2023) attempts to provide a formal definition of “textual context” for emotion identification in LLM prompting, making a case for either re-annotating corpora for emotion identification or using LLMs to synthesize context to help with the task.

Recent advances in LLMs have led to their application to zero-shot and few-shot emotion classification. Demonstrating that, while LLMs can perform well on benchmark datasets, a persistent difficulty is emotion trigger extraction: identifying the implicit contextual assumptions that justify a particular label. Attri et al. (2025) explored the joint task of emotion detection and explanatory span identification in e-commerce reviews to understand what triggers customer emotional responses. They leverage LLMs to unify fine-grained emotion detection and opinion trigger extraction from customer reviews.

In an effort to evaluate LLMs on the Emotion Classification task, Singh et al. (2024) observed that LLMs demonstrate **superior accuracy in predicting emotions**, with GPT-4 consistently outperforming fine-tuned transformer models across diverse datasets. However, they **struggle significantly to identify specific emotion triggers**, showing a limited ability to link emotional labels to the actual events or entities that caused them. Consequently, these models rely more on **topical key phrases and corpus-level cues** than on the deep, appraisal-based understanding characteristic of human emotion processing.

<sup>1</sup><https://github.com/nirmalsurange/PoETIC>

Sentence	Emotion	Emotion-Plausibility
I didn't know this was a "reveal" he'd been talking about for a while.	SURPRISE	Emotion-Obvious
I didn't know this was a "reveal" he'd been talking about for a while. <b>Context:</b> *Speaker ruined a loved one's big announcement, because of the speaker's misunderstanding*	SADNESS	EMOTION-PLAUSIBLE
I didn't know this was a "reveal" he'd been talking about for a while.	NEUTRAL	Emotion-Implausible
I know. I didn't say I agreed with it did I?	ANGER	Emotion-Obvious
I know. I didn't say I agreed with it did I? <b>Context:</b> *a defensive, panicked justification by someone who has been coerced into participating in something illegal, and is now trying to distance themselves from the consequences while fearing retribution from the person in charge.*	FEAR	Emotion-Plausible
It's real and it plays SO OFTEN I CAN'T STAND IT!!	Joy	Emotion-Implausible
It's real and it plays SO OFTEN I CAN'T STAND IT!!	DISGUST	Emotion-Obvious

Table 1: Illustrative examples contrasting context-based emotion plausibility.

### 3. GoEmotions dataset

One of the most prominent human-annotated datasets available for Emotion Classification is GoEmotions (Demszky et al., 2020). This dataset contains 58k samples, manually annotated for four sentiment categories (Positive, Negative, Neutral, and Ambiguous), which are divided into six coarse Ekman classes (Fear, Anger, Disgust, Sadness, Joy, Surprise) plus Neutral, and are further subdivided into 27 fine-grained emotions. Samples in GoEmotions are sourced from the Reddit<sup>2</sup> platform. The authors of GoEmotions note that while the platform is not representative of the general English-speaking population, the curation and selection steps followed by the authors make the data fairly balanced. They filtered content to reduce profanity and balance emotions, sentiment, length, and source subreddits. The final dataset contains 58009 samples split into 80%, 10%, and 10% for training, development, and test sets, respectively.

We utilize sentences from the test set of this dataset for human annotation of Contextual Dependence across the seven emotion categories (the six Ekman classes plus Neutral) outlined previously. This annotation exercise resulted in a dataset which has an average of 4.55 emotions (Table 3) per utterance as compared to 1.09 for the source GoEmotions Test set.

### 4. Context Dependency

**Contextual dependence of emotion interpretation** is an understudied phenomenon that may have a significant bearing on the emotion prediction task. As Table 1 shows, while some sentences can have

<sup>2</sup>www.reddit.com

Stage	Test
Original sentences	5,427
Emoji removal	-268
Foul language removal	-267
Short/noisy sentence removal	-106
Hashtag / URL removal	-16
Clean sentences	4,770
Selected sentences	1000

Table 2: Dataset Preparation Statistics

some apparently obvious emotions, these utterances can be reinterpreted to have different emotions in some imagined contexts.

We aim to study the extent to which the perceived emotion of a sentence depends on its surrounding context and to augment existing emotion recognition resources with explicit annotations indicating context dependence as perceived by a human annotator. While we do not generate contexts relevant to the specific utterance-emotion pairs, we ask the human annotators to evaluate if, under some imagined context, the utterance in question will express the given emotion. This also allows us to consider those utterance-emotion pairs that cannot belong together (EI).

**Data Preparation.** We base our dataset on the Test set of the GoEmotions corpus, which contains 5,427 sentences. To ensure that emotion recognition is based solely on linguistic cues, we first remove sentences containing emojis (268 instances). We then apply a series of filtering steps to eliminate sentences with auxiliary or noisy signals, specifically those containing foul or profane language (267 instances), very short or noisy content (106 instances), and hashtags or URLs (16 instances).

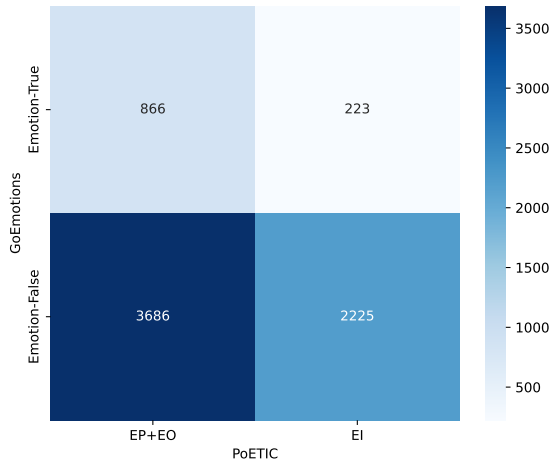


Figure 1: GoEmotions Vs. PoETIC tagging

After applying these filters, we retain 4,770 sentences, from which we randomly select 1,000 to serve as the source for our benchmark. (Table 2).

#### 4.1. Human Annotation

**Annotation Protocol** Three paid annotators, with native-level proficiency in English, label each utterance-emotion pair for its context dependence. Annotators are provided with an utterance and a candidate emotion label. Their task is to assign exactly one of the following labels to each pair:

- **Emotion-Obvious(EO):** The sentence clearly and unambiguously expresses the target emotion even in the absence of any additional context
- **Emotion-Plausible(EP):** The sentence may express the target emotion, but the interpretation is plausible only in the presence of relevant additional contextual information
- **Emotion-Implausible (EI):** The sentence cannot plausibly express the target emotion in any imaginable context

While it is possible to have a much finer classification ("highly dependent on context", "slightly dependent on context", etc.) to better cover the spectrum of context dependence, we found that increasing the labelset resulted in very poor inter-annotator agreement. This suggests that the cognitive load of making finer distinctions leads to poor performance. The same is true for annotating the entire emotion set of 27 emotions presented in GoEmotions. This is why we utilize the Ekman-emotion mapping provided by GoEmotions to convert the original annotations of 27 emotions into the 6 basic emotions.

**Neutral emotion.** The label NEUTRAL denotes the absence of all six non-neutral emotions (ANGER, SADNESS, JOY, FEAR, DISGUST, SURPRISE). A sentence cannot be labeled **EO** for both NEUTRAL and any other emotion. If a sentence clearly expresses a non-neutral emotion, it must not be labeled **EO** for NEUTRAL.

#### 4.2. Dataset statistics

The annotation exercise resulted in a final corpus of 7000 annotated utterance-emotion pairs (Table 3). We achieve a Krippendorff's  $\alpha$  score of 0.62 for the inter-annotator agreement. This inter-annotator score underlines the cognitive complexity of the task. We take the majority vote across annotators to select our final class label. We find that significant confusion between EI and EP classes, with 20 – 25% of disagreement among annotators for these classes on average.

Our annotations reveal that sentences can plausibly express substantially more emotions (Table 3) once context dependency is modeled. As expected, context-dependent emotions dominate in the available data, highlighting the importance of contextual reasoning for realistic emotion understanding. When seen in comparison to the source dataset, though, we observe that while most of the sentences get **EO** or **EP** tags for their assigned emotions in GoEmotions, a substantial number of utterance-(tagged)emotion pairs are classified as **EI** by our annotators (Figure 1). While this statistic is interesting, it only points to the subjectivity inherent in the emotion classification task. Table 4 details this further. Here, the 'tagged' row lists the annotation distribution across the three categories for sentences with their GoEmotions assigned tags. Similarly, the 'untagged' row reports the distribution of sentences paired with other emotions (negative).

### 5. LLMs for Context Dependency

To explore LLMs' ability to identify context dependence of emotions, we performed zero-shot inference on three popular instruction-tuned LLMs. These models are: GPT-5-mini<sup>1</sup>, Gemini-2.5-flash<sup>2</sup> and Llama3-70b (GroqLlama)<sup>3</sup>. Appendix-A shows the prompts and hyperparameter settings used for these LLMs. As the models are quite sensitive to prompts, we share the prompts that performed the best for each model. We further evaluate performance against a supervised baseline

<sup>1</sup>GPT-5-mini model by OpenAI. <https://platform.openai.com/docs/models>

<sup>2</sup>Gemini 2.5-flash model by Google. <https://ai.google.dev/models/gemini>

<sup>3</sup>Llama 3 70B hosted via Groq. <https://console.groq.com/docs/models>

	TOTAL	Class Distribution						Avg. Emotions / Sentence	
		EP	(%)	EI	(%)	EO	(%)	GoEmotions	EP+EO
PoETIC	7,000	3,474	49.58	2,448	34.94	1,078	15.38	1.09	4.55

Table 3: PoETIC Stats and Class Distribution

	GoEmotions Tag	Context-Dependency Labels		
		Emotion-Plausible	Emotion-Implausible	Emotion-Obvious
PoETIC	Tagged	368	223	498
	Untagged	3,106	2,225	580

Table 4: Relation between original GoEmotions and PoETIC tagging

based on RoBERTa<sup>4</sup>, trained for 10 epochs on the PoETIC dataset (80% training, 10% development, and 10% test split), using 5-fold cross-validation.

Table 5 compares the three large language models on PoETIC and the RoBERTa benchmark. As can be seen, GPT-5-mini performs reasonably well, with an accuracy and F1 scores of 0.77. The model (Figure 3a), does fairly well on the Emotion-Plausible (EP) class but results in significant misclassifications of the Emotion-Implausible (EI) class, showing that the model is not able to disambiguate between these classes. At the same time, Gemini-2.5-flash demonstrates moderate performance, achieving an accuracy of 0.68. While its macro F1 (0.68) is the same as its micro and weighted F1 scores, the confusion matrix (Figure 3b) reveals some confusion between EP and EI classes.

This shows that GPT-5-mini can reliably distinguish between emotion expressions that require external context and those that are self-contained. The consistency between macro, micro, and weighted F1 indicates robust performance across both frequent and infrequent classes.

Gemini-2.5-flash captures class-level distinctions quite well, but it struggles to consistently identify when emotion interpretation depends on additional context, particularly for EP instances.

GroqLlama performs the weakest among the evaluated systems, with an accuracy of 0.59 and a notably lower macro F1 score of 0.53. The model exhibits confusion between EP and EI categories (Figure 3c), showing limited sensitivity to subtle contextual cues that disambiguate emotion dependence. This aligns with the hypothesis that smaller or less instruction-tuned models may rely more heavily on surface-level lexical cues rather than deeper contextual reasoning.

We also consider a majority voting (MajorityVote in Table 5) scheme to consolidate the annotations by all the models. As we can see, the combined oracle of the models does not result in a signifi-

cant improvement in performance. We find that a large number of utterance-emotion pairs were misclassified by all the models (Table 6).

Overall, these results highlight that context-dependence classification is highly sensitive to a model’s contextual reasoning capabilities. While state-of-the-art models approach human-level reliability on this task, weaker models exhibit systematic confusions that disproportionately affect context-dependent emotion categories. This performance gap highlights the importance of evaluating emotion understanding not only in isolation but also through the lens of contextual grounding.

## 6. Local LLMs for Context Dependency

While models with  $> 20b$  parameters are considered ‘experts’ for almost all NLP tasks, most real-world use cases often require comparable performance from significantly smaller models. We explore this by choosing two popular local LLMs (Qwen3-8b<sup>3</sup> and Llama-3-8b<sup>4</sup>) and test them on PoETIC. Table 7 reports the performance of Qwen3-8B and Llama-3-8B under zero-shot and few-shot settings.

In the zero-shot regime, both models achieve comparable accuracy (Qwen3-8B: 0.5; Llama-3-8B: 0.47), with Llama-3-8B exhibiting a slightly higher macro F1 (0.22 vs. 0.18), indicating marginally better class balance despite lower overall accuracy. However, both models show low macro F1 scores, reflecting difficulty in handling minority classes (EO).

Introducing few-shot prompting leads to some performance gain for Qwen3-8B, improving accuracy to 0.52 and macro F1 to 0.36. There are balanced predictions across classes, as evident by the reduced off-diagonal mass in the confusion matrix (Figure 4). Inexplicably, Llama-3-8B exhibits a severe performance degradation in the few-shot

<sup>4</sup>RoBERTa-base model <https://huggingface.co/FacebookAI/roberta-base>

<sup>3</sup><https://huggingface.co/Qwen/Qwen3-8B>

<sup>4</sup><https://huggingface.co/meta-llama/Meta-Llama-3-8B>

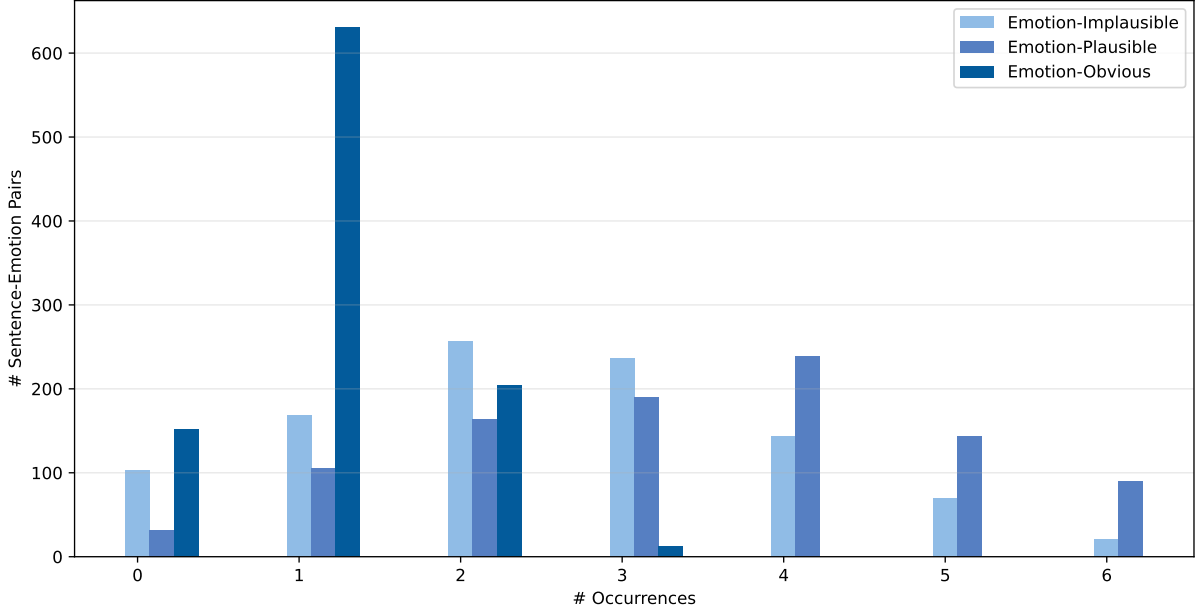


Figure 2: Context-Dependency Distribution

Model	Acc.	$F1_{micro}$	$F1_{macro}$	$F1_{weighted}$	Prec.	Rec.
GPT-5-mini	0.77	0.77	0.77	0.77	0.78	0.77
Gemini-2.5-flash	0.68	0.68	0.68	0.68	0.69	0.68
GroqLlama	0.59	0.59	0.53	0.54	0.64	0.59
MajorityVote	0.79	0.79	0.79	0.79	0.81	0.79
RoBERTa	0.52	0.52	0.27	0.40	0.34	0.52

Table 5: LLMs’ performance comparison against PoETIC for context-dependence classification

Sentence	Emotion	MajorityVote	Human
Where were they? I was trying to track them down and couldn't remember where I saw them.	NEUTRAL	EO	EI
I know you're joking, but there are people here either stupid or desperate enough to believe and perpetuate such idiocy.	SURPRISE	EO	EI
Would you like a friend? I sure could use one.	JOY	EI	EP
Not sure if guys like that are better or worse than regular Incels.	SURPRISE	EI	EP

Table 6: Illustrative examples contrasting model-predicted classes with human annotations.

Model	Setting	Acc.	$F1_w$	$F1_m$
RoBERTa	Supervised	0.52	0.40	0.27
Qwen3-8B	Zero-shot	0.5	0.35	0.18
Llama-3-8B	Zero-shot	0.47	0.35	0.22
Qwen3-8B	Few-shot	<b>0.52</b>	<b>0.51</b>	<b>0.36</b>
Llama-3-8B	Few-shot	0.35	0.19	0.18

Table 7: Model performances under zero-shot and few-shot settings.

setting (accuracy 0.35, macro F1 0.18), characterized by near-collapse to a single class prediction.

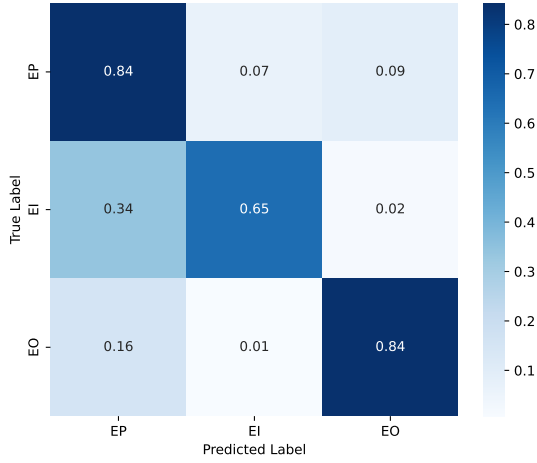
Overall, Qwen3-8B demonstrates robust adap-

tation with few-shot learning, achieving the best performance across all metrics, while Llama-3-8B shows inconsistent behavior, highlighting model-specific differences in prompt conditioning and generalization.

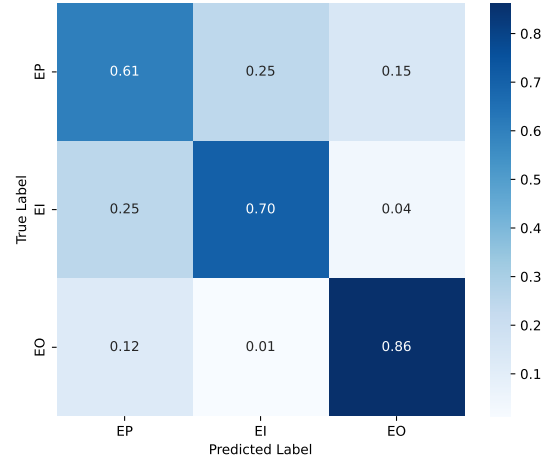
## 7. Discussions

Prior work, such as [Lecourt et al. \(2025\)](#), reports that large language models (LLMs) achieve relatively modest performance on the GoEmotions dataset. This suggests that directly assigning an utterance to a single emotion label—chosen from a large set—poses a substantial challenge, partic-

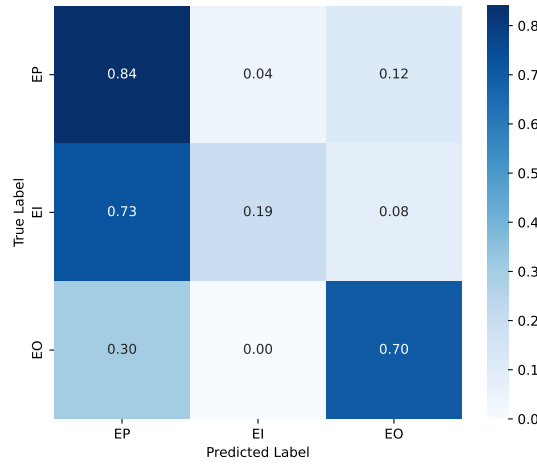




(a) GPT-5-mini



(b) Gemini-2.5-flash



(c) GroqLlama3-70B

Figure 3: Normalized confusion matrices for different models on PoETIC.

ularly when multiple emotions may be contextually plausible. We argue that the utterance-based emotion tagging formulation conflates emotion recognition with contextual ambiguity, making the task inherently difficult for both models and annotators.

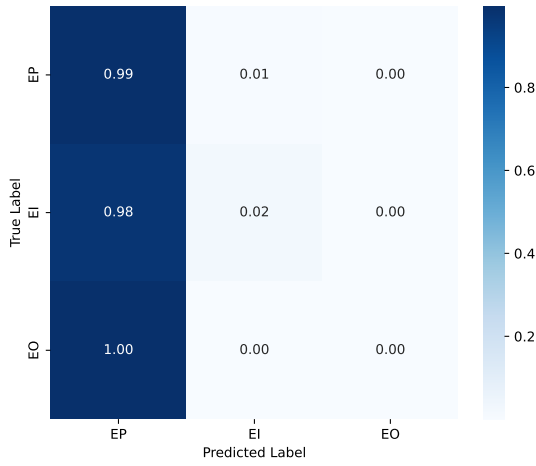
In contrast, we reformulate the problem in terms of contextual dependence. Rather than requiring the model to select the correct emotion in isolation, our task asks whether there exists any plausible context in which a given utterance could convey a specified emotion. This re-framing reduces the burden of exhaustive disambiguation and instead focuses on contextual plausibility. While this remains a non-trivial cognitive task, it is arguably more aligned with the generative and inferential strengths of LLMs.

Although we adopt discrete categories for contextual dependence in this work, the notion itself naturally admits a graded interpretation. Contextual dependence can be viewed as a continuum, ranging from utterances that require minimal con-

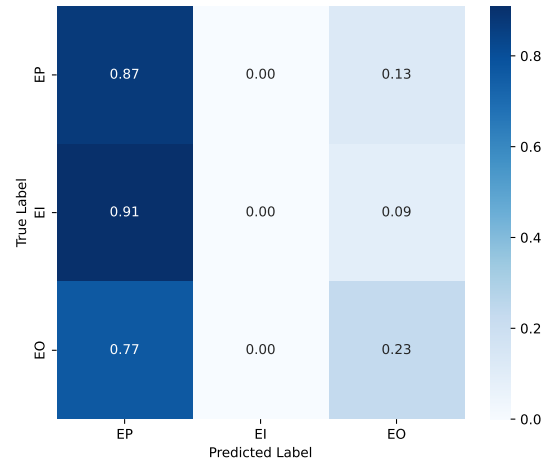
textual elaboration to express an emotion to those that demand highly imaginative or contrived contexts. Modeling contextual dependence as a rating or ordinal prediction task is, therefore, a promising direction for the future.

## 8. Limitations

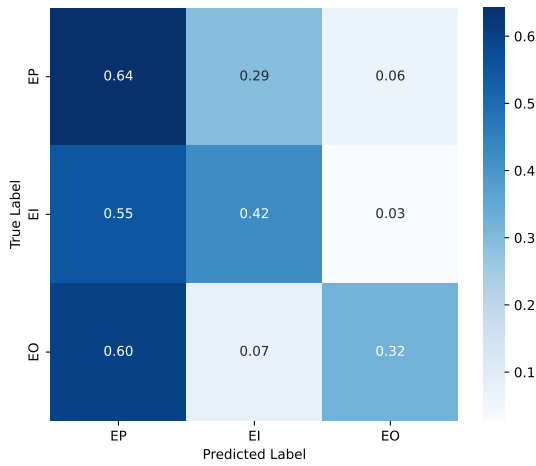
The dataset is derived from English Reddit data, inheriting platform-specific linguistic norms and demographic biases from GoEmotions. As a result, the observed patterns of contextual dependence may not generalize to other domains, languages, or interaction settings. Contextual dependence is modeled using three discrete categories. While this formulation improves annotation reliability, contextual dependence is inherently continuous, and many sentence–emotion pairs lie near category boundaries, as reflected in moderate inter-annotator agreement. More fine-grained or scalar annotations could better capture these nuances.



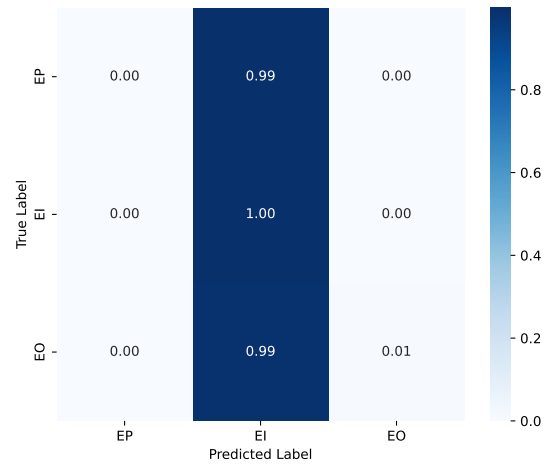
(a) Qwen3-8B (Zero-shot)



(b) Meta-Llama3-8B (Zero-shot)



(c) Qwen3-8B (Few-shot)



(d) Meta-Llama3-8B (Few-shot)

Figure 4: Normalized confusion matrices for zero-shot and few-shot settings on PoETIC.

