Hindawi
Human Behavior and Emerging Technologies
Volume 2022, Article ID 9601630, 23 pages
https://doi.org/10.1155/2022/9601630

Review Article

Emotionally Intelligent Chatbots: A Systematic Literature Review

Ghazala Bilquise,[1] **Samar Ibrahim**,[2] **and Khaled Shaalan**[3]

[1]*Computer and Information Science Department, Higher Colleges of Technology, P.O. Box 15825, Dubai, UAE*
[2]*School of Arts and Sciences, American University in Dubai, UAE*
[3]*Informatics Department, The British University in Dubai, UAE*

Correspondence should be addressed to Ghazala Bilquise; gbilquise@hct.ac.ae

Received 2 July 2022; Revised 4 September 2022; Accepted 13 September 2022; Published 26 September 2022

Academic Editor: Zheng Yan

Conversational technologies are transforming the landscape of human-machine interaction. Chatbots are increasingly being used in several domains to substitute human agents in performing tasks, answering questions, giving advice, and providing social and emotional support. Therefore, improving user satisfaction with these technologies is imperative for their successful integration. Researchers are leveraging Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques to impart emotional intelligence capabilities in chatbots. This study provides a systematic review of research on developing emotionally intelligent chatbots. We employ a systematic approach to gather and analyze 42 articles published in the last decade. The review is aimed at providing a comprehensive analysis of past research to discover the problems addressed, the techniques used, and the evaluation measures employed by studies in embedding emotion in chatbot conversations. The study's findings reveal that most studies are based on an open-domain generative chatbot architecture. Researchers mainly address the issue of accurately detecting the user's emotion and generating emotionally relevant responses. Nearly 57% of the studies use an enhanced Seq2Seq encoding and decoding of the input of the conversational model. Almost all the studies use both the automatic and manual evaluation measures to evaluate the chatbots, with the BLEU measure being the most popular method for objective evaluation.

1. Introduction

The advancement of conversational technologies has led to a massive increase in the integration of chatbots in several domains. A chatbot is a dialog system that interacts with humans in natural language via text and voice or as an embodied agent with multimodal communication [1]. Chatbots are desirable by organizations because they provide proactive service and immediate assistance to consumers and cut operational costs [2]. They are used extensively to automate several tasks such as tracking deliveries, making reservations, requesting flight information, and placing orders. Their 24/7 availability and quick response to general queries make them an appealing solution for organizations. More recently, chatbots are also being used to provide social and emotional support in healthcare and personal lives [3].

Chatbots are the fastest-growing communication channel worldwide across multiple domains [4]. The enormous benefits of integrating chatbots in service and social disciplines lead organizations to invest highly in this technology. However, research indicates that users are still uncomfortable with chatbot communications and prefer interacting with a human agent [2]. Moreover, a review on chatbot usability and user acceptance shows that people prefer natural communication over machine-like interactions and believe that a human can understand them better [5]. The study also reveals that user satisfaction is imperative to successfully integrating and adopting chatbots. Therefore, improving user engagement and satisfaction with chatbot interactions has become crucial to provide a better experience and encourage users to embrace the technology [6].

In the last few years, Artificial Intelligence (AI) and Natural Language Processing (NLP) technologies have been driving the development of chatbots to enable advanced conversational capabilities [7]. Chatbots have evolved from utilizing pattern matching and rule-based models to using

AI-powered deep learning technologies that drastically excel in natural conversation [8]. The advancement in AI and NLP has enabled the development of chatbots that generate dynamic responses that do not exist in the database, thus making the conversation natural. However, despite these technologies, the responses generated by the chatbots are often dull and repetitive, which leads to user disengagement and frustration [9].

Understanding emotion and responding accordingly is the essence of effective communication [10]. Hence, the emerging trend in chatbot development is to create empathetic and emotionally intelligent agents capable of detecting user sentiments and generating appropriate responses [11]. Salovey and Mayer [12] proposed the term emotional intelligence, which refers to identifying, incorporating, comprehending, and controlling emotions. Emotions play a significant role in making or breaking a conversation. Users get frustrated when chatbot responses are irrelevant [13], while a chatbot that verbalizes emotions can enhance the user's mood [14]. Moreover, users often anthropomorphize chatbots, which in turn influences their interaction and behavior [15]. Chatbots that mimic human behavior and emotions lead to increased rapport, higher motivation, and better engagement [16]. Therefore, researchers are investigating ways to improve a chatbot's empathetic and emotional capabilities [17]. Ongoing research focuses on conversational agents capable of perceiving the user's emotion and responding appropriately with emotional cues to better engage users.

1.1. Problem Statement. Investigation into the development of emotionally intelligent chatbots is a recent trend as researchers continue to find better ways to generate human-like empathetic conversations. Although chatbots have existed for decades, the use of AI-driven techniques in empathetic conversational systems is relatively new. This area of research is confronted with several challenges such as the accurate recognition of emotion and emotional state of the user while keeping track of the history of the conversation and generating appropriate responses that are not dull and repetitive. Moreover, emotionally intelligent chatbots that generate diverse responses require a massive dataset [1]. Therefore, it is imperative to gain insights into the datasets used by empirical studies. Furthermore, the performance of a chatbot is measured by various evaluation strategies, making it vital to study the evaluation measures suitable for emotionally intelligent chatbots. Thus, it is imperative to study the state-of-the-art techniques in developing emotionally intelligent chatbots and report the findings to the research community to further the development in this field.

While several systematic reviews on chatbots exist in the literature, these studies differ from our study in their objectives. Some reviews examine chatbot applications and usage in a variety of domains, such as healthcare [3], neuropsychiatric disorders [18], education [19], business sectors [20], and personal assistants [21], with no focus on emotional aspects of the conversation or technical aspects of chatbot development. The study byMohamad Suhaili et al. [22] provides deeper insights into the technical aspects of chatbot development; however, the review is focused on service-oriented chatbots.

There are only a handful of studies that have reviewed empathetic chatbots. A systematic review by Rapp et al. [5] focuses on the human-computer interaction (HCI) perspective of chatbot usage by investigating the usability and user acceptance of human-like chatbots. Our study is differentiated from this study by focusing on the technical aspect of empathetic chatbot development rather than the emotional or psychological aspect of user interaction. The study by Wardhana et al. [23] provides a review of empathetic chatbot development based on the chatbot type, model, and inference techniques. Another study by Pamungkas [24] provides a survey on the approaches to building an empathetic chatbot. Ma et al. [25] survey empathetic dialog systems based on three aspects which include affective dialog, personalization, and knowledge. Notwithstanding their recognized contributions, these studies do not perform a thorough analysis using a systematic approach. Moreover, the past reviews have not provided insights into the challenges and techniques of emotion generation addressed by empirical studies. Furthermore, the studies do not provide researchers with the datasets and evaluation measures that are used in the development of emotion-aware chatbots. Our review offers a novel contribution to the study of emotionally intelligent chatbot development by providing an in-depth analysis of the challenges in emotion generation, techniques used, and evaluation criteria of empathetic chatbots. To the best of our knowledge, there is no systematic review that investigates the development of emotionally intelligent chatbots and their challenges, techniques, and evaluations.

Considering the above factors and the research gap, the objective of this paper is to provide a systematic literature review of the most relevant studies that investigate the development of chatbots enriched with emotional capabilities. We aim to use a methodical approach to discover, categorize, and present our findings on several aspects of emotionally intelligent chatbots and discover the gaps relevant to computer science researchers interested in advancing research in this field. Our study, in particular, is aimed at comparing and contrasting the overall characteristics among the studies, such as chatbot language, the domain of study, and trends. We examine the main problems tackled by researchers in developing emotion-aware conversational agents. We also investigate the techniques and approaches employed by studies in developing chatbots. Lastly, we study the evaluation measures used by the studies to evaluate their solutions. To that effect, our study analyzes contributions in this field and is aimed at answering the following research questions:

RQ1: what are the general characteristics of the studies?

RQ2: what problems are addressed by the studies?

RQ3: what approaches and techniques are employed in chatbot development?

RQ4: what evaluation measures are used to evaluate chatbot performance?

The remaining sections of this study are structured as follows. Section 2 presents the background information on chatbots with an overview of chatbots, chatbot architecture, and the role of emotional intelligence. Section 3 details the

methodology of the systematic review and the phases involved. Section 4 presents the findings of the study. Section 5 presents the discussion of the results. The conclusion, limitations, and further research avenues are presented in Section 6.

2. Background Information

This section presents an overview of chatbots, describing the various classifications used in the literature for describing them. We discuss the significance of emotional intelligence in chatbots, followed by a general chatbot architecture of the different types of chatbots and methods of integrating emotional intelligence in chatbot technology. The following subsections introduce the concepts of chatbot development, in particular incorporating emotion in a chatbot in order to better understand the terminologies and classifications used in the review.

2.1. Overview of Chatbots. Chatbots, also known as conversational agents, are dialog systems that interact with humans in natural language via text and voice or as embodied agents with multimodal communication [1]. A chatbot's primary function is to respond to user requests provided in textual-based or voice-based input. The chatbot processes the user input and generates an appropriate response.

There has been a surge in chatbot development in the last few years, with bot applications manifesting their presence in various domains [26]. Businesses deploy chatbots to provide efficient customer services by responding to customer queries and automating tasks [20]. Chatbots are used for teaching and learning activities, student advising, and administrative tasks [19]. Chatbots have become pervasive for psychiatric care and evaluation of medical diagnoses in the healthcare sector, raising awareness [18, 27]. Chatbots are also popular as social companions [11]. Social chatbots are not designed to accomplish a specific task but rather to engage with humans to fulfill their need for communication and social belonging [28]. Chatbots offer a cost-effective means of delivering services to consumers by eliminating repetitive and time-consuming human-agent communication while enabling the agents to focus on high-end complex tasks [2].

Several taxonomies are used in the literature to classify chatbots. Hussain et al. [29] categorized chatbots based on their purpose as task-oriented and non-task-oriented. The primary function of a task-oriented chatbot is to respond to domain-specific user queries and often perform tasks such as reserving a ticket. A non-task-oriented chatbot interacts with humans in open-ended, domain-specific conversations, also called open-domain chatbots. The primary function of these chatbots is to act as virtual companions where the dialog is open-ended.

