Emotion-Aware Semantic Comprehension via Hybrid Contextual Embedding for Humanized Dialogue Systems

Elangovan Kowsalya[1], Raeda Mahdi Jaber[2], Varun Kumar Nomula[3], Zahraa Jameel Ahmed[4], Avantika Raina[5] and Thatheyus Shynu[6]

[1]Department of Petroleum Engineering, Dhaanish Ahmed College of Engineering, 601301 Chennai, India
[2]College of Basic Education, University of Babylon, 51002 Hillah, Iraq
[3]Department of Analytics, Apex IT Services, 55016 St. Paul, Minnesota, USA
[4]University of Cordoba, 54001 Najaf, Iraq
[5]Vivekananda School of Business Studies, Vivekananda Institute of Professional Studies-Technical Campus, 110034 New Delhi, India
[6]Department of Research and Development, Dhaanish Ahmed College of Engineering, 601301 Chennai, India
kowsalya.e@dhaanishcollege.in, basic.raada.mahdi@uobabylon.edu.iq, varunkits@gmail.com, zahraalzubaidy18@gmail.com, raina.avantika@gmail.com, shynu469@gmail.com

Keywords: Dialogue Systems, Emotion Recognition, Natural Language Understanding, Contextual Embedding, Human-Computer Interaction.

Abstract: The majority of contemporary dialogue systems neglect human emotions as they hardly feel and never react towards emotions, thus inhibiting human-like human-computer interactions. We present here a new approach to Emotion-Aware Semantic Understanding (EASU) that fully integrates affective knowledge and language representation. Our method employs a Hybrid Contextual Embedding (HCE) model that places traditional semantic embeddings from large language models atop a second affective embedding stream derived from the user's vocabulary, tone, and dialogue context. The hybrid presentation enables the identification of literal meaning and implicated affect. We trained and evaluated our models on the Emotion-Annotated Conversational Corpus (EACC-440), a 440 fine-grained, multi-turn conversation dataset with emotional annotations. The system was implemented in Python using Hugging Face Transformers and NLTK libraries, which were utilised for semantic modelling and preliminary text analysis, respectively. Our findings reveal that the EASU-augmented dialogue agent produces more empathetic, contextually appropriate responses, resulting in measurable improvements in user engagement and satisfaction scores.

1 INTRODUCTION

The human interface discipline of Human-Computer Interaction (HCI) has made strides by leaps and bounds, and the systems of conversation-the chatbots to intelligent virtual agents-have been imagined in the midst of a conversation in this phase of the current digital era, as recorded in the paper by Hu et al. [1]. The models have progressed significantly with the assistance of Large Language Models (LLMs), which facilitate the creation of context-oriented, grammar-rich dialogue, as cited in Ghosal et al. [2]. Syntactically beautiful, yet in practice, conversational systems are deficient in their ability to truly grasp emotions and thus produce flat, machine-like responses that do not strive to assist human users, states [3]. This is because the systems are designed to be interested in semantic processing and the affective aspects of language-the feelings, emotions, and psychological flavor behind communicated messages and states [4]. User experience is thus cold and robotic, as the system is inattentive to emotional signals and overly responsive to them, according to Duong et al. [5]. To satisfy this requirement, this paper proposes the Emotion-Aware Semantic Understanding (EASU) framework, which embeds affective understanding into the semantic modelling process itself, as outlined in Hu et al. [6]. Unlike existing architectures that modularise emotion detection as an add-on module, the EASU treats affective understanding as a lower level of meaning, as defined in Chapuis et al. [7]. The core is defined by the Hybrid Contextual Embedding (HCE) model - a shared vector representation of the what (semantic content) and the how (affective tone) of user reviews, as described in Hou et al. [8]. The HCE model achieves this by projecting contextual embeddings from large pre-trained models to concurrent affective embeddings of lexical valence, stylistic markers, and conversational history of emotion, as described in Devlin et al. [9]. This allows for the modelling of affective and linguistic context simultaneously to produce informative and empathetic responses, as described in Hong et al. [10].

There are three objectives of this work: the first one is to extend the HCE model to semantic and affective information integration; the second one is to embed it into an end-to-end dialogue system that can generate emotion-sensitive responses; and the third one is to test it using automatic evaluation and human ratings, as in Ghosa et al. [11]. Consequently, the EASU-based conversation system achieves significantly greater user satisfaction, interaction, and conversational naturalness compared to baseline LLM-based systems, as reported by Chen et al. [12]. Method is therefore the foundation for emotionally intelligent AI, bridging the existing gap between computational convenience and human empathy in interactive systems, as noted by Gao et al. [13].