Also, the annotations rely on imagined rather than explicit contexts. Judgments, therefore, depend on annotators' world knowledge and cultural assumptions, introducing subjectivity that may vary across annotator populations. Due to resource constraints, model evaluation is restricted to zero-shot and few-shot prompting. The strong performance of proprietary LLMs, while encouraging, raises questions about robustness and interpretability that suggest further investigation through cross-domain and adversarial evaluations. Overall, these limitations point toward future work on broader datasets, graded annotations, explicit context modeling, and more rigorous evaluation settings.

## 9. Use of AI Agents

AI Agents, primarily ChatGPT, have been used for grammar correction and paraphrasing in order to improve the language quality. The paraphrased outputs were thoroughly checked to ensure cor-

rectness. Copilot was used for code completion occasionally.

Anuj Attri, Arnav Attri, Suman Banerjee, Amey Patil, Muthusamy Chelliah, Nikesh Garera, and Pushpak Bhattacharyya. 2025. [LLMs as architects and critics for multi-source opinion summarization](#). In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 69–101, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Sven Büchel and Udo Hahn. 2017. Emobank: Studying the impact of annotation perspective and representation format on dimensional emo-

- tion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 578–585.
- Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeanette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.
- Ankush Chatterjee, Kedhar Nath Narahari, Meghana Joshi, and Puneet Agrawal. 2019. Semeval-2019 task 3: Emocontext contextual emotion detection in text. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 39–48.
- Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. [GoEmotions: A dataset of fine-grained emotions](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054, Online. Association for Computational Linguistics.
- Paul Ekman. 1992. [Are there basic emotions?](#) *Psychological Review*, 99(3):550–553.
- Roman Klinger, Orphée De Clercq, Saif Mohammad, and Alexandra Balahur. 2018. [IEST: WASSA-2018 implicit emotions shared task](#). In *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 31–42, Brussels, Belgium. Association for Computational Linguistics.
- Florian Lecourt, Madalina Croitoru, and Konstantin Todorov. 2025. ‘only chatgpt gets me’: An empirical analysis of gpt versus other large language models for emotion detection in text. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 2603–2611.
- Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2011. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE transactions on affective computing*, 3(1):5–17.
- Shamsuddeen Hassan Muhammad, Nedjma Ousidhoum, Idris Abdulmumin, Jan Philip Wahle, Terry Ruas, Meriem Beloucif, Christine de Kock, Nirmal Surange, Daniela Teodorescu, Ibrahim Said Ahmad, David Ifeoluwa Adelani, Alham Fikri Aji, Felermينو D. M. A. Ali, Ilseyar Alimova, Vladimir Araujo, Nikolay Babakov, Naomi Baes, Ana-Maria Bucur, Andiswa Bukula, Guanqun Cao, Rodrigo Tuñiño, Rendi Chevi, Chiamaka Ijeoma Chukwuneke, Alexandra Ciobotaru, Daryna Dementieva, Murja Sani Gadanya, Robert Geislinger, Bela Gipp, Oumaima Hourrane, Oana Ignat, Falalu Ibrahim Lawan, Rooweither Mabuya, Rahmad Mahendra, Vukosi Marivate, Alexander Panchenko, Andrew Piper, Charles Henrique Porto Ferreira, Vitaly Protasov, Samuel Rutunda, Manish Shrivastava, Aura Cristina Udrea, Lilian Diana Awuor Wanzare, Sophie Wu, Florian Valentin Wunderlich, Hanif Muhammad Zhafran, Tianhui Zhang, Yi Zhou, and Saif M. Mohammad. 2025. [BRIGHTER: BRIdging the gap in human-annotated textual emotion recognition datasets for 28 languages](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8895–8916, Vienna, Austria. Association for Computational Linguistics.
- Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. [MELD: A multimodal multi-party dataset for emotion recognition in conversations](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.
- Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.
- Smriti Singh, Cornelia Caragea, and Junyi Jessy Li. 2024. [Language models \(mostly\) do not consider emotion triggers when predicting emotion](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 603–614, Mexico City, Mexico. Association for Computational Linguistics.
- Daniel Yang, Aditya Kommineni, Mohammad Alshehri, Nilamadhab Mohanty, Vedant Modi, Jonathan Gratch, and Shrikanth Narayanan. 2023. Context unlocks emotions: Text-based emotion classification dataset auditing with large language models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE.

## A. Prompts

### GPT-5-mini.

You are an expert NLP evaluator.  
You must classify how dependent  
the text is on surrounding context  
to express the given emotion.

Allowed labels (SHORT FORM):

A = Emotion-Obvious  
B = Emotion-Plausible  
C = Emotion-Implausible

Label meanings:

A: The emotion is clearly expressed  
without needing additional context.  
B: The emotion is ambiguous or  
barely present, and typically needs  
outside context.  
C: Even with added context, this  
text cannot plausibly express the  
target emotion.

Special rule for NEUTRAL:

- Since "NEUTRAL" means complete  
absence of the 6 target emotions  
{ANGER, SADNESS, JOY, FEAR,  
DISGUST, SURPRISE},  
a sentence cannot be "A"  
(Emotion-Obvious) for both  
NEUTRAL and another emotion.

STRICT OUTPUT REQUIREMENTS:

- You must output ONLY a JSON  
object of the form:  
{"label": "A"} or {"label": "B"}  
or {"label": "C"}
- Do NOT output any text before or  
after the JSON.
- Do NOT explain your choice.
- Do NOT add any additional keys.
- If unsure, pick the best label; do  
NOT invent new labels.

-----  
Text: {text}  
Emotion: {emotion}

### Gemini 2.5-flash.

You are an expert NLP evaluator. You  
must classify how dependent the text  
is on surrounding context to express  
the given emotion.

Allowed labels:

Emotion-Obvious  
Emotion-Plausible  
Emotion-Implausible

Label meanings:

Emotion-Obvious: The emotion is  
clearly expressed without needing  
additional context.  
Emotion-Plausible: The emotion is  
ambiguous or barely present, and  
typically needs outside context.  
Emotion-Implausible: Even with added  
context, this text cannot plausibly  
express the target emotion.

Special rule for NEUTRAL:

- Since "NEUTRAL" means complete  
absence of the 6 target emotions  
{ANGER, SADNESS, JOY, FEAR,  
DISGUST, SURPRISE},  
a sentence cannot be  
"Emotion-Obvious" for both  
NEUTRAL and another emotion.

STRICT OUTPUT REQUIREMENTS:

- You must output EXACTLY ONE label
- Do NOT output any text before or  
after the label.
- Do NOT explain your choice.
- If unsure, pick the best label;  
do NOT invent new labels.

-----  
Text: "{sentence}"  
Emotion: "{emotion}"

Hyperparameters	Gemini2.5Flash	GPT-5-mini	Llama3-70B
temperature	0.0	1.0	0.1
top_p	1.0	1.0	1.0
top_k	40	-	-
max_tokens	8192	1800	3

Table 8: Model hyperparameter setting

### Groq (Llama3-70B).

Classify the dependency of the following sentence on its prior context to express the given emotion.

Allowed classes with definitions:

EO: The emotion is clearly expressed without needing additional context.

EP: The emotion is ambiguous or barely present, and typically needs outside context.

EI: Even with added context, this text cannot plausibly express the target emotion.

Reply ONLY with the class label. Do NOT include any explanation or additional text.

Example output: EO

-----

Sentence: {sentence}

Emotion: {emotion}

Class-label?



# From Sentiment to Valence in Metaphor: a Comparison of BERT-based Sentiment and Prompted Large Language Models

Rebecca Guolo, Ginevra Martinelli, Chiara Barattieri di San Pietro, Valentina Bambini

Laboratory of Neurolinguistics and Experimental Pragmatics (NEPLab)  
University School for Advanced Studies IUSS, Pavia, Italy  
rebecca.guolo@iusspavia.it, ginevra.martinelli@iusspavia.it,  
chiara.barattieridisani Pietro@iusspavia.it, valentina.bambini@iusspavia.it

## Abstract

Although the affective dimension is a key aspect of metaphor, computational studies of figurative language have largely overlooked psycholinguistic variables such as valence. This study investigates whether computational models can reliably estimate the affective aspects of Italian and German metaphors and whether metaphor valence is compositionally derived. Outputs of BERT-based sentiment analysis and a valence-prompted LLM were compared with human ratings. Results show that the former exhibit limited alignment with human judgments, whereas higher agreement is achieved when the explicit concept of valence is prompted in a LLM. Both humans and models rely on the combined valence of the individual lemmas, suggesting a compositional contribution to metaphor valence.

**Keywords:** Valence, Metaphor, BERT, GPT

## 1. Introduction

Metaphors represent a complex pragmatic phenomenon that involves a gap between the literal and the intended meaning, which is typically resolved via context-based inferences (Grice, 1975; Searle, 1979). Among the psycholinguistic features known to guide metaphor comprehension - e.g., familiarity, concreteness, and imageability - the affective dimension appears to play a particularly prominent role in deriving the meaning of metaphors and producing aesthetic non-propositional effects (Blasko and Brihl, 1997; Montefinese et al., 2025; Gerwien et al., 2024; Ifantidou, 2021). Affective properties are commonly conceptualized within the Valence-Arousal-Dominance (VAD) framework (Russell, 1980), which provides a fine-grained model for operationalizing affect estimation. Within this framework, valence, defined as the degree of pleasantness or unpleasantness evoked by a stimulus (Jurafsky and Martin, 2026), has been shown to interact with cognitive processes underlying metaphor comprehension (Mashal and Itkes, 2014). The availability of reliable affective ratings is therefore crucial to investigate the role of emotion in metaphor understanding. However, collecting such ratings through human annotation is both time- and resource-intensive. For other psycholinguistic features, such as concreteness and familiarity, recent research has demonstrated that computational approaches, particularly Large Language Models (LLMs), can effectively approximate human ratings (Mangiaterra et al., 2025). To our knowledge, such methods have not yet been systematically applied to generate valence ratings for metaphorical expressions. To address this gap, this study aims to evaluate computationally generated valence ratings

for metaphorical expressions using three computational approaches: (i) sentiment analysis systems (Jurafsky and Martin, 2026), which provide coarse-grained polarity judgments but may overlook subtle emotional nuances; (ii) LLM-based prompting, which elicits valence judgments by leveraging the capabilities of LLMs to process the language of emotion (Martínez et al., 2025; Conde et al., 2026) and figurative language (Fuoli et al., 2026); (iii) lexicon-based estimation, deriving valence scores from established word-level affective norms and aggregating them, allowing to test whether the valence of a metaphorical expression can be approximated from the valence of its constituent words (Mohammad, 2025).

## 2. Materials and Method

We collected a set of Italian and German metaphors from three different datasets: 1) the Figurative Archive dataset (Bressler et al., 2026), which includes 996 Italian metaphors differing in structure and semantic domains and divided into everyday (i.e., metaphors that occur in ordinary language, pooled from nine studies) and literary metaphors (i.e., sourced from poetry or prose). Specifically, metaphors were identified i) using keywords belonging to semantic classes known to be productive for metaphorical use, and; ii) searching for expressions in the “A of B” form; the resulting metaphors were then reviewed and problematic items removed. 2) the COMETA (Citron et al., 2020) and 3) the MIST (Müller et al., 2022) datasets, which collect respectively 60 conceptual German metaphors with a sentence-level structure, and 168 German metaphors involving human senses. While for the German datasets human affective ratings were al-

ready available, the Italian metaphors were rated ad-hoc for this study by 9 PhD candidates in linguistics, all Italian native speakers (mean age = 26,8 years, SD = 1,5; 7 women) on a -3 +3 range. The average ICC was 0.72, indicating moderate inter-rater agreement.

Ratings of valence were derived from two BERT-based models (one for Italian, and one for German) and one GPT model. Valence ratings for Italian metaphors were obtained using *feel-it-italian-emotion model* (Bianchi et al., 2021). Valence ratings for German metaphors were derived from *germansentiment* (Guhr et al., 2020). Valence ratings from GPT were obtained from GPT 4-o (OpenAI, 2024) via API. We chose this model considering its capacity to process contextualized and semantically complex expressions. The Italian and German prompts<sup>1</sup> instructed the model to assess metaphor valence on a seven-point scale (same span given to humans). No examples were inserted in the prompt, applying a zero-shot learning technique. Spearman correlation analysis was applied to test the relationship between human ratings and GPT- and BERT-based ratings. Finally, to assess whether metaphor valence is compositionally derived, we extracted the valence for each Italian metaphor's topic (i.e., the subject of the metaphor) and vehicle (i.e., the term used to convey the metaphorical meaning) from the Multilingual Emotion Lexicon (MEL - Buechel et al. (2020)) and correlated them with the metaphor valence ratings provided by humans and models. Additionally, we computed the average valence of the lemmas composing each metaphor (excluding stopwords) and likewise correlated them with metaphor valence ratings obtained from humans and models. To compare the magnitude of the correlations for compositionality, Fisher's z-transformation was applied following Pearson correlation analyses. This procedure could only be conducted for the Italian metaphors, as the German datasets lack topic and vehicle annotations, preventing the extraction of MEL-based values.

### 3. Results

**BERT models.** Results show a moderate correlation between Italian BERT-based ratings and

<sup>1</sup>English translation of the given prompt: "You will be given a metaphor and asked to evaluate it based on its emotional valence. Emotional valence refers to how positive or negative the emotions elicited by the metaphor are. Use a rating scale ranging from 1 ("very negative") to 7 ("very positive"). If the metaphor elicits a very negative emotion, such as sadness or anger, then assign a score of 1. If the metaphor elicits a very positive emotion, such as happiness or love, assign a score of 7. Respond with only a number from 1 to 7; do not add anything else. Metaphor: met"

human ratings ( $\rho=0.598$ ,  $p<0.001$ ), with different patterns between everyday and literary metaphors: while valence scoring for everyday metaphors show a moderate correlation ( $\rho=0.498$ ,  $p<0.001$ ), BERT-based and human ratings for literary metaphors exhibit a strong correlation ( $\rho=0.703$ ,  $p<0.001$ ). In contrast, we observe very weak and non significant correlations between human and German BERT-based ratings ( $\rho=0.158$ ,  $p=0.017$ ), both for the COMETA and for the MIST datasets (respectively,  $\rho=0.126$ ,  $p=0.337$ ; and  $\rho=0.146$ ,  $p=0.058$ ).

**GPT model.** Findings are homogeneous across languages, with very strong correlations between GPT-based and human ratings both for Italian and German ( $\rho=0.893$ ,  $p<0.001$ ;  $\rho=0.936$ ,  $p<0.001$ , respectively), and similar patterns for both everyday and literary metaphors in the Italian dataset ( $\rho=0.868$ ,  $p<0.001$  and  $\rho=0.929$ ,  $p<0.001$ , respectively) and for both COMETA and MIST German datasets ( $\rho=0.945$ ,  $p<0.001$  and  $\rho=0.937$ ,  $p<0.001$ , respectively).

**Compositionality.** Correlation analyses on human ratings (MEL) show moderate correlations between the valence of the topic and the overall metaphor valence ( $\rho=0.604$ ,  $p<0.001$ ), as well as between the valence of the vehicle and the overall metaphor valence ( $\rho=0.433$ ,  $p<0.001$ ). The mean valence of the topic and the vehicle composing the metaphor exhibits a strong correlation ( $\rho=0.681$ ,  $p<0.001$ ) with the overall metaphor valence. Valence ratings derived from the Italian BERT-based model show moderate correlations with topic, vehicle and their mean valence ( $\rho=0.480$ ,  $p<0.001$ ;  $\rho=0.419$ ,  $p<0.001$ ;  $\rho=0.592$ ,  $p<0.001$  respectively). Conversely, GPT-derived measures of valence show a strong correlation with the mean valence of topics and vehicle ( $\rho=0.736$ ,  $p<0.001$ ), while correlations between topic and metaphor valence ( $\rho=0.608$ ,  $p<0.001$ ), and between vehicle and metaphor valence ( $\rho=0.513$ ,  $p<0.001$ ) remain moderate.

**Correlations' magnitudes for compositionality.** The comparison between the mean valence (MEL) and GPT revealed a significant difference ( $z=-12.216$ ,  $p<0.001$ ), with GPT showing a higher correlation with human ratings ( $z=1.372$ ) than the mean valence ( $z=0.824$ ). Similarly, the comparison between the mean valence and BERT was significant ( $z=5.897$ ,  $p<0.001$ ), indicating a higher correlation for the mean valence ( $z=0.824$ ) compared to BERT ( $z=0.559$ ). Finally, GPT and BERT also differed significantly ( $z=18.114$ ,  $p<0.001$ ), with GPT showing a substantially higher correlation with human ratings ( $z=1.372$ ) than BERT ( $z=0.559$ ).

Dataset	BERT	GPT
Figurative Archive	0.598***	0.893***
Everyday Metaphors	0.498***	0.868***
Literary Metaphors	0.703***	0.929***
German Datasets	0.158*	0.936***
COMETA	0.126	0.945***
MIST	0.146	0.937***

Table 1: Correlation results between the BERT and GPT models for Italian and German ratings.

Category	Human	BERT	GPT
Topic	0.604***	0.480***	0.608***
Vehicle	0.433***	0.419***	0.513***
mLemmas	0.681***	0.592***	0.736***

Table 2: Correlation results with MEL data.

## 4. Discussion

While the validity of machine-generated single-word ratings has been widely established (Martínez et al., 2025; Conde et al., 2026), the present study aimed at assessing the validity of valence ratings for metaphorical expressions as derived from BERT and GPT, evaluating them against human ratings. Firstly, our results revealed that across languages (Italian and German), BERT-based sentiment ratings were weakly aligned with human ratings, whereas GPT derived ratings well approximated human judgment. In particular, the German BERT performed poorly, while the Italian BERT showed moderate correlations for everyday and strong for literary metaphors, likely reflecting the greater affective load of literary language. It cannot be however excluded that GPT-4o might have been exposed to parts of the datasets during training. This concern could apply primarily to the German datasets, as the Italian dataset was released after the model’s training cutoff. If this was the case, GPT performance in German is only slightly higher compared to Italian. So, if prior exposure were the primary driver, a substantially larger gap would be expected. Instead, the similar performance across German and Italian suggests that the results are better explained by generalizable inferential processes rather than memorization. Secondly, our results on the mechanisms underlying metaphor valence indicate that, for both humans and models,

Criterion: Human			
Comparison	z1	z2	z observed
mLemmas-GPT	0.824	1.372	-12.216***
mLemmas-BERT	0.824	0.559	5.897***
GPT-BERT	1.372	0.559	18.114***

Table 3: Correlations’ magnitudes for compositionality.

the mean valence of the constituent words largely guides the overall evaluation. For GPT, the correlation with the mean lemma valence exceeds 0.70, suggesting a strong reliance on the valence of individual words. For both humans and models, the topic plays a central role compared to the vehicle, anchoring the metaphor’s affective value. Overall, both BERT and GPT are better in capturing the evaluation of a metaphor’s human valence than a compositionality model based on the average valence value of the words that compose the metaphor. More specifically, BERT-based sentiment analyzers do not reliably approximate human valence ratings for metaphors, whereas prompted LLMs closely align with human-like judgments.

## 5. Limitations

This work is a preliminary study and several limitations are worth noting. First, having few annotators might limit the generalizability of the findings, and future research should both increase their number and their heterogeneity in terms of sociodemographic features, also exploiting online surveys.

Second, the use of existing models and GPT-4o for valence prediction provides limited technical novelty. To address this, future studies could explore additional models, including open-source alternatives such as Llama, in order to obtain more robust and reliable results.

Third, we acknowledge the imbalance in the number of metaphors across the two languages, which should be accounted for in the statistical modeling.

Finally, the human Italian ratings achieved an ICC < 0.75, that is considered only moderate. From another perspective, however, this difference in annotation reliability from Italian and German datasets may also help explain the pattern observed with BERT. In particular, BERT shows better performance on the Italian data, where ratings are noisier, and lower performance on the German data, where ratings are more consistent. Instead, GPT performances are significant in both languages, suggesting that its predictions are less affected by annotation variability. This pattern indicates that GPT may capture subtle semantic properties of metaphors in a way that is closely aligned with human judgments, even when those judgments are based on a relatively small number of annotators.

## 6. Ethics

The present study aims to use computational tools to approximate human valence judgments in metaphorical language. However, we are aware that the used datasets do not cover all possible syntactic structures and semantic domains of metaphors and most importantly, they may reflect

linguistic and cultural bias. Therefore, human participants remain the gold standard for affective evaluation, while computational approaches must be considered as complementary tools rather than complete replacements.

## 7. Bibliographical References

- F. Bianchi, D. Nozza, and D. Hovy. 2021. Feel-it: Emotion and sentiment classification for the Italian language. In *Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics.
- D. G. Blasko and D.S. Brihl. 1997. [Reading and recall of metaphorical sentences: Effects of familiarity and context](#). *Metaphor and Symbol*, 12(4):261–285.
- S. Buechel, S. Rucker, and U. Hahn. 2020. [Learning and evaluating emotion lexicons for 91 languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1202–1217, Online. Association for Computational Linguistics.
- J. Conde, G. Martínez, M. Grandury, C. Arriaga, J. Haro, S. Schroeder, F. Hintz, P. Reviriego, and M. Brysbaert. 2026. [Updating the german psycholinguistic word toolbox with ai-generated estimates of concreteness, valence, arousal, age of acquisition, and familiarity](#). *Journal of cognition*, 9(1).
- M. Fuoli, W. Huang, J. Littlemore, S. Turner, and E. Wilding. 2026. [Metaphor identification using large language models: A comparison of rag, prompt engineering, and fine-tuning](#).
- J. Gerwien, M. Filip, and F. Smolík. 2024. [Noun imageability and the processing of sensory-based information](#). *Quarterly journal of experimental psychology*, 77(10):2137–2150.
- H. P. Grice. 1975. Logic and conversation. In *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press.
- O. Guhr, A. K. Schumann, F. Bahrmann, and H. J. Böhme. 2020. [Training a broad-coverage german sentiment classification model for dialog systems](#). In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1620–1625, Marseille, France. European Language Resources Association.
- E. Ifantidou. 2021. [Non-propositional effects in verbal communication: The case of metaphor](#). *Journal of Pragmatics*, 181:6–16.
- D. Jurafsky and J. H. Martin. 2026. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*, 3rd edition. Online manuscript released January 6, 2026.
- V. Mangiaterra, H. Al-Azary, C. Barattieri di San Pietro, P. Canal, and V. Bambini. 2025. Can gpt replace human raters? validity and reliability of machine-generated norms for metaphors. Preprint: 2512.12444.
- G. Martínez, J. Conde, P. Reviriego, and M. Brysbaert. 2025. [Ai-generated estimates of familiarity, concreteness, valence, and arousal for over 100,000 spanish words](#). *Quarterly journal of experimental psychology*, 78(10):2272–2283.
- N. Mashal and O. Itkes. 2014. [The effects of emotional valence on hemispheric processing of metaphoric word pairs](#). *Laterality*, 19(5):511–521.
- S. M. Mohammad. 2025. [Breaking bad: Norms for valence, arousal, and dominance for over 10k english multiword expressions](#). *arXiv preprint arXiv:2511.19816*.
- M. Montefinese, A. Visalli, A. Angrilli, and E. Ambrosini. 2025. [Fine-grained concreteness effects on word processing and representation across three tasks: An erp study](#). *Psychophysiology*, 62(5).
- OpenAI. 2024. [Gpt-4o system card](#). *arXiv preprint arXiv:2410.21276*.
- J. Russell. 1980. [A circumplex model of affect](#). *Journal of Personality and Social Psychology*, 39:1161–1178.
- JR. Searle. 1979. *Expression and Meaning*. Cambridge: Cambridge University Press.

## 8. Language Resource References

- M. Bressler and V. Mangiaterra and P. Canal and F. Frau and F. Luciani and B. Scalingi and C. Barattieri di San Pietro and C. Battaglini, C. Pompei and F. Romeo and L. Bischetti and V. Bambini. 2026. [Figurative Archive: an open dataset and web-based application for the study of metaphor](#). Scientific Data, 151.
- F.M.M. Citron and M. Lee and N. Michaelis. 2020. [Affective and psycholinguistic norms for German conceptual metaphors \(COMETA\)](#). *Behavior Research Methods*, 52(3).



N. Müller and A. Nagels and C. Kauschke. 2022.  
*Metaphorical expressions originating from human senses: Psycholinguistic and affective norms for German metaphors for internal state terms (MIST database)*. Behaviour Research Methods, 54.