Adamopoulou and Moussiades [9, 11] classified chatbots based on their response generation method as rule-based, retrieval-based, and generative chatbots. A rule-based chatbot selects a response based on a predefined set of rules. The responses are not dynamic and often repetitive. The strength of a rule-based chatbot lies in its ability to provide precise answers. However, it cannot detect lexical errors and works well when the input message is well formed. Moreover, a rule-based chatbot answers user queries without keeping track of previous responses and is ideal for a question-answer system.

A retrieval-based chatbot fetches responses from a sizeable predefined corpus using keyword matching or machine learning techniques to get the most appropriate response. Personal assistants such as Alexa, Siri, and Google Assistant are retrieval-based as they respond to user requests by retrieving information from a broad range of sources [26]. On the other hand, a generative chatbot generates responses using machine learning techniques, thereby constructing diverse responses by learning from the corpus. The responses are generated by translating input utterances to output data using statistical machine translation and predictive analytics techniques, thus making the conversation natural. A limitation of the generative model is that it requires massive training data. This limitation has led to the development of generative chatbots mainly for open domains since domain-specific conversational data is not readily available [30, 31]. A recent trend is using a hybrid approach by integrating retrieval-based and generative models to create task-oriented chatbots that possess human-like conversational skills to provide a better user experience [31].

The ongoing quest for developing chatbots that mimic humans is evident from its inception. One of the first chatbots, ELIZA and PARRY, was based on pattern matching technology to imitate human responses [26]. Both chatbots used a rule-based approach for generating responses based on keywords limiting the conversation to a predefined set of responses. In 1995, ALICE [32] was developed using Artificial Intelligence Markup Language (AIML) and was more sophisticated in generating human-like responses. Nevertheless, these primitive dialog systems could not keep up with the growing expectations of users in both conversational style and prediction of the user's intent. Chatbots these days are AI-driven and powered by Natural Language Processing (NLP) technologies that are capable of offering sophisticated solutions to meet the language and content expectations of end-users [26].

2.2. Emotionally Intelligent Chatbots. Despite the proliferation of chatbots in our daily lives, recent studies have shown that customers still prefer interacting with humans rather than bots [2]. This resistance is attributed to the poor conversational skills of chatbots which make the interaction unnatural and machine-like leading to frustration and communication breakdown [5]. Furthermore, end-users might be more willing to interact with chatbots if they are enriched with human-like interpersonal qualities [2]. Notwithstanding the limitation of chatbot conversational skills and high end-user expectations, conversational agents are still a desirable solution for reducing operational costs. Therefore, it has become critical for businesses to bridge the gap between customer expectations and chatbot technology.

Emotions play an integral part in an effective conversation. A study by Xu et al. [33] reveals that nearly 40% of customers' interaction with agents on social media is emotional

rather than informational. Several studies have shown that emotionally intelligent conversations lead to a good user experience resulting in fewer communication breakdowns [5]. A qualitative study by Svikhnushina and Pu [34] revealed that users are more likely to engage with emotion-aware chatbots and are eager to have a natural conversational experience with a virtual counterpart. Another study by Ghandeharioun et al. [14] disclosed that emotionally enriched responses by a chatbot could lift a user's mood, thus enhancing customer experience and improving customer relationships. Xiao et al. [31] supported these findings by showing that users are more engaged with chatbots capable of sensing and verbalizing emotions in the conversation. It is evident from these studies that perceiving emotions and responding with an appropriate empathetic reply is crucial to enhancing user satisfaction with chatbot conversations.

A vast amount of ongoing research is dedicated to integrating emotional capabilities in chatbots to enhance their conversational skills. AI-driven chatbots can detect user sentiments in a conversation, thus triggering the chatbot to comprehend the user's emotional state and generate an appropriate response. The following subsection presents an overview of an AI-driven chatbot architecture.

2.3. AI-Driven Chatbot Architecture. Chatbots are composed of several essential components, each playing an indispensable role and working together in a robust system that effectively serves its purpose. These components may be incorporated into text-based or voice-based agents [1]. In most cases, these components are organized in a pipeline based on their order of usage. Figure 1 presents the architecture showing the main components of a chatbot architecture.

2.3.1. Natural Language Processing (NLP). The first component is the Natural Language Processing (NLP) unit that processes the structured input using tokenization, lemmatization, and stemming techniques. Some chatbots apply these techniques to incoming user requests as a preprocessing strategy [22]. An additional Automatic Speech Recognition (ASR) component may exist in voice-based agents that extract text from the audio stream. In addition, the architecture may contain a nonverbal information extraction component, which can detect nonverbal information, like the user's emotions [1].

2.3.2. Natural Language Understanding (NLU). The structured data collected by the NLP unit is passed on to the Natural Language Understanding (NLU) component, which processes the data using various strategies. Usually, in this component, data structures are parsed to understand the user's intent and all particulars associated with that intent [35].

2.3.3. Dialog Manager. The dialog manager component examines the understandable structured data, maintains the dialog framework such as the semantic frame, and encodes the data to determine what action should be taken next. The dialog managers may request clarification from users if the semantic structure is incomplete to ensure that the dialog context is relevant and that all ambiguities are resolved [11]. The dialog manager relies on external or internal sources of data. Internal data sources might be embedded as a template or rules in Artificial Intelligence Markup Language (AIML) to decipher user requests and retrieve responses. Additionally, the chatbot may construct its database internally from scratch or utilize existing databases outfitted with their domains and functions. Alternatively, chatbots may use third-party APIs to obtain external data sources [22].

2.3.4. Natural Language Generator (NLG). Finally, the response generation component, NLG, is based on how the chatbot generates responses. It may use a retrieval-based, rule-based, or generative model. Retrieval- and rule-based models are simple in design and need essential intelligence to select the best response match. However, they have limited usability and flexibility [22]. In comparison, the generative model has incredible flexibility and can handle a variety of domains. However, they can be highly complex and expensive, and they need an extra degree of intelligence.

Researchers who study emotionally intelligent chatbots have adopted the general chatbot architecture. They implemented the neural-based approach, and they use models that enforce emotion-aware characteristics, such as emotion embedding and reinforcement learning models, in addition to encoder-decoder architectures that use Sequence-to-Sequence learning [24].

2.4. Deep Learning in Chatbot Conversations. Artificial neural networks are machine learning algorithms that may be supervised or unsupervised. Deep learning, being an unsupervised machine learning algorithm, can mimic how the human brain develops patterns and employs them for making decisions [29]. There has been an increase in the use of deep learning neural networks in conversational modeling, particularly Recurrent Neural Networks (RNNs), Sequence-to-Sequence (Seq2Seq) networks, and Long Short-Term Memory (LSTM) networks [22].

RNN is an artificial neural network class and a type of recursive artificial neural network. This method saves the output from a layer and feeds that saved output to the new input to forecast the following output. In the context of natural language, RNN captures the inherent sequential nature of words, where the meaning of words is understood through their relationship to the previous words in the sentence. Due to this approach, RNNs are well suited for chatbots since understanding the user input and producing contextually relevant responses is essential [29].

Research in emotionally intelligent chatbots employs encoder-decoder architecture with Seq2Seq learning. The Seq2Seq model utilizes RNN as its architecture, with an encoder processing the input and a decoder producing the output. This model was initially introduced in 2014 as a variation of Ritter's generative model incorporating advancements in deep learning to enhance accuracy [29]. The Seq2Seq model is applied to chatbots to transform input status into output response. It is currently regarded as the industry's best practice for generating responses because Seq2Seq maximizes the likelihood of the response and is

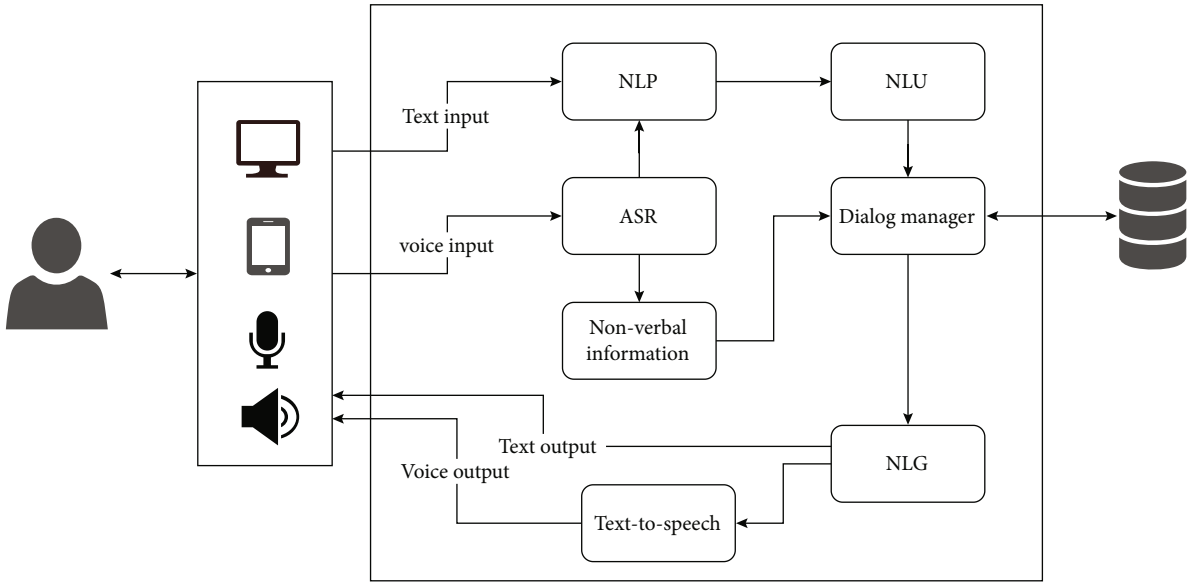

Figure 1: Chatbot architecture.

capable of processing a large amount of data to generate the optimal response [24].

Despite its approximation of a good response, the Seq2-Seq function fails to meet the chatbot's true purpose of simulating human-to-human communication [35]. Therefore, LSTM, a type of RNN, is designed to overcome the long-term dependency problem of RNNs. LSTMs contain memory cells and gates that can retain previous information for long periods where input gates control the data stream, forget gates, and output gates. The LSTM or Gated Recurrent Unit (GRU) is the dominant variant of RNNs used to learn the conversational dataset in these models. An LSTM network outperforms the traditional RNN and other sequence learning networks and replaces these models in learning from experience.

Some studies have implemented LSTM with reinforcement learning tasks to get more generic responses and enable the chatbot to attain long-term conversation effectiveness [35]. In addition to this model, research has shown that the Conditional Variational Autoencoder (CVAE) model can also improve the diversity of responses. In CVAE, a latent variable is used to learn a distribution over possible conversational intents, and greedy decoders are used to generate responses [36].

