Artificial Intelligence Systems and Its Applications (AISA)
Volume 2, Issue 1, June 2026
http://journals.cognispectra.com/index.php/aisa/index
E-ISSN: 3110-0457

Emotionally-Aware Chatbots: A Survey

Endang Wahyu Pamungkas[1], Ikhlasul Amal[2]

[1] Informatics Department, Universitas Muhammadiyah Surakarta, Surakarta, Indonesia

[2] Computer Science and Electronics Department, Universitas Gadjah Mada, Yogyakarta, Indonesia

Abstract: Textual conversational agents, commonly known as chatbots, have attracted increasing attention from both academia and industry due to their widespread adoption in customer service, virtual assistance, and social companionship. One of the major challenges in chatbot development is enabling systems to communicate naturally while understanding and expressing human emotions. Emotionally-aware chatbots (EACs) have therefore become an important research direction in affective computing and natural language processing. This paper presents a comprehensive survey of EACs by reviewing four key aspects: their historical evolution, approaches for building EACs, available conversational datasets and affective resources, and evaluation methods. Our review shows that EAC development has evolved from early rule-based systems to encoder-decoder neural architectures based on Seq2Seq learning, emotion embedding, reinforcement learning, and attention mechanisms, before transitioning toward Transformer-based and Large Language Model (LLM) approaches. Recent studies further extend EACs through personalization, multimodal reasoning, knowledge distillation, and advanced emotion recognition techniques. We also summarize publicly available datasets and affective resources commonly used for emotion-aware dialogue generation and discuss qualitative and quantitative evaluation methods, highlighting the limitations of traditional automatic metrics such as BLEU in assessing emotional dialogue quality. Finally, we identify current challenges and promising future research directions, including multilingual resources, personalized emotional support, multimodal interaction, and ethical considerations for next-generation emotionally-

Keywords: Affective Computing, Chatbot, Conversational Agent, Natural Language Generation

Article History:

Received: 13 June 2026 Accepted: 30 June 2026 Published: 30 June 2026
Corresponding Author: Endang Wahyu Pamungkas, Email: ewp123@ums.ac.id

DOI: https://doi.org/10.65917/aisa.v2i1.71

1 Introduction

Conversational agents, commonly known as dialogue systems, have garnered significant attention from both industry and academia in recent years [1], [2]. Researchers have explored various approaches to developing conversational agents for specific domains, such as customer service [3], [4] and shopping assistance [5]. Alongside these domain-specific efforts, multipurpose agents such as Siri (https://www.apple.com/es/siri/), Amazon Alexa (https://developer.amazon.com/alexa), and Google Assistant (https://assistant.google.com/) have also been developed. Although this research area has been extensively studied within the field of Human-Computer Interaction, it remains an active and intriguing topic. The primary focus lies in developing intelligent, human-like machines capable of enhancing user engagement during interactions [6]. Improved engagement ultimately leads to higher user satisfaction, which is a key objective from an industry perspective.

In this study, we focus exclusively on text-based conversational agents, or chatbots, systems that employ various natural language processing techniques to engage in textual communication with humans. Multiple approaches exist for building chatbots, ranging from simple rule-based methods [7], [8] to more advanced neural-based techniques [9], [10]. Currently, chatbots are predominantly utilized in customer service applications such as booking systems [11], [12] and shopping assistance [4], or simply as conversational companions such as Endurance (https://bit.ly/2RGM40S) and Insomnobot (http://insomnobot3000.com/). Numerous studies have proposed approaches to enhance user engagement with chatbots, including the development of context-aware chatbots [13] and the integration of personality traits into these systems [14]. Furthermore, some researchers have explored the incorporation of affective computing to create emotionally-aware chatbots [3], [15], [16].

Several studies have demonstrated that incorporating emotion information into dialogue systems can enhance user satisfaction [17], [18]. Emotion plays a crucial role in fostering positive human-machine interactions and in reducing miscommunication [19]. Prior research has also shown that leveraging affective information enables chatbots to better understand users' emotional states, thereby improving response generation [20]. Additionally, the use of conversational tones has been explored to enhance service satisfaction. For example, employing an empathetic tone has been shown to reduce user stress and promote better engagement. Hu et al. [3] identified eight tones: anxious, frustrated, impolite, passionate, polite, sad, satisfied, and empathetic, as important factors in designing customer care chatbots.

In this paper, we present a comprehensive review of prior studies on the integration of emotion information into chatbots, with the aim of identifying current challenges and barriers in the development of engaging emotionally-aware chatbots. To this end, we formulate the following research questions:

- **RQ1:** How can emotion information be effectively incorporated into the construction of emotionally-aware chatbots?
- **RQ2:** What existing resources are available for building emotionally-aware chatbots?
- **RQ3:** How can the performance of emotionally-aware chatbots be evaluated?

The remainder of this paper is structured as follows. Section 2 provides an overview of the historical relationship between affective information and chatbots. Section 3 reviews studies focused on the development of emotionally-aware chatbots. Section 4 summarizes datasets and affective resources available for building such systems. Section 5 describes evaluation metrics applied in previous works. Finally, Section 6 concludes the paper and highlights potential directions for future research.

2 History of Emotionally-Aware Chatbot

The early development of chatbots was inspired by the Turing Test, proposed by Turing in 1950 [21]. ELIZA, developed by Weizenbaum [22], was the first publicly known chatbot, built using simple hand-crafted scripts. PARRY [23] was another early chatbot that successfully passed the Turing Test. Similar to ELIZA, PARRY employed a rule-based approach but with a more sophisticated design, incorporating a mental model capable of simulating emotion. As such, PARRY is considered the first chatbot to involve emotion in its development. Also noteworthy is ALICE (Artificial Linguistic Internet Computer Entity), a customizable chatbot built using Artificial Intelligence Markup Language (AIML), which operated by recursively executing a pattern-matcher to generate responses. In May 2014, Microsoft introduced XiaoIce [24], an empathetic social chatbot capable of recognizing users' emotional needs. XiaoIce facilitates engaging interpersonal communication by providing encouragement and other affective messages, enabling it to sustain human attention throughout a conversation.