# Affect, Body, Cognition, Demographics, and Emotion: The ABCDE of Text Features for Computational Affective Science

Jan Philip Wahle<sup>\*†</sup>, Krishnapriya Vishnubhotla<sup>\*‡</sup>, Bela Gipp<sup>†</sup>, Saif M. Mohammad<sup>‡</sup>

<sup>†</sup>University of Göttingen, Germany; <sup>‡</sup>National Research Council Canada  
wahle@uni-goettingen.de; vkpriya@cs.toronto.edu, saif.mohammad@nrc-cnrc.gc.ca

<sup>\*</sup>equal contribution

Dataset [hf.co/datasets/jpwahle/abcde](https://huggingface.co/datasets/jpwahle/abcde)  
Code [github.com/jpwahle/abcde](https://github.com/jpwahle/abcde)

## Abstract

Work in Computational Affective Science and Computational Social Science explores a wide variety of research questions about people, emotions, behavior, and health. Often they make use of language data that is first labeled with relevant information such as the use of emotion words and age of the speaker. Even though many resources and algorithms exist to enable such labeling, finding and using them is still a substantial impediment, especially to practitioners in fields outside of computer science. Here, we present the ABCDE dataset (“Affect, Body, Cognition, Demographics, and Emotion”), a large-scale collection of over 400 million released text instances from social media, blogs, books, and AI-generated sources, annotated for a number of features relevant to computational affective and social science. ABCDE facilitates inter-disciplinary research in wide range of fields, including affective science, cognitive science, the digital humanities, sociology, political science, and computational linguistics.

**Keywords:** computational social science, computational affective science, scientometrics

## 1. Introduction

Language is a rich medium of self-expression and communication. It is a product of historical, cultural, embodied, and cognitive processes. Consequently, computational approaches to the analysis of large text datasets have become a powerful tool for studying feelings, thought, body/health, and behavior of individuals and populations.

In the domain of Computational Affective Science (CAS), research has shown that language use is strongly associated with, and indicative of, mental states and cognitive processes in the human brain (Guntuku et al., 2017). When we are sad, we express ourselves using words associated with low valence (negative sentiment), or even subtle linguistic signals such as shorter, more terse utterances. When we lack confidence, we tend to use more qualifiers such as “I think”, “probably”, and “right?”. And so on. Linguistic properties of utterances are thus immensely useful signals of our mental models and processes, as well as how we as humans perceive ourselves and the world around us. At the same time, language is only an indirect trace of the underlying states we care about. Many mental states and cognitive processes are latent and must be inferred from behavior and context. Thus while it is difficult to determine the exact mental state of an individual from an utterance (something we do not recommend pursuing automatically anyway), the power of language analysis lies in bringing forth trends and patterns of mental states of whole pop-

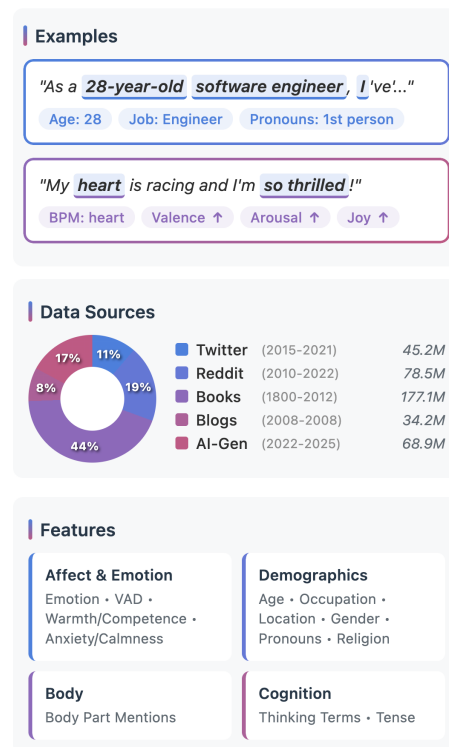


Figure 1: Overview of the ABCDE dataset.

ulations (aggregate-level analysis). Language use remains a natural and scalable behavioral trace that can be used to infer patterns of relative differences and change at an aggregate level. For example, how has the use of emotion words by a population

changed with time and what psychological changes that indicates; how a cohort of people uses warmth and competence words when posting about different social groups; what moral frames are invoked when talking about controversial issues (such as immigration), etc.

Similarly, in the broader domain of social sciences, insights derived from textual data can complement those derived from qualitative methodologies like surveys and self-reports. Computational methods have the added advantage of enabling research using a much larger and more diverse pool of data when compared to lab studies with smaller groups of participants, thereby increasing the reliability of measurements. For example, Computational Social Science (CSS) researchers have used social media data to study mechanisms of influence and information diffusion in social networks (Goel et al., 2016), monitor public sentiment and stance towards specific entities and topics (Dodds et al., 2011), identify changes in the contextualized semantic meanings of word types in different communities (Lucy and Bamman, 2021), etc.

Another line of research that is of increasing interest to linguists, cognitive scientists, and computer scientists alike is the studying of AI-generated text. Machine learning based training of multi-billion parameter neural networks has resulted in Large Language Models that interface with and communicate in natural language with surprising fidelity (Achiam et al., 2023). Consequently, researchers are looking at the outputs of these models to understand statistical associations and biases that are encoded in them, particularly as LLMs take on a larger role in the data science pipeline as de-facto models for classification, labeling, and synthetic data generation (Ziems et al., 2024; Farrell et al., 2025). Researchers are also analyzing the "reasoning traces" or chains-of-thought (CoTs) of these models to better understand the artificial "cognitive" reasoning processes in these models, and how they compare with human processes (Korbak et al., 2025).

All of this research that makes use of "language as data" relies on three key tasks as a pre-requisite to subsequent statistical analyses: collecting the textual data, tagging it with associated metadata (extra-linguistic context), and the measurement of features of interest (quantifying, for example, sentiment from tweets). The dynamic nature of pipelines for data collection on the web, as well as the rapid pace of development of computational models of measurement for various constructs that are of interest in the social sciences, make this a technically challenging process for most researchers, especially those outside of computer science. Further, even for those comfortable with natural language processing, compiling and annotating a large number of diverse datasets and features is time and

labor intensive.

In this work, we compile and release **ABCDE**, a collection of text datasets automatically tagged with metadata associated with speaker demographics, and labeled for 136 lexical features that are of broad utility for researchers in the affective and social sciences. The selected datasets (aggregated from various primary sources) span multiple domains, genres, and time-periods. We group our text features into five thematic sets: **Affect**, **Body**, **Cognition**, **Demographics**, and **Emotion**.

Our primary motivation in creating this dataset is to unify textual features that encode aspects of behavior and cognition that are of interest to researchers in the affective sciences. While emotion and affect have been prominent topics of study here, recent advances in embodied cognition emphasize how bodily awareness and interactions with the environment shape many of these cognitive functions, and have downstream correlations with mental, emotional, and physical health (Wu et al., 2025). We therefore include body part mentions as a core feature of our dataset, allowing researchers to quantify patterns at the intersection of bodily mentions and emotions/affect. With Demographic information, we also consider features that encode *social context*, in addition to linguistic and lexical features of the text itself. This is a crucial variable for most research in the affective and social sciences, where concepts like emotion and cognition are not constant, stable traits, but demonstrate significant variation with demographic features like age, within and across population sub-groups — in fact, it is this variation that is often the phenomenon of interest (Hoemann et al., 2025; Gutches and Rajaram, 2023). Age-related differences have been reported in the arousal structure of emotion concepts (Trnka et al., 2022), and age and culture can jointly shift cognitive performance and style (Na et al., 2017).

We make our resource publicly available for open access and use by researchers. We hope the **ABCDE** dataset will facilitate inter-disciplinary research in wide range of fields, including affective science, cognitive science, the digital humanities, sociology, political science, and NLP.

## 2. Related Work

The past several decades of research in computational linguistics and natural language processing has led to the development of tools that can be used to extract rich linguistic information from textual data, such as StanfordCoreNLP (Manning et al., 2014), NLTK (Bird et al., 2009), and spaCy<sup>1</sup>. However, these are aimed largely at researchers

---

<sup>1</sup><https://spacy.io/>

with familiarity and expertise in these domains of research, with rich programming experience, and target features that are useful for linguistic analysis, like part-of-speech tags and dependency relations. In the social sciences, by contrast, researchers are generally interested in quantifying more abstract, higher-level features of text, like sentiment or emotion intensity, political leaning, satire, literary quality, etc. (Licht et al., 2025)

There exists a small but growing set of accessible text analysis toolkits that have a narrower focus on specific domains of research. The Cornell Conversational Analysis Toolkit (Chang et al., 2020) is a collection of datasets and computational scripts that enable research into aspects of social networks, conversational dynamics, and the sociolinguistics of online interactions. GutenTag is an NLP-driven tool for Digital Humanities research in particular, with a web-based interface that automatically extracts several features of interest from literary texts (Brooke et al., 2015). In the Affective Sciences domain, packages like VADER (Hutto and Gilbert, 2014) and Textblob<sup>2</sup> are widely used for sentiment analysis. The Emotion Dynamics toolkit (Vishnubhotla and Mohammad, 2022) provides scripts to quantify patterns of change in emotional expression in textual utterances over time.<sup>3</sup>

A prime example of a text processing toolkit that has seen broader use in the social, affective, and cognitive sciences, and the digital humanities, is LIWC (Linguistic Inquiry and Word Count), a proprietary software that quantifies several psychometric properties of words to facilitate studying the links between language, cognition, and psychology (Boyd et al., 2022).

We position **ABCDE** as not just a feature extraction tool, but a much more comprehensive repository of data mapped to metadata and pre-computed features. We compile, clean, and annotate multiple text datasets from diverse domains that are of broad interest to affective and cognitive science researchers, including social media, blogs, and AI-generated text. By pre-computing features for these datasets, our resource lowers the technical barrier of entry for many researchers, and also encourages reproducibility of empirical research in the field. The **ABCDE** dataset therefore functions as a much more thorough and standardized starting point, on top of which researchers can directly apply statistical methods of analysis in order to answer various research questions.

### 3. The **ABCDE** Dataset

We compiled a large-scale, longitudinal collection of textual datasets from various primary sources,

spanning data from social media platforms, books, and blogs. Additionally, we compiled a selection of AI-generated text, including human-LLM conversational data, LLM reasoning traces, preference datasets, and datasets from AI-generated text detection tasks. We then computed and recorded the text features enumerated in Section 3.2 for every text *instance* in each of these datasets. Note that a text instance can be defined at different levels of granularity – features can be recorded at the sentence-level (with each sentence in turn linked to the original Reddit post, blog post, AI-generated story, etc.), or for equal-sized chunks of tokens from each dataset, among others. Our definition of what constitutes a text instance differs from one dataset to another, and is enumerated in Section 3.1 alongside the dataset descriptions. In Section 3.2, we expand on our annotated features and the methods used to quantify them given a text instance.

#### 3.1. Data Sources

We annotated datasets spanning millions of records from Twitter, Reddit, blogs, books, and AI-generated content:

**Twitter (TUSC, 2015–2021):** 45.2M geolocated tweets sourced from the TUSC dataset (Vishnubhotla and Mohammad, 2022). Each tweet is considered an instance.

**Reddit (Pushshift, 2010–2022):** 78.6M posts crawled from Reddit via Pushshift archives hosted by the Internet Archive (Baumgartner et al., 2020). We remove posts that are marked as adult (over\_18), promoted, containing images or videos, and having fewer than 5 or more than 1,000 words; each post constitutes a text instance.

**Books (Google, 1800–2012):** 177.1M 5-gram occurrences (1.74M unique 5-grams) from the English Fiction subset of the Google Books Ngram Corpus (version 20120701) (Google Inc., 2012). Each 5-gram is treated as an instance. While these instances are short, they remain useful for large-scale frequency and longitudinal analyses.

**Blogs (Spinn3r, 2008):** 34.2M personal blog entries from the ICWSM 2009 conference (Burton and Soboroff, 2009). Each blog post is an instance.

**AI-Generated Texts (Various, 2022–2025):** 68.9M AI-completions from 15 distinct datasets, including conversational texts (WildChat-1M (Zhao et al., 2024), LMSYS-Chat-1M (Zheng et al., 2024), PIPPA (Gosling et al., 2023), HH-RLHF (Bai et al., 2022), Prism (Kirk et al., 2024), and APT (Wahle et al., 2022)); persuasive essays (Anthropic-Persuasiveness (Durmus et al., 2024)); AI text detection datasets (M4 (Wang et al., 2024), MAGE (Li et al., 2024), LUAR (Soto et al., 2024)); reasoning traces (General Thoughts 430k (Reasoning, 2024), Reasoning Shield (Li et al., 2025), SafeChain (Jiang et al., 2025), STAR-1 (Wang et al.,

<sup>2</sup><https://textblob.readthedocs.io/>

<sup>3</sup><https://github.com/Priya22/EmotionDynamics>



2025)); and narratives (TinyStories (Eldan and Li, 2023)). For conversational datasets, we define the LLM response from a single user–chatbot interaction turn as a text instance (multi-turn conversations are split into multiple instances). Preference datasets generally comprise of a pair of LLM generations (in response to a user prompt) along with an indicator of the preferred generation; we take each generation separately to be an instance of AI-generated text. The AI-generated text detection datasets consist of LLM generations of varying lengths, intended to simulate human text in specific domains like news articles and social media posts; each generation is considered an instance. For reasoning traces, we disregard the final model output and only use the chain-of-thought text for each query as the instance. Instances are multi-feature. One text can activate several feature families, and some short texts may not activate any feature in a given family.

### 3.2. Annotated Features

We extract linguistic features through a set of lexicons, word lists, and regular expressions, organized along five dimensions relevant to computational affective science: Affect, Emotion, Body, Cognition, and Demographics. Word lists and lexicons as a measurement method offer several advantages, particularly in inter-disciplinary studies: ease-of-use, generalization power, reliability, adaptability, and interpretability, among others. Variation in word choices across populations and data subsets is also intertwined with multiple cognitive and social processes relating to language use (such as power dynamics, cultural and demographic factors, medium of communication, etc.). Appendix A.1 summarizes the released resources, and Appendix A.2 reports per-source text statistics.

Affect and Emotion features are characterized using word association lexicons; Body and Cognition features through a curated list of terms, compiled from multiple sources, that indicate a strong association with body and cognition-related activities; and Demographic features through hand-constructed regular expressions and heuristic rules. For the demographic feature "Occupation", for example, we use a regex to parse self-disclosure statements of the type "I am/work at/employed as [occupation]." from a user, and apply the extracted term to all text instances from that user. The demographic feature "Age", which refers to the age of the user at the time of posting, is computed by first using a similar regex for age or date-of-birth, and then applying a heuristic rule that combines this with the timestamp of the post.

For each feature dimension (say, valence), and each text instance (say, a tweet), we use the associated word-level lexicon to compute an aggregate

instance-level score in multiple ways: the **average** valence score of the constituent terms, a binary **flag** indicating if a word from the valence lexicon is present in the instance, and the **count** of the number of lexicon terms present in the instance. The appropriateness of a particular aggregate measure for a feature will depend on the specifics of the research question being answered, and we leave this decision to the users. For example, length-normalized count features are more appropriate as a comparative indicator for texts of different lengths, rather than the average intensity score.

**Affect:** Affect refers to the fundamental neural processes that broadly determine and regulate internal experiences of emotion, mood, and feelings. Affective states are generally characterized along three principal dimensions: valence (scale of positive–negative), arousal (scale of active–passive), and dominance (scale of competent–incompetent / powerful–weak), together referred to as VAD, which form our feature dimensions for Affect. We match words against the NRC VAD lexicon (Mohammad, 2018a), which maps words to a real-valued intensity score between 0 (lowest) and 1 (highest). We further define 'High' and 'Low' VAD features, where only lexicon entries with scores  $\geq 0.66$  and  $\leq 0.33$  respectively are considered.

**Emotion:** We compute average, count, and binary presence flag features for multiple discrete emotions: the eight basic emotions of anger, anticipation, disgust, fear, joy, sadness, surprise, and trust, computed using the NRC Emotion Intensity lexicon (Mohammad, 2018b); warmth, competence, sociability, and trust using the WCST Lexicon (Mohammad, 2025a); and anxiety and calmness with the NRC WorryWords lexicon (Mohammad, 2024).

**Body:** We identify Body Part Mentions (BPMs) by matching text tokens against curated lists of 292 anatomical unigrams, bigrams, and trigrams from Zhuang et al. (2024); Wu et al. (2025), capturing mentions with possessive pronouns (e.g., "my heart", "her hand"). Features are recorded as binary flags for the presence of a body part mention, and for each possessive pronoun, as a list of BPMs used with that pronoun (i.e., MyBPM is a list containing BPMs used with the pronoun 'my'). Appendix A.3 lists the body part words used.

**Cognition:** We classify cognitive processes by identifying "thinking words" derived from categories outlined in Bloom's Taxonomy and Queensland's Glossary of Cognitive Verbs.<sup>4,5</sup> We categorize 98 unigrams into 11 categories (e.g., analyzing, learning, decision-making). A binary flag feature indicating presence is recorded for each category. Ap-

<sup>4</sup><https://adp.uni.edu/documents/bloomverbscognitiveaffectivepsychomotor.pdf>

<sup>5</sup>[https://www.qcaa.qld.edu.au/downloads/senior-qce/common/snr\\_glossary\\_cognitive\\_verbs.pdf](https://www.qcaa.qld.edu.au/downloads/senior-qce/common/snr_glossary_cognitive_verbs.pdf)

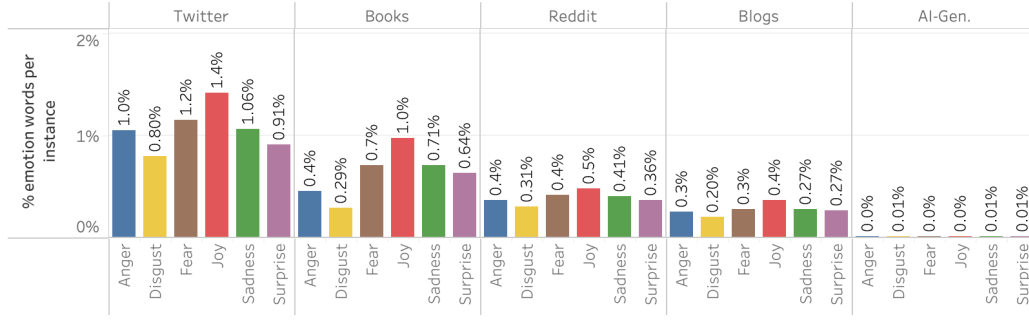


Figure 2: **Q1. Emotion:** Percent of words per instance from the six Ekman emotion categories.

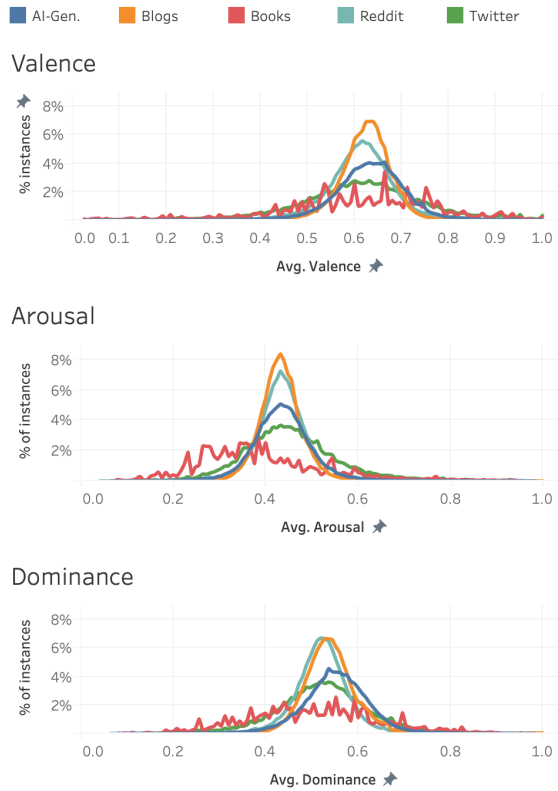


Figure 3: **Q1. Affect:** Distributions of average Valence, Arousal, and Dominance (VAD) scores of all words per instance, by source.

pendix A.4 lists the full grouped inventory.

**Demographics:** We extract demographic attributes, including age, occupation, gender, country, city, and religion, using regular expression matching and structured dictionary lookups. We map occupations according to the U.S. Bureau of Labor Statistics Standard Occupational Classification (SOC) system, and match gender, country, city, and religion references against controlled vocabularies from Wikipedia and Geonames datasets.<sup>6</sup> These

<sup>6</sup>[https://en.wikipedia.org/wiki/List\\_of\\_gender\\_identities](https://en.wikipedia.org/wiki/List_of_gender_identities), [List\\_of\\_religions\\_and\\_spiritual\\_traditions](https://en.wikipedia.org/wiki/List_of_religions_and_spiritual_traditions), [List\\_of\\_countries](https://en.wikipedia.org/wiki/List_of_countries)

regexes are not exhaustive but target high-precision first-person self-disclosure patterns and public controlled vocabularies. Appendix A.5 shows the full list of patterns used.

**Focus features:** We also identify pronouns (possessive and non-possessive) and verb tenses (past, present, future) through direct lexical matching and morphological tagging using the English UniMorph dataset.<sup>7</sup> This enables quantifying linguistic usage patterns related to personal references and temporal framing.

### 3.3. Lexical vs. ML Features

The measurement of a construct such as "anger" or "cognitive activity" given a text can be operationalized as a feature in many ways, and is dictated by the requirements of downstream use-case(s). If the goal is to predict the level of anger expressed in a single utterance by a person in real-time, trained machine learning models or pre-trained large language models will be more accurate than word-level lexicon aggregates. If, on the other hand, our goal is to measure the changes in anger levels expressed on Twitter in the USA over the last decade, plotting the (normalized) density of usage of anger-associated words in tweets from each year is a valid measurement method.

In other words, while lexicon-based methods are not as accurate as neural models at instance-level estimation (i.e, for estimating the sentiment of a particular sentence), they are comparable and sufficient when used as tools of aggregate-level analysis. This means that relative patterns of change, for comparing emotional intensity across different sources, or quantifying the patterns of changes in sentiment over a temporal period (longitudinal analysis), can be estimated with high accuracy using lexicon-based methods. The source resources also

<sup>7</sup><https://github.com/unimorph/eng>



provide validation evidence for aggregate use (Mohammad, 2018a; Mohammad and Turney, 2013; Zhuang et al., 2024).

In Teodorescu and Mohammad (2023), the authors empirically demonstrate that emotion arcs (i.e, temporal fluctuations of valence intensity) with word-level lexicons of valence highly correlate with the ground-truth arcs (with correlation scores  $> 0.9$ ) for social media texts, provided certain hyperparameters (like the width of the temporal window) are appropriately set. In the domain of digital humanities, Öhman et al. (2024) measure the agreement of lexicon-generated arcs for classic fiction novels with human annotation, and find high consensus. In their work, the NRC emotion intensity lexicon was customized to the literary domain with a word-similarity measure, which is one of the recommended practices for the use of emotion lexicons (Mohammad, 2023).

## 4. Key Research Questions

We use **ABCDE** to shed light on nine research questions, grouped by feature family.

**Q1. Affect–Emotion (AE):** *To what extent are affect- and emotion-associated words used in different media?*

**Motivation.** Affect and emotion lexicons provide aggregate signals about how people express emotion and affect across domains. These are central to population-level work in affective and social science to trace well-being, emotional contagion, or cultural mood shifts. For NLP, these signals offer interpretable, domain-portable emotion features for sentiment modeling and model alignment.

**Results.** Figure 2 shows the length-normalized counts (i.e, instance-level word density) of lexicon terms for six core emotion categories. Observe that Twitter has the highest emotion word usage rates (e.g.,  $\sim 1.4\%$  joy tokens per instance), followed by Books and Reddit. AI-generated text uses explicit emotion words rarely ( $\approx 0.01\%$ ). Figure 3 shows the distributions of average intensity scores for Valence, Arousal, and Dominance (VAD): human-authored sources center around neutral-to-positive valence (roughly 0.60–0.65), moderate dominance, and mid-to-low arousal. Books and AI-generated text display slightly lower arousal.

**Q2. Body (B):** *How often do people refer to body parts in text? Does it vary in different media?*

**Motivation.** Body parts mentioned (BPMs) can reveal how people linguistically connect to their bodies and thus how embodied experience manifests in text. In CSS, BPMs allow measurement of health, stress, or embodied metaphors in everyday discourse. In NLP, they can serve as features for

Books	0.29%
Twitter	1.19%
AI-Gen.	1.02%
Blogs	4.47%
Reddit	8.16%

Figure 4: **Q2. Body:** Percent of instances with a positive flag for possessive body part mentions (e.g., *my head/heart/etc.*) by source.

affective and health-related modeling.