2.5. Emotionally Intelligent Chatbot Technology. It is crucial to select preprocessing steps carefully when building an emotionally intelligent chatbot, as different preprocessing techniques suit different contexts. For example, the NLP process is primarily used to collect, tokenize, and parse information. Parsing is a technique that implements algorithms where the input is deconstructed according to a predefined rule, such as left-right or bottom-up [37].

Embedding techniques are commonly used in emotionally intelligent chatbot technologies. The embedding model transforms the input text data into a numerical form that is easily understood by the machine [24]. Various embedding methods exist, such as character embedding, word embedding, and sentence embedding. Word embedding is a compact vector representation of words in the lower-dimensional space. It is possible to represent words and phrases with matrices that produce massive sets of data as the size of the input increases, such as a bag of words and Term Frequency-Inverse Document Frequency (TF-IDF) [22]. Word2Vec and BERT models are the two popular word embedding models used in neural networks, which can also be used for emotion and semantic embedding. These models strive to maximize conditional probabilities for better word matching [37].

In terms of semantic relations among linguistic concepts, the Valence, Arousal, and Dominance (VAD) [38] space is widely used as the primary source of structure since it accounts for about 70% of the variance in meaning. VAD ratings have also been used in empathetic tutoring, sentiment analysis, and other affective computing applications. The three standard dimensions of emotion are Valence (the pleasantness of a stimulus), Arousal (the intensity of emotion produced by the stimulus), and Dominance (the degree of power produced by the stimulus). There are three levels of emotion intensity in these words: very low (e.g., dull), moderate (e.g., watchdog), and very high (e.g., insanity) [39].

The artificial neural-based approach is extensively used to develop emotionally intelligent chatbots. Artificial neural network-based chatbots apply both the retrieval-based and generative approaches for producing responses. However, the research trend is heading towards generative approaches [24] as it offers diverse responses. This paper explores the research studies investigating AI technologies to generate emotionally intelligent responses to report state-of-the-art techniques.

3. Research Methodology

This study explores existing literature on the development of emotionally intelligent chatbots by adopting the systematic review framework of Kitchenham and Charters [40]. This framework was chosen because it defines the guidelines for conducting reviews in the technical field instead of other frameworks like Tranfield et al. [41] that are more oriented towards qualitative studies in the medical field. A rigorous theoretical framework is essential to guiding the comprehensive data collection and inquiry methods required for our investigation. Moreover, the methodical process ensures the reliability of our findings. The systematic literature review guidelines by Kitchenham and Charters [40] outline a thorough method for collecting, analyzing, and documenting findings from secondary data sources. We aim to answer our research questions following this methodology to uncover the latest trends and technologies to develop emotionally intelligent chatbots.

The review process is divided into three phases: planning the review, conducting the review, and reporting the results. Each phase is further subdivided into several steps, each of which is described in the sections below.

3.1. Planning the Review. In recent years, a vast amount of research has been conducted to improve user satisfaction with chatbot conversations by detecting user sentiments and generating appropriate emotional responses. Therefore, it is crucial to provide researchers with the current state of the art regarding emotionally intelligent chatbots, including the techniques used to embed emotions in computer-generated responses, the datasets used, and the evaluation processes adopted to measure the performance of the chatbots.

To begin our systematic review, we start with the planning phase that defines the search strategy and the inclusion/exclusion criteria and identify the data sources used for selecting the articles of the study. Finally, we describe the quality assessment checklist for assessing the quality of the articles and set a threshold for their inclusion.

3.1.1. Search Strategy. The primary aim of the search criteria is to investigate the latest advances in the development of emotionally intelligent chatbots. To that effect, we conducted a preliminary search of existing literature and systematic reviews to understand our study's context, keywords, and scope. We used the Population, Intervention, Comparison, Outcome, and Context (PICOC) method outlined by Petticrew and Roberts [42] as a guideline to define our research directions. In this regard, our study's population relates to the main keywords and their derivatives with similar connotations for emotionally driven chatbots, such as conversation agents, virtual or digital assistants for chatbots, and empathy or feelings for emotion. We used these keywords to define the search string for the search process presented in Section 3.2.1. The intervention in our study refers to the search context [42]. We used the identified keywords to filter studies that meet our objectives: Emotional, Chatbot, Conversational agent, and virtual assistant. In the comparison step of PICOC, we consider all possible approaches, models, development, algorithms, and evaluation metrics in developing emotionally intelligent chatbots. The outcome determines our data coding requirements and results, including the knowledge of techniques used in developing emotionally intelligent chatbot solutions and the problems addressed, the datasets, and the evaluation metrics used. Finally, we define the context as only empirical studies related to emotionally intelligent chatbot development.

3.1.2. Inclusion/Exclusion Criteria. Selecting articles for the review led us to outline essential criteria that define the characteristics of the studies included in the study. Table 1 summarizes the inclusion/exclusion criteria applied for selecting the articles. First, empirical studies related to the development of chatbots with emotion-embedded responses were included. Second, only peer-reviewed journal and conference papers were included in the study, thereby excluding books, book chapters, and reviews. Third, only articles published in the English language were included to eliminate the bias that may result from poor translation. Finally, the study period was determined to be between 2011 and 2022, as chatbot development with the integration of AI techniques has emerged in recent years. A ten-year period is sufficient to view the emotionally intelligent chatbot research trend.

3.1.3. Data Sources. Various data sources were considered to retrieve relevant publications for this study, ranging from general to computer science topics. Accordingly, the search utilized the following six digital databases: Scopus, IEEE Xplore, ProQuest, ScienceDirect, ACM Digital Library, and EBSCO. Furthermore, we also used a manual snowballing method to identify additional relevant studies by exploring references of all selected primary studies.

3.1.4. Quality Assessment Checklist. Quality assessment is crucial in systematic reviews to ensure the validity of the results and reduce bias that may be caused due to the inclusion of less robust studies [43]. Furthermore, the quality assessment also provides more detailed inclusion/exclusion criteria [40].

To ensure a rigorous assessment of the included articles in our review, we developed a quality assessment checklist consisting of eleven questions presented in Table 2. We considered the elements essential to our data extraction and coding phases, such as relevance to our study, clear identification of the problem statement, and validity of the results. Furthermore, we also considered the source's credibility, which we evaluated using the ranking of the journal/conference and the number of citations of the study.

3.2. Conducting the Review. We implement the plan by searching and retrieving the articles in this phase. The articles were retrieved in Jan 2022. The articles were further screened using the inclusion/exclusion criteria and quality assessment checklist described in Section 2.

3.2.1. Search Process. An extensive range of search strategies was used to retrieve the studies from the identified databases

Table 1: Inclusion/exclusion criteria.

Inclusion criteria	Exclusion criteria
Must be an empirical study on the development of chatbots	Studies that are qualitative or not related to chatbot development
Must involve emotion detection in input and generation of appropriate emotional response	Studies that do not consider emotion in chatbot conversation
Must be a peer-reviewed journal or conference paper	Book, book chapters, reviews, or articles in the press
Must be written in English	Papers written in a language other than English
Must be published between 2011 and 2022	Papers published prior to 2011

Table 2: Quality assessment checklist.

#	Question
Q1	Is the study relevant to our research?
Q2	Are the research aims and contributions clearly identified?
Q3	Is the problem statement clear?
Q4	Is the experimental setup described adequately?
Q5	Are the techniques/methods clearly explained and analyzed?
Q6	Are the results compared to previous studies/baseline?
Q7	Is sufficient data used for the evaluation of the model?
Q8	Is the proposed technique evaluated using established metrics?
Q9	Is the conclusion explained clearly and linked to the purpose of the study?
Q10	Is the source of the article credible (published in a ranked venue)?
Q11	Has the study been cited in other publications?

to raise the probability of identifying highly relevant studies. We used logical operators **AND** and **OR** by combining the keywords identified in the planning process. Furthermore, the search was performed on the title, abstract, and keywords to ensure that relevant studies were not left out. The following is the search query syntax used in all the identified databases: ("*chat bot*" *OR* "*chatbot*" *OR* "*talkbot*" *OR* "*talk bot*" *OR* "*personal assistant*" *OR* "*virtual assistant*" *OR* "*digital assistant*" *OR* "*conversational agent*") *AND* ("*emotional*" *OR* "*emotion*" *OR* "*emotions*" *OR* "*empathy*" *OR* "*sentiment*" *OR* "*feeling*").

In addition to the automated search, we also performed the manual snowballing search as detailed in the planning process. The results of the search are presented in Table 3. A total of 2219 results were retrieved with the highest number of studies from Scopus because it is generic and sources publications from all domains.

After retrieving the search results, we performed a bibliometric analysis of the results to analyze the research areas. Figure 2 shows the visualization of the terms in the results, constructed using VOSviewer [44]. The diagram presents the significance and interconnections between the frequently occurring terms extracted from the abstract, title, and keyword search results. The size of the shape and the label associated with the term determines its importance. The color of the terms determines the clusters in the visualization. Each cluster represents terms related to each other in that group. Moreover, the distance between the clusters represents the relatedness of the clusters.

The visualization of the terms in the extracted studies reveals several clusters. This shows that there are various dimensions of studies on empathetic chatbots from the perspective of usage, applications, usability and user experience, and chatbot development. The clusters are tightly overlapped, indicating that several aspects of the studies are interrelated. The clusters show that the current research trends on emotionally intelligent chatbots are on chatbot response generation, chatbot effectiveness, chat evaluation, and usability. Considering only the clusters with highly weighted terms, four main clusters can be seen in the visualization. The first and central cluster (red) includes the following keywords: *emotional intelligence*, *emotional*, *conversational agent*, *research*, and *emotional response*. This cluster implies that research is active in this area and related to chatbot empathetic response generation. The second cluster (purple) contains keywords such as *effectiveness*, *conversational agent*, *framework*, *patient*, *problem*, and *technique*, which entails that the area of research in this cluster is about chatbot usage and effectiveness. In the third cluster (green), the significant keywords are *input utterance*, *factor*, *generation model*, and *human evaluation*, which implies that research is related more to evaluating chatbot technology. Finally, in the fourth cluster, the main keywords are *consistency*, *performance*, *human*, *affect*, and *technique*, indicating that research in this cluster is about chatbot performance (light blue). Examining these clusters provides an idea of where research findings are located for better analysis and discussion of the studies.