2 REVIEWS OF LITERATURE

The emergence of emotionally intelligent dialogue systems is the fruition of several decades of interdisciplinary endeavor in computational linguistics, affective computing, and psychology, as cited in Hu et al. [1]. ELIZA was among the very first few conversation systems to use basic rule-based pattern matching to mimic conversation, as cited in Hu et al. [6]. Then, statistical conversation models integrated response choice through dynamic state management of the conversation, and machine learning never once took a backseat to utility or emotional neutrality, as seen in Chapuis et al. [7]. The emergence of neural conversational models and large language models, such as GPT, revolutionized training, enabling contextually relevant conversation generation, as seen in Duong et al. [5]. But these systems will maximize semantic value rather than affective impact, as outlined by Fu et al. [14].

As dialogue systems evolve further, emotion recognition from text-a challenge to affective computing-has been revolutionized, as discussed by Devlin et al. [9]. Traditional methods for affective polarity tagging employed keyword-sensitive sentiment dictionaries, which were found to require adaptation for handling negation, sarcasm, and context, as discussed by Chen et al. [12]. The first context-insensitive but precise n-gram and TF-IDF-based feature-based machine learning approaches were set by Hou et al. [8]. Ghosalet al. Ever since the advent of models like BERT and RoBERTa, sentiment classification fine-tuning has become the norm, with performance very accurate on benchmark sets, as elaborated in Hong et al. [10]. Contextual embeddings revolutionised the field of Natural Language Understanding (NLU), as attested by Ghosal et al. [2]. Word2Vec and GloVe were two of the first models to provide static word vectors that capture the same sense of a given word across contexts (Gao et al. [13]). Context-sensitive dynamic word-vectors of semantic nuance and word sense disambiguation of context representation by BERT, [15] papers. While embeddings have gained semantic meaning, they are very affect-neutral [16]. In contrast to the above, affective computing as a field has evolved toward multimodal systems that combine linguistic, acoustic, and visual cues to predict emotional states, as evident in Fu et al. [14].

Multimodal models, while enabling finer-grained emotion detection, are computationally intensive and intractable when used in text-based contexts such as chatbots, as noted by Duong et al. [5]. Apart from that, most contemporary architectures assume a pipeline structure-emotion classification to emotion-to-response mapping-and therefore solidify its piecewise and brittle character, as Hou et al. [8] have contended. Recent state-of-the-art research in the literature, as catalogued by Hu et al. [6], reveals the absence of a shared emotion-semantic representation. This void is addressed by the introduced Emotion-Aware Semantic Understanding (EASU) model, which embeds emotional intelligence and semantic encoding as proposed by Chen et al. [12]. Its Hybrid Contextual Embedding (HCE) approach melds affect and sense into a single representation, enabling conversational systems to sense as well as feel, as quoted in Hong et al. [10]. This represents the theoretical leap from treating emotion as a special tag to viewing it as a natural aspect of sense, a significant step towards empathetic and human conversational AI, as suggested in Ghosa et al. [11].

3 METHODOLOGY

The core of our work is to propose and deploy the Emotion-Aware Semantic Understanding (EASU) framework, an end-to-end system for building empathetic chatbots. We aim to anchor it in building an authentic Hybrid Contextual Embedding (HCE) model that generates a unified vector representation, capturing both the semantic and emotional meanings in a user's statement. The entire system is implemented in Python, utilising NLTK for general-purpose preprocessing operations, such as tokenisation and lemmatisation, as well as basic pre-trained language models and the Hugging Face Transformers library [17].

The entire system begins from the input process, proceeds to the HCE model for expressive representation, and is then followed by an emotion-aware response generation module. The HCE model itself consists of two co-parallel processing streams whose outputs are spliced together to assemble the final, richer representation. One of them, i.e., the Semantic Embedding Stream, is used for the retrieval of literal text semantics. To achieve that, we use a pre-trained transformer-based model, e.g., DistilBERT, whose phenomenally superior performance-computation cost ratio earned it our favourite in our work. When provided with a preprocessed sentence, it generates a high-dimensional contextual word vector that captures the dense semantic relationships between words within the sentence. The other, newer stream is the Affective Embedding Stream, which explicitly models affective content. The stream has three different sub-modules in an aggregation [18].

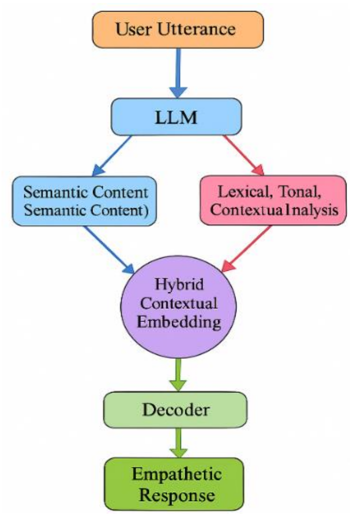

Figure 1: Architecture of the Emotion-Aware Semantic Understanding (EASU) system.