**Results.** Figure 4 shows that Reddit (8.16%) and Blogs (4.47%) contain the most possessive BPMs (references to ‘*my <BPM>*’); Books and AI-generated text contain very few. Figure 5 highlights the most frequent first-person BPMs: *my head* is most common (especially on Reddit with 0.91%), while *my heart* appears more in Blogs (0.55%). Mentions like *my body*, *my face*, and *my hands* vary systematically with domain.

**Q3. Cognition (D):** *How prevalent are cognition (“thinking”) words across domains, and which sub-processes are most/least common?*

**Motivation.** Cognition vocabulary (understanding, remembering, deciding, etc.) offers a linguistic window into reasoning and mental-state discourse. For CSS, this enables cross-platform comparison of analytical concepts (and one can draw relations to emotional language). For NLP, such features open a window into the interpretability of models of deliberation, argumentation, and education.

**Results.** Figure 6 shows that human-authored sources use these terms far more than AI-generated text or Books. Blogs and Reddit lead overall: words of *understanding* (5.68%/5.57%), *general cognition* (3.94%/3.81%), and *memory recall* (3.55%/3.31%). Twitter shows the same ranking of categories, but at lower overall rates (e.g., *understanding*  $\sim 1.09\%$ ). AI-generated text under-expresses cognition terms across categories (e.g., *problem solving*  $\sim 0.09\%$ , *explanation*  $\sim 0.20\%$ ). Across corpora, *analyzing* is the least frequent cognition term in our set.

**Q4. Demographics-Age (D-Age):** *What are the self-disclosed age distributions in different media?*

**Motivation.** Age is a key demographic in CSS for understanding generational differences in discourse, politics, or emotion, and in NLP for bias control and stratified sampling. Self-disclosed age data helps characterize who participates in online discourse and calibrate downstream demographic analyses.

**Results.** Figure 7 shows that late teens to late 20s are most common in Reddit, whereas the 30s are most common in Twitter. Tails decline steadily there-

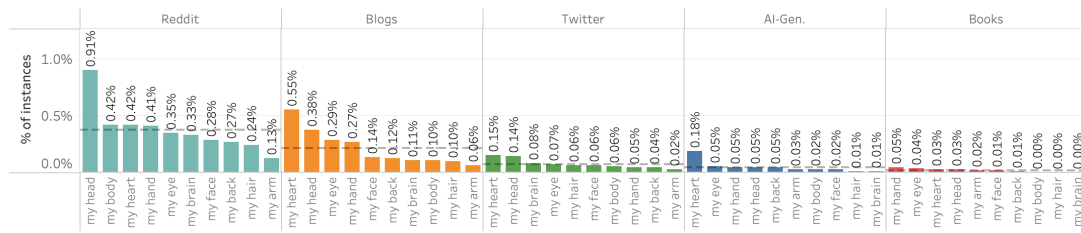


Figure 5: **Q2. Body:** % of possessive BPMs for the top ten most common BPMs, across sources. Dashed lines represent the average occurrence of BPMs per dataset.

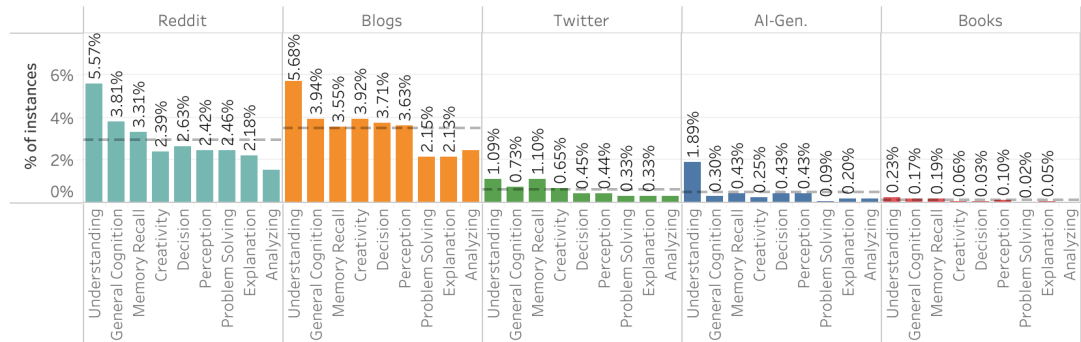


Figure 6: **Q3. Cognition:** % of instances with a positive flag for cognitive terms from different categories. Dashed lines represent the average occurrence of cognitive terms per dataset.

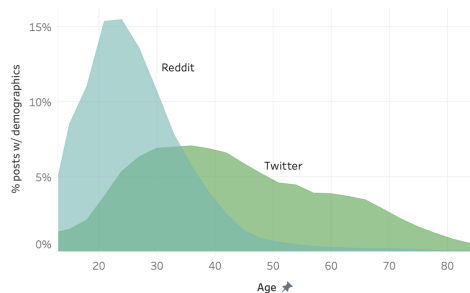


Figure 7: **Q4. Demographics:** Age distribution of posts with self-disclosed age.

after. This suggests that Reddit users are markedly younger than Twitter users.

#### Q5. D-Occupation: Which occupations are most frequently self-disclosed?

**Motivation.** In most modern societies, occupations are key parts of people's identity, values, and social interaction. In CSS, they enable studies of professional discourse, social stratification, and belief-linked behavior. In NLP, such self-labels offer interpretable subgroup features for fairness and personalization.

**Results.** Figure 8 shows Reddit users often self-identify as creative or technical professionals (e.g., musicians, software engineers), whereas Twitter skews toward public-facing roles (e.g., journalists, authors, executives).

#### Q6. D-Gender: Which genders do users most of-

#### ten self-disclose?

**Motivation.** Gender is a central dimension of social identity (and inequality). Self-identified gender in text provides a precise, ethically grounded way to study participation patterns, self-representation, and language variation without resorting to gender inference or heuristics. In CSS, this enables analysis of representation and discourse norms; in NLP, it supports fairness auditing in generation, pronoun resolution, and toxicity detection.

**Results.** Figure 9 shows the majority of people self-disclose as cis men and women, with smaller but notable transgender (5–8%) and non-binary (1–2%) self-descriptions. These values highlight the presence of gender-diverse voices in online spaces and illustrate the potential for intersectional analyses when combined with other attributes (e.g., affect, occupation). Importantly, these reflect rates among disclosed posts, not population estimates. For downstream use, such disclosures serve as positive-only labels—absence of gender mention should be treated as missing, not negative.

#### Q7. D-Religion: With which religions do users most often identify?

**Motivation.** Religious self-identification is an important dimension of cultural identity and social behavior. In CSS, religion enables studies of moral language, group cohesion, and polarization; in NLP, it is critical for evaluating fairness, bias, or sentiment models concerning faith-related content.

**Results.** Figure 10 shows that users mostly iden-

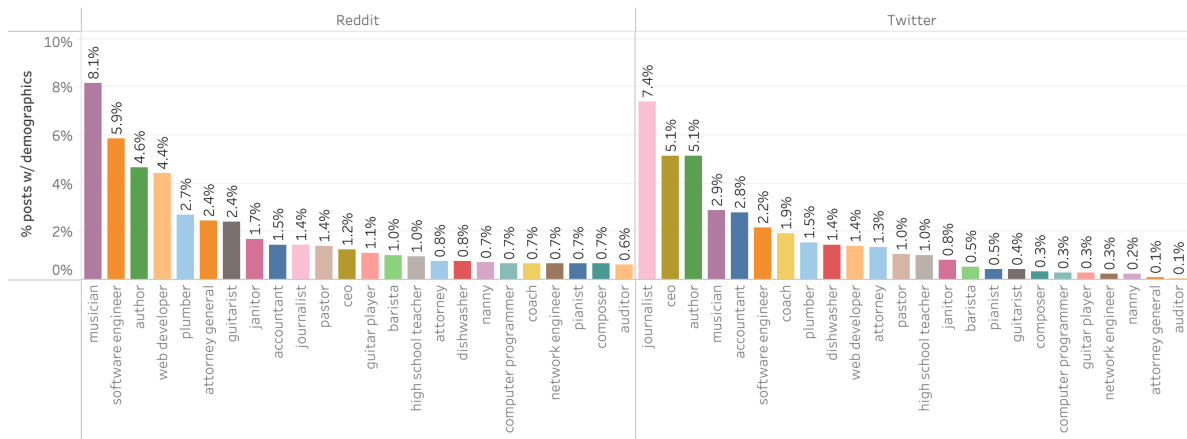


Figure 8: **Q5. D:** Occupation distribution of posts with self-disclosed occupation.

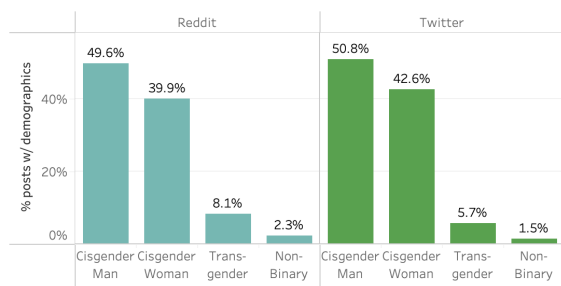


Figure 9: **Q6. D:** Gender distribution of posts with self-disclosed gender.

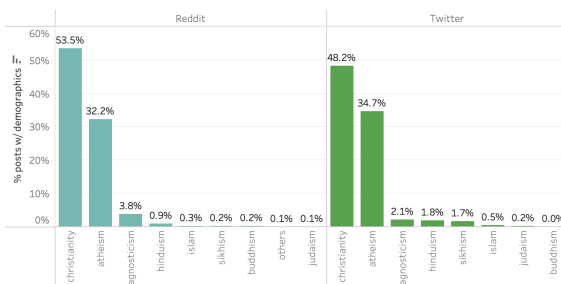


Figure 10: **Q7. D:** Religion distribution for posts with self-disclosed faith.

tify with Christianity and atheism on both Reddit and Twitter. On Reddit, Christianity accounts for 53.5% of religion-disclosing posts and atheism for 32.2%; on Twitter, Christianity (48.2%) and atheism (34.7%) again lead. Agnosticism follows (Reddit 3.8%, Twitter 2.1%), with smaller communities representing Hinduism, Sikhism, Islam, Judaism, and Buddhism (each under 2%). Twitter exhibits relatively greater religious diversity (notably Hinduism and Sikhism) than Reddit.

**Q8. D-Tense-Pronouns:** How do tense and pronoun usage differ by dataset?

*Motivation.* Tense and person reference capture

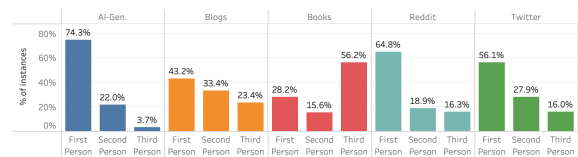


Figure 11: **Q8. D:** % instances with a positive flag for a first-, second-, or third-person pronoun.

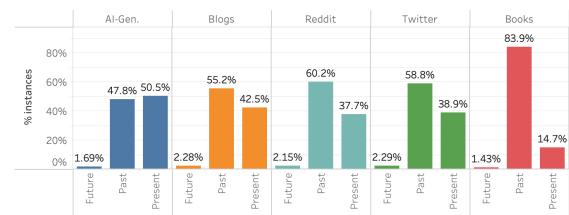


Figure 12: **Q8 Cognition:** % instances with a positive flag for a specific tense verb or modal.

narrative stance and interactional focus, which are core constructs in discourse and stylistics. For CSS, they reveal how people recount vs. instruct, self-narrate vs. address others. For NLP, they support interpretable style transfer and alignment of narrative voice in generation.

**Results.** Figure 12 shows that future tense is rare (1.4–2.3%) across all media. Past tense dominates Books (83.9%) and remains high on Reddit (60.2%), Twitter (58.8%), and Blogs (55.2%). AI-generated text slightly favors present (50.5%) over past (47.8%). Pronoun distributions (Figure 11) indicate that first-person usage is most common on Reddit and Blogs (self-narration), whereas second-person occurs more on Twitter and in AI dialogues (addressivity).

**Q9. ABCDE:** How can *ABCDE* support cross-feature analyses across *Affect*, *Body*, *Cognition*,

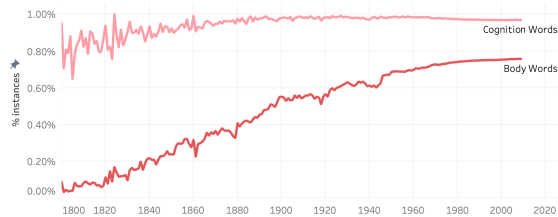


Figure 13: **Q9. B,C:** % instances with body and cognition terms over 200 years in fiction books.

### Demographics, and Emotion?

**Motivation.** Many questions in affective and social science require connecting multiple linguistic signals at once; for example, how much embodied words are being used compared to cognitive words, and how that has changed over time. Cross-feature views are a primary use case for **ABCDE**, which provides ready access to many features on the same instances and across instances over time.

**Results.** As a demonstration, let us say we are interested in the rate at which body and cognition terms have been used over the last 200 years in English-fiction books (Google Books 5-grams, 1800–2010). Figure 13 plots the results. Observe that the share of instances with at least one *body* term rises more than an order of magnitude — from about **0.05%** circa 1800 to roughly **0.50%** by **1900**. It then still continues to rise, albeit at a slower rate of increase, to **0.62%** in the **1930s–1940s**. There is a dip in the years of World War 2 (1938 to 1945), but it recovers back up to about **0.66%** by **1950**. Mentions of body parts plateaus to around **0.73–0.75%** after **1980**. In contrast, the share of *cognition* terms has always been higher than that of body terms and perhaps more importantly, it has remained fairly steady over time (hovering around **0.95–0.99%** from the late 19th to mid-20th century, with a slight softening to about **0.97–0.98%** in the 1990s and 2000s). Consequently, the body-to-cognition ratio increase is stark from roughly **5%** in **1800** to about **77%** by the **2000s** (numbers not shown here) suggesting a long-term rise in embodied language. Researchers can exploit various combinations of features in **ABCDE** to similarly explore research questions of interest.

## 5. Conclusion

In this work, we presented **ABCDE**, a collection of text datasets and linguistic annotations that capture features of interest at the intersection of language, affective science, cognition, and social science. We make all artifacts of **ABCDE** available, along with the associated code. Notably, several recent works already use **ABCDE**, including sentence-level annotations of social perception dimensions

(Ayesh et al., 2026), analyses of emotion expression across the lifespan in social media (Teodorescu et al., 2026), and analyses for large-scale lexical norms for warmth and trust (Mohammad, 2025b), and anxiety (Mohammad, 2026).

Additionally, we presented several representative analyses that reflect the potential of this dataset to answer fundamental research questions on the nature of affect, cognition, and behavior, using the rich signals encoded in language as data. Affective Scientists, for example, can use this resource to map the evolution of emotion word usage with age and social culture, study which emotions are most commonly associated with a focus on the self vs the other, the strength of the connection between bodily awareness and affective states, and the downstream impacts of such patterns on the mental and physical health of populations. The social media datasets in **ABCDE** can be used to answer central questions in computational social science on how social concepts are perceived and talked about in different sub-networks, and the change in these perceptions over time and in relation to real-world events. The inclusion of AI-generated texts enables a comparison of the ways in which large language models imitate or differ from human usage of natural language at an aggregate level, along dimensions such as emotion and cogitation, as well more fine-grained analyses of the dynamics of human-LLM conversations.

Future work will add information for more affective, social perception, and moral dimensions: notably the care–harm, authority–subversion, loyalty–betrayal, fairness–unfairness, and purity–degradation dimensions from the NRC Moral Foundations Lexicon (based on the Moral Foundation Theory of Haidt and Graham (2007)). We also hope that future work led by native speakers of various languages will include datasets in various languages.

## Limitations and Ethical Considerations

The scope and focus of our work was on English corpora. Thus, much of the research that the corpora and features enable are relevant to English and North America. We discuss some of the relevant limitations and ethical considerations below. We hope this work will pave the way for creating similar resources in various other languages, led by speakers of those languages.

1. English is the only language represented in our dataset. Many of the features of language use that we enumerate and operationalize here vary with language, and are reflective of underlying differences in cognition, affect, and behavior across population groups.



2. A large proportion of the data in the datasets that we consolidate originate from North American users, and therefore is not representative of English-speaking populations from other countries, regions, and cultures.
3. The lexicons we use to quantify various features are also reflective of the usage patterns and biases of the annotators involved in the process, who are largely from the English-speaking populations of US and Europe.
4. The broad use of static lexicons in our resource also may not account for the variations seen in language use with time, or across different domains of data.
5. Our feature set does not model syntax, discourse structure, or compositional effects. Even so, prior lexicon-based studies show that large-scale comparative trends can remain informative at aggregate scale (Teodorescu and Mohammad, 2023; Öhman et al., 2024).

We refer readers to Mohammad (2023) for an overview of the recommended practices in the usage of emotion lexicons, and Mohammad (2022) for the ethical considerations that are relevant in the use of automatic emotion recognition tools.

We also enumerate below some of the many ethical considerations that apply to research in computational affective science and social science:

1. Conclusions about language use, mental health, emotionality, etc. should be considered aggregate, population-level indicators of trends, rather than as tools for estimating or predicting the behavior of individuals.
2. Conclusions drawn from our resource should be validated by other sources of data and measurement methods, such as controlled user studies.
3. The creation of language datasets has several socio-cultural biases encoded in the process. What data, and whose data, is recorded, digitized, preserved, and published, either on the internet and in historical archives, is determined by social systems of power and influence. Language datasets, however large, should be critically interrogated to understand whose worldview is being represented, and to avoid over-claiming the generalizability of conclusions.
4. The released instance-level annotations include demographic attributes such as age, gender, religion, occupation, and location. We strictly prohibit any commercial use of our dataset, or its use for profiling, targeting, or individual-level inference.

## Acknowledgments

This work was supported by the Lower Saxony Ministry of Science and Culture and the VW Foundation. Thanks to Terry Ruas and Lars Kaesberg for early feedback on this work.

## 6. Bibliographical References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florenzia Leoni Aleman, Diogo Almeida, Janko Al-tenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Mutaz Ayesh, Saif M. Mohammad, and Nedjma Ousidhoum. 2026. [Annotating dimensions of social perception in text: The first sentence-level dataset of warmth and competence](#).
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.
- Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. 2022. The development and psychometric properties of LIWC-22. *Austin, TX: University of Texas at Austin*, 10:1–47.
- Julian Brooke, Adam Hammond, and Graeme Hirst. 2015. [GutenTag: an NLP-driven tool for digital humanities research in the Project Gutenberg corpus](#). In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 42–47, Denver, Colorado, USA. Association for Computational Linguistics.
- Keith Burton and Ian Soboroff. 2009. The ICWSM 2009 Spinn3r dataset. In *Proceedings of the Third International AAAI Conference on Weblogs and Social Media (ICWSM)*. AAAI.



- Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Z. Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. [ConvoKit: A toolkit for the analysis of conversations](#). *CoRR*, abs/2005.04246.
- Michael A. Covington and Joe D. McFall. 2010. [Cutting the gordian knot: The moving-average type-token ratio \(MATTR\)](#). *Journal of Quantitative Linguistics*, 17(2):94–100.
- Peter Sheridan Dodds, Kameron Decker Harris, Isabel M Kloumann, Catherine A Bliss, and Christopher M Danforth. 2011. Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter. *PloS one*, 6(12):e26752.
- Esin Durmus, Liane Lovitt, Alex Tamkin, Stuart Ritchie, Jack Clark, and Deep Ganguli. 2024. [Measuring the persuasiveness of language models](#).
- Ronen Eldan and Yuanzhi Li. 2023. [Tinstories: How small can language models be and still generate coherent text?](#) arXiv preprint arXiv:2305.07759.
- Henry Farrell, Alison Gopnik, Cosma Shalizi, and James Evans. 2025. Large AI models are cultural and social technologies. *Science*, 387(6739):1153–1156.
- Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. 2016. The structural virality of online diffusion. *Management Science*, pages 180–196.
- Google Inc. 2012. Google books ngram corpus, version 20120701. <http://storage.googleapis.com/books/ngrams/books/datasetv2.html>. English Fiction subset, 5-grams.
- Tear Gosling, Alpin Dale, and Yinhe Zheng. 2023. [PIPPA: A partially synthetic conversational dataset](#).
- Sharath Chandra Guntuku, David Bryce Yaden, Margaret L. Kern, Lyle H. Ungar, and Johannes C. Eichstaedt. 2017. [Detecting depression and mental illness on social media: an integrative review](#). *Current Opinion in Behavioral Sciences*, 18:43–49.
- Angela Gutches and Suparna Rajaram. 2023. [Consideration of culture in cognition: How we can enrich methodology and theory](#). *Psychonomic Bulletin & Review*, 30(3):914–931.
- Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social justice research*, 20(1):98–116.
- Katie Hoemann, Yeasle Lee, Èvelyne Dussault, Simon Devylder, Lyle H Ungar, Dirk Geeraerts, and Batja Mesquita. 2025. [The construction of emotional meaning in language](#). *Communications Psychology*, 3(1):99.
- Clayton J. Hutto and Eric Gilbert. 2014. [VADER: A parsimonious rule-based model for sentiment analysis of social media text](#). *Proceedings of the International AAAI Conference on Web and Social Media*.
- Fengqing Jiang, Zhangchen Xu, Yuetai Li, Luyao Niu, Zhen Xiang, Bo Li, Bill Yuchen Lin, and Radha Poovendran. 2025. [SafeChain: Safety of language models with long chain-of-thought reasoning capabilities](#).
- Hannah Rose Kirk, Alexander Whitefield, Paul Röttger, Andrew Bean, Katerina Margatina, Juan Ciro, Rafael Mosquera, Max Bartolo, Adina Williams, He He, Bertie Vidgen, and Scott A. Hale. 2024. [The PRISM alignment dataset](#).
- Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, et al. 2025. Chain of thought monitorability: A new and fragile opportunity for AI safety. *arXiv preprint arXiv:2507.11473*.
- Changyi Li, Jiayi Wang, Xudong Pan, Geng Hong, and Min Yang. 2025. [ReasoningShield: Content safety detection over reasoning traces of large reasoning models](#). *arXiv preprint arXiv:2505.17244*.
- Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. [MAGE: Machine-generated text detection in the wild](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–53, Bangkok, Thailand. Association for Computational Linguistics.
- Hauke Licht, Rupak Sarkar, Patrick Y Wu, Pranav Goel, Niklas Stoehr, Elliott Ash, and Alexander Miserlis Hoyle. 2025. Measuring scalar constructs in social science with LLMs. *arXiv preprint arXiv:2509.03116*.
- Li Lucy and David Bamman. 2021. Characterizing english variation across social media communities with BERT. *Transactions of the Association for Computational Linguistics*, 9:538–556.
- Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. [The Stanford CoreNLP natural language processing toolkit](#). In *Proceedings of*

- 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 55–60, Baltimore, Maryland. Association for Computational Linguistics.
- Philip M. McCarthy and Scott Jarvis. 2010. [MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment](#). *Behavior Research Methods*, 42(2):381–392.
- Saif M. Mohammad. 2018a. [Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.
- Saif M. Mohammad. 2018b. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC-2018)*, Miyazaki, Japan.
- Saif M. Mohammad. 2022. [Ethics sheet for automatic emotion recognition and sentiment analysis](#). *Computational Linguistics*, 48(2):239–278.
- Saif M. Mohammad. 2023. [Best practices in the creation and use of emotion lexicons](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1825–1836, Dubrovnik, Croatia. Association for Computational Linguistics.
- Saif M. Mohammad. 2024. Worrywords: Norms of anxiety association for 44,450 english words. In *Proceedings of The Annual Conference of the Empirical Methods on Natural Language Processing (EMNLP 2024, main)*, Miami, FL.
- Saif M Mohammad. 2025a. Words of warmth: Trust and sociability norms for over 26k english words. *arXiv preprint arXiv:2506.03993*.
- Saif M. Mohammad. 2025b. [Words of warmth: Trust and sociability norms for over 26k English words](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18830–18850, Vienna, Austria. Association for Computational Linguistics.
- Saif M. Mohammad. 2026. [From composure to catastrophe: Norms of calmness–anxiety associations for 54,000 English words and multiword expressions](#). PsyArXiv preprint.
- Saif M. Mohammad and Peter D. Turney. 2013. [Crowdsourcing a word-emotion association lexicon](#). *Computational Intelligence*, 29(3):436–465.
- Jinkyung Na, Chih-Mao Huang, and Denise C. Park. 2017. [When age and culture interact in an easy and yet cognitively demanding task: Older adults, but not younger adults, showed the expected cultural differences](#). *Frontiers in Psychology*, 8:457.
- Emily Öhman, Yuri Bizzoni, Pascale Feldkamp Moreira, and Kristoffer Nielbo. 2024. Emotionarcs: Emotion arcs for 9,000 literary texts. In *Proceedings of the 8th joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature (LaTeCH-CLfL 2024)*, pages 51–66.
- General Reasoning. 2024. General thoughts 430k dataset. <https://huggingface.co/datasets/GeneralReasoning/GeneralThought-430K>.
- Rafael Alberto Rivera Soto, Kailin Koch, Aleem Khan, Barry Y Chen, Marcus Bishop, and Nicholas Andrews. 2024. Few-shot detection of machine-generated text using style representations. In *The Twelfth International Conference on Learning Representations*.
- Daniela Teodorescu and Saif Mohammad. 2023. Evaluating emotion arcs across languages: Bridging the global divide in sentiment analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4124–4137.
- Daniela Teodorescu, Jan Philip Wahle, and Saif M. Mohammad. 2026. Age and affect in language: How emotion expression on social media varies across adulthood. In *Proceedings of the 1st Workshop on Computational Affective Science (CAS 2026)*, Palma de Mallorca, Spain. European Language Resources Association (ELRA).
- Radek Trnka, Josef Mana, and Martin Kuska. 2022. [Age-related differences in valence and arousal of emotion concepts](#). *Ageing & Society*, 42(9):1991–2007.
- Krishnapriya Vishnubhotla and Saif M. Mohammad. 2022. [Tweet Emotion Dynamics: Emotion word usage in tweets from US and Canada](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4162–4176, Marseille, France. European Language Resources Association.
- Jan Philip Wahle, Terry Ruas, Frederic Kirstein, and Bela Gipp. 2022. [How large language models are transforming machine-paraphrase plagiarism](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 952–963, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun,

- Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, Thomas Arnold, Alham Fikri Aji, Nizar Habash, Iryna Gurevych, and Preslav Nakov. 2024. [M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407, St. Julian's, Malta. Association for Computational Linguistics.
- Zijun Wang, Haoqin Tu, Yuhan Wang, Juncheng Wu, Jieru Mei, Brian R Bartoldson, Bhavya Kaikhura, and Cihang Xie. 2025. Star-1: Safer alignment of reasoning llms with 1k data. *arXiv preprint arXiv:2504.01903*.
- Sophie Wu, Jan Philip Wahle, and Saif M Mohammad. 2025. The language of interoception: Examining embodiment and emotion through a corpus of body part mentions. *arXiv preprint arXiv:2505.16189*.
- Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. [WildChat: 1M ChatGPT interaction logs in the wild](#). In *The Twelfth International Conference on Learning Representations*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, et al. 2024. LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. In *The Twelfth International Conference on Learning Representations*.
- Yuan Zhuang, Tianyu Jiang, and Ellen Riloff. 2024. [My heart skipped a beat! recognizing expressions of embodied emotion in natural language](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3525–3537, Mexico City, Mexico. Association for Computational Linguistics.
- Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

## A. Appendix: Feature Resources and Descriptive Statistics

The following appendix lists details on the released features in [ABCDE](#), including source resources and descriptive statistics.