Table 3: Search results.

Database	Search results
Scopus	1003
IEEE Xplore	115
ProQuest	191
ScienceDirect	487
ACM Digital Library	272
EBSCO	121
Snowballing	30
Total	**2219**

3.2.2. Article Selection. In this phase, we applied the inclusion/exclusion criteria to screen the retrieved articles for eligibility following the PRISMA [45] framework. This framework provides a detailed guideline and structured approach to screen the documents. The steps of the screening process are outlined in Figure 3.

First, we removed the duplicate records. Then, we applied the inclusion/exclusion criteria to ensure that only relevant articles were included. Each author independently performed a title and abstract screening of the studies to remove irrelevant articles. At this stage, most of the articles ($n = 1671$) were excluded as they did not match the inclusion criteria. As discovered in the network analysis of the search terms, most of the articles were related to usability and user acceptance of chatbots. These articles were excluded as they did not contribute to the context of our study. Next, we performed a full-text screening of the remaining articles ($n = 325$) to assess relevance and eligibility. Each author performed this step independently by equally dividing the studies to be reviewed. In cases where the eligibility was unclear, the authors discussed resolving the discrepancy. Finally, a quality assessment was performed of the remaining articles ($n = 57$) after the full-text screening. In this step, the authors screened the articles initially screened by the other to reduce bias and ensure that each article was reviewed twice. Finally, 42 studies were included in the systematic literature review.

3.2.3. Quality Assessment. Each author performed the quality assessment independently using the assessment checklist presented in Table 2 using a scale from 0 to 1, where 1 represents that the criteria are wholly met, 0.5 represents partially met, and 0 represents not met. We assigned one point to the article having at least two citations per year regarding the number of citations. Table 4 presents the detailed quality assessment of the included articles, showing that all included articles are of good quality. It must be noted that the quality assessment is a means to determine whether the selected article is relevant to the contribution of this study, with no attempt to criticize any of the studies and their findings.

3.2.4. Data Analysis and Coding. The objective of this phase is to accurately record all findings of the study by collecting metadata from the primary studies included in the review. The metadata relates to the research questions of our study. We conducted a thorough data analysis of all the relevant features identified in the planning phase to accomplish this task. The metadata analysis includes various characteristics that are essential to answering our research questions, such as the characteristics of the study in terms of publication type and year. We also examine the technical aspects of the study, such as the chatbot's language of development, the emotions detected and used, the problems addressed, the technique used for the development and evaluation measures of the chatbot, and the dataset used for evaluation.

3.3. Reporting the Review. The final phase of the systematic review presents the study results. After a detailed and in-depth analysis of the metadata extracted from the full-text review, we present our findings in Section 4 to answer each research question.

4. Results

This section presents the results obtained from the meta-analysis and in-depth review of the included articles with reference to our research questions. We analyzed 42 journal and conference papers published in the span of 10 years to determine the state-of-the-art technologies used to develop emotionally intelligent chatbots. The following subsections present the results of each research question.

4.1. RQ1: What Are the General Characteristics of the Studies? This subsection presents the general characteristics of the reviewed articles. We analyzed the distribution of the studies by the year of publication, the region of study, the source type (journals vs. conference papers), the interface language, the chatbot type, and the domain of study. These characteristics provide an overview of the development trend of emotionally intelligent chatbots.

4.1.1. Source of the Articles. Figure 4 shows the distribution of studies by source type. Most of the studies included in the review are peer-reviewed journals ($n = 25$), while conference papers ($n = 17$) constitute 40% of the studies. The overall distribution of the papers is balanced. Moreover, all the sources of the studies were verified in the quality assessment phase to ensure content validity and reduce bias resulting from inaccurate or poorly reported results.

4.1.2. Publication Year. Figure 5 presents the distribution of the reviewed articles by year of publication. It is evident from the graph that there is significant interest in emotion-aware chatbots over time. The graph also reveals a sharp increase in the investigation of emotionally intelligent chatbots in 2018. This may be attributed to the technological advancement of conversational technologies and a sudden surge of chatbot usage in 2016, referred to as the chatbot "tsunami" by Grudin and Jacques [26]. The Sequence-to-Sequence model [46] published by Google became a basis for most neural conversational agents, leading to the proliferation of generative chatbot studies.

4.1.3. Chatbot Type. Figure 6 presents the distribution of studies by chatbot type. A chatbot may be text-based,

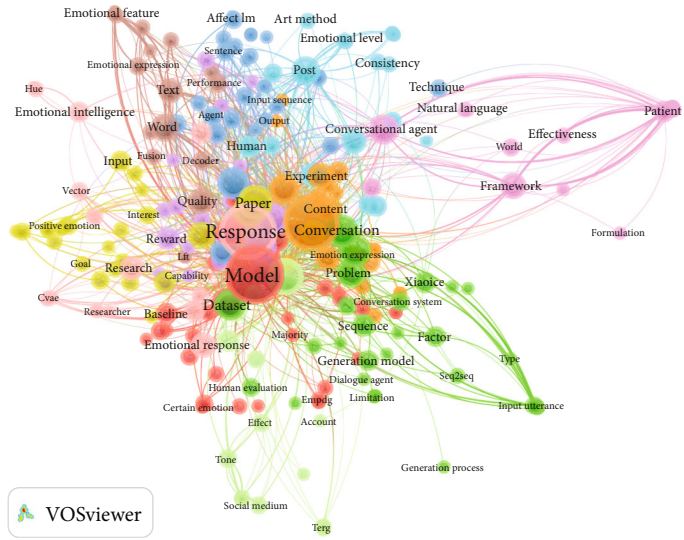

Figure 2: Bibliometric analysis of search results.

voice-based, or multimodal. The majority of the work in developing emotionally intelligent dialog systems is on text-based chatbots. Text-based chatbots are more favored than other forms due to the increased use of messaging technologies. Moreover, the chatbots need to be trained to produce an appropriate response. It is easier to train chatbots to generate text rather than speech, as the training data for text is more readily available than speech data.

4.1.4. Domain of Study. Figure 7 shows that most emotionally intelligent chatbots have been developed for the open domain. These chatbots specialize in natural and emotionally rich conversations without focusing on specific topics. High development in this area is due to the lack of conversational dataset availability. Moreover, the conversational dataset must be labeled with emotion as a preprocessing step before using it in the conversational dataset.

4.1.5. Chatbot Language and Region of Study. Figure 8 reveals that English and Chinese are the two most predominant interface languages to develop emotionally intelligent chatbots. Chang and Hsing [47] support our finding and claim that due to the popularity of social media in China, the Chinese language will soon be one of the most prevailing languages online. Figure 9 further shows that most of the studies originated from China. This finding shows that China has played a leading role in developing empathetic and emotion-aware chatbots since 2018. Moreover, one of the first emotionally intelligent chatbots, XiaoIce, developed by Microsoft, is vastly used in China to provide emotional support to users [48]. These findings reveal the popularity of chatbots in Chinese culture and that China is taking a lead role in developing AI technologies.

4.2. RQ2: What Problems Are Addressed in the Chatbot Development? After conducting an in-depth analysis of the studies, we identified seven main problems addressed by all the studies. Figure 10 shows an overview of the main problems and how studies have addressed the problems. This section describes each problem and the approaches employed to resolve the problem.

4.2.1. Response Diversity. The studies highlight the limitation of the Seq2Seq model [46] as it produces dull and meaningless responses. Therefore, several studies ($n = 8$) tackle the challenge of generating diverse responses that are emotionally relevant. Asghar et al. [39] argued that neural conversational models do not capture the complexity of emotions and often result in short and ambiguous responses. They used a heuristic search algorithm to ensure diversity in generated responses. Multiple studies employed a CVAE-based model to generate varied emotional responses [36, 49–52]. Yao et al. [52] argue that a chatbot must generate diverse responses for the same input to simulate a human-like conversation. Their model uses a latent space variable and six emotion categories to generate multiple responses that generate multiple emotionally consistent responses.

Similarly, Liu et al. [36] also generate several responses and select the most appropriate one based on grammar, meaning, and emotional score. Zhang et al. [53] argued that an intervention mechanism is needed to improve response diversity. They consider the input emotion and model a responder state and topic preference to generate diverse responses.

4.2.2. Content Relevance. Several studies ($n = 9$) focus on content relevance to achieve a natural conversation in human-computer dialog systems. Both Srinivasan et al. [54] and Sun et al. [55] used reinforcement learning with a reward function to ensure that the responses were content-specific and emotionally relevant. Several studies embedded topics and emotions in the decoder to generate appropriate responses that are emotionally appropriate [53, 55–57]. Huo et al. [49] augmented the encoder-decoder with a topic-aware decoder to enhance the content relevance of the response. They differentiated words in output as

