Why ‘I’m Fine’ Doesn’t Always Mean Fine: Understanding Emotion Beyond Text

Biswajit Roy (mroy.biswajit@gmail.com)[1], Sidharth Ranjan[1]

[1]School of Artificial Intelligence and Data Science, Indian institute of Technology, Jodhpur

Language is the primary vehicle for communicating emotion, yet sentences built from words alone are often ambiguous in their emotional meaning. The same utterance "I'm fine" may signal contentment, suppressed distress, or sarcastic resignation depending on how it is spoken and what the speaker's face shows. How much emotional meaning is systematically lost when language is processed as text alone remains unquantified in a psychologically grounded framework.Following the Circumplex Model of Affect [1], we hypothesize that speech and facial expression resolve emotional ambiguity that text leaves unspecified, measurable by how closely multimodal predictions align with Russell's 28 labelled emotion coordinates in a two-dimensional Valence–Arousal (VA) space. We refer to this failure of text as "affective underspecification": a sentence in isolation cannot be reliably placed at a fixed emotional coordinate, and its true position only becomes recoverable when combined with the voice and face that deliver it. We substantiate our hypothesis using CMU-MOSEI [2] (23,000 utterances) and IEMOCAP [3] (5,000 dyadic utterances with continuous VA labels), evaluated against Russell's coordinate map as ground truth, within frameworks of multimodal language comprehension [5] and embodied meaning construction [6].

Method. We train the model in three stepwise configurations: text-only, bimodal (text + speech or text + facial expression), and trimodal (all three), to directly measure how much emotional meaning each added modality recovers. Each modality is encoded into a 128-dimensional representation via modality-specific BiGRU encoders with attention pooling: GloVe 840B word embeddings for text (2.2 million words mapped to semantic vectors), openSMILE eGeMAPSv02 acoustic features for speech (pitch, energy, speaking rate and voice quality capturing how something is said rather than what is said), and FaceMesh landmark coordinates for facial expression (468 geometric points tracking muscle movement, eye openness and mouth shape). A cross-modal Transformer fuses these into a shared two-dimensional Valence–Arousal projection aligned with Russell's coordinate space. Models are trained and evaluated on CMU-MOSEI [2] and IEMOCAP [3], with performance reported in Concordance Correlation Coefficient for valence (CCC-V) and arousal (CCC-A) against Russell's ground truth coordinates.

Results. The full trimodal model achieves CCC-V = 0.623 and CCC-A = 0.470 on CMU-MOSEI, outperforming all text-only and bimodal variants [Table 1]. Fig. 1 The utterance "I am so hurt" is correctly placed in the Sad quadrant (VA = [-1.14, 0.17]) text alone succeeds here, because the emotional meaning is transparent. Fig. 2 shows where text fails: "What's that fly doing in my soup?" is mispredicted as Neutral (VA = [+0.18, 0.28]) by the text-only model, while the ground truth is Angry/Stressed (VA = [−1.11, +1.21]). The complaint structure of the sentence misleads the model. Fig. 3 shows a case of deeper ambiguity: "I can't believe you did this" carries multiple possible emotions like anger, excitement, or sadness and text alone cannot decide. Only when voice and facial expression are added does the prediction reach the correct ground truth coordinate. This pattern is also observed across full conversational sequences from IEMOCAP, where text-only predictions consistently lag behind the true emotional trajectory, while the trimodal model closely follows the correct ground truth throughout.

Conclusion. Our findings demonstrate that affective underspecification, the inability of text alone to place a sentence at a fixed emotional coordinate, is a systematic and measurable property of language, not an edge case. Text discards the very signals that make emotional meaning recoverable: the tone and rhythm of the speaker's voice, and the expression on their face. This work directly demonstrates where text-only language processing fails and how each added modality closes the gap to the correct emotional ground truth [4]. Our results confirm that this underspecification is only resolved when language is read alongside the voice and face that deliver it, with direct implications for how we model emotional language comprehension, empathy, and human communication. Future work will integrate our model into conversational language models, moving towards cognitively grounded NLP systems that understand not just what is said, but how it is meant.

Figures and Tables.