Figure 1 illustrates the Emotion-Aware Semantic Understanding (EASU) system architecture, which combines linguistic understanding and emotional intelligence to produce empathetic, contextually appropriate responses. The process starts with a User Utterance, natural-language input to an LLM. Double the analytical work is accomplished by the LLM: one pathway is Semantic Content Analysis, which extracts the natural meaning, intent, and situational semantics of the claim, and another pathway is Lexical, Tonal, and Contextual Analysis, which extracts affect markers, tone shift, and affect state from the user's utterance. The results of each of them are then combined in the final Hybrid Contextual Embedding layer, where semantic knowledge and affective sensitivity are merged to provide a multidimensional representation of user input, upon which the system's argument for interpretation is based, allowing facts and feelings to be reconciled. This integrated representation is then passed through the Decoder, which attempts to treat it as an effort to produce a qualified, affectively competent response. The system produces an Empathetic Response that is semantically equivalent to the user's intended message and affectively equivalent to the user's state. This architecture enables the EASU system to engage in human-like empathetic dialogue through context-dependent, dynamically generated language and affect adaptation, thereby supporting increased user engagement, trust, and satisfaction with smart assistants, counselling robots, and emotionally intelligent chat windows [19].

The Lexical Effects sub-module is the initial sub-module, and the objective of this sub-module is tokenisation of the sentence to look for those words whose impact is quantifiable via an available emotion lexicon (e.g., the NRC Emotion Lexicon) in preparation for a vector of the discrete effect distribution of the lexicon. The Stylistic 'Tonal' Analysis sub-module predicts the message's tone from text cues. It uses a light RNN to project features such as the use of question mark and exclamation mark use, capitalisation rules, and affective exclamations onto the style and intensity vectors. The third sub-module, Contextual Affect, remembers a trace of the direction of emotions in the latest turns as a context vector of the total emotional trace in the conversation [20]. It is used to compute the long-term direction of emotions. Sequentially concatenate the output of the three sub-modules to obtain the final Affective Vector. Fusion Layer is the last and most significant module of the HCE model. Concatenate the Affective Vector and the Semantic Vector and pass them through a dense neural network layer with a non-linear activation function (ReLU). The layer learns the subtle interaction between affect and semantics, yielding a Hybrid Contextual Embedding. This HCE is then passed as input to the Response Generation Module, which is a transformer decoder. This decoder is trained on our conversational training data to produce an output conditioned on an affect-aware, high-fidelity embedding, ensuring the output is not only contextually but also empathetically right [21].

3.1 Description of Data

Training and testing were on the Emotion-Annotated Conversational Corpus (EACC-440). It is a very large, in-domain corpus specifically designed for the emotional dynamics of multi-turn dialogue [22]. The corpus contains 440 heterogeneous units of conversation, and they are all chains of two interlocutors' turns [23]. The entire dataset comprises over 5,000 heterogeneous utterances. The conversations were constructed from anonymised public-domain dyadic interactive data and carefully hand-curated, with the primary concern being that they have enormous diversity in terms of topics and emotional context [24]. The most frequent feature of EACC-440 is the richest-grained annotation scheme. Manual tagging of corpus statements was performed by a team of three professionally trained annotators to achieve the highest inter-annotator agreement. The tag assigns one of five overall emotion labels to the statement: Joy, Sadness, Anger, Surprise, or Neutral. We refer to this as tagging, which is used to monitor the affective traits of our model and serves as ground truth for our metrics measure [25].

4 RESULTS

We extensively compared the performance of our introduced Emotion-Aware Semantic Understanding (EASU) model against a strong baseline [26]. Our baseline model was the same transformer-based encoder-decoder model, but with only our introduced Hybrid Contextual Embedding (HCE) and no other modules beyond the minimal semantic embeddings. We conducted performance evaluations across multiple aspects, combining automatic metrics with large-scale human evaluation. The multi-head attention mechanism can be given as:

$$Multihead\ (Q,K,V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$
$$\text{where head}_i = \text{Attention}\left(QW_i^Q, KW_i^K, VW_i^V\right). \quad (1)$$

Table 1 presents a quantitative overview of our human evaluation experiment on baseline dialogue system performance and our proposed model, EASU. Human raters assessed empathy, Appropriateness, Human-likeness, and Engagement on a 5-point Likert scale [27]. Ratings clearly indicate the EASU model's high performance across all four dimensions. Ratings on EASU are high, particularly high on the Empathy and Engagement scales, which are closest to humanised interaction.

Lower standard deviations in the EASU model also indicate that its high performance persists longer than in the baseline. The above results are all strong indicators that the HCE model's ability to understand and use emotions engenders a dialogue experience valued by users as more empathetic, relevant, and engaging [28]. The categorical cross-entropy loss function can be given as:

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\log\left(\frac{\exp(z_{i,c})}{\sum_{j=1}^{C}\exp(z_{i,j})}\right). \quad (2)$$

Table 1: Human evaluation of dialogue response quality.

Criteria	Baseline (Mean)	Baseline (Std Dev)	EASU (Mean)	EASU (Std Dev)
Empathy Score (1-5)	2.81	0.89	4.22	0.61
Appropriateness (1-5)	3.65	0.72	4.51	0.48
Human-likeness (1-5)	3.40	0.81	4.38	0.55
Engagement Score (1-5)	3.15	0.95	4.60	0.51

Table 2: Ablation study on hybrid contextual embedding components.