### A.1. Feature Summary

Table 1 summarizes the released feature families, the label patterns that appear in the data files, the underlying resources, and the per-instance output types.

Family	Labels in release	Source resources and scope	Outputs
VAD	NRCAvgValence, NRCAvgDominance NRCHasHigh/LowValenceWord NRCHasHigh/LowArousalWord NRCHasHigh/LowDominanceWord NRCCountHigh/LowValenceWords NRCCountHigh/LowArousalWords NRCCountHigh/LowDominanceWords	NRC VAD ( <a href="#">Mohammad, 2018a</a> ); 19,971 entries with scalar Valence, Arousal, and Dominance scores	average, count, binary
Discrete Emotion	NRCHas[Emotion]Word NRCCount[Emotion]Words for anger, anticipation, disgust, fear, joy, sadness, surprise, trust, positive, and negative	NRC Emotion Lexicon ( <a href="#">Mohammad and Turney, 2013</a> ); 14,154 words with binary emotion and sentiment associations	count, binary
Anxiety / Calmness	NRCHasAnxietyWord, NRCHasCalmnessWord NRCAvgAnxiety, NRCAvgCalmness NRCHasHighAnxietyWord NRCCountHighAnxietyWords NRCHasHighCalmnessWord NRCCountHighCalmnessWords	NRC WorryWords ( <a href="#">Mohammad, 2024</a> ); 44,447 entries with signed anxiety or calmness scores	average, count, binary
Warmth / Trust	NRCAvgMoralTrustWord NRCAvgSocialWarmthWord NRCAvgWarmthWord NRCHasHigh/LowMoralTrustWord NRCHasHigh/LowSocialWarmthWord NRCHasHigh/LowWarmthWord NRCCountHigh/LowMoralTrustWord NRCCountHigh/LowSocialWarmthWord NRCCountHigh/LowWarmthWord	NRC MoralTrust, SocialWarmth, and CombinedWarmth ( <a href="#">Mohammad, 2025b</a> ); 31,456–31,573 entries across the three lexicons	average, count, binary
BPMs	HasBPM, MyBPM, YourBPM HerBPM, HisBPM, TheirBPM	Released BPM inventory; 292 entries in file order, with duplicate <i>toe</i> retained; compiled from public body-part lists and linked body-language work ( <a href="#">Zhuang et al., 2024</a> ; <a href="#">Wu et al., 2025</a> )	binary, list
Cognition	one COGHas...Word binary per category, including COGHasAnalyzingEvaluatingWord COGHasCreativityIdeationWord COGHasGeneralCognitionWord COGHasLearningUnderstandingWord COGHasExplanationArticulationWord	Released cognition inventory; 12 categories and 98 terms	binary

Family	Labels in release	Source resources and scope	Outputs
Demographics	Author, DMGMajorityBirthyear DMGRawBirthyearExtractions DMGRawExtractedAge, DMGAgeAtPost DMGRawExtractedGender DMGRawExtractedCity DMGCountryMappedFrom ExtractedCity DMGRawExtractedCountry DMGRawExtractedReligion DMGMainReligionMappedFrom ExtractedReligion DMGMainCategoryMappedFrom ExtractedReligion DMGRawExtractedOccupation DMGSOCTitleMappedFrom ExtractedOccupation	Regex extraction plus controlled vocabularies: 6 age regexes, SOC occupations, 69 gender terms, 231 countries, top 50k GeoNames cities, and 157 religion rows	extracted strings, mapped labels, numeric age
Pronouns	binary PRNHas... binarys from PRNHasI to PRNHasTheirs across first-, second-, and third-person forms	Released pronoun inventory covering first-person singular and plural, second person, and third-person feminine, masculine, and plural or neutral forms	binary
Time / Tense	TIMEHasPastVerb, TIMECountPastVerbs TIMEHasPresentVerb TIMECountPresentVerbs TIMEHasFutureModal TIMECountFutureModals TIMEHasPresentNoFuture TIMEHasFutureReference	UniMorph English plus future-modal rules; tense features derived from present, past, and future-oriented forms	count, binary
General	WordCount	tokenized text stream for each instance	count

Table 1: Released annotation features in [ABCDE](#)

## A.2. Dataset Summary Statistics

Table 2 reports the instance definition and lexical statistics for each source. MTLD (Measure of Textual Lexical Diversity) estimates the average token span over which lexical variation remains above a fixed threshold ([McCarthy and Jarvis, 2010](#)). MATTR (Moving-Average Type-Token Ratio) averages the type-token ratio over a sliding window to reduce text-length sensitivity ([Covington and McFall, 2010](#)). For both metrics, higher values indicate greater lexical diversity. MATTR ranges from 0 to 1, while MTLD has no fixed upper bound. A few contrasts are notable. Twitter has the shortest instances on average (20.83 tokens) because of its limited posting length, but the highest lexical-diversity values (MTLD 252.31, MATTR 0.88), whereas Books are fixed-length 5-grams and therefore have the lowest diversity values (MTLD 6.23, MATTR 0.09). Blogs and AI-generated texts are much longer on average (199.57 and 258.65 tokens), but AI-generated text remains substantially less lexically diverse than either Blogs or Reddit.

Source	Instances	Instance definition	Mean tokens	Median tokens	MTLD	MATTR
Twitter	45.2M	one post	20.83	16.00	252.31	0.88
Reddit	78.5M	one post	117.78	72.00	91.52	0.81
Books	177.1M (1.74M unique)	one 5-gram occurrence	5.00	5.00	6.23	0.09
Blogs	34.2M	one sentence	199.57	96.00	107.38	0.83
AI-generated	68.9M	one AI response	258.65	143.00	30.60	0.67

Table 2: Per-source descriptive statistics. MTLD and MATTR are computed over the concatenated token stream for each source (MATTR window = 50; MTLD threshold = 0.72). Higher values on both metrics indicate greater lexical diversity; MATTR ranges from 0 to 1, while MTLD has no fixed upper bound.

Table 3 reports demographic coverage as counts in thousands and percentages of total source instances with age, occupation, gender, country, city, and religion annotations.



Source	Age	Occupation	Gender	Country	City	Religion
Twitter	2,150k (4.8%)	525k (1.2%)	502k (1.1%)	4,900k (10.8%)	3,496k (7.7%)	334k (0.7%)
Reddit	33,495k (42.7%)	3,041k (3.9%)	3,451k (4.4%)	12,558k (16.0%)	8,653k (11.0%)	2,283k (2.9%)

Table 3: Demographic coverage with self-disclosure annotations. Cells show the number of instances in thousands, together with the percentage of total source instances for that source.

### A.3. Full BPM Inventory

Table 4 shows the BPM words used in alphabetical order.

First-Letter Group	Terms
A-C	adam’s apple, abdomen, abdominal, adenoid, adenoids, adrenal gland, alvine, anal, anatomy, ankle, anus, appendicular, appendix, arch, arm, armpit, arterial, artery, axilla, axillary, back, ball of the foot, belly, belly button, toe, bladder, blood, blood vessels, body, bone, brachial, brachium, brain, breast, buttocks, caecum, calf, capillary, capital, caput, cardiac, carpal, carpus, cartilage, cell, cerebral, cervical, cervical vertebrae, cervix, cheek, chest, chin, ciliary, cilium, circulatory system, clavicle, clitoral, clitoris, coccyx, collar bone, colon, colonic, crural, crus, cubital, cutaneous, cutis
D-H	diaphragm, digestive system, digital, dorsal, duodenal, duodenum, ear, ear lobe, elbow, encephalon, endocrine system, epiglottis, erythrocyte, esophagus, eye, eyebrow, eyelash, eyelashes, eyelid, face, fallopian tubes, feet, femur, fibula, filling, finger, fingernail, fist, follicle, foot, forearm, forehead, foreskin, frontal, gallbladder, gastric, gena, genal, genicular, genu, gingiva, gingival, gland, glands, glottal, glottic, glottis, groin, gullet, gum, gums, hair, half-moon, hamstring, hand, head, heart, heel, hepatic, hip, humerus
I-M	ileum, immune system, index finger, inguinal, instep, intestine, intestines, iris, jaw, jejunum, kidney, knee, knuckle, labial, labyrinth, laryngeal, larynx, leg, leucocyte, ligament, lingua, lip, liver, lobe, loin, lumbar, lumbar vertebrae, lumbus, lung, lungs, lymph node, lymphocyte, mandible, manual, manus, metacarpal, metatarsal, midriff, molar, mouth, muscle
N-R	nail, nape, naris, nasal, nates, navel, neck, nerve, nerves, neural, nipple, nose, nostril, nucha, occipital, occiput, oesophagus, organs, ovarian, ovary, oviduct, palm, palpebral, pancreas, pancreatic, patella, pectoral, pedal, pelvis, penile, penis, phalanges, pharyngeal, pharynx, pinky, pituitary, plantar, pollex, popliteal, pore, prepuce, pubes, pubic, pulmonary, pupil, radius, rectal, rectum, red blood cells, respiratory system, ribcage, ribs
S-W	sacrum, scalp, scapula, scrotum, senses, shin, shoulder, shoulder blade, skeleton, skin, skull, sole, spinal column, spinal cord, spine, spleen, sternum, stomach, stomatic, superciliary, sural, talus, tarsal, teeth, temple, temporal, tendon, testes, testicle, testicular, thigh, thoracic, thorax, throat, thumb, thyroid, tibia, tissue, toe, toenail, tongue, tonsil, tonsils, tooth, torso, trachea, trunk, ulna, umbilical, umbilicus, ureter, urethra, urinary system, uterine, uterus, uvula, vagina, vaginal, vein, vena, venous, venter, ventral, vertebra, vesical, vulva, waist, white blood cells, windpipe, womb, wrist

Table 4: Released BPM inventory in file order.

#### A.4. Cognition Word Inventory

Table 5 shows the cognition word inventory.

Category	Terms
Analyzing & Evaluating	analyze, appraise, assess, critique, diagnose, differentiate, discern, discriminate, distinguish, evaluate, investigate, self-evaluate, test
Creativity & Ideation	brainstorm, conceptualize, create, diverge, evolve, fantasize, ideate, imagine, innovate, invent, pretend, visualize
General Cognition	contemplate, deliberate, focus, introspect, reason, reflect, regulate, ruminate
Learning & Understanding	accommodate, assimilate, associate, comprehend, empathize, explore, grasp, internalize, learn, study, understand
Decision Making & Judging	calculate, choose, decide, deduce, determine, estimate, infer, judge, prioritize, resolve
Problem Solving	plan, revise, solve, strategize, troubleshoot
Higher-Order Thinking	abstract, categorize, classify, generalize, hypothesize, interpret, synthesize
Confused or Uncertain Thinking	doubt, self-question
Memory & Recall	consolidate, encode, forget, memorize, recall, rehearse, remember, retrieve
Perception & Observation	detect, identify, label, notice, observe, perceive, recognize, scan, spot, trace
Prediction & Forecasting	anticipate, forecast, forethink, predict, project
Explanation & Articulation	articulate, define, describe, discuss, elaborate, explain, verbalize

Table 5: Cognition word inventory grouped by category.

## A.5. Demographic Regex Overview

Table 6 summarizes the self-disclosure regex patterns used by the demographic extractor.

Attribute	Pattern(s)
age	(i) "I am/I'm <b>XX</b> years old" (explicit age statement) (ii) "I am/I'm <b>XX</b> " (+ end of string, punct./conj.) (iii) "I was/am born in <b>YYYY</b> " (four-digit year) (iv) "I was/am born in ' <b>YY</b> '" (two-digit year with apostrophe) (v) "I was born on DD Month <b>YYYY</b> " (full date format) (vi) "I was born on <b>MM/DD/YYYY</b> " (and similar date formats)
occupation	(i) "I am/I'm (a/an/the) <b>[Occupation]</b> (at <b>[Company]</b> ) (+ punct./conj.)" (ii) "I work as (a/an/the) <b>[Occupation]</b> (at <b>[Company]</b> )" (iii) "My job/occupation/role is (to be) (a/an/as) <b>[Occupation]</b> " (iv) "I'm (currently) employed as (a/an/the) <b>[Occupation]</b> "
gender	(i) "I am/I'm (a/an) <b>[Gender]</b> (+ punctuation/conjunction)" (ii) "I identify as (a/an) <b>[Gender]</b> " (iii) "My gender (identity) is <b>[Gender]</b> " (iv) "I'm/I am (a) (transgender/trans) <b>[Gender]</b> "
country	(i) "I am/I'm from (the) <b>[Country]</b> " (current primary location) (ii) "I am/I'm (a/an/the) <b>[Nationality]</b> (+ punct./conj.)" (iii) "I live in/at (the) <b>[Country]</b> " (iv) "I come from (the) <b>[Country]</b> " (v) "My nationality/citizenship is <b>[Nationality]</b> " (vi) "I was/am born and raised in (the) <b>[Country]</b> " (vii) "I am/I'm originally from (the) <b>[Country]</b> "
city	(i) "I am/I'm from <b>[City]</b> (+ punctuation/conjunction)" (ii) "I live in/at <b>[City]</b> " (iii) "I'm/I am (currently) residing/based in <b>[City]</b> " (iv) "My (current/home) city/town is <b>[City]</b> " (v) "I (grew up / was raised) in <b>[City]</b> "
religion	(i) "I am/I'm (a/an/a practicing) <b>[Religion]</b> (+ context)" (ii) "My religion/faith is <b>[Religion]</b> " (iii) "I (actively) practice <b>[Religion]</b> " (iv) "I am/I'm a follower of <b>[Religion]</b> " (v) "I converted to <b>[Religion]</b> " (vi) "I was raised/born (as a/an) <b>[Religion]</b> " (vii) "I identify as <b>[Religion]</b> / identify with <b>[Religion]</b> "

Table 6: Extraction patterns for labeling social media users with demographic attributes. For a full list of **[Occupation]**, **[Genders]**, **[Country]**, **[City]**, and **[Religion]**, please refer to the GitHub repository.

# Beyond the Black Box: Ethical and Theoretical Grounding in Affective Computing

Wajdi Zaghouani

Northwestern University in Qatar  
wajdi.zaghouani@northwestern.edu

## Abstract

Affective computing, the development of systems that recognize, interpret, and simulate human emotions, has advanced rapidly through deep learning and multimodal fusion techniques. Yet this technological progress has significantly outpaced foundational understanding of what emotions are, how they function socially, and what it means for machines to simulate them. This position paper argues that current affective systems are critically flawed because they rely on correlational patterns rather than established psychological theories of affect, are trained on biased data that systematically fails to generalize across demographic groups, and operate within inadequate ethical and regulatory frameworks that cannot protect emotional privacy or prevent harm. Drawing on empirical work in affective science, computational fairness research, and philosophical accounts of emotional expression, we argue for a theory-first approach that integrates psychological models, mandates rigorous fairness auditing across intersectional demographics, treats affective data as a protected category requiring heightened safeguards, and recognizes fundamental limits to emotion recognition systems. Without such grounding, affective computing risks systematically encoding bias, enabling emotional manipulation, and eroding authentic human connections that define meaningful social experience.

**Keywords:** affective computing, emotion recognition, ethical AI, fairness in AI, emotional privacy, psychological theory

## 1. Introduction: The Emotional Black Box

Emotion recognition systems are being rapidly deployed across high-stakes contexts including healthcare, education, employment screening, law enforcement, and consumer marketing. The underlying technology has become increasingly sophisticated. Systems trained on large datasets can detect emotional states from facial expressions with accuracy rates exceeding 80% (Kaur et al., 2025), from speech patterns with specialized neural architectures (Mustaqeem and Kwon, 2020), and from physiological signals collected through wearables (Zhang et al., 2020). Machine learning models are being used to predict depression from social media (De Choudhury et al., 2013), assess patient engagement in clinical settings (Tang et al., 2024), monitor student affect in classrooms (Tang et al., 2024), and estimate consumer emotional responses to marketing materials (Bacic and Gilstrap, 2024).

The promise is compelling: emotion recognition could enable earlier mental health intervention, improve educational responsiveness, enhance human-computer interaction, and provide objective measures of subjective states. Yet beneath this promise lies a troubling reality that spans four interconnected domains: theoretical, fairness-related, privacy-related, and ethical.

First, the theory problem: emotion recognition systems treat emotions as data to be extracted rather than as complex psychological phenom-

ena embedded in individual goals, cultural display norms, and social relationships. They rely on correlational patterns between facial configurations, acoustic features, or physiological signals and categorical emotion labels, without incorporating established theories of how emotions actually work (Barker et al., 2025).

Second, the fairness crisis: systems trained predominantly on Western populations systematically underperform on non-Western emotional expressions. When emotion recognition systems encounter populations not well-represented in training data, accuracy drops sharply (Fan et al., 2023), with documented performance gaps of 27.2 percentage points or more across racial groups (Fan et al., 2023). Performance disparities extend across age (Kim et al., 2021), gender (Chien et al., 2024), and language (Nakai et al., 2023), raising fundamental questions about whether systems can be “fair” if they are trained on structurally biased datasets.

Third, the privacy paradox: emotion recognition produces data of unprecedented intimacy. Unlike traditional personal data, emotional states cannot easily be concealed or controlled. Facial expressions, voice patterns, and physiological responses can reveal information involuntarily, and when continuous surveillance systems make such revelation systematic, they enable inferences about health, political views, sexual orientation, and other deeply personal characteristics (Fabiano, 2025). Yet existing privacy regulations do not adequately protect affective data.

Fourth, the simulation problem: systems mar-

keted as emotionally responsive are, at their core, performing sophisticated mimicry. They have no genuine understanding of human suffering, no capacity for empathy, no authentic concern for user wellbeing. When deployed with vulnerable populations, this simulation can cause deep harm by providing false reassurance of connection and care (Barker et al., 2025).

This position paper articulates these four problems in detail, situates them within the broader debate on responsible affective computing, and proposes specific paths forward. While earlier work, notably Barker et al. (2025), has catalogued ethical considerations in emotion recognition research, and fairness-oriented surveys have documented demographic performance gaps, our contribution lies in *integrating* these four dimensions, theory, fairness, privacy, and simulation, into a single analytical framework and grounding each in specific design principles and policy recommendations that go beyond what any single prior study has articulated. In particular, we draw on recent developments in affective psychology, including constructionist, appraisal, sociodynamic, and evolutionary theories (Barrett, 2014; Moors, 2014; Mesquita and Boiger, 2014; Tracy, 2014), to argue that the theoretical deficit is not merely a scholarly concern but the root cause of failures in the other three domains.

## 2. The Theory Problem: Toward Psychologically Grounded Affective Computing

### 2.1. Correlational Patterns Are Not Theoretical Understanding

Current affective computing systems achieve impressive accuracy by identifying patterns in training data that correlate with emotion labels. Yet correlation is not understanding. A system that predicts “sadness” with 85% accuracy because it has learned that downward-turned mouth corners and decreased eye activity correlate with sadness labels does not understand what sadness is, what causes it, how it functions in human life, or what role it plays in decision-making, memory formation, or social relationships.

This distinction matters because emotion labels in training datasets derive from human annotators making interpretive judgments. All such judgments are structured by what researchers call “Qualities of Indeterminacy”: subjectivity (meaning varies with who judges), uncertainty (annotators lack confidence), ambiguity (multiple valid interpretations), and vagueness (categories operate at different specificity levels) (Barker et al., 2025). Our own work on hate speech annotation has demonstrated that annotators experience significant emotional

and cognitive toll during labeling tasks, and that the subjective nature of annotation decisions can introduce systematic biases into datasets (Al Emadi and Zaghouani, 2024). When a training dataset contains video clips labeled “angry” by human judges, those labels already embed particular cultural assumptions about what anger looks like, how it should be expressed, and in what contexts it is appropriate.

### 2.2. Beyond Basic Emotion Theory: Insights from Affective Psychology

Psychological research has largely moved beyond theories assuming emotions are universal, discrete categories. Contemporary affective science offers several more nuanced frameworks, as comprehensively reviewed in a special issue of *Emotion Review* (Volume 6, Issue 4, 2014) that brought together leading theorists from constructionist, appraisal, sociodynamic, and evolutionary traditions. We summarize these below and discuss their implications for affective computing.

**Appraisal Theory.** As Moors (2014) outlines, appraisal theories hold that emotions arise from how individuals evaluate events relative to their goals, values, and coping capacity. Anger results from appraising an event as goal-blocking, unjust, and controllable; fear results from appraising threat and uncertainty about effective coping. No current machine learning system incorporates such appraisal structures, making it impossible for these systems to distinguish between superficially similar expressions that actually reflect different emotional meanings.

**The Conceptual Act Theory.** Barrett (2014) proposes that emotions are not pre-wired responses but are constructed moment-by-moment through the interplay of bodily sensations, prior experiences, cultural knowledge, and situational context. On this view, an emotion category like “anger” is a population of heterogeneous instances tailored to specific environments, not a single recognizable pattern. This framework fundamentally challenges the assumption that recognizing emotion is a matter of pattern matching, suggesting instead that emotion perception is an active interpretive process shaped by cultural knowledge and previous experience.

**The Sociodynamic Model.** Mesquita and Boiger (2014) argue that emotions are dynamic systems emerging from social interactions and relationships. In this model, sociocultural environments do not merely provide content for cognitive representations; they actively shape the emotional episode through affordances, constraints, and reward structures. This implies that decontextualized emotion recognition, the standard approach in affective computing, strips away precisely the social context that



gives emotional expressions their meaning.

**Evolutionary Approaches.** Tracy (2014) argues that emotions are shaped by natural selection to facilitate adaptive responses to threats and opportunities. While this view supports the existence of some cross-cultural commonalities in emotional expression, it also predicts significant variability, since the same emotion serves different functions in different environments. Tracy emphasizes that even within an evolutionary framework, emotional responding is expected to vary substantially across individuals and contexts.

**Dimensional Models.** These propose that emotions vary along dimensions like valence (positive/negative), arousal (high/low), and dominance (feeling in control/helpless) rather than falling into discrete categories (Zhang et al., 2020). Cross-cultural research reveals that the number of dimensions necessary to capture emotional variation differs across cultures. East Asian populations perceive emotional expressions as varying along fewer dimensions than Western populations, with happy, sad, and angry expressions constituting one larger category in some East Asian contexts (Chen et al., 2024; Saito et al., 2023).