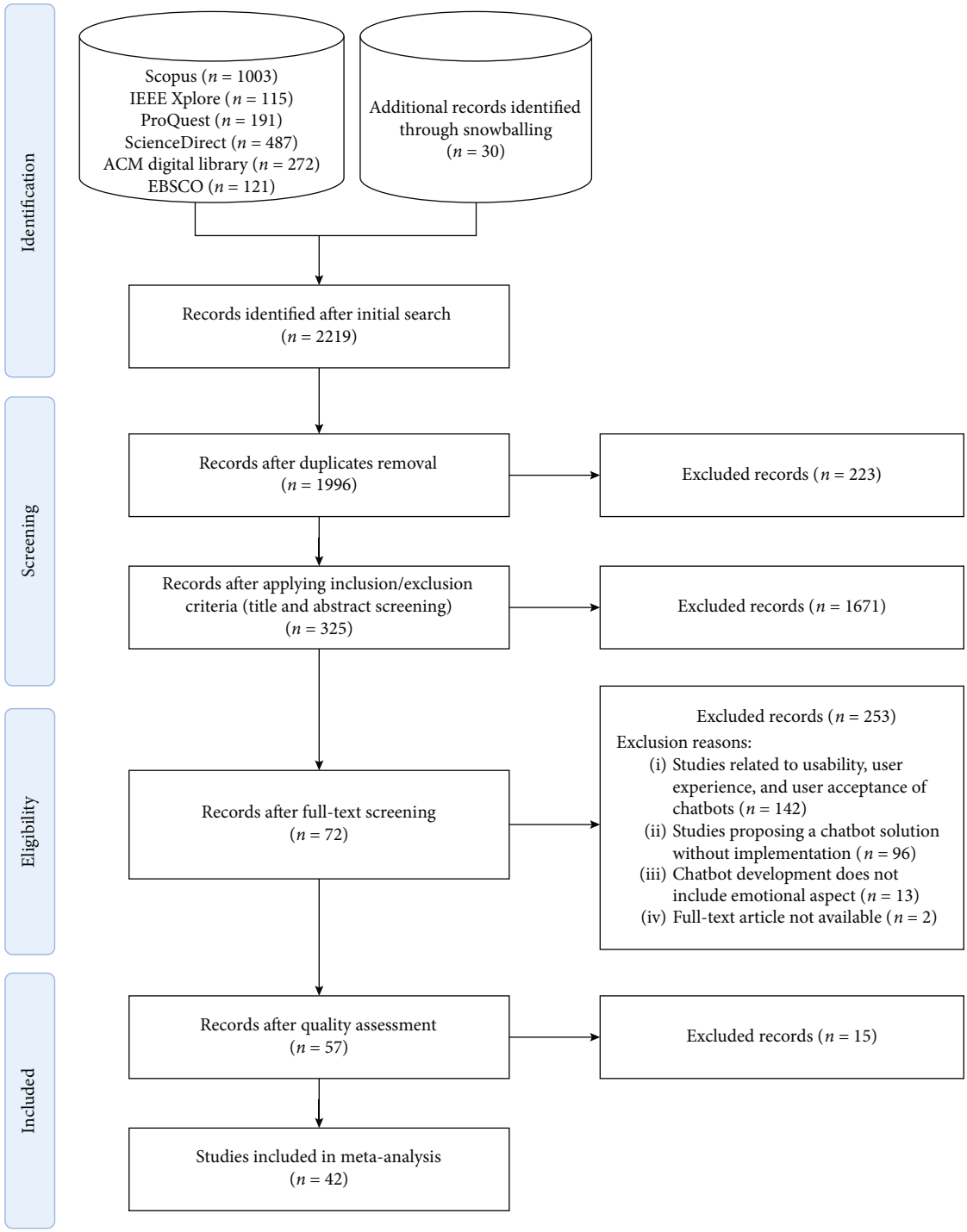

Figure 3: PRISMA flowchart.

emotion-related, keywords, and familiar words. In two separate studies, Wei et al. [58] and Wei et al. [59] focused on generating emotionally intelligent and content-relevant responses by embedding semantics and emotions in the input.

4.2.3. Poor Emotion Capture. A large number of studies ($n = 14$) focus on accurately detecting the emotion of the input message. While some studies predicted the input emotion using a classifier [60, 61], several studies argue that emotions are complex and cannot be captured by a coarse-grained emotion label. To that effect, some studies predict the emotion by applying the principle of Valence and Arousal (VA) to embed affective meaning for each word in the input message [47, 62, 63]. Other studies built on the previous work and embedded each input word with a three-dimensional emotion embedding based on Valence, Arousal, and Dominance (VAD) [38] to achieve a more

Table 4: Quality assessment.

SNo	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11	Total	%
1	1	1	1	1	1	1	1	1	1	1	0	10	91%
2	1	1	1	1	1	0.5	1	0.5	1	1	1	10	91%
3	1	1	1	1	1	1	1	1	1	1	0.5	10.5	95%
4	1	1	1	1	1	1	1	1	1	1	0.5	10.5	95%
5	1	1	1	1	1	1	1	1	1	0	1	10	91%
6	1	1	1	1	1	1	1	1	1	1	1	11	100%
7	1	1	1	1	1	0.5	1	1	1	1	1	10.5	95%
8	1	1	1	0.5	0	1	1	0.5	1	0.5	1	8.5	77%
9	1	1	1	1	1	1	1	1	1	1	1	11	100%
10	1	1	1	1	1	0	1	0.5	1	1	1	9.5	86%
11	1	1	1	1	0.5	0.5	1	1	1	0.5	1	9.5	86%
12	1	1	1	1	1	1	1	0	1	1	1	10	91%
13	1	1	1	1	0.5	1	1	1	1	1	0.5	10	91%
14	1	1	1	1	1	1	1	1	1	1	1	11	100%
15	1	1	1	1	1	1	1	1	0.5	1	1	10.5	95%
16	1	1	1	1	1	1	1	1	1	1	0	10	91%
17	1	1	1	1	1	1	1	1	1	1	0.5	10.5	95%
18	1	1	1	1	1	0	1	1	1	1	1	10	91%
19	1	1	1	1	1	1	1	1	1	1	1	11	100%
20	1	1	1	1	1	1	1	0.5	1	1	1	10.5	95%
21	1	1	1	1	1	1	1	1	1	1	1	11	100%
22	1	1	0.5	1	1	1	1	1	1	1	1	10.5	95%
23	1	1	1	1	1	1	1	1	1	1	0.5	10.5	95%
24	1	1	1	0.5	0.5	0.5	1	1	1	1	1	9.5	86%
25	1	1	0	1	1	1	1	1	1	1	1	10	91%
26	1	1	1	1	0.5	0.5	1	1	1	1	1	10	91%
27	1	1	0.5	1	1	0.5	1	0	1	1	1	9	82%
28	1	1	1	1	1	1	1	1	1	1	1	11	100%
29	1	1	1	1	1	1	1	1	1	1	0	10	91%
30	1	1	1	1	1	0	0.5	0.5	1	1	1	9	82%
31	1	1	1	1	1	1	1	1	0.5	1	1	10.5	95%
32	1	1	1	1	1	1	1	1	1	1	1	11	100%
33	1	1	1	1	0.5	0.5	1	1	0.5	1	0	8.5	77%
34	1	1	1	0.5	0.5	1	1	1	1	1	1	10	91%
35	1	1	1	1	1	1	1	1	1	1	1	11	100%
36	1	1	0.5	1	1	1	1	1	1	0	1	9.5	86%
37	1	1	1	1	1	1	1	1	1	1	0.5	10.5	95%
38	1	1	1	1	1	1	1	1	1	1	1	11	100%
39	1	1	1	1	1	1	1	1	1	0	1	10	91%
40	1	1	1	1	1	1	1	1	1	1	1	11	100%
41	1	1	0	1	1	1	1	1	1	1	1	10	91%
42	1	1	1	1	1	1	1	1	1	1	1	11	100%

fine-grained emotion detection [36, 39, 51, 64, 65]. Li et al. [66, 67] argue that words in messages are usually connected and show that capturing the connections of words enables a deeper understanding of the user's emotion.

Some studies focus on detecting the user's emotional state rather than just predicting the sentiment from a single utterance. Lin et al. [68] identify the user's emotional state by employing a tracker that determines the various emotional aspects of the input. They use multiple decoders to respond to each emotional category and generate an appropriate response. Hasegawa et al. [69] argue that natural conversation is achieved only when the user's emotional state is

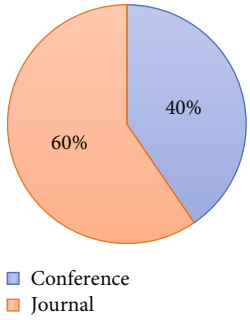

Figure 4: Distribution of studies by article source.

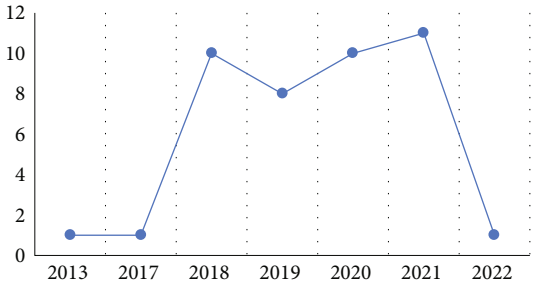

Figure 5: Distribution of studies by publication year.

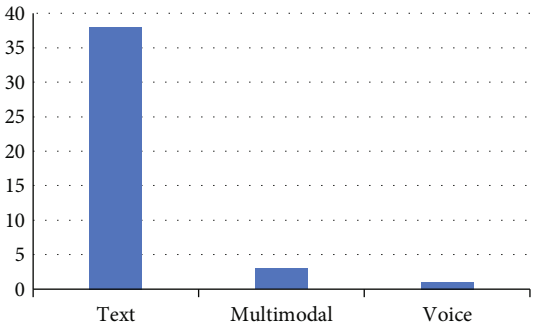

Figure 6: Distribution by chatbot type.

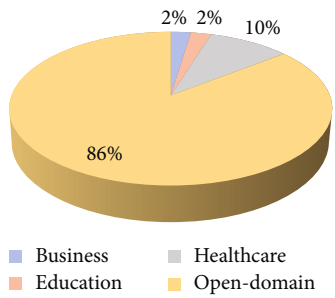

Figure 7: Distribution by domain of study.

predicted from historical conversational utterances rather than a single utterance. They generate the response based on a predicted target emotion using past utterances. Similarly, Li et al. [50] also utilize conversational data to generate more relevant responses.

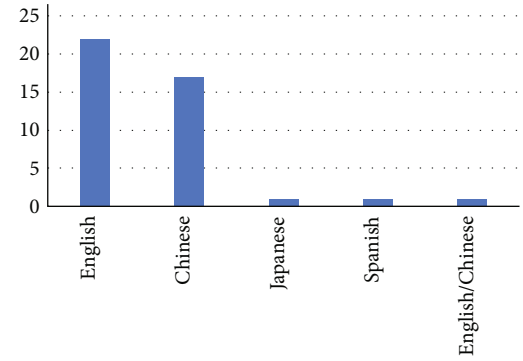

Figure 8: Distribution by chatbot interface language.

On the other hand, Li et al. [66, 67] argue that it is crucial to understand the reason behind the user's emotion and develop a chatbot that elicits the emotional cause by asking appropriate questions. They generate the response based on the chat history and the identified cause. Qiu et al. [70] track the user's emotional state using a transition network. They model the dynamic emotion flow to predict emotions based on past utterances and generate the most appropriate response.

4.2.4. Irrelevant Emotional Responses. Several studies argue that emotions generated using an NLP chatbot are often not emotionally relevant and attempt to alleviate the problem by controlling the emotion exhibited in the response. Several studies control the generated response by embedding a target emotion in the response generator module [8, 61, 67–69, 71, 72]. Zhou et al. [61] use internal and external memory to generate explicit emotional words in the response. Niu and Bansal [72] conditioned the response generator to generate polite, rude, or neutral responses.