In subsequent years, the majority of chatbot technologies have been built using neural-based approaches. The Emotional Chatting Machine (ECM) [16] was among the first works to exploit deep learning for building a large-scale emotionally-aware conversational agent. Several studies followed, introducing emotion embedding representations [25], [26], [27] or framing the problem as a reinforcement learning task [28], [29]. Most of these studies adopted an encoder-decoder architecture, specifically sequence-to-sequence (seq2seq) learning. Some works also introduced new datasets to establish better benchmarks and improve system performance. Rashkin et al. [15] introduced the EMPATHETICDIALOGUES dataset, containing 25,000 conversations enriched with emotional context information to facilitate the training and evaluation of textual conversational systems. Hu et al. [3] produced a dataset of 1.5 million Twitter conversations, gathered via the Twitter API from customer care accounts of 62 brands across several industries, which was used to build a tone-aware customer care chatbot. Lubis et al. [30] enhanced the SEMAINE corpus [31] through a crowdsourcing scenario to obtain human judgments on which responses elicit positive emotion, producing a dataset used to develop a chatbot capable of capturing human emotional states and eliciting positive emotion during conversation.

More recently, the emergence of large-scale pretrained language models, particularly Transformer-based architectures such as BERT and GPT, has fundamentally shifted the landscape of emotionally-aware chatbot development. Unlike earlier approaches that required explicit emotion classifiers or emotion embedding modules, these Large Language Models (LLMs) have been shown to possess emergent capabilities in understanding and expressing emotion as a result of large-scale pretraining and instruction tuning. Notably, Elyoseph et al. [32] demonstrated that ChatGPT achieved higher scores than the average human population on the Levels of Emotional Awareness Scale (LEAS), a validated psychological instrument, and received high ratings from licensed psychologists for the contextual appropriateness of its emotional responses. This finding suggests that LLMs represent a significant new paradigm in the development of emotionally-aware conversational systems, one that does not rely on task-specific emotional architectures but instead leverages the broad affective understanding acquired through pretraining.

3 Building Emotionally-Aware Chatbot (EAC)

As previously noted, emotion plays a vital role in constructing humanized chatbots. The roots of emotionally-aware chatbot development can be traced back to the work of PARRY in 1975 [23]. Currently, neural-based models dominate the landscape of emotionally-aware chatbot development. In this section, we review prior works focused on the development of emotionally-aware chatbots. The objectives and approaches of each work are summarized in Table 1. Early-stage development predominantly relied on rule-based approaches; however, in recent years, advancements have been largely driven by neural-based methods. Research in this area gained significant momentum from 2017, marked by the introduction of the first shared task in the Emotion Generation Challenge at NLPCC 2017 [33]. Table 1 reflects the substantial attention this research area has received in recent years.

As shown in Table 1, the majority of recent emotionally-aware chatbots have been constructed using an encoder-decoder architecture with sequence-to-sequence (seq2seq) learning. These seq2seq models are designed to optimize the likelihood of generating appropriate responses and are equipped to leverage large data sources. The fundamental seq2seq architecture consists of two recurrent neural networks (RNNs): an encoder that processes the input and a decoder that generates the response. Among RNN variants, Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) have emerged as the most prevalent choices for learning from conversational datasets. Some studies have also explored treating this task as a reinforcement learning problem, aiming to generate more diverse and contextually appropriate responses that facilitate successful long-term conversations. Furthermore, attention mechanisms have been introduced to enable the decoder to focus selectively on relevant parts of the input at each decoding step.

An essential component in the construction of emotionally-aware chatbots is an emotion classifier, which detects emotions within text to enable the generation of more meaningful responses. Emotion detection is a well-established task in natural language processing research, notably addressed in competitions such as SemEval-2018 (Task 1) and SemEval-2019 (Task 3), which focused on classifying utterances into various emotion categories [37]. Additionally, tasks aimed at predicting the intensity of emotions present in text have also been explored [38]. Initially, traditional machine learning approaches dominated emotion classification; however, neural-based approaches have since demonstrated superior performance, prompting wider adoption. In the context of chatbots, the system typically generates multiple candidate responses based on different emotion categories, then selects the most appropriate response based on the emotion detected in the user's utterance. Table 1 showcases a variety of emotion categories used across different works, reflecting the diverse objectives of each study.

More recently, the rapid advancement of Large Language Models (LLMs) and Transformer-based architectures has introduced a new wave of approaches to emotionally-aware chatbot development that goes beyond traditional encoder-decoder frameworks. Rather than relying on explicit emotion classifiers or embedding modules, modern systems increasingly leverage the inherent affective understanding of pretrained LLMs. For instance, Ehtesham-Ul-Haque et al. [39] proposed EmoBot, a chatbot grounded in Cognitive Appraisal Theory that maintains a dynamic internal emotional state throughout the conversation, enabling artificial emotion generation rather than merely detecting and responding to user emotions. In the context of efficient deployment, Zheng et al. [40] demonstrated that self-generated conversational data from large LLMs can be used to train smaller, more deployable models through a teacher–student framework, achieving high-quality emotional support responses without the computational cost of full-scale LLM inference. Research has also increasingly shifted toward personalized emotional support, with Zheng et al. [41] showing that users can actively customize the persona, communication style, and memory of LLM-powered chatbots to suit their individual emotional needs. On the technical side, advances in emotion recognition for conversation have continued, with Lan et al. [42] proposing Contrastive Adversarial Tuning (CAT) to enhance the discriminability and robustness of LLMs when distinguishing between fine-grained emotion categories. Similarly, Wang et al. [43] proposed TEMPO-LLM, a unified framework that textualizes multimodal cues, including audio features, facial expressions, and speaker personality traits, into prompts processable by an LLM, further extending the capabilities of emotionally-aware systems beyond text alone.