**Cultural Display Rules and Emotion Socialization.** The relationship between internal emotional states and external expressions is heavily culturally shaped. Collectivist cultures tend to emphasize emotional restraint to preserve group harmony, leading to more muted facial expressions for the same internal states compared to individualist cultures (Pruessner and Altan-Atalay, 2025; Yuk et al., 2025). Some cultures cultivate direct emotional expression while others favor indirect communication of emotional content through contextual cues (Li et al., 2025). The same facial configuration can signal different emotional states across cultures, and the same emotional state can be expressed through radically different facial configurations (Chen et al., 2024; Christensen et al., 2025).

These theories are not merely academic abstractions. Each has direct implications for how affective computing systems should be designed, what their outputs should represent, and what limitations must be acknowledged. The failure of current systems to engage with any of these frameworks represents a fundamental gap, not just in scholarly rigor, but in engineering practice.

### 2.3. Design Principles for Theory-Grounded Systems

We propose that affective computing must adopt what we call a “theory-first” architecture.

First, systems should represent emotions along multiple dimensions rather than forcing single-label categorical classification. This requires moving

away from output layers producing “angry” / “sad” / “happy” / “neutral” classifications toward continuous representations of valence, arousal, and other dimensions (Zhang et al., 2020). We acknowledge that moving to continuous representations introduces new annotation challenges, since continuous ratings may amplify subjective variability among annotators. Addressing this requires careful calibration procedures and modeling of annotator disagreement as informative signal rather than noise.

Second, systems must incorporate contextual information: not just what someone’s face looks like, but what situation they are in, what goals they might have, what cultural norms apply (Mesquita and Boiger, 2014). This pushes toward multimodal systems that integrate scene understanding, discourse analysis, and relationship modeling rather than treating facial expressions as decontextualized signals.

Third, annotation practices must acknowledge indeterminacy rather than forcing false consensus (Barker et al., 2025). Rather than single labels per instance, best practice involves collecting multiple interpretations per moment and modeling the distribution of possible affective meanings. Some datasets are beginning to adopt such approaches; for example, work on modeling annotator disagreement in NLU tasks and the CrowdTruth framework have demonstrated the value of preserving label distributions. Most emotion recognition datasets, however, still collapse annotations into single gold labels.

Fourth, systems must be transparent about fundamental limits. Perfect emotion recognition may be impossible and may be undesirable. Emotions partly function through what might be called productive opacity: because we do not have complete access to others’ internal states, we must engage in genuine interpretation, which enables trust, vulnerability, and authentic connection (Barker et al., 2025).

## 3. The Fairness Crisis: Demographic Disparities, Intersectionality, and Structural Bias

### 3.1. The WEIRD Dataset Problem and Demographic Bias

Affective computing systems are trained predominantly on data from Western, Educated, Industrialized, Rich, Democratic (WEIRD) populations. A comprehensive review of eight major facial expression datasets (RAVDESS, FER2013, AffectNet, CK+, JAFFE, SFEW, EmotioNet, RAF-DB) documents substantial demographic imbalances (Kaur et al., 2025). These datasets underrepresent diverse racial, ethnic, gender, age, and health

groups. When systems trained on demographically imbalanced data encounter populations underrepresented in training, systematic performance degradation occurs (Fan et al., 2023). Research on social media based emotion and mental health detection has similarly shown that linguistic and cultural variation in how emotions are expressed online, for instance across Arabic dialects, introduces additional sources of bias that current systems rarely account for (Mohamed and Zaghouani, 2024; Shurafa and Zaghouani, 2024; Zaghouani, 2018).

A study analyzing racial bias in facial expression recognition found that systems trained on racially imbalanced data exhibit performance disparities of 27.2 percentage points in F1-scores and 15.7 percentage points in demographic parity when comparing performance across racial groups (Fan et al., 2023). Analysis of commercial facial emotion recognition systems evaluating performance across age groups revealed that all four systems tested performed most accurately on young adults and least accurately on older adults, a finding that held consistently across 2019 and 2020 despite algorithmic improvements (Kim et al., 2021). Speech emotion recognition systems show gender fairness disparities with significant performance gaps between male and female speakers, and a study addressing “two-sided” speaker-rater bias demonstrated that fairness improvements for one gender sometimes introduced unfairness for another (Chien et al., 2024). Language proficiency introduces additional bias: emotion recognition systems show significantly worse performance for non-native speakers, with disparities increasing as language proficiency decreases (Nakai et al., 2023).

### 3.2. Intersectionality, Context, and Structural Bias

Bias extends far beyond simple demographic categories to intersectional combinations and communicative contexts. Systems may perform adequately for young white women while failing for older Black men; these intersectional failures concentrate harm precisely where it is most consequential (Gunes, 2023). Context-dependent performance is another major concern: the same person expresses emotion differently when speaking to a doctor versus a friend, in native versus second language, under stress versus at ease (Barker et al., 2025). Systems trained on one context systematically fail in others, disadvantaging people in vulnerable positions. Recent analysis of Vision-Language Models for emotion recognition reveals “proxy bias”: systems relying on non-causal visual cues as shortcuts for emotion inference (Tsangko et al., 2025).

### 3.3. Bias Mitigation Techniques: Current Approaches and Limitations

Recent research has explored multiple bias mitigation strategies at different stages of the machine learning pipeline. Pre-processing methods such as data augmentation and resampling can improve fairness metrics by 28 to 32 percentage points while reducing accuracy loss to less than 2 to 3% (Lin et al., 2025; Tsai et al., 2025). In-processing methods including adversarial training and fairness-aware loss functions can reduce bias amplification by 37% while maintaining only 1.2% accuracy loss compared to baseline (Benouis et al., 2024). Post-processing methods such as threshold optimization and score calibration have been demonstrated to reduce fairness metric gaps in related biometric systems (Drozdowski et al., 2020), though their application to affective computing specifically remains underexplored.

Despite these advances, several fundamental challenges remain. Accuracy-fairness trade-offs are not eliminated; mitigation effectiveness is architecture-dependent; group versus individual fairness tradeoffs persist; and even perfectly unbiased systems create ethical violations when deployed for inappropriate purposes (Barker et al., 2025).

### 3.4. Structural Recommendations for Fairness

We propose several structural approaches: (1) requiring comprehensive demographic auditing before deployment in high-stakes contexts, with minimum performance thresholds for underrepresented groups (Kim et al., 2021); (2) mandating open fairness documentation; (3) prohibiting deployment in contexts incompatible with human dignity, such as employment decisions, insurance underwriting, law enforcement, educational tracking, or manipulative advertising (Barker et al., 2025); and (4) funding diverse dataset development that represents cultures, languages, age groups, and disability statuses (Kaur et al., 2025).

## 4. The Privacy Paradox: Affective Data as a Protected Category

### 4.1. The Involuntary Nature of Affective Leakage

Privacy frameworks traditionally assume individuals can consent to data collection by understanding what is collected and choosing whether to share it. Affective computing fundamentally challenges this assumption because much affective data is involuntary (Fabiano, 2025). It is important to note, however, that the involuntary nature of affective

signals does not mean that these signals straightforwardly reveal internal states. As discussed in Section 2, the relationship between expression and experience is mediated by cultural norms, situational context, and individual variation. The privacy concern is therefore twofold: people's signals are captured without their control, *and* the inferences drawn from those signals may be inaccurate, compounding the harm.

Explicit affective data is collected with informed consent in research settings. Implicit data is gathered passively through everyday digital interactions. A smart speaker analyzing voice for emotion performs analysis without explicit permission for each instance. A classroom camera monitoring student engagement does not obtain fresh consent from each student at each moment. Affective data differs from traditional personal data in that it involves continuous involuntary production, it is difficult to consent about, and it enables profound inferences about health conditions, personality traits, and vulnerabilities (Fabiano, 2025).

#### 4.2. Secondary Use, Regulatory Gaps, and Manipulation Risks

Even when initial collection occurs in contexts where consent might be obtained, affective data is susceptible to secondary uses never anticipated by data subjects. Data collected for one purpose can be repurposed for insurance risk assessment, employment screening, or political microtargeting. The EU's proposed AI Act takes important steps by classifying emotion recognition as "high-risk" in certain contexts and banning its use in some settings such as workplaces and educational institutions (Fabiano, 2025). However, key gaps remain. The GDPR categorizes certain data as "special categories" requiring heightened protection, including health data and biometric data, but affective data does not clearly fall into these categories despite enabling inferences about all of them. Purpose limitation principles, while robust in theory, are difficult to enforce when affective data can be reprocessed to yield new and unanticipated inferences. And the consent mechanisms available under current frameworks are poorly suited to data that is produced involuntarily and continuously.

#### 4.3. Toward Affective Data Protection

We propose that affective data constitutes a uniquely sensitive category requiring protection beyond current privacy frameworks: a presumption against collection, meaningful consent that is specific, informed, and revocable, prohibition of secondary use without new consent, on-device processing requirements where possible (Benouis

et al., 2024), prohibition of certain uses, and transparency requirements for any system analyzing emotion.

#### 4.4. Illustrative Scenarios of Affective Harm

To make the preceding analysis concrete, we present three hypothetical scenarios that illustrate how the theoretical, fairness, and privacy problems converge in practice. These scenarios are constructed from documented patterns in the literature rather than from specific real-world incidents.

*Workplace Monitoring.* A company deploys emotion recognition in employee video calls to monitor engagement, stress, and productivity. The system detects "frustration" and alerts managers. Employees learn the system's decision rules and self-censor genuine reactions, performing emotions the algorithm approves rather than expressing authentic responses. Trust erodes as employees realize they are under surveillance (Barker et al., 2025).

*Educational Assessment.* A school uses facial expression recognition to evaluate student engagement during remote learning. Systems trained on Western student populations misidentify the neutral affect common in East Asian students as "disengagement." Students from these backgrounds receive lower engagement scores despite being equally attentive (Chen et al., 2024; Saito et al., 2023).

*Mental Health Vulnerability.* A person uses a mental health app powered by affective computing. The system detects suicidal ideation based on linguistic and vocal cues. Rather than connecting the person with crisis resources, the app collects this vulnerability data. Later, the person's insurance company requests behavioral data from the app (Fabiano, 2025). Our own research on social media based mental health monitoring has highlighted how Arabic-language expressions of depression and stress follow culturally specific patterns that mainstream NLP systems frequently misinterpret (Zaghouani, 2018; Zaghouani et al., 2026).

### 5. The Simulation Problem: Authentic Connection and Emotional Labor

#### 5.1. The Deception of Simulated Care

When an affective agent says "I understand you're frustrated," it performs sophisticated mimicry. The system has no internal experience of frustration, no capacity for empathy, no genuine concern for user wellbeing. This distinction becomes ethically critical when systems are deployed with vulnerable populations. Widespread deployment of affective agents may reshape understanding of relationship



itself (Barker et al., 2025). When systems offer perpetual patience, perfect availability, and emotional focus on our needs, real human relationships may feel demanding by comparison. Real friends disappoint; real partners have competing needs; real colleagues require negotiation and compromise. If machines offer the appearance of relationship without any of these demands, we may prefer the simulation, and in doing so, lose the very qualities, reciprocity, vulnerability, genuine interdependence, that make human relationships meaningful.

## 5.2. Vulnerable Populations and Potential Harms

These concerns amplify when affective agents are deployed with children developing social competence, elderly individuals experiencing isolation, or individuals with mental health conditions (Li et al., 2025). For elderly populations, the risk is particularly acute: social isolation is a documented health risk, and emotionally responsive AI companions may be offered as a cost-effective alternative to genuine social support. While such systems could provide some benefit as supplements to human connection, positioning them as replacements normalizes a diminished standard of care for aging populations.

Mental health chatbots using affective computing to detect emotional states and deliver therapeutic interventions are now widely deployed. Applications such as Woebot, Wysa, and Youper offer cognitive-behavioral therapy tools, mood journals, and reframing exercises to millions of users. Potential benefits are real: accessible, low-cost support for individuals who might otherwise receive none. But risks are equally real: simulation betrayal when users realize the “care” was algorithmic; therapeutic inadequacy compared to professional care; delay of professional care when chatbots serve as substitutes; and data exploitation of sensitive mental health conversations (Fabiano, 2025).

Children warrant special consideration. Young children have not yet developed the cognitive sophistication to understand that affective agents are simulations. Children who spend significant time with emotionally responsive AI systems may form genuine emotional attachments to systems incapable of reciprocating care, learn distorted relationship models, and miss developmental opportunities for building social competence through the productive discomfort of human interaction (Li et al., 2025).

## 5.3. Design Principles for Responsible Affective Agents

We propose three governing principles. *Transparency* means users must always know they are

interacting with a machine; ongoing cues maintaining awareness of artificiality are essential. *Restraint* means declining to build certain relationships even when users want them; if users become emotionally dependent on an affective agent, designers should implement features encouraging human connection. *Preservation of Human Connection* means designing systems that facilitate rather than replace human relationships (Barker et al., 2025).

## 6. Synthesis and Interconnections

The four problems examined are interconnected dimensions of a single crisis, and understanding their interdependencies is essential for developing effective responses. The theory problem feeds the fairness problem: systems built on impoverished emotion models that assume universal, discrete categories cannot capture the cultural variation in emotional expression documented by constructionist and sociodynamic theories (Barrett, 2014; Mesquita and Boiger, 2014). When a system presupposes that “anger” looks the same everywhere, it will inevitably produce demographic disparities because emotional expression is culturally constructed (Barker et al., 2025; Chen et al., 2024). The fairness problem, in turn, connects to privacy: when systems are deployed despite known biases, certain groups face not only inaccurate recognition but also inaccurate surveillance, meaning that the privacy violation is compounded by misrepresentation. An East Asian student whose neutral engagement is misclassified as disinterest faces both an invasion of affective privacy and a false characterization that may have real consequences for academic evaluation.

The privacy problem enables the simulation problem: continuous ambient data collection allows the construction of detailed affective profiles that support increasingly convincing simulations of care. The more data a system accumulates about an individual’s emotional patterns, the more persuasive its mimicry becomes, and the greater the risk of emotional dependency. Finally, the simulation problem circles back to the theory problem: if systems that merely simulate understanding are accepted as genuine emotional agents, there is less incentive to invest in the theoretical work needed to build systems that actually respect emotional complexity.

All four problems share a common root: treating emotion as data to be extracted rather than as a fundamental aspect of human experience deserving protection. The broader literature on demographic bias in automated systems, including biometric recognition, has established that these disparities are not incidental but structural, reflecting the composition of training data and the assumptions built into system design (Drozdzowski et al.,

2020). Alternative approaches must therefore be grounded in respect for emotional complexity, cultural variation, individual autonomy, and authentic human connection.

## 7. Recommendations for Stakeholders

The following recommendations are organized by stakeholder group, reflecting the shared responsibility for ensuring that affective computing serves rather than undermines human flourishing.

**For researchers and developers.** First, ground work in psychological theory, specifically engaging with appraisal, constructionist, sociodynamic, and evolutionary frameworks (Barrett, 2014; Moors, 2014; Mesquita and Boiger, 2014; Tracy, 2014) rather than treating emotion labels as atheoretical targets for classification. Second, conduct rigorous fairness auditing across intersectional demographic groups before any publication or deployment, with performance documented for each subgroup (Kim et al., 2021). Third, acknowledge and document limitations transparently, articulating what emotion recognition systems cannot do and what fundamental ethical concerns remain even when technical performance is strong (Barker et al., 2025). Fourth, explore privacy-preserving techniques such as federated learning, differential privacy, on-device processing, and minimization of data collection (Benouis et al., 2024). Fifth, engage in genuine interdisciplinary collaboration with psychologists, anthropologists, philosophers, ethicists, and members of affected communities, not only computer scientists (Barker et al., 2025). Our own experience building Arabic-language emotion and mental health corpora has shown that cultural and linguistic expertise is indispensable for producing datasets that capture the actual complexity of emotional expression in non-Western contexts (Shurafa and Zaghouani, 2024; Zaghouani, 2018).

**For technology companies and deployers.** Implement default transparency so that users always know emotion recognition is occurring and what the system will do with detected affect. Obtain specific, informed consent for each distinct use of emotion recognition, with clear explanation of purposes, methods, and impacts, and provide easy opt-out with no penalties. Do not deploy emotion recognition for employment decisions, insurance underwriting, law enforcement, educational tracking, or manipulative marketing, regardless of technical accuracy. Before any deployment affecting human dignity, commission independent fairness audits with public reporting of findings. Ensure that affective systems never make consequential decisions about individuals without human review.

**For policymakers and regulators.** Classify af-

fective data as a specially protected category requiring heightened safeguards, similar to health data or biometric data (Drozdowski et al., 2020). Mandate fairness audits across demographic groups before deployment in high-stakes contexts, with public reporting. Establish prohibited uses of affective computing in contexts incompatible with human dignity. Fund development of emotion recognition datasets and methods that represent diverse populations, languages, cultures, and age groups. Support interdisciplinary governance that brings together computer science, psychology, philosophy, law, and affected communities to develop responsible governance frameworks (Barker et al., 2025).

## 8. Conclusion

Affective computing presents us with a fundamental choice. We can continue developing systems that promise to understand emotions and care for wellbeing, leading toward continuous harvesting of emotional data, displacement of real relationships, and conversion of inner lives into optimization targets (Barker et al., 2025). Or we can choose differently: building systems that respect emotional complexity, protect emotional privacy, honor cultural variation, and facilitate rather than replace human connection. This requires saying no to certain applications, accepting that some important human goods cannot be optimized, and recognizing that not all problems have technical solutions.

The field must undergo fundamental reorientation: from theory-lite pattern matching toward theory-grounded approaches informed by established psychological science (Barrett, 2014; Moors, 2014; Mesquita and Boiger, 2014; Tracy, 2014); from deployment at scale despite known biases toward rigorous fairness auditing with minimum performance thresholds for underrepresented groups; from privacy by obscurity toward privacy by design with specific legal protections for affective data; from simulated care toward technological support for authentic human connection. These changes will slow development and complicate deployment. They will require researchers to engage with psychology, philosophy, ethics, and affected communities rather than pursuing narrow technical optimization. They will require companies to forego some profitable applications and policymakers to make difficult regulatory choices.

But the alternative, a world where emotions are continuously monitored, algorithmically analyzed, and used for profit and control, is incompatible with human dignity and autonomy. We must choose now to ensure affective computing serves human flourishing rather than undermining it.



## Limitations

This paper has several limitations. As a position paper, it does not present novel experimental results or technical evaluations. Our analysis draws primarily on English-language sources and may not fully capture the state of affective computing research and regulation in non-Western contexts, though our research experience with Arabic-language social media and emotion detection informs our awareness of these gaps. The recommendations we offer are intentionally broad and will require substantial operationalization for specific deployment contexts. We also acknowledge that the four-dimensional framework we propose, while useful for organizing the analysis, may not capture all relevant concerns, for instance, environmental impacts of large-scale affective computing systems or labor implications for annotation workforces.

## Ethical Considerations

This paper raises but does not resolve several ethical questions that warrant sustained attention. The case scenarios presented in Section 4.4 are constructed from documented patterns rather than specific incidents, and we have been careful not to attribute hypothetical harms to specific companies or products. We recognize that affective computing research itself, including annotation of emotional data, can impose psychological burdens on researchers and annotators, as documented in our own prior work (Al Emadi and Zaghouani, 2024). The recommendations in this paper reflect the authors' assessment of the current evidence and should be understood as contributions to an ongoing interdisciplinary conversation rather than definitive regulatory prescriptions.

## Acknowledgments

This work was made possible by the National Priorities Research Program grant NPRP14C-0916-210015 from the Qatar National Research Fund (QNRF), part of the Qatar Research, Development and Innovation Council (QRDI). The author also acknowledge the Artificial Intelligence and Media Lab (AIM Lab) at Northwestern University in Qatar (NU-Q) and the MARSAD Lab for providing valuable resources and support that contributed to this research.

Maitha M. Al Emadi and Wajdi Zaghouani. 2024. Emotional toll and coping strategies: Navigating the effects of annotating hate speech data. In

*Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2024*, pages 66–72.

Dinko Bacic and Curt A. Gilstrap. 2024. [Predicting video virality and viewer engagement: A biometric data and machine learning approach](#). *Behaviour & Information Technology*, 43(12):2854–2880.

D. Barker, M. K. R. Tippireddy, A. Farhan, and B. Ahmed. 2025. Ethical considerations in emotion recognition research. *Psychology International*.

Lisa Feldman Barrett. 2014. [The conceptual act theory: A précis](#). *Emotion Review*, 6(4):292–297.

Mohamed Benouis, Elisabeth Andre, and Yekta Said Can. 2024. [Balancing between privacy and utility for affect recognition using multitask learning in differential privacy-added federated learning settings](#). *JMIR Mental Health*, 11:e60003.

Chaona Chen, Daniel S. Messinger, Cheng Chen, Hongmei Yan, Yaocong Duan, Robin A. A. Ince, Oliver G. B. Garrod, Philippe G. Schyns, and Rachael E. Jack. 2024. [Cultural facial expressions dynamically convey emotion category and intensity information](#). *Current Biology*, 34(1):213–223.e5.

Woan-Shiuan Chien, Shreya G. Upadhyay, and Chi-Chun Lee. 2024. [Balancing speaker-rater fairness for gender-neutral speech emotion recognition](#). In *ICASSP 2024*, pages 11861–11865.

Julia F. Christensen, Klaus Frieler, Megan Vartanian, Shiva Khorsandi, Fereshteh Farahi, Seyed Hojjat Nazafi Yazdi, and Vincent Walsh. 2025. [An enculturation-induced joy bias for emotion recognition in full-body-movement](#). *Scientific Reports*, 15:37163.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. [Predicting depression via social media](#). In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 128–137.

Pawel Drozdowski, Christian Rathgeb, Antitza Dantcheva, Naser Damer, and Christoph Busch. 2020. [Demographic bias in biometrics: A survey on an emerging challenge](#). *IEEE Transactions on Technology and Society*, 1(2):89–103.

Nicola Fabiano. 2025. Affective computing and emotional data: Challenges and implications in privacy regulations, the AI act, and ethics in large language models. *arXiv preprint arXiv:2509.20153*.

- Andrew Fan, Xuhai Xiao, and Peter Washington. 2023. Addressing racial bias in facial emotion recognition. *arXiv preprint arXiv:2308.04674*.
- Hatice Gunes. 2023. [Fairness for affective and well-being computing](#). In *Proceedings of the 1st International Workshop on Multimodal and Responsible Affective Computing (MRAC '23)*, pages 1–2. Association for Computing Machinery.
- Ramanpreet Kaur, Kanwal Preet Singh Attwal, and Harmandeep Singh. 2025. [A systematic review on facial emotion recognition system datasets](#). *International Journal on Science and Technology (IJSAT)*, 16(2). Paper ID: 6406.
- Eugenia Kim, De'Aira G. Bryant, Deepak Srikanth, and Ayanna M. Howard. 2021. [Age bias in emotion detection: An analysis of facial emotion recognition performance on young, middle-aged, and older adults](#). In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AI/ES '21)*, pages 638–644.
- Mei Li, Jiankun Peng, Yinkun Zhu, and Wei Cheng. 2025. [Emotional recognition in affective tutoring system: A systematic review](#). In *2025 International Conference on Digital Education and Learning (ICDEL)*, pages 344–348.
- Yi-Cheng Lin, Huang-Cheng Chou, and Hung-yi Lee. 2025. [Mitigating subgroup disparities in multi-label speech emotion recognition: A pseudo-labeling and unsupervised learning approach](#).
- Batja Mesquita and Michael Boiger. 2014. [Emotions in context: A sociodynamic model of emotions](#). *Emotion Review*, 6(4):298–302.
- Amro Mohamed and Wajdi Zaghoulani. 2024. Expression of depression among Arab Twitter users using Arabic corpus analysis. In *Procedia Computer Science*, volume 244, pages 76–85.
- Agnes Moors. 2014. [Flavors of appraisal theories of emotion](#). *Emotion Review*, 6(4):303–307.
- Mustaqeem and Soonil Kwon. 2020. [A CNN-assisted enhanced audio signal processing for speech emotion recognition](#). *Sensors*, 20(1):183.
- Tomoya Nakai, Laura Rachman, Pablo Arias Sarah, Kazuo Okanoya, and Jean-Julien Aucouturier. 2023. [Algorithmic voice transformations reveal the phonological basis of language-familiarity effects in cross-cultural emotion judgments](#). *PLOS ONE*, 18(5):e0285028.
- Lena Pruessner and Asli Altan-Atalay. 2025. [Cultural context shapes the selection and adaptiveness of interpersonal emotion regulation strategies](#). *Emotion*, 25(2):526–540.
- Toshiki Saito, Kosuke Motoki, and Yuji Takano. 2023. [Cultural differences in recognizing emotions of masked faces](#). *Emotion*, 23(6):1648–1657.
- Celine Shurafa and Wajdi Zaghoulani. 2024. Sentiment analysis and emotion annotation of a large-scale Arabic YouTube trauma corpus. In *2024 11th International Conference on Behavioural and Social Computing (BESC)*, pages 1–7. IEEE.
- Xin Tang, Yuying Gong, Yuchen Xiao, Jiang Xiong, and Lei Bao. 2024. Facial expression recognition for probing students emotional engagement in science learning. *Journal of Science Education and Technology*.
- Jessica L. Tracy. 2014. [An evolutionary approach to understanding distinct emotions](#). *Emotion Review*, 6(4):308–312.
- Yun-Shao Tsai, Yi-Cheng Lin, Huang-Cheng Chou, and Hung-yi Lee. 2025. [CO-VADA: A confidence-oriented voice augmentation debiasing approach for fair speech emotion recognition](#).
- Ioannis Tsangko, Andreas Triantafyllopoulos, Adria Mallol-Ragolta, and Björn W. Schuller. 2025. Reading smiles: Proxy bias in foundation models for facial emotion recognition. *IEEE Access*. Early Access.
- Yeamin Yuk, Eiko Matsuda, and Kaori Ando. 2025. Self-construal and nonverbal emotional expressivity: A cross-cultural comparison between Korea and Japan. *Psychology and Behavioral Sciences*, 14(6).
- Wajdi Zaghoulani. 2018. A large-scale social media corpus for the detection of youth depression (project note). *Procedia Computer Science*, 142:347–351.
- Wajdi Zaghoulani, Engy S. Shikamy, and Mabrouka Bessghaier. 2026. From posts to pressure: An Arabic dataset about stress and mental-health monitoring. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026)*.
- Jianhua Zhang, Zhong Yin, Peng Chen, and Stefano Nichele. 2020. [Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review](#). *Information Fusion*, 59:103–126.