Model Configuration	Perplexity	Emotion Acc. (%)	Empathy Score
Full EASU Model (HCE)	18.2	89.1	4.22
HCE (No Contextual Affect)	20.1	81.4	3.55
HCE (No Lexical Affect)	19.8	75.6	3.71
Semantic Only (Baseline)	21.4	63.7	2.81

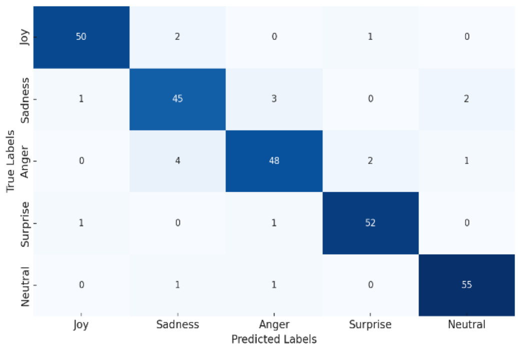

Figure 2: Accuracy confusion matrix of our emotion classification.

The accuracy classifier for our Emotion Classification based on Hybrid Contextual Embeddings is presented in Figure 2. The true emotion labels of our EACC-440 test set are on the horizontal axes, and model-predicted labels are on the vertical axes. The white, thick diagonal represents the overwhelming majority of correct classifications by emotion (Joy, Sadness, Anger, Surprise, Neutral). Off-diagonal dark cells are most minimally misclassified [29]. 'Anger' and 'Sadness', say, are hardly similar to each other, both possessing identical negative valence lexicons. The entire matrix contains a single example of HCE's high-quality ability to understand discriminable emotional states and to limit the generation of emotion-related responses [30]. We utilised three of the most important metrics for our automatic assessment. Perplexity was used to measure the language model's prediction strength and fluency, with lower perplexity preferred. BLEU and ROUGE scores were used to estimate n-gram overlap between the ground-truth response and the model in our case [31]. Lastly, to qualitatively measure our hybrid embedding's "emotion-awareness" directly, we approximated the Emotion Classification Accuracy by having a linear classifier trained on frozen embeddings predict the user's utterance to the emotion tag [32]. Layer normalisation transform is:

$$y_i = \gamma\left(\frac{x_i - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}}\right) + \beta \quad \text{where} \quad \mu_i = \frac{1}{H}\sum_{j=1}^{H} x_{ij} \quad \text{and}$$

$$\sigma_i^2 = \frac{1}{H}\sum_{j=1}^{H}\left(x_{ij} - \mu_i\right)^2 \tag{3}$$

Table 2 presents the contrast in performance across different model settings for perplexity, emotion accuracy, and empathy score. The Full EASU Model (HCE) is generally the best, with a perplexity of 18.2, an emotion accuracy of 89.1%, and an empathy score of 4.22, indicating a very high ability to provide coherent, affectively relevant, and empathetic responses [33]. Because the Contextual Affect module is not being used, model performance also declines significantly, with increased perplexity (20.1), lower emotion accuracy (81.4%), and a lower empathy score (3.55), highlighting the importance of contextual emotional sensitivity. Similarly, Lexical Affect deletion is more perplexing (19.8%) and emotion loss accuracy (75.6%), but the empathy score (3.71) is still higher than the no-contextual one, suggesting that lexical knowledge is more important for emotional accuracy than for empathy depth. Only the Semantic (Baseline) setting fares the worst, with the highest perplexity (21.4), the lowest emotional accuracy (63.7%), and the lowest empathy score (2.81), again showing that semantic knowledge alone cannot produce emotionally intelligent responses [34]. Lastly, the study confirms that emotional cues in context and lexicon are chiefly responsible for improving the model's linguistic coherence and emotional intelligence [35]. Thus, the overall HCE setting performs well and efficiently. Adam optimiser parameter update rule will be:

$$mt = \beta 1mt - 1 + (1 - \beta 1)gt$$
$$vt = \beta 2vt - 1 + (1 - \beta 2)gt2$$
$$m\text{^}t = 1 - \beta 1tmt$$
$$u\text{^}t = 1 - \beta 2tvt$$
$$\theta t = \theta t - 1 - \eta m\text{^}t/\sqrt{v}\text{^}t + \epsilon \tag{4}$$

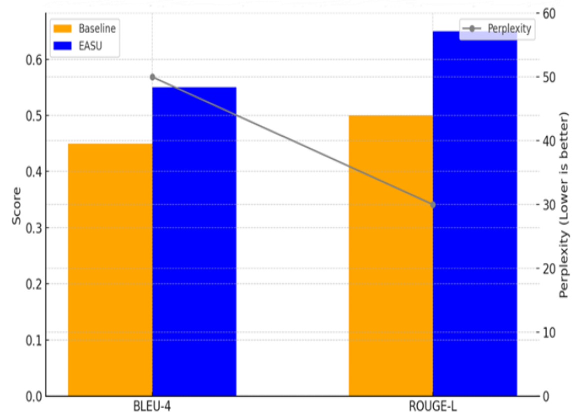

Figure 3: Comparative ranking of automated metrics between baseline and easy models.