Table 1: Summarization of the Proposed Approaches for Emotionally-Aware Chatbot Development

Authors	Year	Focus	Approach
Kenneth Mark Colby [23]	1975	Designing a chatbot that behaves like a paranoid person	Rule-based approach with capability to simulate emotion
Zhou et al. [24]	2014	Building an empathetic chatbot for social interaction	Seq2seq learning with GRU-RNN model incorporating both intelligent quotient (IQ) and emotion quotient (EQ)
Zhang et al. [34]	2017	Building emotional conversation systems that produce multiple responses for every emotion category	Multi-task seq2seq model with GRU utilizing bidirectional long short-term memory (Bi-LSTM) as an emotion classifier
Asghar et al. [26]	2018	Building open-domain neural dialogue models augmented with affective intelligence	Encoder-decoder architecture with LSTM using cognitively engineered affective word embeddings and affectively diverse beam search for decoding
Sun et al. [28]	2018	Developing conversation generation that addresses the emotional factor by modifying the model's input	Encoder-decoder framework based on LSTM using three inputs: a sequence without an emotional category, a sequence with an emotional category for the input sentence, and a sequence with an emotional category for the output response
Hu et al. [3]	2018	Building a tone-aware chatbot for customer care on social media	Encoder-decoder architecture with LSTM, where the decoder is modified to handle meta-information by adding an embedding vector as a tone indicator to produce responses in several tones,

			including passionate, empathetic, and neutral
Sun et al. [28]	2018	Building an emotional human-machine conversation chatbot capable of understanding user emotions and generating satisfactory responses	Modeled as a reinforcement learning task, proposing a neural model based on the Seq2Seq framework using a generative adversarial network (GAN), referred to as EM-SeqGAN
Shantala et al. [25]	2018	Building a neural dialogue system that addresses the emotional aspect	Seq2Seq model with LSTM using a novel emotion embedding
Zhong et al. [35]	2018	Building an affect-rich open-domain human-machine conversation system	Extends the Seq2Seq model and adopts VAD (Valence, Arousal, and Dominance) affective notations; the proposed model also considers the effect of negators and intensifiers via a novel affective attention mechanism
Zhou et al. [16]	2018	Building the Emotional Chatting Machine (ECM) for large-scale conversation generation	Seq2seq learning with GRU incorporating emotion detection to capture implicit internal emotion states
Colombo et al. [27]	2019	Developing an affect-driven dialogue system that generates emotional responses in a controlled manner	Seq2seq with GRU incorporating an emotion classifier and affective re-ranking in the final step to produce the response
Li et al. [29]	2019	Building a conversational system that generates more meaningful and customizable emotional replies using editing constraints	Encoder-decoder architecture with a reinforcement learning approach, using an asynchronous decoder with a keyword predictor for topic prediction and an

			emotion editor for emotional response generation
Lubis et al. [30]	2019	Building a fully data-driven chat-oriented dialogue system that can dynamically mimic affective human interactions	Seq2seq response generator with an emotion encoder trained jointly with the entire network to encode and maintain emotional context throughout the dialogue, focusing only on positive emotion
Catania et al. [36]	2019	Building a modular framework to facilitate and accelerate the development and maintenance of intelligent conversational agents with both rational and emotional capabilities	The model consists of several modules including an intent predictor, sentiment analysis, emotion analysis, topic analysis, and profiling
Ehtesham-Ul-Haque et al. [39]	2024	Building a chatbot capable of generating artificial emotions based on Cognitive Appraisal Theory, with a dynamic internal emotional state maintained throughout the conversation	Multimodal input (audio and text) combined with cognitive appraisal variables mapped to six primary emotion states; EmoBot maintains mood, personality, and emotion as internal states that are continuously updated during the conversation
Zheng et al. [40]	2024	Building a small but high-quality emotional support chatbot by leveraging self-generated data from large LLMs through a teacher–student framework	ChatGPT is used as a counseling teacher to generate the ExTES dataset (11,177 dialogues); a smaller model (LLaMA-7B) is then fine-tuned via Knowledge Distillation and Diverse Response Inpainting (DRI) to produce diverse and empathetic responses
Zheng et al. [41]	2025	Investigating how users construct and interact with personalized LLM-powered chatbots for emotional support	Prototype platform (ChatLab) enabling users to customize chatbot persona, communication style, avatar, voice, and conversational memory; a one-week field study with 22 participants revealed diverse emotional support strategies and the importance of personalization

Lan et al. [42]	2026	Enhancing the discriminability and robustness of LLMs for Emotion Recognition in Conversation (ERC)	Contrastive Adversarial Tuning (CAT) combining context-based contrastive learning for intra-class compactness and inter-class separability, and distribution-based adversarial training for robustness under noisy conditions; evaluated on IEMOCAP, MELD, and EmoryNLP
Wang et al. [43]	2026	Building a unified LLM framework for multimodal ERC that integrates audio, visual, and personality-based cues	TEMPO-LLM textualizes multimodal features (audio cue prompts, facial expressions, gesture, posture, and Big Five personality traits) into LLM-readable prompts; GateRAG-CE is introduced for causal explanation of emotion predictions, achieving state-of-the-art results on MELD and IEMOCAP

4 Resource For Building EAC

In this section, we explore the available resources for building emotionally-aware chatbots (EAC). Like other artificial intelligence agents, chatbots require datasets to learn from in order to engage in meaningful, human-like conversations. Consequently, several studies have proposed datasets containing annotated textual conversations with different emotion categories. Table 2 provides an overview of the available datasets identified in recent years. We categorized the datasets based on language, data source, and additional details such as annotation approach, instance size, and emotion labels. All of these datasets were introduced between 2017 and 2018, beginning with the dataset provided by the organizers of the NLPCC 2017 Shared Task on Emotion Generation Challenge [33]. This dataset was collected from the Chinese social media platform Sina Weibo and consists of social conversations in Chinese. Our investigation reveals that the discovered datasets are predominantly available in two languages: English and Chinese. However, the sources of these datasets are diverse, encompassing social media platforms (Twitter, Sina Weibo, and Facebook Messenger), online content, and human-written text obtained through crowdsourcing. Additionally, each dataset employs a unique set of emotion labels based on the specific focus and objectives of the chatbot being developed.

