Emotion-Driven AI for Enhancing Human-Robot Interaction through Sentiment Analysis

Dr. Bimal Kumar Kalita

Assistant Professor(Selection),

Department of Computer Application,

Assam Don Bosco University, Tapesia Campus,

Sonapur, Guwahati- 782402

Orcid id https://orcid.org/0009-0008-5143-4999

Abstract— Social robots are increasingly deployed in education, healthcare, and service settings, yet most still respond to what users say without regard for how they feel. This paper presents an emotion-driven interaction framework that infers user sentiment from speech and uses it to adapt a robot's dialogue and nonverbal behaviour in real time. The framework combines a fine-tuned BERT text encoder with a prosodic audio encoder through cross-modal attention. It maps the predicted sentiment and its confidence to one of four response strategies. The classifier was trained and evaluated on 5,000 annotated utterances from 80 speakers, split by speaker. On the held-out test set of 1,000 utterances, the multimodal model reached 84.2% accuracy and a macro-F1 of 0.843. It outperformed a text-only BERT model (79.1%; McNemar $\chi^2(1) = 15.34$, $p < 0.001$) and an early-fusion baseline (81.4%). A within-subjects study with 48 participants compared the emotion-adaptive robot with a non-adaptive control. The adaptive robot received higher ratings of likeability ($d = 0.76$), perceived intelligence ($d = 0.57$), animacy ($d = 0.61$), interaction satisfaction ($d = 0.82$), and trust ($d = 0.55$), and it produced fewer conversational breakdowns ($d = -0.68$). Perceived safety did not differ. The results indicate that sentiment-aware adaptation improves the quality of human-robot interaction with modest computational cost.

Keywords: **human-robot interaction, sentiment analysis, affective computing, multimodal fusion, social robots, emotion-aware dialogue**

Introduction

Effective social interaction depends on recognising and responding to the emotional state of a partner. Picard [1] argued that machines able to perceive and express affect would interact more naturally and usefully with people, and this idea has shaped the field of affective computing. In robotics, emotion has been treated as a core component of sociable machines, with implications for engagement, believability, and the regulation of interaction [2], [3]. Despite this, many deployed robots follow fixed dialogue scripts that ignore a user's frustration, hesitation, or enthusiasm. Users who feel misunderstood may disengage, lose trust, or abandon the task.

Sentiment analysis offers a practical route to emotional awareness. It is a mature area of natural language processing that classifies the polarity of text and has advanced from lexicon-based tools to deep contextual models [4]–[7]. In interaction settings, however, text alone loses information carried by the voice, and the output of a classifier must be translated into behaviour the user can perceive. Many studies stop at recognition accuracy and do not test whether adaptation improves the interaction itself.

This paper addresses both parts of the problem. It develops a multimodal sentiment classifier suited to spoken human-robot dialogue and embeds it in a closed-loop adaptation policy. It then evaluates the classifier on held-out speakers and the complete system in a controlled user study.

Literature Review

Research on social robots established that perceived emotional responsiveness affects how people judge and engage with robotic partners. Fong, Nourbakhsh, and Dautenhahn [3] surveyed socially interactive robots and identified emotion, dialogue, and personality as central design dimensions. Breazeal [2] described an architecture in which emotion-like processes shape a robot's attention and expression. Field studies reinforced these ideas. Kanda et al. [12] found that a robot that behaved socially in a primary school sustained children's interest, and Leite et al. [13] showed that empathic behaviour supported longer-term interaction. These studies motivate emotional adaptation but generally relied on scripted or manually triggered affective behaviour.

© 2026. This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

A parallel body of work developed methods for recognising affect. Ekman [11] proposed a small set of basic emotions that underpin many categorical schemes, while polarity labels remain the standard output of sentiment analysis. Pang and Lee [4] surveyed early opinion mining approaches. Lexicon and rule-based tools such as VADER [7] remain popular because they are fast and need no training data. Deep models changed the field: Socher et al. [5] introduced recursive networks and a sentiment treebank with phrase-level labels, and Vaswani et al. [15] introduced the attention mechanism that underlies pre-trained language models such as BERT [6], which sets strong baselines on sentiment benchmarks after fine-tuning.