Several other studies argued that a predefined label to condition the response generator suffers from poor quality of response [59], and furthermore, it cannot be assumed that the output emotion must be the same as the input emotion. To this effect, some studies attempted to generate more dynamic responses. Zhang et al. [73] generate multiple responses for six emotional categories and select the most appropriate response based on rankings. Similarly, Colombo et al. [64] use two Seq2Seq models to generate several responses and rank them based on emotion to get the most appropriate response. Zhou et al. [74] add an additional emotion classifier model for the responses over multiple emotional distributions, generating two types of responses, one for the specified emotion and one unspecified.

4.2.5. Lack of Emotionally Labeled Conversational Datasets. One of the challenges of developing a chatbot using machine learning is that it requires a massive dataset for training. While several conversational datasets are available for the open domain, datasets labeled with emotions are not readily available. Therefore, several studies resorted to classifying conversational data using a dynamic classifier as a preprocessing technique. Few studies tackled the challenge of the lack of a publicly available labeled corpus of conversational

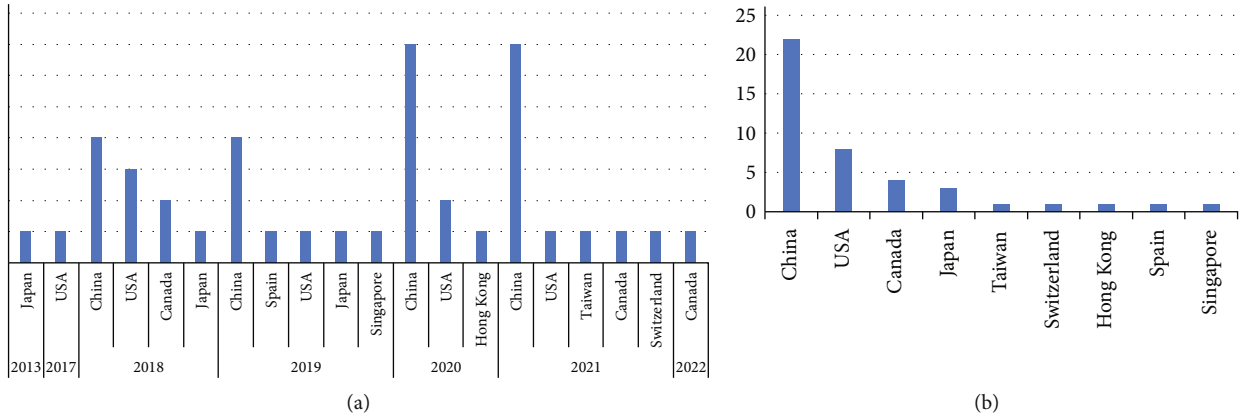

Figure 9: Distribution by region of study.

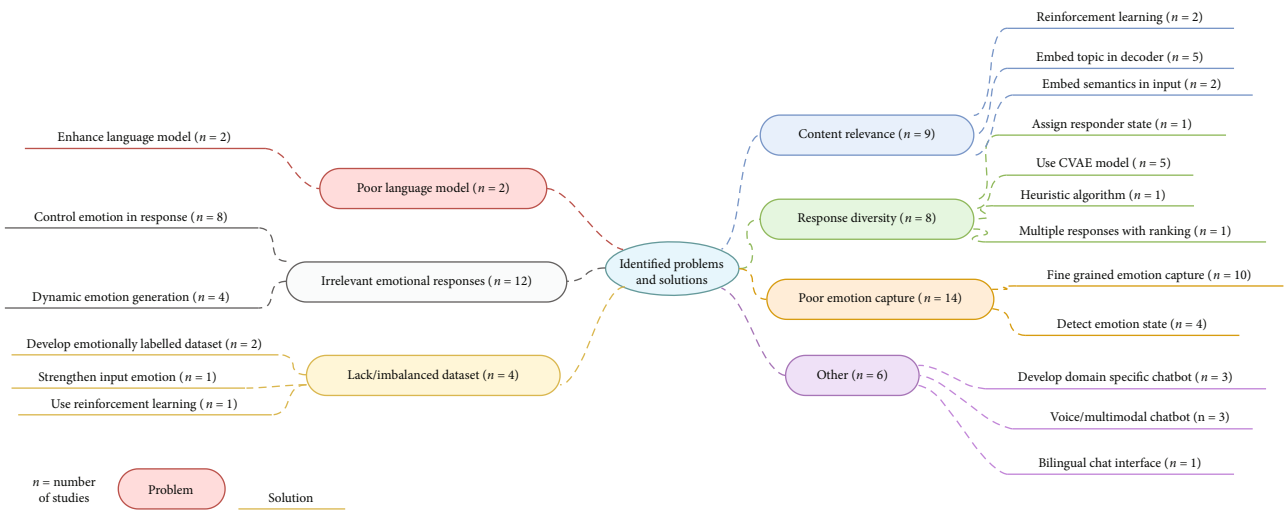

Figure 10: Mind map of problems and solutions.

data. Rashkin et al. [17] developed an empathetic dataset of 25 thousand labeled conversations and tested it against well-known neural models. Zhou and Wang [75] generated a labeled dataset from Twitter using emojis as labels to depict the emotion of the input. Their dataset consisted of 64 emotional labels. Song et al. [76] argued that an emotionally labeled dataset of conversations is usually imbalanced, which leads to incorrect predictions. To alleviate the issue, they explicitly embedded emotional words in the input to increase the strength of the emotion. On the other hand, Srinivasan et al. [54] used reinforcement learning to address the unavailability of supervised training data.

4.2.6. Poor Language Model. Two studies addressed the problem of a weak language model for emotional responses. Ghosh et al. [77] extended the LSTM language model trained in a conversational speech corpus to generate text enriched in emotion. In another study, Casas et al. [60] attempt to understand the context and implicit emotions expressed in input data to generate empathetic responses. To that effect, they developed an enhanced language model for empathetic responses.

4.2.7. Other. While previous studies addressed challenges in enhancing an emotionally intelligent chatbot to enable perceiving emotion and generating appropriate responses, some studies focused on other novel areas. We classified it into three main areas.

(1) Domain-Specific Chatbot. Some studies addressed the problem specific to a domain. For example, Adikari et al. [78] stated that previous chatbots in the healthcare sector mainly focused on question-answer systems. They developed a rule-based chatbot that detects patient emotion using NLP techniques and generates a response using a template. In another study based on the healthcare domain, Wang et al. [79, 80] developed a chatbot that provides timely responses to users seeking emotional support. Hu et al. [8] claim that previous chatbots in customer care focused solely on grammar and syntax. They highlight the significance of emotional

intelligence in customer care and develop a chatbot that integrates tones in responses by embedding target tones (empathetic or passionate) in output.

(2) Voice/Multimodal Chatbot. Few studies investigated emotionally intelligent voice-based and multimodal chatbots. Griol et al. [81] enhance communication in virtual educational environments by integrating emotion recognition in social interaction with multiple modalities. The study utilizes user profile data and context information from the dialog history to generate emotionally appropriate responses. Hu et al. [82] claim that emotion recognition in vocal responses is novel and explores emotion regulation in voice-based conversations. Their model comprehends the input emotion using acoustic cues and generates emotional responses by integrating emotional keywords in the generated response.

(3) Bilingual Chatbot Interface. Wang et al. [79] use a bilingual decoding algorithm that captures the contextual information and generates emotional responses in two languages. The model employs two decoders to generate primary and secondary language responses.

It is essential to note that some of the problems identified are also applicable to chatbots that are not emotionally intelligent; however, the development of these chatbots faces additional complexities. For example, all chatbots are confronted with the challenge of generating diverse and relevant responses. However, the additional challenge for emotionally intelligent chatbots is to ensure that the diverse and relevant response matches the emotion of the interlocutor. On the other hand, several challenges are specific to empathetic chatbots such as accurate detection of emotion, generation of emotional response, and lack of emotionally labeled datasets.

4.3. RQ3: What Approaches and Techniques Are Employed in Chatbot Development? This section discusses the various approaches and techniques used in the studies to develop an emotionally intelligent chatbot. Figure 11 presents a taxonomy that classifies the major adopted models and divides them into four categories relating to response generation techniques: Seq2Seq model, rule-based model, CVAE-based model, and other models. These studies further used three different approaches to detect emotion in the input and response: lexicon-based, machine-based, and hybrid method that combines both types of learning. Lexicon-based learning and machine-based learning are two distinct emotion detection techniques used in emotionally intelligent chatbots; i.e., one captures the emotion using a dictionary, and the second captures the emotion by training a classifier. In contrast, the hybrid model adopts both these techniques in emotion detection. The taxonomy diagram reveals that lexicon-based learning is the most used method by studies that address the problem of capturing emotions accurately. The machine learning approach enables the detection of emotion in a more coarse-grained approach.

4.3.1. Response Generation Models

(1) Seq2Seq-Based Model. Nearly 50% of the studies ($n = 24$) developed an emotionally intelligent chatbot using a Seq2-Seq model, in which a query is represented by one sequence of words and the response by another sequence. Studies have been conducted to extend the model and improve the performance of Seq2Seq and address the limitation of having dull and meaningless responses by generating an appropriate emotional response.

(2) CVAE Model. Some studies ($n = 6$) adopt the CVAE approach to develop an emotionally intelligent chatbot to generate diverse and affective responses and overcome limitations created by adopting the Seq2Seq model. CVAE allows a more diverse response generator, but syntax and grammar errors are compromised to a certain extent.

(3) Rule-Based Model. Only two reviewed studies use a rule-based approach to develop emotionally intelligent chatbots, using a hybrid approach to combine lexicons and machine learning to achieve the desired results. The first study extracts individual emotions from patient conversations using NLP techniques based on a psychological emotion model proposed by Plutchik that sets up an emotion dictionary from a variety of pretrained language models such as Word2Vec and GloVe Bag of Words. Furthermore, they use AI techniques and multiple classifiers to detect the group's emotions. Both kinds of emotions, the group and individual emotions, are used to capture the emotion expression sequence. A rule-based system is used to generate responses based on negative emotions expressed by patients to predict and generate an automated personalized empathetic alert [78]. The second study uses a rule-based approach to detect, predict, and build a statistical response generator based on an utterance's tags. The training data were automatically obtained from Twitter, in which a classifier is trained to predict and generate specific emotions based on conversational history [69].