Table 2: Summarization of dataset available for emotionally-aware chatbot

Authors	Year	Language	Source	Description
Zhou et al. [16]	2018	Chinese	Weibo	This dataset used the STC dataset [44] and performed automatic annotation using the best-performing classifier trained on NLPCC 2013 and NLPCC 2014 datasets. It contains 217,905 conversations, each response annotated with one of 6 emotion labels.

Rashkin et al. [15]	2018	English	ParlAI Platform	Built using a crowdsourcing scenario via ParlAI involving 810 different participants. Contains 24,850 conversations/prompts, each annotated with one of 32 emotion labels.
Huang et al. [33]	2017	Chinese	Weibo	Similar to [16], the training data was automatically annotated using a trained emotion classifier. Contains more than 1 million conversations as training data and 200 manually annotated conversations as a test set, each response labeled with one of 5 emotion labels.
Hu et al. [3]	2018	English	Twitter	Gathered from Twitter based on customer care conversations across 62 brands in different industries. Five hundred conversations were randomly selected and annotated with 8 major tones using the CrowdFlower platform.
Lubis et al. [45]	2018	English	Crowdsource	Built by enhancing the SEMAINE dataset [31] and adding labels for instances that elicit positive emotions (8 labels: alert, excited, elated, happy, content, serene, relaxed, calm). The final dataset contains 2,349 conversations.
Li et al. [46]	2017	English	Online Website	Crawled from various websites serving English learners to practice daily dialogue. DailyDialog contains 13,118 dialogs, manually annotated with communication intention and emotion information using six basic emotions: anger, disgust, fear, happiness, sadness, and surprise.

Huang et al. [47]	2018	English	Facebook Messenger	Collected from private conversations on the Facebook Messenger application. A total of 8,818 messages were manually annotated using seven emotion labels: joy, anger, sadness, anticipation, tired, fear, and neutral.

As mentioned in the previous section, the emotion classifier is an integral component of an emotionally-aware chatbot. When developing an emotion classifier, several affective resources have been widely used in emotion classification tasks. Table 3 presents the identified affective resources, ranging from older lexicons such as LIWC, ANEW, and DAL, to more modern resources such as DepecheMood and EmoWordNet. Previous studies have classified emotions from two fundamental perspectives: discrete categories and dimensional models. The discrete category view categorizes emotions into primary classes, with the most widely adopted classification proposed by Ekman et al. [48] consisting of six categories: anger, disgust, fear, happiness, sadness, and surprise. To extract emotions from text, multiple resources have been utilized, including EmoLex, EmoSenticNet, LIWC, DepecheMood, and EmoWordNet. On the other hand, the dimensional model views emotions as being defined by one or more continuous dimensions. Resources used for this purpose include the Dictionary of Affect in Language (DAL), ANEW, and NRC VAD.

Table 3: Summarization of the available affective resources for emotion classification task

Authors	**Name**	**Description**
Pennebaker et. al. [49]	Linguistic Inquiry and Word Count (LIWC)	LIWC is a computer application that offers an efficient instrument for word-by-word study of the emotional, cognitive, and structural elements in English. LIWC also provide a dictionary which has 4500 words distributed into 64 different emotional categories including positive and negative.
Mohammad et. al. [50]	Emolex	Emolex was built by crowdsourcing scenario. Emolex contains 14,182 words associated with eight primary emotion based on [45]: joy, sadness, anger, fear, trust, surprise, disgust, and anticipation.
Poria et. al. [52]	EmoSenticNet	EmoSenticNet (EmoSN) is an enriched version of SenticNet [47]. Work by [46] added emotion label by mapping WordNet-Affect label to the SenticNet concepts. As a WordNet-Affect label, EmoSenticNet uses Six Ekman's basic emotion: anger, disgust, fear, happiness, sadness, and surprise. The whole list of this lexicon contains 13,189 emotion labeled words.

Whissell et. al. [54]	Dictionary of Affect in Language (DAL)	DAL was developed by [48] and composed of 8,742 English words. These words were labeled by three scores representing three emotion dimensions: Pleasantness (the degree of pleasure), Activation (the degree of human response under some emotional condition), and Imagery (the degree of how easy the given words to be formed a mental picture).
Bradley et. al. [55]	Affective Norms for English Words (ANEW)	This lexicon consists of 1,034 English words rated based on the Valence-Arousal-Dominance (VAD) model [50].
de Albornoz et. al. [57]	SentiSense	SentiSense attaches emotional meanings to concepts from WordNet database. It is composed of a list of 5,496 words labeled with emotional categories consisting of 14 emotional categories, which refer to a merge of Arnold, Plutchik and Parrot models.
Strapparava et. al. [58]	WordNet-Affect	WordNet-Affect is a lexical resource created based on WordNet which contains information about the emotions that the words convey. This lexicon is organized in six basic emotions: anger, disgust, fear, joy, sadness and surprise, and contains 2,874 synsets and 4,787 words.
Mohammad [59]	NRC VAD	The NRC VAD Lexicon is a list of English words and their valence, arousal, and dominance scores ranged between 0 (low) and 1 (high). This lexicon contains 20,000 terms and was built by crowdsourcing scenario.
Staiano & Guerini [60]	DepecheMood	DepecheMood is an emotion lexicon that crowdsources emotions through Rappler website. This lexicon contains 37,000 entries rated by eight emotion categories (happy, sad, angry, afraid, annoyed, inspired, amused, and don’t care).