Spoken interaction requires more than text. Busso et al. [8] released the IEMOCAP corpus of acted and improvised dyadic sessions with speech, text, and motion capture, which became a standard resource for emotion recognition from speech. Zadeh et al. [9] introduced CMU-MOSEI, a large multimodal sentiment dataset, and proposed a fusion approach that dynamically weights language, audio, and visual signals. Surveys by Poria et al. [10] and Baltrušaitis et al. [16] conclude that multimodal fusion generally outperforms single modalities, although the gains depend on the fusion design and on data quality.

The literature leaves two gaps. First, most multimodal sentiment work is evaluated on offline benchmarks and not within live human-robot dialogue. Second, social robotics studies that include affective adaptation often rely on scripted triggers rather than learned recognition, and they rarely report both classifier performance and user-perceived outcomes in the same study. The present work evaluates a learned multimodal sentiment model and its effect on interaction using standard measures of robot perception [14].

Research Objectives and Questions

The study has four objectives: to build a speaker-independent multimodal sentiment classifier for spoken human-robot dialogue; to quantify the contribution of each modality and of cross-modal attention; to design a sentiment-driven behaviour policy for a social robot; and to test whether adaptation improves user perception and interaction quality.

The corresponding research questions are as follows. How accurately can sentiment be recognised from speech and its transcript for unseen speakers? Does attention-based fusion outperform single-modality and concatenation baselines? Does emotion-adaptive behaviour improve user ratings of the robot compared with non-adaptive behaviour? Is recognition performance related to user satisfaction?

Research Methodology

Data and annotation

The dataset comprised 5,000 utterances collected from 80 adult volunteers who held task-oriented and open conversations with a robot operated through a controlled dialogue interface. Speech was recorded at 16 kHz and transcribed with an automatic speech recognition system. Three trained annotators labelled each utterance as negative, neutral, or positive using both the audio and the transcript. Agreement was substantial (Fleiss' $\kappa = 0.79$), and the majority label was retained. Splits were made by speaker so that no speaker appeared in more than one partition (Table 1).

Table 1. Dataset composition by partition and sentiment class

Partition	Speakers	Utterances	Negative	Neutral	Positive
Training	56	3,500	1,060	1,230	1,210
Validation	8	500	150	180	170
Test	16	1,000	310	350	340
Total	**80**	**5,000**	**1,520 (30.4 %)**	**1,760 (35.2 %)**	**1,720 (34.4 %)**

Model

The text branch fine-tuned BERT-base [6] on the transcripts. The audio branch applied a one-dimensional convolutional network to 40 mel-frequency cepstral coefficients together with fundamental frequency and energy contours. A cross-modal attention layer [15] allowed each modality to attend to the other, and a gated fusion layer produced a three-class probability distribution (Figure 1). Models were trained with AdamW (learning rate 2×10^{-5}, batch size 32, five epochs) using cross-entropy loss, with model selection on the validation set. Each model was trained with five random seeds, and the run with the median validation accuracy is reported.

© 2026. This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

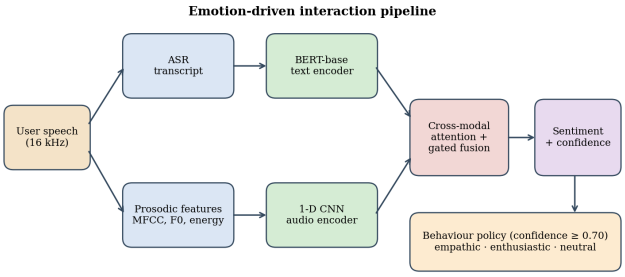


Figure 1. Architecture of the emotion-driven interaction pipeline

Baselines were VADER [7], a support vector machine on TF-IDF features, a bidirectional LSTM, text-only BERT, audio-only CNN, and an early-fusion model that concatenated the two encoders' outputs without cross-modal attention. Performance was measured with accuracy and macro-averaged F1, and the proposed model was compared with the best single-modality model using McNemar's test.