(4) Other Approaches. Ten studies ($n = 10$) utilize approaches that do not fall into the previous categories. Chen et al. [6] use an encoder-decoder architecture in which the semantic and multiresolution emotional contexts are encoded. In addition, they implement 2-CNN-based semantics with an emotional discriminator used to capture fine-grained emotion using NRC emotion vocabulary for response generation. Wu et al. [83] use encoders-decoders that create emotional label datasets to generate various emotional responses. The model by Lin et al. [68] consists of an emotion detector that uses a transformer encoder and an empathetic listener. The model utilizes an independently parameterized transformer decoder with a metalistener to fuse listeners' information and produce an empathetic response. Casas et al. [60] used a pretrained DeepMoji DailyDialog dataset to build an emotion classifier using a labeled training set to predict emotional states in text-based messages. Furthermore, Griol et al. [81] combine information from the user profiles with emotional content extracted from

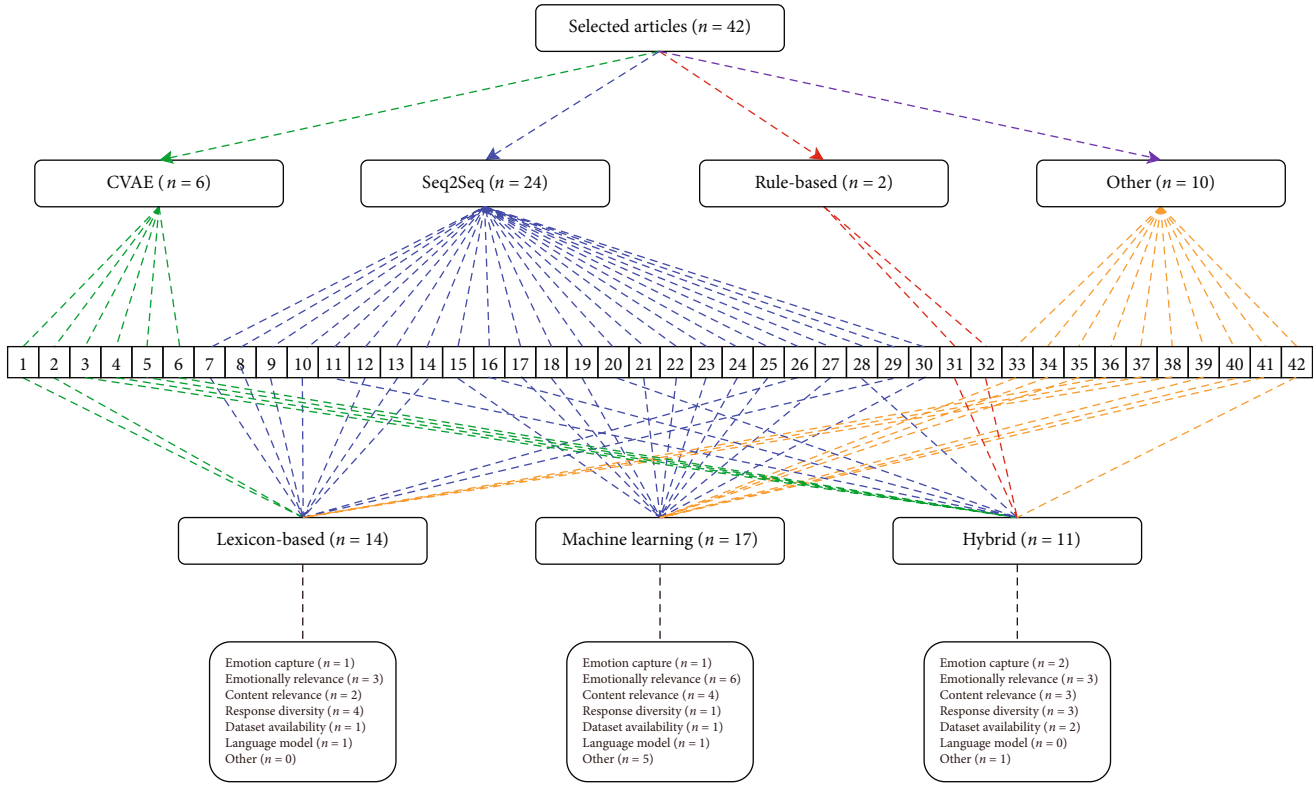

Figure 11: Taxonomy of the selected articles.

the user's utterances and apply an emotional recognizer in the dialog manager to choose an adapted system response. Rashkin et al. [17] use a generative pretrained transformer and an emotion classifier trained on the DailyDialog (DD) dataset to predict emotional states using an encoder-decoder model. However, Sun et al. [55] use a topic class embedding based on the LDA vector that generates a topic keyword. An emotion embedding vector generates the emotion keyword by reinforcement learning to generate accurate emotional responses.

4.3.2. Input Emotion Detection Models

(1) Lexicon-Based Learning. In several studies ($n = 14$), lexicon-based learning models are primarily used to detect and embed emotions to develop emotionally intelligent chatbots. Asghar et al. [39] and Zhong et al. [65] adopt the 3D semantically augmented affective space VAD (Valence, Arousal, and Dominance) [38] paired with an external cognitively engineered affective dictionary in order to implement emotion embedding techniques to enhance emotion diversity. Furthermore, using the bidirectional Seq2Seq model with a reinforcement framework that provides rewards and adopting the VAD affective space to append embedding emotion values would enable better emotion detection and allow to overcome limitations and generate an appropriate emotional response [54]. On the other hand, other studies apply emotion embedding using the VA vector based on two dimensions of emotion: Valence and Arousal. Valence measures the positivity or negativity of emotion, whereas Arousal measures emotion detection and activity. This study is based on neural networks using a dialog corpus that reflects a positive emotion elicitation strategy [62, 63].

Chang and Hsing [47] propose a two-layered BiLSTM-based model where word embeddings are constructed by encoding forward and backward sequences of characters into a continuous latent space. They capture the emotion enriched with semantic representations to provide a capture of more fine-grained emotions. Furthermore, Ghosh et al. [77] used the Linguistic Inquiry and Word Count (LIWC) text analysis program based on a dictionary. Each word is assigned an LIWC category in which the categories were selected based on their association with social, affective, and cognitive. They use the text analysis program to identify keywords within a text and extract emotions and features. Another study applied the LDA model to derive a topic dictionary and specify the topic related to emotion, and this technique would overcome the limitation of a supervised labeled dataset [56].

(2) Machine-Based Learning. Many studies ($n = 17$) used a machine learning approach for emotion classification solely. Several studies use the Seq2Seq-based model with a GRU to improve the Seq2Seq model and improve detecting and generating response consistency [48, 58, 59, 61]. Some studies use a dynamic classifier and a BiLSTM to train the dataset to better capture emotion [53, 72]. In addition, many studies use the Seq2Seq attention model based on deep RNN and pair it with a GRU to target the specific emotion-attention

[59]. GRU-RNN is an extension of a neural generator based on gated neural networks by adding three additional cells (refinement, adjustment, and output cells) to capture, control, and produce appropriate sentences [53]. Another study used a multilayer encoder-decoder extended with a Generative Adversarial Network (GAN). The discriminator output data are used as rewards for reinforcement learning, pushing the system to generate dialogs that are most similar to human dialogs [7]. Hu et al. [8] implemented a tone-aware model based on LSTM by adding an indicator vector capable of controlling the tones of generated conversations that allowed embedding target tones of empathy and passion into chatbot responses. Moreover, Niu and Bansal [72] developed a model consisting of 2 layers of the BiLSTM decoder, followed by a convolution layer with reinforcement rewards and trained for polite and rude labels that employ an LSTM-CNN politeness classifier to generate a polite response.

(3) Hybrid Model. Several studies ($n = 5$) apply a hybrid model to overcome limitations created by adopting only one approach. For example, in addition to using the VAD lexicon vector presentation as an emotion embedding technique, many studies use Bidirectional LSTM (BiLSTM). This affective classifier can train a Seq2Seq network in an encoder-decoder setting to label the sentences according to their emotional content [64]. Likewise, Peng et al. [57] increase the emotion intensity by pairing the VA lexicon-based emotion model variations of autoencoders that produce sentences containing a given sentiment or tense using an emotion classifier. This classifier can increase the intensity of emotional expression and identify or capture emotion and intensify emotions that do not include any sentiment. Similarly, Song et al. [76] paired the LDA topic model with a classifier BiLTSM, and Huang et al. [71] used a BiLSTM with LIWC (Linguistic Inquiry and Word Count) dictionary to be trained on the dataset.

4.4. RQ4: What Evaluation Measures Are Used to Evaluate Chatbot Performance? This section describes the datasets used for evaluating the chatbot performance and the different evaluation metrics used by the studies.

4.4.1. Datasets. A conversational dataset is required to evaluate the performance of a chatbot. Moreover, the dataset must be labeled with emotional tags to feed the encoder with emotional input and train the decoder to generate appropriate output.

Most studies have used conversational datasets from various sources, including social media and online websites, as shown in Table 5. The most popular datasets used are Weibo, followed by Twitter, which are open-domain conversational datasets. Only one study used a domain-specific conversational dataset for the healthcare domain [78]. Since none of these datasets are labeled with emotions, researchers have used a machine learning, lexicon-based, or hybrid approach to label the conversations with emotions. Several selected studies have used the NLPCC2013, NLPCC2014, and NLPCC2017 as corpora for labeling. These corpora can only be used in an open domain where the chatbot is not task-oriented-based. Some studies lack conversational and emotionally labeled datasets because of the limited publicly available datasets for training and evaluating the classifier systems, which poses a significant challenge [17]. They propose a new methodology for empathetic dialog generation and introduce a novel dataset of conversations grounded in an emotional context. Table 5 provides details about the various datasets.

4.4.2. Evaluation Measures. This section describes the methods used by the reviewed studies to measure the overall performance of emotionally intelligent chatbots in generating emotional responses. Almost all the studies have used both the automatic and manual evaluation methods to measure the effectiveness of their solution. In the automatic method, a test set is used to evaluate the model by comparing the generated responses with the existing responses using well-known metrics. The studies also used an automated method to measure the accuracy of emotion classification. Moreover, most studies use automated metrics to compare results against a baseline and other standard models. Several studies compared their models against a Seq2Seq baseline approach. On the other hand, the manual method employs humans to rate the responses against specified criteria.

(1) Automatic Evaluation. Table 6 summarizes the metrics used in an automated method. It shows the evaluation metrics for the response generation and classifies the input data. BLEU (Bilingual Evaluation Understudy) is the most common metric to evaluate emotionally intelligent chatbot responses. It is derived from a precision tool that automatically compares machine translation efficiency with human translation. BLEU is used to estimate the overlap between the generated and target responses. Thus, it measures how well the emotional response has been developed. However, BLEU's low correlation with human judgment is not suitable for measuring conversation generation [61].