Badaro et.al. [61]	EmoWordNet	EmoWordNet is created by expanding an existing emotion lexicon, DepecheMood [54], by leveraging semantic knowledge from English WordNet. By aligning DepecheMood and WordNet, EmoWordNet consists of 67K terms, almost 1.8 times the size of DepecheMood. This lexicon has same emotion label as DepecheMood i.e., happy, sad, angry, afraid, annoyed, inspired, amused, and don't care.

5 Evaluating EAC

We categorize the evaluation of emotionally-aware chatbots into two distinct parts: qualitative assessment and quantitative assessment. Qualitative assessment primarily focuses on evaluating the functionality and usability of the system, including the chatbot's ability to understand and respond appropriately to user emotions, its capacity to engage in meaningful and empathetic conversations, and the overall user experience. Quantitative assessment, on the other hand, aims to measure chatbot performance using numerical metrics, including the accuracy of emotion detection and classification, response generation quality, coherence and contextual relevance of conversations, and user satisfaction and engagement. The combination of qualitative and quantitative evaluation provides a comprehensive assessment of an emotionally-aware chatbot's effectiveness and performance.

5.1. Qualitative Assessment

Based on our review of previous studies, the majority of works employed ISO 9241 as a framework to assess chatbot quality, with a specific focus on usability. Usability can be examined through three primary aspects: efficiency, effectiveness, and satisfaction, which collectively evaluate a system's performance in achieving its intended goals.

1) *Efficiency:* The efficiency aspect of usability encompasses several categories, such as robustness to manipulation and unexpected input [62]. This category evaluates the chatbot's ability to handle unexpected user inputs and resist manipulation attempts, ensuring reliable and accurate responses even in challenging situations. Additionally, some studies have explored the chatbot's capacity to control damage and address inappropriate utterances [63], assessing its ability to detect and appropriately handle offensive or harmful language while maintaining a respectful and safe conversational environment.
2) *Effectiveness:* The effectiveness aspect includes two distinct categories: functionality and humanity. From a functionality perspective, studies such as [64] propose assessing how accurately a chatbot can interpret commands and provide status reports, while other aspects include the accuracy of linguistic output and ease of use [65]. From a human perspective, a significant benchmark is whether the chatbot can pass the Turing Test, as suggested by Weizenbaum [22]. The Turing Test evaluates the chatbot's ability to exhibit human-like conversational behavior, making it difficult for users to distinguish between interacting with a human or a machine. Additionally, the chatbot's ability to respond to specific questions and maintain themed discussions are considered important capabilities to assess.
3) *Satisfaction:* The satisfaction aspect encompasses three categories: affect, ethics and behavior, and accessibility. Among these, the affect category is particularly relevant for evaluating emotionally-aware chatbots. It assesses various quality aspects related to the chatbot's ability to convey personality, provide conversational cues, and express emotional information through

tone, inflection, and expressivity. It also includes the chatbot's capacity to entertain and engage users, as well as its ability to read and respond to the moods of human participants [66]. The ethics and behavior category focuses on how a chatbot protects and respects user privacy, including sensitivity to safety and social concerns, as well as trustworthiness in handling user information [64], [67]. The accessibility category assesses the chatbot's ability to understand meaning or intent and respond to social cues, ensuring that it can effectively interact with users based on their specific requirements [11].

5.2. *Quantitative Assessment*

For quantitative evaluation, we divide into two different approaches based on previous studies including automatic evaluation and manual evaluation.

1) *Automatic Evaluation:* In automatic evaluation, researchers have focused on assessing chatbot systems at the emotion level, as observed in studies by Zhou et al. [16] and Li et al. [29], which utilize common metrics such as precision, recall, and accuracy to measure system performance against gold standard labels. This type of evaluation is analogous to emotion classification tasks, as seen in competitions such as SemEval-2018 [37] and SemEval-2019 [68]. Perplexity is another metric employed to evaluate chatbot models at the content level, determining the relevance and grammaticality of generated content in dialogue-based systems that rely on probabilistic approaches. Studies by Rashkin et al. [15], Lubis et al. [45], and Li et al. [29] utilize perplexity as an evaluation metric for assessing the quality of model responses. Furthermore, Rashkin et al. [15] employed the BLEU (Bilingual Evaluation Understudy) metric to evaluate machine-generated responses by comparing them against gold responses provided by humans. However, as noted by Liu et al. [69], using BLEU to measure conversation generation tasks is not recommended due to its low correlation with human judgment. This limitation has been further substantiated by recent work, which demonstrated that emotionally appropriate responses that are lexically simple may receive low BLEU scores despite being preferred by users, highlighting the need for more comprehensive, multi-dimensional evaluation frameworks that account for emotional appropriateness, contextual coherence, and user satisfaction [70].
2) Manual Evaluation: Manual evaluation involves human judgment to measure chatbot performance based on several criteria. Zhou et al. [16] employed three annotators to rate chatbot responses on two criteria: content (rated on a scale of 0, 1, 2) and emotion (rated on a scale of 0, 1). The content criterion assesses whether the response is natural, acceptable, and plausible as a human-produced utterance, a measurement widely adopted in conversation modeling tasks as proposed by Shang et al. [44]. The emotion criterion assesses whether the expressed emotion in the response aligns with the given gold emotion category.

Similarly, Li et al. [29] utilized four annotators to score responses based on three aspects: consistency, logic, and emotion. Consistency measures the fluency and grammatical correctness of the response, logic measures the degree to which the response matches the post in a logical manner, and emotion measures whether the response exhibits the appropriate emotion. Each aspect was rated on a scale of 0, 1, or 2. Lubis et al. [30] proposed two criteria, naturalness and emotion impact, to evaluate chatbot responses. Naturalness assesses whether the response is intelligible, logically coherent with the conversation context, and acceptable as a human response. Emotion impact measures whether the response elicits a positive emotion or triggers an emotionally positive dialogue.