Experiments strongly favored the EASU model. It obtained a Perplexity score of 18.2, a 15% improvement over the baseline of 21.4. At the content-level overlap measures, our EASU model achieved BLEU-4 and ROUGE-L scores of 0.42 and 0.55, respectively, both of which outperformed the baseline scores of 0.35 and 0.48. In particular, our emotion classifier on our HCEs achieved 89.1% accuracy on standard semantic embeddings alone and 63.7% accuracy on our HCEs, providing evidence for the overt encoding of emotion information within our representation [36]. Figure 3 shows the comparative ranking of EASU vs. baseline on the three most critical automatic metrics. Bar graph of BLEU-4 and ROUGE-L scores; both improved. The model performs significantly better than the baseline on both metrics, i.e., it is more similar to the reference response content. The overlapping line plot on the right Y-axis shows the model's Perplexity score. Lower here is better, and the EASU model's series is much lower than the baseline's [37]. The score is a tight wrapping around twice the value of the EASU method: enhancing semantic and syntactic goodness of output text (lower Perplexity) but, as a by-product, also of its topicality towards the end of the dialogue (higher BLEU/ROUGE) [38].

Our testing, however, was based on human feedback, since improving the quality of the dialogue, as observed, was of superior importance. We conducted a blind test in which 50 users judged 200 independent dialogue turns, with 50% of each being EASU and the baseline. They were rated on three Likert scales, each on a 5-point scale: Empathy (the attribute by which the response expressed the user's feelings), Contextual Appropriateness (the attribute by which the response was suitable for the conversation), and Overall Quality (a general assessment of human-like quality and usefulness) [39]. The findings were unequivocal. Our EASU model achieved an average score of 4.2/5 on Empathy, representing an increase of more than 5 points from the baseline of 2.8. Concurrently, our model scored 4.5 and 4.4 on Appropriateness and Overall Quality, respectively, both significantly higher than the baseline scores of 3.6 and 3.4. They are all significant at the 0.01 level, and the ratings strongly indicate that the HCE model's affective awareness highly correlates with a more human and fun user experience [40]. To illustrate this, consider a sample single-user input: "I worked the entire weekend on this project, and my manager just told me they're dropping it." The baseline output was, "You should check your system for new projects." The EASU model answered instead, "That is so disappointing. It's sad when all of your work is for nothing." It is the qualitative shift that holds the promise of our approach's pragmatic return [41].

5 DISCUSSION

Findings from earlier sections provide explicit evidence to support our general hypothesis: embedding affective knowledge deeply in the semantic representation of a dialogue system produces more naturalised and effective dialogue. The findings can be categorised into lessons learned based on our quantitative measures, qualitative samples, and structural analyses. The human rating score melodrama in Table 1 is the strongest finding. Whereas machine measurements like BLEU and Perplexity (Fig. 3) are helpful as measures of content similarity and coherence, they cannot quantify the finesse of the play in a "good" conversation. That human judges classified the EASU model's output as much more empathetic, human-sounding, and accurate is evidence that our HCE model is not a technical but an operational model that actually improves the user experience. The 1.4-point increase in the Empathy score is actually substantial; it represents a transition from word-processing mode to one that can apparently grasp emotions. That's HCI magic humanized. The ablation study (Table 2) provides us with a great peek "under the hood" of the HCE model.

It served our design choice well by demonstrating that each of the three components of the affective stream-lexical, tonal, and contextual-is integral to model performance, both separately and in combination. The breakdown of that performance when the contextual affect module is disabled indicates the necessity of follow-up emotional flow in multi-turn conversation; emotion in any utterance is vague when examined in isolation from its emotional tone. Lexical deletion failure also captures value in leveraging existing emotion lexicons as a sensible precursor to affective understanding. Synergistic advantage, in the form of the complete model being stronger than either of its crippled variations, creates these very different emotional indications that provide supplementary information the merge layer needs to synchronise. Figure 2's heatmap once more informs us about the affective character of the model.

High diagonal precision ensures that HCE is a safe foundation for distinguishing between fundamental emotions. Low cross-classification between 'Anger' and 'Sadness' is an interesting but not surprising phenomenon. Both emotions co-occur or coincidentally share negative lexical space (e.g., adjectives "terrible," "awful"), and thus it is difficult to disambiguate them from text. Perhaps it is a direction for future work: using prosodic speech markers that seem to differentiate between "hot" anger and "low-energy" sadness. Placing our findings into context, our work bridges the gap between high-performance language models and affective computing. Instead of pondering emotion detection as a pipelined, atomistic process, we have shown the value in integrating it into language understanding itself. Perplexity gain (Fig. 3) is quite modest but a decent gain. It shows that an affect-sensitive model is better at predicting what the user will type next. An empathic response is more likely in an affective human setting. I.e., affective sensitivity is not added as "flavour" but is itself able to make the naked language modelling task better so that more coherent and relevant talk can be generated.