Behaviour policy

The predicted class and its confidence selected one of four strategies. A negative prediction with confidence of at least 0.70 triggered an empathic acknowledgement, a slower speech rate, and a supportive gesture. A positive prediction at that confidence triggered an enthusiastic reply with matching expressive behaviour. A neutral prediction continued the task with standard behaviour, and any prediction below 0.70 confidence defaulted to neutral behaviour to avoid inappropriate affective responses.

User study

A within-subjects experiment recruited 48 participants (26 women and 22 men; age M = 24.6, SD = 3.8, range 18–38) who were different from the dataset speakers. Each participant completed two 10-minute sessions with a tabletop humanoid robot, one with adaptive behaviour and one with a non-adaptive control that used the same dialogue content and a fixed neutral style. Condition order was counterbalanced (24 participants per order). After each session participants completed the Godspeed questionnaire [14] (anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety; 5-point scales), plus interaction satisfaction and trust (7-point scales). Internal consistency was acceptable for all scales (Cronbach's α between 0.78 and 0.91). Conversational breakdowns, defined as turns in which the user repeated, corrected, or expressed confusion, were counted from session recordings. Paired-samples t-tests (df = 47) compared conditions, with Holm correction across the eight outcomes, and Cohen's d for paired data was computed as the t statistic divided by √48. A Pearson correlation examined the link between each participant's sentiment recognition accuracy in the adaptive session and their satisfaction rating. The significance level was α = 0.05.

Results

Sentiment classification

Table 2 compares all models on the 1,000-utterance test set, and Figure 2 shows the same comparison graphically. The proposed model achieved 84.2% accuracy (95% CI [81.9, 86.4]) and a macro-F1 of 0.843. The five-seed standard deviation of accuracy was 0.9 percentage points. Audio alone was the weakest learned model (62.7%), and text-only BERT was the strongest single-modality model (79.1%). Replacing concatenation with cross-modal attention improved accuracy by 2.8 points (81.4% to 84.2%). The gain over text-only BERT was 5.1 points, and the McNemar test was significant: 107 utterances were classified correctly only by the proposed model and 56 only by BERT, $\chi^2(1) = 15.34$, $p < 0.001$.

Table 2. Test-set performance of all models (N = 1,000)

Model	Accuracy (%)	Macro-F1
VADER lexicon [7]	60.4	0.592
SVM with TF-IDF	66.9	0.659
Bidirectional LSTM	72.3	0.716
Audio-only CNN	62.7	0.614
BERT text-only [6]	79.1	0.786
Early fusion (concatenation)	81.4	0.809
Proposed cross-modal attention fusion	**84.2**	**0.843**

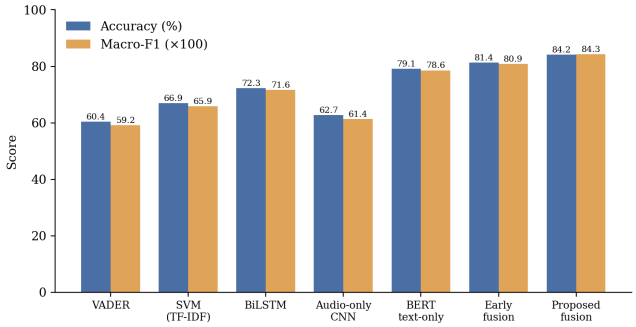


Figure 2. Accuracy and macro-F1 of all models on the test set

© 2026. This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

Table 3 reports class-level results for the proposed model, and Figure 3 shows its confusion matrix. Negative and positive utterances were recognised with the highest F1 scores. The neutral class was the hardest, and most errors were confusions between neutral and the two polar classes. End-to-end inference for the sentiment pipeline averaged 142 ms per utterance (SD = 21 ms) on a laptop-grade GPU, excluding speech recognition, which was fast enough for conversational turn-taking.