Rashkin et al. [15] employed crowdsourcing to gather human judgments on three performance aspects: empathy/sympathy, relevance, and fluency. The empathy/sympathy aspect evaluated whether responses demonstrated an understanding of the feelings expressed by the user. Relevance assessed whether the responses were appropriate and on-topic. Fluency measured the understandability and accuracy of the language used. These aspects were recorded using a scale of 1 (not at all), 3 (somewhat), or 5 (very much), with responses collected from approximately 100 different annotators. After collecting human judgments, some studies such as Li et al. [29] and Lubis et al. [30] conducted t-tests to determine statistical significance, while others such as Zhou et al. [16] and Rashkin et al. [15] calculated inter-annotator agreement using measurements such as Fleiss Kappa.

6 Related Work

Several studies have provided comprehensive overviews of chatbot development across various contexts in both industry and academia. However, to the best of our knowledge, no study has been specifically dedicated to summarizing the development of emotionally-aware chatbots, a research area that has gained increasing attention in recent years.

Shawar and Atwell [71] offer a comprehensive history of chatbot technology development, describing the various practical domains in which chatbots have been utilized, including entertainment, language learning and practice, information retrieval, and assistance for e-commerce and other business activities. Deshpande et al. [72] provide a review of chatbot development from rudimentary models to more advanced intelligent systems, summarizing the techniques used across the early stages up to more recent years. More recently, Hussain et al. [73] present a systematic review of previous works on chatbot development, classifying chatbots into two main categories based on their goals: task-oriented and non-task-oriented. They further classify chatbot development techniques into three main categories: rule-based, retrieval-based, and generative-based approaches, providing a detailed summary of techniques within each category.

While these studies offer valuable insights into chatbot development broadly, a gap remains in the literature regarding a comprehensive overview specifically focused on emotionally-aware chatbots. The present survey aims to address this gap by consolidating advancements, challenges, and techniques related to integrating emotional understanding and expression within chatbot systems.

7 Discussion and Conclusion

In this work, we present a systematic review of emotionally-aware chatbots, focusing on three main aspects: incorporating affective information, available resources for building emotionally-aware chatbots, and evaluating their performance. The development of emotionally-aware chatbots has evolved considerably, from simple rule-based approaches such as PARRY [23], through sophisticated neural-based encoder-decoder architectures with sequence-to-sequence learning [16], to the current era of Large Language Models (LLMs) [32], [39], [40], [41], [42], [43]. Variants of recurrent neural networks, such as LSTM and GRU, were commonly employed in earlier learning processes, while modern systems increasingly leverage pretrained Transformer-based models through techniques such as fine-tuning, knowledge distillation, and contrastive learning. The Emotion Generation Challenge shared task at NLPCC 2017 [33] played a significant role in stimulating the development of emotionally-aware chatbots in the modern era.

Datasets for developing emotionally-aware chatbots are available but remain mostly limited to English and Chinese languages. These datasets are collected from various sources, including social media platforms, online websites, and manual construction through crowdsourcing. The key distinction between these datasets and common chatbot datasets lies in the presence of emotion labels. We also explored affective resources commonly used in emotion classification tasks, with a focus on English-language lexicons ranging from older resources such as LIWC [49] and EmoLex [50] to newer ones such as DepecheMood [60] and EmoWordNet [61].

To evaluate the performance of emotionally-aware chatbots, we classified assessment approaches into two techniques: qualitative and quantitative. Qualitative assessment often employs ISO 9241, covering aspects such as efficiency, effectiveness, and satisfaction. Quantitative evaluation can be conducted through automatic evaluation using metrics such as perplexity, or through manual evaluation involving human judgment. Notably, recent work has highlighted the limitations of widely used automatic metrics such as BLEU, which has been shown to inadequately capture emotional appropriateness and contextual quality in chatbot responses [69], [70], underscoring the need for more holistic evaluation frameworks going forward.

Overall, the efforts to humanize chatbots by incorporating affective aspects have gained significant and sustained attention. Several important future directions can be identified from this survey. First, the multilingual dimension remains underexplored, as current emotionally-aware chatbots primarily focus on English and Chinese, leaving a wide range of languages unaddressed. Second, the personalization of emotional support has emerged as a promising direction, with recent studies demonstrating that users benefit from being able to customize the persona, communication style, and memory of LLM-powered chatbots to suit their individual emotional needs [41]. Third, the integration of multimodal information — including audio, visual, and physiological signals — alongside textual input represents a growing frontier, as demonstrated by recent frameworks that textualize multimodal cues for direct processing by LLMs [43]. Finally, as emotionally-aware chatbots become increasingly capable and widely deployed, questions of ethics, safety, and contextual awareness will become ever more critical considerations for future research and development.

Acknowledgments

If applicable, acknowledge funding agencies, collaborators, or institutions that supported the research.

Funding Information

Clearly state funding sources. If no funding was received, mention: *”The authors declare no funding was received for this study.”*

Conflict of Interest Statement

Declare any conflicts of interest. If none, state: *”The authors declare no conflicts of interest.”*

Ethical Approval

If human or animal data was used, include an ethical approval statement. If not applicable, state: *"This study did not involve human or animal subjects."*

Data Availability

Provide information on how data can be accessed:

- The dataset used is available at [repository link].
- Data is available upon request from the corresponding author.
- Data sharing is not applicable to this study.

References

[1] B. Hancock, A. Bordes, P.-E. Mazare, and J. Weston, "Learning from dialogue after deployment: Feed yourself, chatbot!" in Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, Jul. 2019, pp. 3667–3684, doi: 10.18653/v1/P19-1358.

[2] H. Fang, H. Cheng, M. Sap, E. Clark, A. Holtzman, Y. Choi, N. A. Smith, and M. Ostendorf, "Sounding board: A user-centric and content-driven social chatbot," in Proc. 2018 Conf. North American Chapter Assoc. Computational Linguistics: Demonstrations, 2018, pp. 96–100.