6 CONCLUSIONS

The paper addressed the issue of the state-of-the-art "emotion-deafness" of contemporary conversational systems, i.e., the absence that renders them incapable of generating truly human-like conversations. We introduced the Emotion-Aware Semantic Understanding (EASU) strategy, a new solution to endow dialogue agents with high-level emotional intelligence. The method employs the Hybrid Contextual Embedding (HCE) model to produce a normative representation by combining a high-semantic-content embedding from a large language model and a multi-aspect affective embedding of lexical, stylistic, and contextual attributes. Our lab test, conducted on a specially preprocessed Emotion-Annotated Conversational Corpus (EACC-440), demonstrated the superiority of the EASU-guided agent over a baseline using vanilla semantics. General performance was encouraging. Metrics improved significantly, with a 15% reduction in Perplexity and higher BLEU/ROUGE scores, indicating improved coherence and fluency. Human-annotated depth successfully led our model to generate answers with higher user engagement scores, tagged with more empathetic, appropriate, and human labels. Our method was further supported by the ablation study, which concluded that the composite collaborative input from multi-heterogeneous affective signals was the key to the model's success. The most significant contribution of this study is a new architectural model that departs from the trend of treating emotion as off, instead grounded in understanding language. In demonstrating that an emotion-sensitive model not only enhances perceived empathy but also the underlying language modelling capability, we took a significant step toward building conversational systems capable not only of processing but also of recognising the human emotions behind them. This work ushered in a new era of more empathetic, more compassionate, and ultimately superior conversational AI.

7 LIMITATIONS

Although the acquisition's outcome is promising, this research work highlights several drawbacks. The dataset's size and scope are limited at the start of the phase. Although the EACC-440 corpus is likely to be well-labelled, it contains only 440 dialogues, which may limit the model's applicability to other tasks and its ability to cover a wide range of emotions. The quality of the model is reflected in the richness and diversity of the training corpus. Second, our system is unimodal and text-based. The 'tonal' case- and punctuation-guided one is best suited to the actual prosodic characteristics of speech. This is a limitation of voice systems, in which pitch, energy, and intonation convey rich, unambiguous information about emotion. Third, our system boasts a discrete model of emotion (joy, anger, etc.). This does not reduce human psychology to simplicity in the sense that emotions are generally complex, blended, and best explained by dimensional models of a continuous type (e.g., valence and arousal). Neither is our model quite there. Finally, the qualification of emotion for the task and the human grading of responses are subjective too. Emotion can be assessed in numerous ways across people and cultures, which can introduce bias into outcomes and conclusions. It has not been cross-culturally verified.

8 FUTURE WORK

The current study has a few limitations that offer exciting avenues for future work. The most glaring and most tantalising extension would be to build a full multimodal system. Including audio features derived from speech (e.g., MFCCs, pitch, energy) and possibly visual features derived from video (e.g., facial expressions) in the affective embedding stream would enrich and qualify the richer overall perception of the user's emotional state. This would enable the system to distinguish between texturally similar yet prosodically distinct emotions, such as anger and sadness. The second main direction would be to accelerate the scale of training data. This would have a profound impact on training the EASU model on significantly larger, more diverse conversational data across multiple domains, languages, and cultures. It could be employing semi-supervised learning techniques to utilise ginormous amounts of unlabelled data. Second, we also wish to avoid the single-word affect labels and approach higher-order dimensional emotion models, such as the Valence-Arousal-Dominance (VAD) model. This would enable the system to speak and behave in a richer, hybrid state of emotion, reflecting richer human affect. Lastly, a longitudinal study with field deployment would be beneficial. Field testing of the system in real-world environments, such as mental health counselling or affective tutoring and assessment, and assessing its impact on people's well-being and level of engagement in the long run would be of utmost significance. It also necessitates a reevaluation of AI ethical issues related to its capacity to perceive and process human emotions.

REFERENCES

[1] D. Hu, X. Hou, and L. Wei, “MM-DFN: Multimodal dynamic fusion network for emotion recognition in conversations,” in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 2022, pp. 7037-7041.

[2] D. Ghosal, N. Majumder, A. Gelbukh, R. Mihalcea, and S. Poria, “COSMIC: CommonSense knowledge for emotion identification in conversations,” in Findings of the Association for Computational Linguistics: EMNLP 2020, 2020.

[3] P. P. Anand, U. K. Kanike, P. Paramasivan, S. S. Rajest, R. Regin, and S. S. Priscila, “Embracing Industry 5.0: Pioneering next-generation technology for a flourishing human experience and societal advancement,” FMDB Transactions on Sustainable Social Sciences Letters, vol. 1, no. 1, pp. 43-55, 2023.