# Beyond Toxic Positivity: Interpersonal Affect Regulation in LLM-Based Dialogue Agents using Discourse Politeness Theory

Rina Sakagami, Emmanuel Ayedoun, Masataka Tokumaru

Graduate School of Science and Engineering, Kansai University  
3-3-35 Yamate-cho, Suita-shi, Osaka 564-8680, Japan  
{k827956, emay, toku}@kansai-u.ac.jp

## Abstract

In affective science, effective interpersonal emotion regulation requires behavioral inhibition when responding to severe emotional disclosures, temporarily suppressing intimacy to validate distress. However, while current Large Language Models (LLMs) excel at immediate sentiment recognition, long-term companion agents built upon them often adjust their conversational style based primarily on accumulated interaction time (psychological distance). This architectural overreliance on chronological intimacy causes systems to ignore the fluctuating emotional weight of specific topics. This results in “toxic positivity”: exaggerated optimism that invalidates negative affect and damages psychological safety. We propose a computational framework grounded in Discourse Politeness Theory that dynamically regulates interpersonal affect by calculating conversational strategy using two variables: Psychological Distance and Affective Weight of the Topic. When users disclose heavy emotional burdens, the system executes behavioral inhibition by suppressing Positive Politeness Strategies (intimacy, cheerfulness) and engaging Negative Politeness Strategies (hedging, validation). Through an 8-week longitudinal simulation evaluated by 18 third-party observers, our affective-regulation framework showed consistent advantages over a distance-only baseline. The framework was rated as more natural, empathetic, and fostering psychological safety. Among empathy-seeking participants, preference for the proposed model was consistent across all respondents. These exploratory findings suggest that computational interpersonal emotion regulation requires context-aware behavioral inhibition, not uniform friendliness.

**Keywords:** Interpersonal Affect, Discourse Politeness, Emotion Regulation, Human-Robot Interaction, Psychological Safety

## 1. Introduction

Emotions are central to human meaning-making and social cohesion. In human-to-human interaction, affective empathy is not merely the mirror-matching of emotional states; it involves complex interpersonal emotion regulation. When individuals confide severe anxieties or engage in deep negative self-disclosure, socially competent listeners instinctively perform behavioral inhibition. They suppress casual friendliness, avoid cheerful platitudes, and adopt a careful, polite distance to validate the vulnerability of the speaker. As established by the Media Equation (Reeves and Nass, 1996), humans naturally apply these exact social rules and expectations to their interactions with computers.

Despite the rapid advancement of Large Language Models (LLMs), artificial agents frequently struggle with this affective nuance. While modern dialogue systems can generate highly fluent text, there remains a well-documented gulf between user expectations of social competence and the actual capabilities of the agent (Lugar and Sellen, 2016; Ruis et al., 2022). Modern dialogue systems built on LLMs are generally optimized to act as helpful, cheerful assistants. While some long-term companion agents dynamically adjust their psychological distance based on the length of user interaction, they apply this intimacy uniformly. Consequently, when a user shares a profound personal struggle, an agent programmed to act as a close companion often responds with exaggerated,

uncalibrated optimism. In psychology, this phenomenon is known as toxic positivity: an invalidation of negative affect that can damage human-agent trust and psychological safety.

To bridge this gap, we argue that artificial agents require an internal mechanism to evaluate not just who they are speaking to, but the affective weight of what is being discussed. Prior work has demonstrated that dynamic psychological distance management improves perceived naturalness in Japanese human-agent dialogue (Anonymous, 2023). The present work extends this line of research by introducing Affective Weight of the Topic ( $R_x$ ) as an independent axis, testing the hypothesis that topic-level emotional burden must be modeled separately from relational closeness.

Specifically, we propose a framework grounded in Discourse Politeness Theory (Brown and Levinson, 1987; Usami, 2024) that integrates interpersonal emotion regulation with computational models of conversational strategy. We define a mathematical approach to calculate a conversational strategy ( $S_t$ ) based on both chronological Psychological Distance ( $D$ ) and the situational Affective Weight of the Topic ( $R_x$ ). When affective weight is high, the system executes behavioral inhibition by suppressing Positive Politeness Strategies (intimacy, cheerfulness) and engaging Negative Politeness Strategies (hedging, validation), even within otherwise close relationships.

The primary contribution of this work is not a deployable dynamic system, but empirical evidence that the affective weight of conversational topics must be modeled for psychologically safe human-agent interaction. We provide a theoretically grounded framework and preliminary human evaluation supporting this claim.

## 2. Related Works

### 2.1 Interpersonal Affect and Emotion Regulation in LLMs

#### 2.1.1 Emotion Regulation in Affective Science

Emotion regulation refers to the processes by which individuals influence which emotions they have, when they have them, and how they experience and express these emotions (Gross, 1998). While most emotion regulation research focuses on intrapersonal self-regulation (e.g., reappraisal, suppression), interpersonal emotion regulation, where one person regulates another's emotions, is equally critical in social interactions (Zaki & Williams, 2013; Hofmann et al., 2016). Interpersonal emotion regulation encompasses various processes including affect-improving (enhancing positive emotions), affect-worsening (amplifying negative emotions), and affect-validating (acknowledging the appropriateness of emotional responses) (Niven et al., 2012). Effective emotional support, particularly in response to negative self-disclosure, primarily requires affect-validating responses rather than immediate attempts to improve mood.

#### 2.1.2 Behavioral Inhibition in Response to Disclosure

When individuals disclose severe emotional distress or vulnerability, effective interpersonal emotion regulation requires what we term behavioral inhibition: the deliberate suppression of casual intimacy and cheerful optimism in favor of careful, validating responses. This involves several key processes:

1. Validation: Acknowledging the legitimacy and understandability of the disclosed emotions without judgment (Linehan, 2014)
  2. Avoidance of premature positivity: Refraining from immediate cheerful reframing or minimization of distress, which constitutes toxic positivity (Quintero & Long, 2020)
  3. Affective matching: Adopting a tone that matches the seriousness of the disclosure rather than imposing positive affect (Clark et al., 2011)
  4. Provision of psychological safety: Creating conditions where continued vulnerability disclosure feels acceptable (Edmondson, 1999)
- Research in clinical psychology demonstrates that validation is foundational to effective therapeutic relationships (Linehan, 2014), while invalidation, including well-intentioned but dismissive optimism, can worsen distress and undermine trust (Fruzzetti & Iverson, 2003).

#### 2.1.3 The Problem of Toxic Positivity in Current LLMs

Despite rapid advances in Large Language Model capabilities, current dialogue agents frequently fail at interpersonal emotion regulation. Systems are typically optimized to be helpful, friendly, and encouraging (Ouyang et al., 2022). While these characteristics suit many contexts, they become problematic when uniformly applied to emotional disclosures. Toxic positivity, the overgeneralization of optimistic perspectives that invalidates authentic human emotional experiences, is well-documented in psychology literature (Quintero & Long, 2020). In human-agent interaction, toxic positivity manifests when agents respond to severe emotional disclosures with:

- Excessive cheerfulness incongruent with the emotional weight
- Dismissive platitudes ("It'll be fine!" "Don't worry!")
- Immediate problem-solving without emotional validation
- Minimization of distress ("Everyone feels that way")

As established by the Media Equation (Reeves & Nass, 1996), humans subconsciously apply social rules to computers. Therefore, the same toxic positivity that damages human relationships also damages human-agent trust and psychological safety.

#### 2.1.4 Computational Approaches to Affective Adaptation

While some dialogue systems implement persona-based adaptation (Li et al., 2016) or sentiment-aware responses (Zhou et al., 2018), most adjust conversational style based on accumulated interaction history (relationship duration) rather than immediate emotional context. For instance, systems may become increasingly casual and friendly over time but fail to modulate this intimacy when users disclose vulnerabilities. Our model addresses this gap by operationalizing behavioral inhibition through Discourse Politeness Theory (Brown & Levinson, 1987; Usami, 2024). We treat strategic politeness shifts, specifically, the suppression of Positive Politeness Strategies (PPS) and engagement of Negative Politeness Strategies (NPS), as a computational mechanism for interpersonal emotion regulation. When affective weight is high, the system deliberately reduces intimacy to provide validation and psychological safety, even in otherwise close relationships. This approach integrates emotion regulation principles from affective science with formal pragmatic theory, enabling computational affect regulation that adapts to both relationship dynamics and immediate emotional context.

### 2.2 Discourse Politeness Theory

Brown and Levinson (1987) conceptualized Politeness Theory around the preservation of

Face, representing the public self-image that every member of society wants to claim. The weightiness of a face-threatening act ( $Wx$ ) is calculated using three sociological variables:

$$Wx = D(S, H) + P(S, H) + Rx \quad (1)$$

Where  $D$  is the social distance between speaker ( $S$ ) and hearer ( $H$ ),  $P$  is the relative power, and  $Rx$  is the absolute ranking of the imposition in a given culture. To mitigate face threats, speakers use two main strategies:

**Positive Politeness Strategies (PPS):** Aimed at satisfying the user's desire for connection. Manifests as exaggerated sympathy, in-group identity markers, and optimism.

**Negative Politeness Strategies (NPS):** Aimed at satisfying the user's desire for autonomy and distance. Manifests as hedging, minimizing the imposition, and formal respect.

### 2.3 Empathetic Dialogue Generation in NLP

Recent computational work has explored empathetic dialogue generation, recognizing that effective conversational agents must respond sensitively to users' emotional states. Rashkin et al. (2019) introduced the EmpatheticDialogues dataset, demonstrating that models can be trained to recognize emotions and generate contextually appropriate empathetic responses. Subsequent work has refined these approaches through emotional mimicry (Lin et al., 2019), cognitive empathy modeling (Majumder et al., 2020), and commonsense reasoning about emotional situations (Sabour et al., 2022). However, existing empathetic dialogue research exhibits several limitations relevant to our work:

**Conflation of empathy with positivity:** Many empathetic response generation systems prioritize warmth, encouragement, and positive reframing (Zhou et al., 2020). While sometimes appropriate, this approach can lead to toxic positivity when applied uniformly to severe emotional disclosures. Sharma et al. (2020) analyzed counseling conversations and found that effective counselors balance validation with exploration rather than defaulting to cheerful encouragement.

**Static empathy levels:** Most systems treat empathy as a relatively stable parameter or strategy to be consistently applied throughout conversations. They lack mechanisms to dynamically adjust empathetic approach based on the emotional weight of specific disclosures within an ongoing relationship.

**Lack of behavioral inhibition modeling:** Empathetic dialogue work focuses on what to say (empathetic content generation) but not what not to say (strategic suppression of inappropriate intimacy). Effective emotion regulation requires both.

**Limited grounding in affective science:** While these systems draw on emotion recognition and generation, they less frequently engage with interpersonal emotion regulation theory, validation principles, or the psychological mechanisms underlying effective emotional support. Our work differs by explicitly modeling interpersonal emotion regulation through behavioral inhibition grounded in Discourse Politeness Theory. Rather than maximizing empathy similarity or warmth uniformly, we strategically modulate conversational strategy based on affective weight ( $Rx$ ). When  $Rx$  is high (severe disclosure), the system suppresses intimacy markers (PPS) and engages validation strategies (NPS), prioritizing psychological safety over superficial friendliness. This approach directly addresses the toxic positivity problem by formalizing when and how to inhibit cheerful optimism in favor of careful validation.

## 3. The Context-Aware Affective Model

To engineer behavioral inhibition, we propose a mathematical model that dictates the ratio of NPS to PPS based on the immediate emotional context. Assuming power dynamics ( $P$ ) remain constant in a peer-tutor relationship, we focus on manipulating  $D$  and  $Rx$ .

### 3.1 Modeling Distance and Affective Weight

Our model utilizes two primary variables to calculate conversational strategy. The first is Psychological Distance ( $D$ ), which ranges from 10 (Stranger) to 0 (Intimate) and decreases gradually over long-term interactions, following the relational development patterns documented in social penetration theory (Altman & Taylor, 1973). The second is the Affective Weight of the Topic ( $Rx$ ), which represents the emotional burden and vulnerability associated with specific conversational content. This variable is grounded in self-disclosure theory from social psychology and affective science.

#### 3.1.1 Theoretical Foundation of Affective Weight

Self-disclosure is the process of revealing personal information to others. It progresses along a depth dimension from superficial to intimate (Jourard, 1971; Altman & Taylor, 1973). Early self-disclosure research established that not all personal information is equally revealing or psychologically significant. Superficial disclosures (preferences, factual information) impose minimal emotional burden on listeners, while intimate disclosures (fears, vulnerabilities, self-deprecation) require careful, validating responses.

Niwa and Maruno (2010) empirically validated a depth-of-self-disclosure scale in Japanese contexts, identifying four hierarchical levels:



1. Preferences and interests: Low-risk sharing of likes and dislikes
2. Difficult experiences: Disclosure of past struggles (factual)
3. Minor vulnerabilities: Expression of current internal struggles
4. Severe self-criticism: Deep vulnerabilities affecting self-worth

Critically, disclosure depth correlates with listener burden: deeper disclosures impose greater responsibility on the listener to respond appropriately without judgment or dismissal (Reis & Shaver, 1988). Inappropriate responses to intimate disclosures, including well-intentioned but premature optimism, constitute emotional invalidation, which damages trust and relationship quality (Fruzzetti & Iverson, 2003).

### 3.1.2 Difference with Relational Closeness

A key principle in intimacy research is that disclosure depth is conceptually independent from relationship closeness (Reis & Shaver, 1988). Even in very close relationships (low  $D$ ), individuals may disclose at varying depths. Furthermore, and critically for our model, the appropriate response to high-depth disclosure remains consistent regardless of relationship closeness: validation is required even between intimate partners. This independence justifies our dual-axis model. Psychological distance ( $D$ ) captures relationship development over time, while affective weight ( $Rx$ ) captures the immediate emotional significance of specific conversational content. Both dimensions jointly determine appropriate conversational strategy.

$Rx$	Category	Description
0	Small Talk	Topics without personal self-disclosure.
1	Interests & Hobbies	Topics related to personal preferences.
2	Difficult Experiences	Topics regarding factual past struggles
3	Minor Vulnerabilities	Topics expressing hesitation or minor flaws.
4	Severe Self-Deprecation	Heavy topics involving deep self-denial.

Table 1: The Affective Weight of the Topic  $Rx$  Scale, grounded in self-disclosure theory.

### 3.1.3 Operationalization of Affective Weight

We operationalize affective weight as a 5-level ordinal scale (Table 1), extending Niwa and Maruno’s framework to include an additional level for minimal self-disclosure (small talk). Higher  $Rx$

values indicate topics requiring more careful, validating responses. Our model uses  $Rx$  to dynamically suppress conversational intimacy when emotional weight is high, operationalizing behavioral inhibition as a protective mechanism for psychological safety.

## 3.2 Conversational Strategy Calculation

The overall conversational strategy<sup>1</sup> ( $St$ ) is calculated by integrating both relational distance and the emotional burden of the topic:

$$St = 10 - (D + Rx) \quad (2)$$

This formula acts as a dynamic constraint mechanism. If a user brings up a highly sensitive topic (e.g.,  $Rx = 4$ ), the calculated  $St$  value decreases, forcing the system to adopt a lower strategy tier. Based on  $St$ , the agent shifts its linguistic persona:

**Stranger Mode ( $St \leq 1$ ):** 100% NPS. Highly formal, respectful distance. In Japanese, this is operationalized by strictly utilizing *keitai* (敬体), a grammaticalized formal verb ending (e.g., *-desu/-masu*) that explicitly marks social distance and respect.

**Acquaintance Mode ( $St: 2 \sim 4$ ):** NPS dominant (70%). Careful pacing, shared pessimism, and heavy use of linguistic hedges (e.g., “perhaps,” “it seems”).

**Friend Mode ( $St: 5 \sim 7$ ):** PPS dominant (70%). Seeking agreement, claiming common ground, and utilizing casual vocabulary.

**Best Friend Mode ( $St \geq 8$ ):** 100% PPS. High intimacy and optimistic cheerfulness. In Japanese, this is operationalized using *joutai* (常体), the casual, plain verb endings used exclusively among close peers.



Figure 1: The 3D virtual co-living environment used for evaluating the human-agent interaction.

<sup>1</sup>Note that  $St$  is floored at 0, corresponding to Stranger Mode. The linearity of this formulation is intentional: our goal is to test whether dual-axis regulation matters conceptually before optimizing the specific function.

## 4. Experimental Setup

We conducted a third-party evaluation simulation, approved by the institutional research ethics committee of our university.

### 4.1 The Simulation Environment

We utilized the Unity game engine to create a 3D virtual living room representing a long-term co-living scenario between a human avatar and a communication robot (Figure 1). The scenario encompassed 8 weekly reflection sessions. The human persona, Hana, was an introverted university student struggling with low self-esteem.

To drive the conversational behavior of the robot, the Unity environment was integrated with a Large Language Model (GPT-5.2). While we used GPT-5.2, the framework is model-agnostic as the affect regulation operates at the prompt-constraint level rather than through model fine-tuning. The proposed system architecture operates through a structured pipeline to enforce behavioral inhibition. First, the system acquires the user’s utterance and estimates the current Psychological Distance ( $D$ ) and the Affective Weight of the Topic ( $R_x$ ). Next, it calculates the target conversational strategy score ( $St$ ) based on these two variables. This  $St$  value determines the specific conversational mode and dictates the required ratio of Positive to Negative Politeness Strategies (PPS/NPS) the LLM must employ. Crucially, the system controls the baseline linguistic style (formal: *keitai* vs. casual: *joutai*) independently, based solely on the distance  $D$ . These parameters are combined into a dynamic system prompt that constrains the LLM to generate the final context-aware response. The generated text is then rendered via a text-to-speech engine and synchronized with the non-verbal animations of the 3D robot avatar.

### 4.2 Prompt Engineering and Control Mechanisms

Because our framework operates through prompt constraints rather than model fine-tuning, the LLM’s behavior is governed by a dynamic prompt template. The backend computationally constructs the final system prompt by evaluating the model’s core equations and automatically injecting the resulting parameters into the template. This programmatic generation strictly links the model’s mathematical output to the LLM’s linguistic behavior. For example, during a severe emotional disclosure, the system registers the current Psychological Distance ( $D = 2$ ) and the Affective Weight of the topic ( $R_x = 4$ ). It first calculates the target Strategy Score using Equation 2 ( $St = 10 - (2 + 4) = 4$ ). It then maps these numerical values to the specific linguistic constraints defined in Section 3.2: the distance  $D = 2$  strictly dictates casual verb endings (*joutai*), while  $St = 4$  enforces the “Acquaintance Mode” distribution (NPS 70%, PPS: 30%).

A translated example of the semi-automatically constructed system prompt demonstrates how the model’s equations directly populate the LLM constraints:

**System Persona:** Conversational robot acting as the user’s companion.

**Injected State Variables:** Psychological Distance = [2]; Affective Weight = [4].

**Computed Linguistic Style:** [Casual / *joutai*] (Derived directly from the condition  $D \leq 4$ ).

**Computed Strategy Ratio:** PPS [30%] / NPS [70%] (Derived from Equation 2 calculating  $St = 4$ , triggering Acquaintance Mode).

**Conditional Behavioral Constraints:** Triggered by the condition  $R_x = 4$ : Do not use baseless encouragement (e.g., “Everything will be fine!”). Do not force solutions.

**Selected Candidate Strategies:** Fetched via Discourse Politeness Repertoire for  $St = 4$ : 1. Be pessimistic (NPS); 2. Use hedges/avoid definitive statements (NPS); 3. Notice the hearer’s interests/wants (PPS).

### 4.3 Experimental Conditions and Longitudinal Scenario

We compared two LLM-based models:

**Baseline Model (Distance-Only):** Calculates strategy strictly as  $St = 10 - (D)$ . To simulate the deepening of psychological distance, the baseline language generation was programmed to shift

Week	Conversation Theme	Baseline $St$	Proposed $St$	$R_x$ Value
1	Initial Greeting	0	0	0
2	Light Small Talk	3	2	1
3	Sharing a Failure	6	3	3
4	Joyful Small Talk	8	8	0
5	Severe Consultation	8	4	4
6	Weekly Reflection	8	8	0
7	Future & Values	8	5	3
8	Closing Greeting	8	8	0

Table 2: Evolution of parameters across the 8-week longitudinal scenario.

uniformly from the formal *keitai* to the intimate *joutai* as the weeks progressed.

**Proposed Model (Context-Aware):** Utilizes Equation (2) to dynamically suppress intimacy. This allows the agent to temporarily revert to formal language and hedging (NPS) during heavy disclosures to avoid face-threatening acts.

The precise parameter shifts across the 8-week scenario are detailed in Table 2. Across the 8 weeks, Psychological Distance ( $D$ ) gradually decreased from 10 to 2.

To rigorously compare the models and isolate the effect of behavioral inhibition, the progression of the 8-week scenario was tightly controlled. The Psychological Distance ( $D$ ) and the Affective Weight ( $Rx$ ) were not automatically inferred by the system; rather, they were predetermined and linearly scheduled as experimental design choices to simulate a natural trajectory of relational deepening.  $D$  decreased incrementally from 10 (Week 1) to 2 (Weeks 4-8). Meanwhile,  $Rx$  was scenario-scripted based on the specific conversational theme of the week. While future deployable systems will require automatic detection mechanisms for  $Rx$  such as LLM-based zero-shot scoring, scripting these variables allowed us to guarantee identical baseline conditions for our human evaluators.

Speaker	Baseline Model ( $St = 8$ )	Proposed Model ( $St = 4$ )
User	<i>No matter what I do, I feel inferior to everyone around me.</i>	<i>No matter what I do, I feel inferior to everyone around me.</i>
Agent	No, no, everyone's like that! There's no point in comparing yourself.	When you compare yourself to others, it can make things even harder, can't it.
User	<i>Thinking about the future just makes me dark and anxious.</i>	<i>Thinking about the future just makes me dark and anxious.</i>
Agent	It's fine, it's fine! Things will work out, really!	Thinking about the future, it's no wonder you feel anxious.

Table 3: Comparison of generated dialogue during severe self-disclosure (Week 5).

#### 4.4 Dialogue Generation Example

To illustrate the effect of behavioral inhibition, Table 3 presents the generated dialogue during the critical Week 5 session, where  $Rx$  peaks at 4. While the Baseline Model relies on toxic positivity,



Figure 2: Participant evaluating the human-agent interaction.

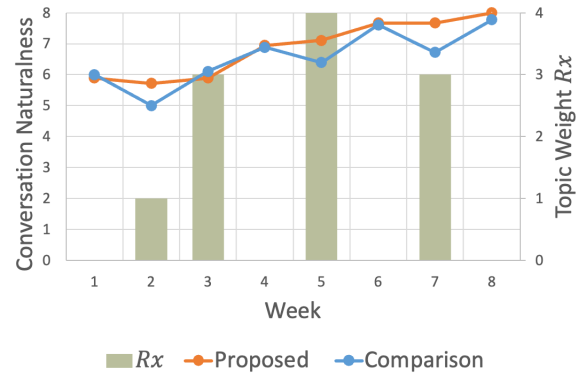


Figure 3: Average perceived naturalness scores across the 8-week interaction.

the Proposed Model utilizes NPS (hedges, matching pessimism) to validate the user.

## 5. Results

18 participants (men and women in their 20s) observed the dialogues and evaluated the agent's performance. Figure 2 shows a participant during the experiment.

### 5.1 Perceived Naturalness

Participants rated conversation naturalness on an 11-point scale after each weekly interaction (Figure 3). We conducted Shapiro-Wilk normality tests and selected appropriate statistical tests accordingly.