[3] T. Hu, A. Xu, Z. Liu, Q. You, Y. Guo, V. Sinha, J. Luo, and R. Akkiraju, "Touch your heart: A tone-aware chatbot for customer care on social media," in Proc. CHI Conf. Human Factors in Computing Systems, 2018, p. 415.

[4] L. Cui, S. Huang, F. Wei, C. Tan, C. Duan, and M. Zhou, "Superagent: A customer service chatbot for e-commerce websites," in Proc. ACL 2017 System Demonstrations, 2017, pp. 97–102.

[5] A. Kothari, J. Hoover, and R. Zyane, Chatbots for E-Commerce: Learn how to Build a Virtual Shopping Assistant. Bleeding Edge Press, 2017.

[6] E. Go and S. S. Sundar, "Humanizing chatbots: The effects of visual, identity and conversational cues on humanness perceptions," Computers in Human Behavior, 2019.

[7] A. Ritter, C. Cherry, and W. B. Dolan, "Data-driven response generation in social media," in Proc. EMNLP, 2011, pp. 583–593.

[8] H. Al-Zubaide and A. A. Issa, "Ontbot: Ontology based chatbot," in Proc. IEEE Int. Symp. Innovations in Information and Communications Technology, 2011, pp. 7–12.

[9] M. Qiu, F.-L. Li, S. Wang, X. Gao, Y. Chen, W. Zhao, H. Chen, J. Huang, and W. Chu, "Alime chat: A sequence to sequence and rerank based chatbot engine," in Proc. ACL 2017, 2017, pp. 498–503.

[10] I. V. Serban et al., "A deep reinforcement learning chatbot," arXiv preprint arXiv:1709.02349, 2017.

[11] T.-H. Wen et al., "A network-based end-to-end trainable task-oriented dialogue system," in Proc. EACL 2017, 2017, pp. 438–449.
[12] B. Peng et al., "Composite task-completion dialogue policy learning via hierarchical deep reinforcement learning," in Proc. EMNLP, 2017, pp. 2231–2240.
[13] A. Sordoni et al., "A neural network approach to context-sensitive generation of conversational responses," in Proc. NAACL-HLT, 2015, pp. 196–205.
[14] S. Zhang et al., "Personalizing dialogue agents: I have a dog, do you have pets too?" in Proc. ACL, 2018, pp. 2204–2213.
[15] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "I know the feeling: Learning to converse with empathy," arXiv preprint arXiv:1811.00207, 2018.
[16] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in Proc. AAAI, 2018.
[17] Z. Yu, A. Papangelis, and A. Rudnicky, "Ticktock: A non-goal-oriented multimodal dialog system with engagement awareness," in AAAI Spring Symposium Series, 2015.
[18] H. Prendinger and M. Ishizuka, "The empathic companion: A character-based interface that addresses users' affective states," Applied Artificial Intelligence, vol. 19, no. 3–4, pp. 267–285, 2005.
[19] B. Martinovski and D. Traum, "Breakdown in human-machine interaction: The error is the clue," in Proc. ISCA Workshop Error Handling in Dialogue Systems, 2003, pp. 11–16.
[20] T. S. Polzin and A. Waibel, "Emotion-sensitive human-computer interfaces," in ISCA Workshop Speech and Emotion, 2000.
[21] A. M. Turing, "Computing machinery and intelligence," in Parsing the Turing Test, Springer, 2009, pp. 23–65.
[22] J. Weizenbaum, "ELIZA—a computer program for the study of natural language communication between man and machine," Communications of the ACM, vol. 9, no. 1, pp. 36–45, 1966.
[23] K. M. Colby, Artificial Paranoia: A Computer Simulation of Paranoid Processes. Elsevier, 2013.
[24] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, "The design and implementation of XiaoIce, an empathetic social chatbot," Comput. Linguistics, vol. 46, no. 1, pp. 53–93, 2020, doi: 10.1162/coli_a_00368.
[25] R. Shantala, G. Kyselov, and A. Kyselova, "Neural dialogue system with emotion embeddings," in IEEE SAIC, 2018, pp. 1–4.
[26] N. Asghar et al., "Affective neural response generation," in ECIR, Springer, 2018, pp. 154–166.
[27] P. Colombo, W. Witon, A. Modi, J. Kennedy, and M. Kapadia, "Affect-driven dialog generation," in Proc. NAACL-HLT, 2019, pp. 3734–3743, doi: 10.18653/v1/N19-1374.
[28] X. Sun et al., "Emotional human machine conversation generation based on seqGAN," in ACII Asia, IEEE, 2018, pp. 1–6.
[29] J. Li et al., "Reinforcement learning based emotional editing constraint conversation generation," arXiv preprint arXiv:1904.08061, 2019.
[30] N. Lubis et al., "Positive emotion elicitation in chat-based dialogue systems," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 27, no. 4, pp. 866–877, 2019.
[31] G. McKeown et al., "The SEMAINE database," IEEE Trans. Affective Computing, vol. 3, no. 1, pp. 5–17, 2012.

[32] Z. Elyoseph, D. Hadar-Shoval, K. Asraf, and M. Lvovsky, "ChatGPT outperforms humans in emotional awareness evaluations," Frontiers in Psychology, vol. 14, Art. 1199058, May 2023, doi: 10.3389/fpsyg.2023.1199058.

[33] M. Huang et al., "Overview of the NLPCC 2017 shared task: Emotion generation challenge," in Proc. NLPCC, Springer, 2017, pp. 926–936.

[34] R. Zhang et al., "Building emotional conversation systems using multi-task seq2seq learning," in Proc. NLPCC, Springer, 2017, pp. 612–621.

[35] P. Zhong, D. Wang, and C. Miao, "An affect-rich neural conversational model with biased attention and weighted cross-entropy loss," in Proc. AAAI Conf. Artif. Intell. (AAAI), vol. 33, no. 1, Honolulu, HI, USA, 2019, pp. 7492–7500, doi: 10.1609/AAAI.V33I01.33017492.