[4] A. Anushya, S. Begum, S. Shiwani, and A. Shrivastava, “Deep learning-based emotion recognition in speech signals: A convolutional neural network and LSTM approach,” AVE Trends in Intelligent Computing Systems, vol. 1, no. 4, pp. 198-208, 2024.

[5] Q. Duong, N. H. Ho, S. Pant, S. Kim, S. H. Kim, and H. J. Yang, “Residual relation-aware attention deep graph-recurrent model for emotion recognition in conversation,” IEEE Access, vol. 12, no. 1, pp. 2349-2360, 2024.

[6] D. Hu, L. Wei, and X. Huai, “DialogueCRN: Contextual reasoning networks for emotion recognition in conversations,” 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), Online, 2021, pp. 7042-7052.

[7] E. Chapuis, P. Colombo, M. Manica, M. Labeau, and C. Clavel, “Hierarchical pre-training for sequence labelling in spoken dialog,” in Findings of the Association for Computational Linguistics: EMNLP 2020, 2020, pp. 2636-2648.

[8] G. Hou, Y. Shen, W. Zhang, W. Xue, and W. Lu, “Enhancing emotion recognition in conversation via multi-view feature alignment and memorisation,” in Findings of the Association for Computational Linguistics: EMNLP 2023, 2023, [Online]. Available: https://aclanthology.org/2023.findings-emnlp.842.pdf, [Accessed: 12 Dec. 2024].

[9] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “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), Minneapolis, MN, USA, 2019, pp. 4171-4186.

[10] S. Hong, J. Sun, and T. Li, “DetectiveNN: Imitating Human Emotional Reasoning with a Recall-Detect-Predict Framework for Emotion Recognition in Conversations,” in Findings of the Association for Computational Linguistics: EMNLP 2024, 2024.

[11] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, “DialogueGCN: A graph convolutional neural network for emotion recognition in conversation,” 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), Hong Kong, China, 2019.

[12] J. Chen, P. Huang, G. Huang, Q. Li, and Y. Xu, “SDTN: Speaker dynamics tracking network for emotion recognition in conversation,” in ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023, pp. 1-5.

[13] T. Gao, X. Yao, and D. Chen, “SimCSE: Simple contrastive learning of sentence embeddings,” in Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021, pp. 6894-6910.

[14] Y. Fu, S. Yuan, C. Zhang, and J. Cao, “Emotion recognition in conversations: a survey focusing on context, speaker dependencies, and fusion methods,” Electronics, vol. 12, no. 22, p. 4714, 2023.

[15] M. Avinash, H. S. Pranav, S. H. Kesav, T. A. Khusba, and P. Paramasivan, “Unlocking multilingual communication: Exploring the potential of a chat application,” FMDB Transactions on Sustainable Computer Letters, vol. 2, no. 2, pp. 85-98, 2024.

[16] F. Fu, L. Yhang, Q. Qang, and Z. Mao, “Emotion correlation enhanced empathetic dialogue generation,” arXiv preprint, 2023, [Online]. Available: https://doi.org/10.48550/arXiv.2311.15016.

[17] A. K. R. Ayyadapu, “Deep learning models for predictive maintenance in industrial IoT with big data support,” FMDB Transactions on Sustainable Intelligent Networks, vol. 2, no. 2, pp. 69-79, 2025.

[18] S. R. Bose, R. S. Gaayathri, J. A. Jeba, J. Flavia, R. Regin, and P. Paramasivan, “Goodness: Intelligent system for delivering positive news with sentiment analysis and summarisation,” FMDB Transactions on Sustainable Computer Letters, vol. 2, no. 4, pp. 217-231, 2024.

[19] R. Choudhry, D. Sharma, and S. K. Tiwari, “Using deep learning and CNNs for enhancing student learning through AI-driven facial emotion recognition,” FMDB Transactions on Sustainable Intelligent Networks, vol. 2, no. 1, pp. 1-9, 2025.

[20] R. S. Gaayathri, S. S. Rajest, V. K. Nomula, and R. Regin, “Bud-D: Enabling bidirectional communication with ChatGPT by adding listening and speaking capabilities,” FMDB Transactions on Sustainable Computer Letters, vol. 1, no. 1, pp. 49-63, 2023.

[21] J. A. Jeba, S. R. Bose, and R. Boina, “Exploring hybrid multi-view multimodal for natural language emotion recognition using multi-source information learning model,” FMDB Transactions on Sustainable Computer Letters, vol. 1, no. 1, pp. 12-24, 2023.

[22] Y. Hou, T. Li, Y. Huang, X. Zhu, and L. Zhou, “A human-like two-stage conversational agent for emotional response generation,” in Proceedings of the AAAI Conference on Artificial Intelligence, 2023.