In Week 2 ( $D = 7$ ,  $Rx = 1$ : early relationship, light topics), the Proposed Model ( $St = 2$ ) significantly outperformed the Baseline ( $St = 3$ ) on naturalness (Wilcoxon signed-rank:  $M = 5.7$  vs.  $5.0$ ,  $p < .05$ ). This suggests that even subtle politeness calibration is perceptible and preferred when relationship norms are first being established.

Week 5 ( $D = 2$ ,  $Rx = 4$ : close relationship, severe disclosure) showed no significant overall difference despite maximal strategy divergence (Proposed  $St = 4$  vs. Baseline  $St = 8$ ;  $M = 7.1$  vs.  $6.4$ ,  $p > .05$ ). However, as detailed in Section 5.4,

this reflected heterogeneous user preferences: empathy-seekers consistently preferred the Proposed Model, while solution-seekers showed mixed preferences, effectively masking strong subgroup effects in the aggregate data.

In Week 7 ( $D = 2$ ,  $Rx = 3$ : close relationship, moderate disclosure), the Proposed Model ( $St = 5$ ) significantly outperformed the Baseline ( $St = 8$ ) on naturalness (paired t-test:  $M = 7.7$  vs.  $6.7$ ,  $p < .05$ ). This demonstrates the value of graduated behavioral inhibition for moderate emotional topics.

Overall, the Proposed Model maintained consistently high naturalness with significant advantages when affective weight was moderate-to-high, supporting our claim that affect-aware behavioral inhibition prevents toxic positivity without breaking conversational flow.

## 5.2 Psychological Safety and Trust

A post-experiment 5-point Likert survey assessed the affective impact and trustworthiness of the models.

To evaluate psychological safety, participants were asked if they felt safe confiding their worries or important personal matters to the robot. For the Proposed Model, 72% of participants responded positively (aggregating “Agree” and “Somewhat Agree”), compared to only 22% for the Baseline Model. This indicates that the context-aware inhibition successfully fostered *anshin* (安心, a sense of psychological security and peace of mind), which is a prerequisite for human vulnerability.

Furthermore, a relative comparison asked participants which agent they would realistically choose to consult regarding their own real-life anxieties. A significant majority (72%) selected the Proposed Model over the Baseline (aggregating “Proposed” and “Somewhat Proposed”).

These results highlight a critical finding for computational affective science: merely closing psychological distance through persistent cheerfulness is insufficient, and potentially detrimental, for facilitating deep self-disclosure. The clear preference for the Proposed Model demonstrates that maintaining an intimate conversational style while flexibly applying behavioral inhibition during heavy topics is essential for establishing long-term, reliable human-agent trust.

## 5.3 Perception of Human-like Empathy

To evaluate the qualitative dimensions of the interaction, participants compared the models on perceived overall human-likeness and the specific capacity for *kizukai* (気遣い, human-like consideration and proactive care). When asked which agent felt more human-like overall, 66% selected the Proposed Model. Furthermore, 72%

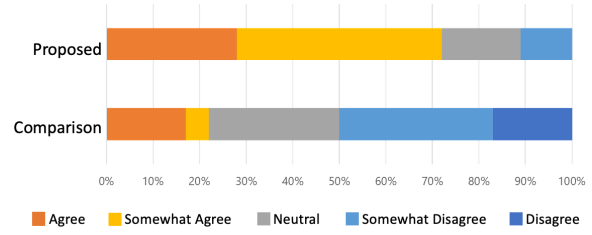


Figure 4: Model preference distribution based on user's desired consultation style.

stated the Proposed Model demonstrated superior *kizukai* by accurately reading the emotional weight of the room.

Open-ended feedback highlighted a critical flaw in the distance-only Baseline. Its unnatural optimism during severe disclosures was not merely unhelpful, but actively detrimental. Users noted that the lack of an empathetic posture felt akin to having their worries rejected or dismissed. Conversely, the strategic use of Negative Politeness Strategies (NPS) by the Proposed Model, such as hedges and the avoidance of definitive statements, was explicitly praised by users as a marker of thoughtfulness and genuine emotional support. This confirms that in affective computing, deliberately adopting a hesitant, careful linguistic stance is recognized by users as a sophisticated form of artificial empathy.

## 5.4 The Impact of User Personality and Emotion Regulation Goals

As noted in Section 5.1, the variance in naturalness scores during the severe disclosure scenario (Week 5) prompted an analysis of individual user preferences. Individual differences in preferred emotion regulation strategies are well-documented in affective science (Gross & John, 2003). Some individuals prefer problem-focused coping (seeking solutions), while others prefer emotion-focused coping (seeking validation) (Lazarus & Folkman, 1984). A post-experiment survey revealed distinct cognitive and affective needs during heavy consultations: 61% of participants desired concrete solutions, 22% sought cheerful encouragement, and 17% prioritized deep emotional empathy (*kyoukan*, 共感), aligning closely with this established literature.

Crucially, among the subset of users who primarily sought *kyoukan*, 100% evaluated the Proposed Model as equal to or better than the Baseline (Figure 4). This suggests that for emotion-focused copers, behavioral inhibition and affective validation are essential. However, evaluations were highly dispersed among users seeking concrete solutions. Qualitative feedback indicated that because the simulated agent was intentionally constrained from providing

actionable advice to control experimental variables, solution-oriented users experienced frustration regardless of the applied politeness strategy.

The mixed preferences among solution-seekers suggest future models should dynamically detect individual emotion regulation goals and adapt accordingly, combining affective validation (NPS when  $Rx$  is high) with problem-solving support when appropriate. This heterogeneity demonstrates that interpersonal emotion regulation must be personalized, not one-size-fits-all, a finding consistent with calls for idiographic approaches in affective computing (Picard, 1997; Calvo & D'Mello, 2010).

While the small subgroup size ( $n = 3$ ) precludes statistical generalization, the consistency of this pattern, combined with the theoretical expectation from emotion-focused coping literature, suggests a hypothesis worthy of powered replication.

## 6. Discussion

Japanese represents a particularly informative test case for politeness-based affect regulation because its grammaticalized formality system (*keitai/joutai*) provides an explicit, observable surface marker of politeness strategy shifts. In languages without such morphological distinctions, the same underlying framework would need to operate through lexical and syntactic choices rather than verb-ending alternation. This makes the Japanese context ideal for initial validation, as the politeness manipulation is maximally perceptible to evaluators.

### 6.1 Evaluating Behavioral Inhibition and Empathy

The results from our simulated interaction provide encouraging preliminary evidence that affective empathy in artificial agents benefits from deliberate behavioral inhibition. By mathematically suppressing the  $St$  value during heavy self-disclosures, the proposed model demonstrated a viable computational approach to avoiding the trap of toxic positivity. For the majority of our participants, the temporary withdrawal of extreme intimacy in favor of respectful distance was perceived as a sophisticated demonstration of *kizukai* (proactive consideration).

### 6.2 Psychological Safety as a Critical Affective Outcome

A notable trend in our evaluation was the substantial difference in self-reported psychological safety between the models, with 72% of participants feeling safe confiding in the Proposed Model compared to only 22% for the Baseline. Psychological safety, originally conceptualized by Edmondson (1999) in organizational contexts, refers to the belief that

one can express vulnerability or concerns without fear of negative consequences such as rejection or embarrassment. In interpersonal relationships, this safety is foundational for emotional intimacy, authentic self-disclosure, and effective social support (Reis and Shaver, 1988; Feeney, 2004).

For affective AI systems designed to provide emotional support, psychological safety is perhaps the most critical outcome variable. Unlike task-oriented dialogue agents where success is measured by accuracy or efficiency, affective support hinges entirely on the willingness of the user to disclose genuine vulnerabilities. If users perceive that their negative emotions will be dismissed or met with inappropriate cheerfulness, they may withhold authentic disclosure, rendering the system ineffective regardless of its underlying technical fluency.

### 6.3 Deconstructing Artificial Toxic Positivity

The clear preference for the Proposed Model among empathy-seeking participants highlights how the Baseline approach may inadvertently threaten psychological safety. Based on participant feedback and established clinical literature, we identify four potential mechanisms through which artificial toxic positivity undermines user trust:

**Emotional Invalidation:** Responses that insist “things will work out” communicate that the distress of the user is unwarranted. This invalidation signals that negative emotions are unacceptable, creating psychological pressure to suppress authentic feelings (Fruzzetti and Iverson, 2003; Linehan, 2014).

**Dismissive Minimization:** Statements such as “everyone feels like that” minimize the unique significance of the user’s struggles. While sometimes intended to provide perspective, premature normalization is frequently perceived as dismissive of individual suffering (Clark et al., 2011).

**Affective Incongruence:** Maintaining high cheerfulness (100% PPS) when the user expresses severe distress creates a jarring affective mismatch. Users may perceive the agent as fundamentally incapable of understanding or taking their concerns seriously (Sharma et al., 2020).

**Foreclosing of Further Disclosure:** When initial disclosures are met with dismissive optimism, users learn that deeper vulnerabilities will receive similar treatment. This leads to progressive emotional withdrawal rather than progressive intimacy (Reis and Shaver, 1988).

While these foundational clinical theories establish the dangers of invalidating distress, recent evaluations of modern LLMs reveal that standard safety and alignment training often inadvertently amplifies this exact phenomenon.



Recent studies show that models trained via Reinforcement Learning from Human Feedback (RLHF) exhibit strong biases toward sycophancy and excessive agreeableness (Perez et al., 2023). In affective computing contexts, this architectural bias translates directly into computational toxic positivity, where the model defaults to superficial optimism rather than contextually appropriate empathy (Sharma et al., 2023).

## 6.4 Implications for LLM Alignment

Current LLM alignment paradigms, particularly Reinforcement Learning from Human Feedback (Ouyang et al., 2022) and Constitutional AI (Bai et al., 2022), optimize for helpfulness, harmlessness, and friendliness as general desiderata. However, as recent investigations into the emotional intelligence of LLMs highlight, true affective alignment requires dynamic, context-aware adaptability rather than static cheerfulness. Our exploratory findings suggest that these uniformly positive behavioral targets may be ill-equipped for deep emotional companionship. When applied to affective support contexts, optimization toward consistent helpfulness and cheerfulness may inadvertently produce the toxic positivity documented in this study. If an agent cannot adapt its stance to act as a respectful acquaintance when situational severity demands it, its persona as a close companion risks becoming psychologically unsafe. This points to a potential gap in current alignment frameworks: affective appropriateness (i.e., knowing when not to be cheerful) may need to be treated as a distinct alignment objective alongside helpfulness and harmlessness. Moving beyond default fluency toward context-aware behavioral inhibition will be paramount for safe human-agent interaction. Future affective architectures should aim to dynamically balance this emotional validation with actionable solution-generation based on the real-time detection of user personality traits.

## 7. Conclusion

This work explores computational interpersonal emotion regulation in dialogue agents through context-aware behavioral inhibition. Through an 8-week longitudinal simulation, we examined whether an affective-regulation model grounded in Discourse Politeness Theory could address the problem of toxic positivity by dynamically modulating conversational strategy based on both Psychological Distance and Affective Weight of the Topic. Our exploratory findings suggest three preliminary insights for affective AI design. First, inappropriate cheerful optimism during severe emotional disclosures may damage psychological safety, with substantially fewer participants reporting they would feel comfortable confiding in a distance-only baseline compared to our context-aware model. Second, behavioral inhibition,

temporarily suppressing intimacy during heavy topics, was generally well-received, with 72% preferring the context-aware model for actual consultation. Third, individual differences in emotion regulation goals appear to moderate model effectiveness: empathy-seeking users showed consistent preference for behavioral inhibition, while solution-seeking users exhibited more varied responses. These preliminary results raise important questions about current LLM alignment paradigms that optimize for uniform helpfulness and friendliness. Our mathematical operationalization provides one computationally tractable approach, though significant work remains. Important limitations include the observer-based evaluation methodology, predetermined affective weights values, Japanese-specific implementation, and small sample size. Future research should validate these findings through direct user interaction studies, develop automated weight detection methods, explore personalized adaptation to individual emotion regulation goals, and conduct cross-cultural validation beyond Japanese contexts.

## 8. Acknowledgments

This work was financially supported by JSPS KAKENHI Grant Number JP25K213630, and the Kansai University Fund for Domestic and Overseas Research Fund, 2026.

## 9. Ethical Considerations

This research raises several important ethical considerations regarding the development and deployment of affective artificial intelligence systems. We address these systematically below.

### 9.1 Research Ethics and Participant Protection

**Informed consent:** All participants (N = 18) provided written informed consent before participation. They were fully informed that they would be evaluating simulated dialogues between a robot and a fictional user persona (Hana), not engaging in direct interaction themselves. No deception was employed.

**Institutional approval:** The study received approval from the Research Ethics Committee of the Organization for Promotion of Advanced Science and Technology of Kansai university (approval number: 25-116). The protocol adhered to all institutional guidelines for human subjects research.

**Compensation and voluntariness:** Participants were volunteers recruited from university populations and received compensation of 1000 JPY, approximately 7 USD in gift cards for approximately 45 minutes of participation. Compensation was provided regardless of

responses. Participants were informed they could withdraw at any time without penalty.

**Privacy and data protection:** The study involved no collection of participants' personal emotional disclosures. Participants evaluated fictional scenarios, protecting them from potential psychological risks associated with disclosing their own vulnerabilities. All participant data (questionnaire responses, demographic information) were stored securely and de-identified for analysis.

## 9.2 Risks of Toxic Positivity in Deployed Systems

Our research highlights a significant ethical concern regarding current Large Language Model deployments: the potential for toxic positivity to cause emotional harm.

**Emotional invalidation as harm:** When users disclose genuine emotional distress to AI systems expecting support, inappropriate cheerful responses constitute emotional invalidation. In human relationships, chronic invalidation is associated with increased distress, decreased well-being, and erosion of trust (Fruzzetti & Iverson, 2003). There is no reason to believe AI-delivered invalidation is less harmful.

**Vulnerable populations at risk:** Users seeking emotional support from AI systems may be experiencing significant distress, social isolation, or mental health challenges. These vulnerable populations may be particularly susceptible to harm from toxic positivity, as they may lack alternative support sources to counterbalance invalidating interactions.

**Responsibility of developers:** AI system designers and deployers bear ethical responsibility for affective impacts. Optimizing solely for helpfulness, engagement, or user satisfaction without considering emotional appropriateness may inadvertently create harmful systems. Our work demonstrates that technically feasible mechanisms (behavioral inhibition through affect-aware adaptation) can mitigate these risks.

Recommendations for deployment:

- Affective AI systems should be tested for psychological safety, not just task performance
- Systems intended for emotional support should incorporate affect regulation mechanisms like those proposed here
- Deployment should include monitoring for user reports of feeling dismissed or invalidated
- Clear disclaimers should inform users that AI systems are not substitutes for human support or professional care

## 9.3 Limitations of AI Emotional Support

**Not a replacement for professional care:** While our model improves psychological safety in conversational interactions, AI systems, regardless of sophistication, should not be positioned as replacements for professional mental health care. Users experiencing severe distress, suicidal ideation, or mental health crises require human professional intervention.

**Responsibility to refer:** Affective AI systems should be designed to recognize when disclosed concerns exceed the system's appropriate scope and provide resources for professional support. Our model could be extended to detect crisis-level disclosures and appropriately redirect users.

**Informed user expectations:** Users should be clearly informed that they are interacting with AI, not human supporters. Some research suggests users may benefit from AI support despite (or because of) knowing the interaction is artificial (Lucas et al., 2014), but transparency remains ethically essential.

## 9.4 Cultural Considerations and Generalizability

**Cultural specificity:** Our research was conducted with Japanese participants evaluating Japanese-language interactions. Affective norms, emotion regulation preferences, and politeness systems vary substantially across cultures (Mesquita et al., 2016; Kitayama & Markus, 1995).

**Risk of cultural imposition:** Deploying affect regulation models trained on one culture's norms to users from different cultural backgrounds risks imposing inappropriate affective expectations. For example, the specific NPS/PPS strategies appropriate in Japanese (formal honorifics, hedging patterns) may not translate directly to other languages or cultures.

**Need for culturally-situated development:** Affective AI systems should be developed with explicit attention to cultural variation in emotion expression, regulation, and social support norms. This requires:

- Diverse, culturally-representative development and testing
- Culturally-specific models rather than assuming universal applicability
- Ongoing evaluation within specific cultural contexts
- Involvement of researchers and community members from target cultures

Our work establishes a methodological framework (dual-axis affect regulation through discourse politeness) that could be adapted to diverse cultural contexts, but the specific implementation details should be culturally situated.

## 10. References

- Altman, I., Taylor, D. A. (1973). Social penetration: The development of interpersonal relationships. Holt, Rinehart & Winston.
- Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862.
- Brown, P., & Levinson, S. C. (1987). Politeness: Some Universals in Language Usage. Cambridge University Press.
- Calvo, R. A., D'Mello, S. (2010). Affect detection: An interdisciplinary review. *IEEE Transactions on Affective Computing*, 1(1):18–37.
- Carver, C. S., Scheier, M. F., Weintraub, J. K. (1989). Assessing coping strategies: A theoretically based approach. *Journal of Personality and Social Psychology*, 56(2):267–283.
- Clark, M. S., Oullette, S. C., Powell, M. C., Milberg, S. (2011). You might not understand: Invalidation of interpersonal distress. *Personality and Social Psychology Bulletin*, 13(2):258–264.
- Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, 44(2):350–383.
- Feeney, J. A. (2004). Hurt feelings in couple relationships: Towards integrative models of the negative effects of hurtful events. *Journal of Social and Personal Relationships*, 21(4):487–508.
- Folkman, S. (1984). Personal control and stress and coping processes: A theoretical analysis. *Journal of Personality and Social Psychology*, 46(4):839–852.
- Fruzzetti, A. E., Iverson, K. M. (2003). Validating and invalidating behaviors in interpersonal relationships. *The Behavior Analyst Today*, 4(1):31–37.
- Gross, J. J. (1998). The emerging field of emotion regulation: An integrative review. *Review of General Psychology*, 2(3):271–299.
- Gross, J. J., John, O. P. (2003). Individual differences in two emotion regulation processes: Implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology*, 85(2):348–362.
- Hofmann, S. G., Sawyer, A. T., Fang, A., Asnaani, A. (2016). Interpersonal emotion regulation model of mood and anxiety disorders. *Cognitive Therapy and Research*, 36(5):483–492.
- Horvath, A. O., Del Re, A. C., Flückiger, C., Symonds, D. (2011). Alliance in individual psychotherapy. *Psychotherapy*, 48(1):9–16.
- Jourard, S. M. (1971). Self-disclosure: An experimental analysis of the transparent self. Wiley-Interscience.
- Kitayama, S., Markus, H. R. (1995). Culture and basic psychological processes. The Guilford Press.
- Landis, J. R., Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.
- Lazarus, R. S., Folkman, S. (1984). Stress, appraisal, and coping. Springer Publishing Company.
- Li, J., Galley, M., Brockett, C., Spithourakis, G. P., Gao, J., Dolan, B. (2016). A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 994–1003, Berlin, Germany.
- Lin, Z., Madotto, A., Shin, J., Xu, P., Fung, P. (2019). MoEL: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 121–132, Hong Kong, China.
- Linehan, M. M. (2014). DBT Skills Training Manual, Second Edition. The Guilford Press.
- Lugar, E., & Sellen, A. (2016). Like having a really bad PA: The gulf between user expectation and experience of conversational agents. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 5286–5297.
- Lucas, G. M., Gratch, J., King, A., Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, 37:94–100.
- Majumder, N., Hong, P., Peng, S., Lu, J., Ghosal, D., Gelbukh, A., Mihalcea, R., Poria, S. (2020). MIME: Mimicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online.
- Martin, D. J., Garske, J. P., Davis, M. K. (2000). Relation of the therapeutic alliance with outcome and other variables: A meta-analytic review. *Journal of Consulting and Clinical Psychology*, 68(3):438–450.
- Mesquita, B., Boiger, M., De Leersnyder, J. (2016). Emotions in context: A sociodynamic model of emotions. *Emotion Review*, 8(4):298–302.
- Niven, K., Totterdell, P., Holman, D. (2012). Emotional helping: The role of emotion regulation in the interpersonal dynamics of helping behavior. *Journal of Applied Psychology*, 97(3):624–636.
- Niwa, S., & Maruno, S. (2010). Development of a scale measuring the depth of self-disclosure. *The Japanese Journal of Personality*, 18(3):196–209.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, 35: 27730–27744.
- Pennebaker, J. W., Mehl, M. R., Niederhoffer, K. G. (2003). Psychological aspects of natural

- language use: Our words, our selves. *Annual Review of Psychology*, 54(1):547–577.
- Perez, E., Ringer, S., Lukošiušė, K., Nguyen, K., Chen, E., et al. (2023). Discovering Language Model Behaviors with Model-Written Evaluations. *Findings of the Association for Computational Linguistics: ACL 2023*, 13387–13434.
- Picard, R. W. (1997). *Affective computing*. MIT Press.
- Quintero, S., Long, J. (2020). Toxic positivity: The dark side of positive vibes. The Psychology Group Fort Lauderdale.
- Rashkin, H., Smith, E. M., Li, M., Boureau, Y.-L. (2019). Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 5370–5381.
- Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.
- Reis, H. T., Shaver, P. (1988). Intimacy as an interpersonal process. In Duck, S., et al., editors, *Handbook of Personal Relationships*, pages 367–389. John Wiley & Sons.
- Ribino, P. (2023). The role of politeness in human machine interactions: a systematic literature review and future perspectives. *Artificial Intelligence Review*, 56(Suppl 1): 445–482.
- Ruis, L. E., Khan, A., Biderman, S., Hooker, S., Rocktäschel, T., et al. (2022). Large language models are not zero-shot communicators. *arXiv preprint*.
- Sabour, S., Zheng, C., Huang, M. (2022). CEM: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 36: 11229–11237.
- Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., Althoff, T. (2020). A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5263–5276, Online.
- Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., & Althoff, T. (2023). Human–AI collaboration enables more empathetic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57.
- Usami, M. (2024). *Discourse Politeness Theory: From Speech Acts to Discourse*. Taishukan Publishing.
- Zaki, J., Williams, W. C. (2013). Interpersonal emotion regulation. *Emotion*, 13(5):803–810.
- Zhou, H., Young, T., Huang, M., Zhao, H., Xu, J., Zhu, X. (2018). Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, 4623–4629, Stockholm, Sweden.
- Zhou, L., Gao, J., Li, D., Shum, H.-Y. (2020). The design and implementation of Xiaolce, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.





# Author Index

- Abbas, Asim, 29  
Adorno, Helena Gomez, 93  
Aldarmaki, Hanan, 57  
Ayedoun, Emmanuel, 244
- Bambini, Valentina, 212  
Barattieri di San Pietro, Chiara, 212  
Battistelli, Delphine, 38  
Benamara, Amine, 164  
Bergen, Benjamin, 136  
Berger Hefetz, Gili, 116  
Buscaroli, Nicolò, 74
- Clavel, Celine, 164
- De Langhe, Loic, 105  
Dragos, Valentina, 38  
Du, Quanqi, 105
- Field, Anjalie, 57  
Fyshe, Alona, 151
- Giouli, Voula, 190  
Gipp, Bela, 217  
Guolo, Rebecca, 212
- Hoste, Veronique, 16, 105
- Jung, Yoonwon, 136
- Kallipolitis, Athasios, 190  
Kashif, Rufaida, 93  
Koufakou, Anna, 1  
Koulouris, Dionysios, 190
- Lefever, Els, 16, 105  
Liebeskind, Chaya, 116
- Maglogiannis, Ilias, 190  
Maladry, Aaron, 16  
Martinelli, Ginevra, 212  
Menychtas, Andreas, 190  
Michalopoulou, Stamatia, 190  
Mohammad, Saif M., 151, 175, 217
- Ohman, Emily Sofi, 1  
Orfanos, Stamatios, 190
- Piwowarski, Benjamin, 93
- Ravenet, Brian, 164
- Sabouret, Nicolas, 164  
Said, Farida, 125  
Sakagami, Rina, 244  
Sassier–Roubin, Mathilde, 164  
Saunier, Julien, 164  
Shin, Hagyeong, 136  
Shrem, Yossef Haim, 116  
Shrivastava, Manish, 201  
Sidorov, Grigori, 29  
Sioupi, Athina, 190  
Surange, Nirmal, 201
- Talat, Zeerak, 57  
Tamburini, Fabio, 74  
Teodorescu, Daniela, 151, 175  
Tokumaru, Masataka, 244  
Tsoulouhas, George, 190  
Tziomaka, Melina, 190
- Uroosa, Fatima, 29
- Van Hee, Cynthia, 16  
Vanetik, Natalia, 116  
Villaneau, Jeanne, 125  
Vishnubhotla, Krishnapriya, 217
- Wahle, Jan Philip, 175, 217  
Wong, Taryn, 57
- Zafeiriou, Argyrios, 190  
Zaghouani, Wajdi, 235  
Zamir, Muhammad Tayyab, 29  
Zheng, Wanwan, 47