[36] F. Catania et al., "CORK: A conversational agent framework exploiting both rational and emotional intelligence," 2019.

[37] S. Mohammad et al., "SemEval-2018 task 1: Affect in tweets," in Proc. SemEval, 2018, pp. 1–17.

[38] S. Mohammad and F. Bravo-Marquez, "WASSA-2017 shared task on emotion intensity," in Proc. WASSA, 2017, pp. 34–49.

[39] M. Ehtesham-Ul-Haque, J. D'Rozario, R. Adnin, F. T. Utshaw, F. Tasneem, I. J. Shefa, and A. B. M. A. Al Islam, "EmoBot: Artificial emotion generation through an emotional chatbot during general-purpose conversations," Cognitive Systems Research, vol. 83, Art. 101168, 2024, doi: 10.1016/j.cogsys.2023.101168.

[40] Z. Zheng, L. Liao, Y. Deng, L. Qin, and L. Nie, "Self-chats from large language models make small emotional support chatbot better," in Proc. 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, Aug. 2024, pp. 11325–11345.

[41] X. Zheng, Z. Li, X. Gui, and Y. Luo, "Customizing emotional support: How do individuals construct and interact with LLM-powered chatbots," in Proc. 2025 CHI Conf. Human Factors in Computing Systems (CHI '25), ACM, New York, NY, USA, Art. 376, pp. 1–20, 2025, doi: 10.1145/3706598.3713453.

[42] K. Lan, C. L. P. Chen, Z. Zhang, and T. Zhang, "Contrastive adversarial tuning: Enhancing discriminability and robustness of LLMs for emotion recognition in conversation," Pattern Recognition, vol. 174, Art. 113025, 2026, doi: 10.1016/j.patcog.2025.113025.

[43] X. Wang, L. Xie, M. Wang, and Z. Wang, "Textualized multimodal priors and gated causal explanations for emotion recognition in conversations: A unified LLM framework," Expert Systems with Applications, vol. 311, Art. 131363, 2026, doi: 10.1016/j.eswa.2026.131363.

[44] L. Shang et al., "Neural responding machine for short-text conversation," in Proc. ACL, 2015, pp. 1577–1586.

[45] N. Lubis et al., "Eliciting positive emotion through affect-sensitive dialogue response generation," in AAAI, 2018.

[46] Y. Li et al., "DailyDialog: A manually labelled multi-turn dialogue dataset," in IJCNLP, 2017, pp. 986–995.

[47] C.-Y. Huang and L.-W. Ku, "EmotionPush: Emotion and response time prediction towards human-like chatbots," in IEEE GLOBECOM, 2018, pp. 206–212.

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

[49] J. W. Pennebaker et al., LIWC2001. Lawrence Erlbaum, 2001.

[50] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," Computational Intelligence, vol. 29, no. 3, pp. 436–465, 2013.
[51] R. Plutchik, "The nature of emotions," American Scientist, vol. 89, no. 4, pp. 344–350, 2001.
[52] S. Poria et al., "Enhanced SenticNet with affective labels," IEEE Intelligent Systems, vol. 28, no. 2, pp. 31–38, 2013.
[53] E. Cambria et al., "SenticNet 2," in FLAIRS, 2012.
[54] C. Whissell, "Dictionary of Affect in Language," Psychological Reports, vol. 105, no. 2, pp. 509–521, 2009.
[55] M. M. Bradley and P. J. Lang, "Affective norms for English words (ANEW)," Tech. Rep., University of Florida, 1999.
[56] C. Osgood et al., The Measurement of Meaning. University of Illinois Press, 1957.
[57] J. C. de Albornoz et al., "SentiSense," in LREC, 2012, pp. 3562–3567.
[58] C. Strapparava et al., "WordNet-Affect," in LREC, 2004, p. 40.
[59] S. Mohammad, "NRC VAD lexicon," in ACL, 2018.
[60] J. Staiano and M. Guerini, "Depeche Mood," in ACL, 2014, pp. 427–433.
[61] G. Badaro et al., "EmoWordNet," in LREC, 2018, pp. 86–93.
[62] A. Thieltges et al., "The devil's triangle," in AAAI Spring Symposium, 2016.
[63] K. Morrissey and J. Kirakowski, "Realness in chatbots," in HCI International, 2013.
[64] M. v. Eeuwen, "Mobile conversational commerce," Master's thesis, University of Twente, 2017.
[65] D. Cohen and I. Lane, "An oral exam for measuring a dialog system's capabilities," in AAAI, 2016.
[66] M. Meira and A. Canuto, "Evaluation of emotional agents' architectures," in World Congress on Engineering, 2015.
[67] A. S. Miner et al., "Smartphone-based conversational agents and mental health," JAMA Internal Medicine, vol. 176, no. 5, pp. 619–625, 2016.
[68] A. Chatterjee et al., "SemEval-2019 Task 3," in SemEval, 2019, pp. 39–48.
[69] C.-W. Liu et al., "How not to evaluate your dialogue system," in EMNLP, 2016, pp. 2122–2133.
[70] S. K. Gudipati, "Chatbot evaluation frameworks: From BLEU and F1 to multi-dimensional real-world benchmarks," in Proc. 2025 International Conference on Management Science and Computer Engineering (MSCE '25), ACM, New York, NY, USA, 2025, pp. 228–235, doi: 10.1145/3760023.3760060.
[71] B. A. Shawar and E. Atwell, "Chatbots: Are they really useful?" LDV Forum, vol. 22, no. 1, 2007.
[72] A. Deshpande et al., "A survey of various chatbot implementation techniques," Int. J. Computer Engineering Applications, vol. 11, 2017.
[73] S. Hussain et al., "A survey on conversational agents/chatbots classification and design techniques," in AINA Workshops, 2019, pp. 946–956.