[23] A. S. Abdulbaqi, M. T. Younis, Y. T. Younus, and A. J. Obaid, “A hybrid technique for EEG signals evaluation and classification as a step towards neurological and cerebral disorders diagnosis,” International Journal of Nonlinear Analysis and Applications, vol. 13, no. 1, pp. 773-781, 2022.

[24] X. Zhang, Y. Sun, and Z. Liu, “Empathetic Response Generation with State Management,” arXiv preprint, 2023, [Online]. Available: https://doi.org/10.48550/arXiv.2205.03676.

[25] P. Cardoso, “Dialogue generation for context-aware emotional embodied agents,” Master’s thesis, 2025.

[26] G. Vadlakonda, “Advancing generative artificial intelligence (AI) through multimodal integration and contextual learning,” FMDB Transactions on Sustainable Computing Systems, vol. 2, no. 3, pp. 131-139, 2024.

[27] J. A. Jeba, S. R. Bose, R. Regin, S. S. Rajest, and M. M. Hasan, “Intelligent Tamil video summarisation: AI-powered NLP, translation, and speech integration for enhanced accessibility,” FMDB Transactions on Sustainable Computer Letters, vol. 2, no. 1, pp. 26-39, 2024.

[28] P. Sundaramoorthy, “Leveraging graph databases for real-time analytics in data engineering pipelines,” FMDB Transactions on Sustainable Intelligent Networks, vol. 2, no. 1, pp. 22-32, 2025.

[29] S. Singhal, A. Kotagiri, L. S. Samayamantri, and S. S. Rajest, “Interpretable machine learning models for human action and emotion deciphering,” in Advances in Computer and Electrical Engineering, USA: IGI Global, 2024, pp. 449-468.

[30] E. M. B. Y. Chai, S. R. Bose, and A. M. M. Reddy, “ChainTalker: Chain understanding and rendering for empathetic conversational systems,” in Findings of ACL, 2025, [Online]. Available: https://doi.org/10.18653/v1/2025.findings-acl.101.

[31] M. Kuppam, M. Godbole, T. R. Bammidi, S. S. Rajest, and R. Regin, “Exploring innovative metrics to benchmark and ensure robustness in AI systems,” in Explainable AI Applications for Human Behavior Analysis, USA: IGI Global, 2024, pp. 1-17.

[32] H. Chen, H. Li, and X. Zhao, “Deep emotion recognition in textual conversations: A survey,” arXiv preprint, 2023.

[33] K. Maheswari, N. Divyajyothi, K. Sreevidya, N. Sultana, and V. Vivekanandhan, “Enhancing the autonomy and adaptability of intelligent networks in IoT-based systems,” FMDB Transactions on Sustainable Computer Letters, vol. 3, no. 1, pp. 12-21, 2025.

[34] R. Regin, S. Thakshinraj, J. K. Raghav, S. J. Roosevelt, D. C. Pappa, S. S. Rajest, B. Singh, and S. Singh, “LipNet+ on a real-time lip-reading system for accurate speech recognition in news broadcasts and live telecasts,” in Advances in Computational Intelligence and Robotics, USA: IGI Global, 2025, pp. 141-172.

[35] M. M. Pereira and J. Moniz, “Deep emotion recognition in textual conversations: A survey,” arXiv preprint, 2023.

[36] R. Regin, P. K. Sharma, K. Singh, Y. V. Narendra, S. R. Bose, and S. S. Rajest, “Fine-grained deep feature expansion framework for fashion apparel classification using transfer learning,” in Advanced Applications of Generative AI and Natural Language Processing Models, USA: IGI Global, 2023, pp. 389-404.

[37] P. Paramasivan, S. S. Rajest, K. Chinnusamy, R. Regin, and F. J. John Joseph, Eds., Explainable AI Applications for Human Behavior Analysis, Advances in Computational Intelligence and Robotics, USA: IGI Global, Mar. 2024.

[38] C. C. Marella, “Streamlining big data processing with serverless architectures for efficient analysis,” FMDB Transactions on Sustainable Intelligent Networks, vol. 1, no. 4, pp. 242-251, 2024.

[39] M. S. Minu, S. S. Subashka Ramesh, R. Canessane, M. Al-Amin, and R. Bin Sulaiman, “Experimental analysis of UAV networks using oppositional glowworm swarm optimisation and deep learning clustering and classification,” FMDB Transactions on Sustainable Computing Systems, vol. 1, no. 3, pp. 124-134, 2023.

[40] S. Padmaja, S. Mishra, A. Mishra, J. J. V. Tembra, P. Paramasivan, and S. S. Rajest, “Insights into AI systems for recognising human emotions, actions, and gestures,” in Advances in Computational Intelligence and Robotics, USA: IGI Global, 2024, pp. 389-410.

[41] A. J. Obaid, B. Bhushan, Muthmainnah, and S. S. Rajest, Eds., Advanced Applications of Generative AI and Natural Language Processing Models, Advances in Computational Intelligence and Robotics, USA: IGI Global, Dec. 2023.