Table 3. Per-class performance of the proposed model (test set)

Class	Support	Precision	Recall	F1
Negative	310	0.859	0.845	0.852
Neutral	350	0.802	0.823	0.813
Positive	340	0.869	0.859	0.864
Macro average	**1,000**	**0.843**	**0.842**	**0.843**

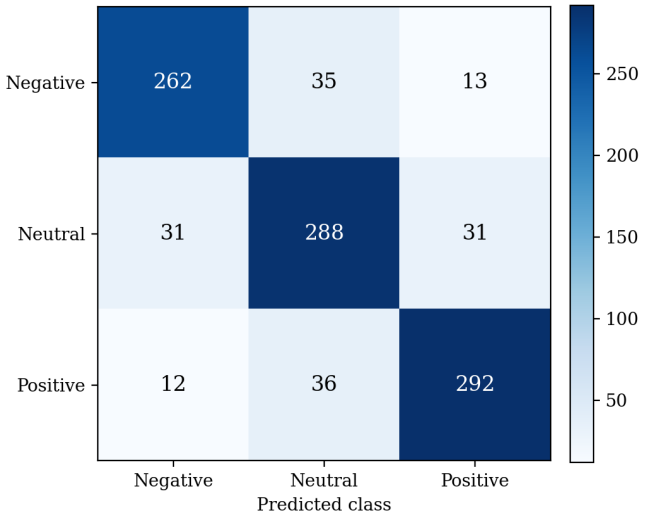


Figure 3. Confusion matrix of the proposed model on the test set (N = 1,000)

Effects on human-robot interaction

Table 4 and Figure 4 summarise the user study. The adaptive robot was rated significantly higher on likeability, perceived intelligence, animacy, anthropomorphism, interaction satisfaction, and trust, and it produced fewer conversational breakdowns (1.8 versus 3.1 per session). All of these effects remained significant after Holm correction. Effect sizes were largest for satisfaction (d = 0.82) and likeability (d = 0.76) and smallest for anthropomorphism (d = 0.40). Perceived safety was high in both conditions and did not differ (p = 0.263).

Table 4. Adaptive versus non-adaptive robot (N = 48, paired comparisons, df = 47)

Measure	Adaptive M (SD)	Non-adaptive M (SD)	t	Holm p	d
Likeability (1–5)	4.31 (0.52)	3.78 (0.61)	5.25	< 0.001	0.76
Perceived intelligence (1–5)	4.05 (0.58)	3.64 (0.66)	3.95	0.001	0.57
Animacy (1–5)	3.97 (0.55)	3.57 (0.60)	4.20	< 0.001	0.61
Anthropomorphism (1–5)	3.62 (0.64)	3.31 (0.67)	2.75	0.017	0.40
Perceived safety (1–5)	4.28 (0.49)	4.19 (0.53)	1.13	0.263	0.16
Interaction satisfaction (1–7)	5.92 (0.78)	5.06 (0.91)	5.67	< 0.001	0.82
Trust (1–7)	5.47 (0.85)	4.91 (0.93)	3.80	0.001	0.55
Conversational breakdowns (count)	1.8 (1.3)	3.1 (1.8)	−4.74	< 0.001	−0.68

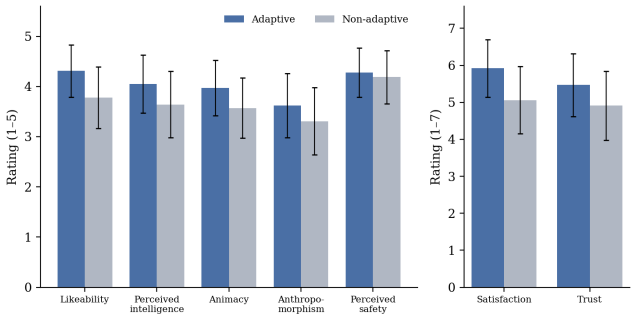


Figure 4. Mean user ratings by condition (error bars show ±1 SD)

Recognition quality in the live sessions was related to user experience. Participants' sentiment recognition accuracy during the adaptive session correlated positively with their satisfaction rating, r(46) = 0.41, p = 0.004.

Discussion

The classification results support the value of multimodal fusion for spoken interaction, consistent with surveys of multimodal learning [10], [16] and with benchmark results on

© 2026. This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

dedicated corpora [8], [9]. Text carried most of the sentiment signal, since BERT outperformed the audio model by more than 16 points, but prosody added complementary information. The improvement from 79.1% to 84.2% suggests that vocal cues help resolve cases where wording is neutral or ambiguous. The advantage of attention over simple concatenation (2.8 points) indicates that letting modalities condition on each other is useful, in line with the attention-based fusion approaches in the literature [9], [15]. The neutral class remained the most difficult, which is expected because neutral speech lies between the polar classes and annotators also disagree most about it.