Figure 1: Clear Emotion, Correct Prediction

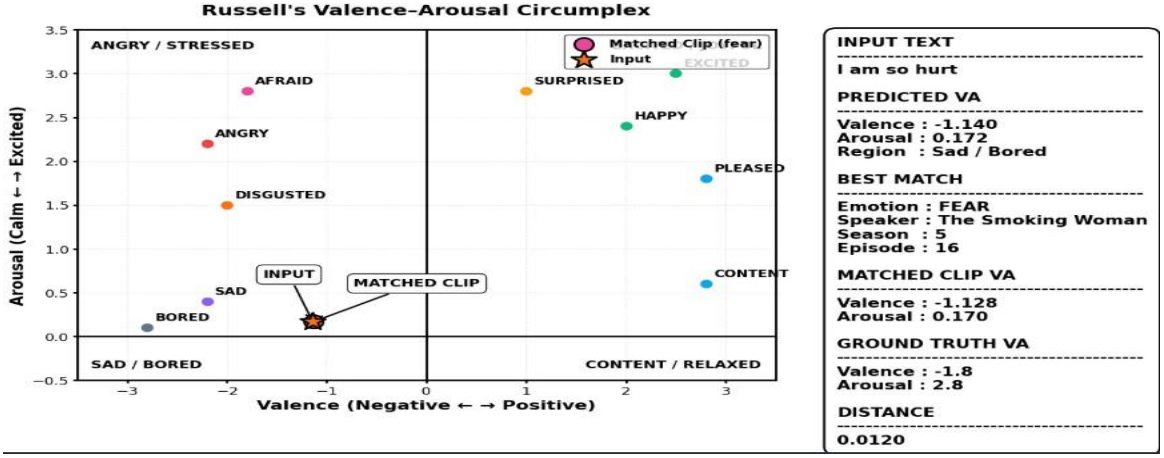

Figure 2: predicted Neutral, Truth was Angry

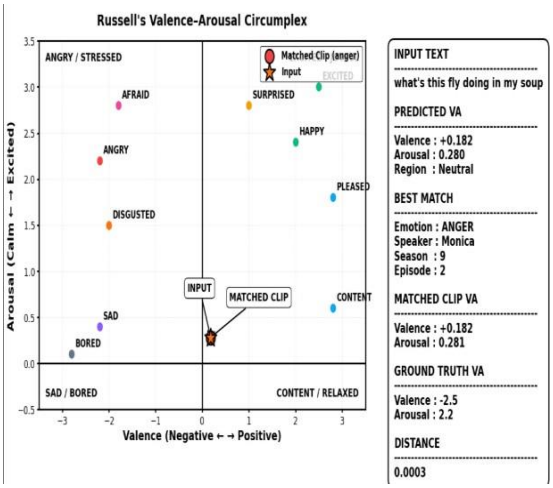

Figure 3: A Sentence, Infinite Meanings

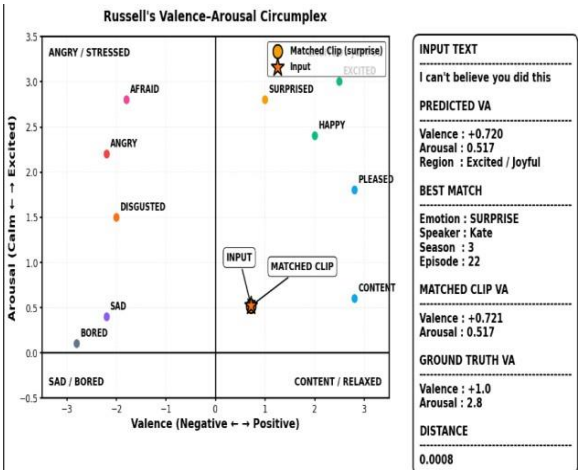

Table 1: From Words to Full Meaning: Modality Contribution.

Configuration	CCC-V	CCC-A	Notes
Text only	0.312	0.276	Fails on ambiguous utterances
Audio only	0.381	0.408	Good arousal, weak valence
Text + Audio	0.478	0.447	Reduces arousal errors
Text + Audio + Video	0.623	0.470	Best, all modalities combined

References.

[1] Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology,* 39(6), 1161–1178.

[2] Zadeh, A., et al. (2018). CMU-MOSEI: Multimodal corpus of sentiment intensity and subjectivity analysis. *arXiv:1606.06259.*

[3] Busso, C., et al. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation,* 42(4), 335–359.

[4] Bänziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech Communication,* 46(3–4), 252–267.

[5] Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review,* 9(3), 558–565.

[6] Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences,* 22(4), 577–660.