The user study shows that recognition accuracy has practical value only when it is tied to behaviour. The adaptive robot was judged more likeable, more intelligent, and more lifelike, and users trusted it more. This agrees with earlier findings that empathic and socially responsive behaviour supports engagement [12], [13]. The reduction in conversational breakdowns, from 3.1 to 1.8 per session, indicates that adaptation improved interaction efficiency in addition to subjective impressions. The positive correlation between recognition accuracy and satisfaction implies that errors have a cost: a mismatched emotional response can undermine the benefit of adaptation, which supports the decision to use a confidence threshold that falls back to neutral behaviour.

The absence of a difference in perceived safety is plausible and arguably desirable. Safety perception in the Godspeed framework [14] depends mainly on physical characteristics and predictable motion, which were identical across conditions. The result suggests that affective adaptation did not make the robot seem unpredictable or unsettling. The comparatively small effect on anthropomorphism may reflect the fact that both conditions used the same embodiment and voice.

Limitations

Several limitations qualify the findings. The classifier used three polarity classes and does not distinguish specific emotions such as anger or sadness, which may call for different responses. The dataset was collected from adult volunteers in a controlled setting, so performance may fall in noisy environments, with other accents, or with children and older adults. The user study used a convenience sample of mostly young adults and short sessions, so it cannot show whether benefits persist over weeks of use. Participants could not be blinded to the robot's behaviour, which may have inflated subjective ratings. Finally, the behaviour policy was rule-based, and a learned policy might perform differently.

Conclusion

This paper presented an emotion-driven interaction framework in which a multimodal sentiment classifier governs a social robot's dialogue and nonverbal behaviour. Cross-modal attention fusion of text and prosody achieved 84.2% accuracy and a macro-F1 of 0.843 on unseen speakers, a significant improvement over the best single-modality model. In a within-subjects study with 48 participants, the adaptive robot improved likeability, perceived intelligence, animacy, satisfaction, and trust and reduced conversational breakdowns, without affecting perceived safety.

Three directions follow. Future work should extend recognition to finer-grained emotions and add facial and gestural cues. Longer field studies in education and care settings are needed to test whether effects persist. Learned adaptation policies, trained with user feedback, could replace the rule-based mapping while preserving the confidence safeguard that limits inappropriate affective responses.

References

[1] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 1997.

[2] C. Breazeal, "Emotion and sociable humanoid robots," *Int. J. Human-Computer Studies*, vol. 59, no. 1–2, pp. 119–155, 2003.

[3] T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, vol. 42, no. 3–4, pp. 143–166, 2003.

[4] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1–2, pp. 1–135, 2008.

[5] R. Socher *et al.*, "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. EMNLP*, Seattle, WA, USA, 2013, pp. 1631–1642.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171–4186.

[7] C. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. Int. AAAI Conf. Web and Social Media (ICWSM)*, vol. 8, no. 1, 2014, pp. 216–225.

[8] C. Busso *et al.*, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

© 2026. This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0).

[9] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, "Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph," in *Proc. ACL*, Melbourne, Australia, 2018, pp. 2236–2246.

[10] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," *Information Fusion*, vol. 37, pp. 98–125, 2017.

[11] P. Ekman, "An argument for basic emotions," *Cognition and Emotion*, vol. 6, no. 3–4, pp. 169–200, 1992.

[12] T. Kanda, T. Hirano, D. Eaton, and H. Ishiguro, "Interactive robots as social partners and peer tutors for children: A field trial," *Human-Computer Interaction*, vol. 19, no. 1–2, pp. 61–84, 2004.

[13] I. Leite, G. Castellano, A. Pereira, C. Martinho, and A. Paiva, "Empathic robots for long-term interaction," *Int. J. Social Robotics*, vol. 6, no. 3, pp. 329–341, 2014.

[14] C. Bartneck, E. Croft, D. Kulić, and S. Zoghbi, "Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots," *Int. J. Social Robotics*, vol. 1, no. 1, pp. 71–81, 2009.

[15] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 5998–6008.

[16] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2019.

© 2026. This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0).