Perplexity is another way to evaluate how well a selected model generates an emotional response—the lower the perplexity score, the better the generation performance. Another measure is the Distinct-1 grams and Distinct-2 grams that measure the diversity of the response. As a result, words with many repetitions are penalized, and sentences with many Distinct-n grams are rewarded. These metrics are devoted exclusively to the property of a given sentence and require no reference to ground truth [22].

Accuracy, F1-score, precision, and recall are the most common metrics for measuring emotion classification. Accuracy is the percentage of correctly predicted outcomes divided by the total amount of predictions [22]. F1-score is also used to assess machine learning models (or classifiers) as an alternative to accuracy. It measures how well the classifier balances precision and recall. In addition, it measures how the classifier balances between detecting or capturing the precise emotion and recalling it. Finally, data accuracy in a dialog indicates the number of times the data is aligned with the topic discussed. On the other hand, recall measures

Table 5: Datasets used by studies.

Datasets	# of studies	Studies
Weibo	15	[7, 47–49, 52, 55–57, 59, 61, 73, 74, 79, 83, 84]
Twitter	11	[8, 36, 48, 50, 51, 53, 68, 69, 74, 75, 82]
Fisher English Training Speech Corpus	3	[77, 82, 83]
DailyDialog labeled dataset	4	[17, 36, 60, 79]
Cornell Movie-Dialogs	4	[39, 54, 64, 65]
Conversations Support Group: dataset from Kaggle	1	[78]
Durban and Reddit	1	[84]
VoxCeleb	1	[82]
SEMAINE dataset	5	[17, 62, 63, 77, 83]
X-EMAC	1	[67]

Table 6: Automatic evaluation metrics.

Evaluation type	Metric	# of studies	Studies
Evaluation of generated responses	BLEU	24	[17, 39, 47, 49–53, 55, 57–59, 64, 65, 68, 69, 72–76, 79, 83]
	Perplexity	20	[17, 36, 48, 49, 51, 52, 54, 55, 60–67, 72, 73, 75, 79]
	Distinct-1 grams Distinct-2 grams	12	[36, 50, 53, 56–59, 64, 67, 74, 76, 80]
	ROUGE	5	[39, 52, 54, 59, 66]
	METEO	4	[39, 48, 52, 66]
Evaluation of emotions	F1	5	[47, 60, 71, 78, 80]
	Precision	9	[17, 47, 48, 60, 67, 69, 71, 78, 84]
	Recall	8	[47, 48, 60, 67, 69, 71, 78, 84]
	Accuracy	22	[8, 17, 48–52, 55, 57, 58, 61, 66, 68, 71–73, 75, 76, 78, 80, 81, 84]

the number of replies that the chatbot can group into appropriate topics through human-computer interaction [22].

(2) Human Evaluation. Using human evaluators is another way to measure the performance of emotionally intelligent chatbots. Although automatic evaluation is more efficient and has fewer overheads than human evaluation, it does not consider whether the generated emotional response is appropriate and natural. Human evaluation is usually measured on a Likert scale. Several studies employed the Amazon Mechanical Turk (MTurk) participants ($n = 5$) for evaluation. Multiple studies ($n = 14$) used Fleiss' kappa test to measure the annotator's agreements and their consistency in rating [39]. Table 7 summarizes the evaluation criteria used for human evaluation.

5. Discussion

5.1. Chatbot Interface Language. Chinese and English are the most popular chatbot interface languages used by researchers to develop emotionally intelligent chatbots. The conversational datasets for these languages are retrieved from Twitter and Weibo. Only one study proposed the development of a bilingual chatbot [79]. In a multicultural environment, this is an essential solution where a chatbot must be able to converse in the user's preferred language. This is an avenue open for further research and exploration.

5.2. Dataset Availability. A vast majority of research studies focus on developing an emotionally intelligent chatbot for an open domain, whereas only a few have focused on the closed domain, using a rule-based approach for generating responses. And only one of the reviewed studies sourced a domain-specific dataset for healthcare [78]. A generative chatbot that synthesizes human-like natural responses requires a massive dataset for training [39]. The unavailability of domain-specific conversational datasets is the main reason for the research gap in this field. A ripe area for exploration for researchers is the development of domain-specific datasets for education, business, and more as they can provide appealing solutions for empathetic customer service chatbots, advising chatbots, and more.

Moreover, the conversational datasets used for open-domain chatbots are not emotionally labeled. The reviewed studies have used extensive preprocessing of the datasets retrieved from Twitter and other datasets to extract conversations and classify them further with labels. However, an issue with this approach is that the dataset is usually imbalanced, and the classification is usually prone to errors. Rashkin et al. [17] addressed this challenge by developing a dataset of emotionally labeled conversations. The dataset

Table 7: Manual evaluation criteria.

Evaluation criteria	# of studies	Studies
Emotion accuracy	3	[49, 53, 78]
Response emotion quality and specificity	4	[47, 57, 60, 83]
Response emotion reflection and expression	8	[55–57, 60, 74, 79, 80, 83]
Response emotion diversity	8	[8, 36, 39, 58, 59, 61, 64, 76]
Response emotion appropriateness	10	[8, 36, 39, 48, 58, 59, 61, 64, 69, 82]
Response empathetic emotion intensity	7	[8, 17, 66–68, 82, 84]
Emotion intensity	7	[8, 17, 66–68, 82, 84]
Response grammatical correctness	11	[36, 39, 48, 50, 51, 58, 59, 64, 76, 77, 80]
Response user preference	1	[64]
Response naturalness	4	[7, 50, 54, 80]
Response coherence	5	[7, 39, 54, 68, 80]
Response fluency	4	[17, 49, 68, 80]
Response relevance	13	[17, 36, 48, 49, 52, 57, 61, 66–68, 72, 74, 80, 84]
Response consistency	9	[7, 50–52, 54–56, 73, 84]
Response logic	5	[55–57, 62, 63]
Response intelligible	2	[62, 63]
Response context	1	[72]
Response politeness	1	[72]

consists of 25k conversational utterances. This is another area of research that needs further exploration where researchers may investigate the development of more emotionally labeled datasets to be used as the gold standard in the open domain.

5.3. Encoder-Decoder Model. Several studies use techniques to enhance the previously adopted model for developing emotionally intelligent chatbots, i.e., extending Seq2Seq to overcome its dull and meaningless response limitations. Many studies use a bidirectional classifier that is trained using an emotionally labeled dataset to develop the model [64]. However, the limitation of such models is that conversational models based on neural networks cannot capture the complexities of emotions and produce short and unclear responses. More recent researchers have utilized the CVAE model to alleviate this problem and generate diverse emotional responses. The studies have demonstrated that CVAE can solve this problem and increase the diversity of responses. Additionally, it overcomes the dullness and meaninglessness of Seq2Seq. However, it impacts the syntax of the responses [36]. A further area for exploration by researchers is to enhance the CVAE model to make it more robust to syntax errors.

5.4. Emotion Detection and Embedding. The primary focus of most studies was to accurately detect the input emotion or the user's emotional state and generate appropriate affective responses. Several studies indicate that emotions are complex and cannot be captured accurately by a classifier [47, 62, 63]. By adopting a lexicon-based learning approach and using VAD vector spaces where each word is embedded with emotion, it is possible to overcome the inability of classifiers to detect fine-grained emotion [39]. The taxonomy diagram (Figure 11) shows that mainly lexicon-based approaches are used by studies that address the challenge of emotion capture. Only four studies have attempted to capture the user's emotional state from multiple historical utterances. Connecting the meanings and emotions from previous utterances is essential to comprehend the user's emotional state and foster a continuous conversation. This is still an unexplored area and requires further investigation.

5.5. Voice-Based/Multimodal Chatbots. All of the chatbots included in the review are text-based. Another area for exploration and further research is the development of voice-based and multimodal chatbots that are domain-specific.

5.6. Hybrid Chatbots. Finally, there are no studies investigating generative emotionally intelligent chatbots that are task-oriented. Non-task-oriented chatbots are usually rule-based because they provide precise information but at the same time suffer from machine-like responses. A task-oriented emotionally intelligent chatbot could assist the user in accomplishing a task, such as making a reservation, placing an order, and providing advising information, while embedding empathy in the conversation to eliminate user frustration and provide a good user experience. Moreover, such a chatbot could trigger human intervention if required by determining the user's emotional state [6, 47, 74].

6. Conclusion

This section includes a summary of the paper and its significance, limitations, and new directions for future research.

Recent technological advances have made chatbots increasingly feasible to deliver information to various

Table 8: Studies included in the review.

#	Study	Purpose
1	[78]	To build a chatbot that captures the emotions of patients during interaction and accordingly updates human therapists to provide timely care
2	[39]	To generate affective responses in an open-domain chatbot by using a three-method approach in an LSTM conversational model
3	[60]	To develop an empathetic chatbot that generates responses based on the user's emotional state and the context of the message
4	[47]	To incorporate emotional content into the response generation process to make chatbot responses more emotionally sound
5	[64]	To produce an affect-driven dialog system that generates multiple diverse emotional responses and ranks them based on emotion
6	[77]	To extract the affect category of the input text using the Linguistic Inquiry and Word Count (LIWC) and generate grammatically correct responses embedded with emotion
7	[81]	To develop an embodied conversational agent that responds based on user profiles and emotional content
8	[69]	To predict the emotional state of the sender based on historical responses and accordingly generate an emotionally appropriate response
9	[82]	To build a voice-based conversational agent that embeds responses with emotion
10	[8]	To develop a novel tone-aware chatbot that generates toned responses to user requests on social media
11	[71]	To embed emotions in the dialog based on input emotion and to tackle the problem of generic responses that are not emotionally intelligent
12	[49]	To develop a topic-aware emotional response generation (TERG) model, which can not only exactly generate desired emotional response but also perform well in topic relevance
13	[56]	To embed emotion and topic in the input data to generate meaningful and emotionally relevant responses
14	[66]	To develop and evaluate a multiresolution adversarial model that generates more empathetic responses
15	[50]	To elicit a topic-coherent response embedded with emotion using a loss function to predict the corresponding word in every generation step
16	[67]	To develop an online empathetic chatbot influenced by emotion information using large-scale empathetic conversational datasets to detect the user's emotion or ask questions for self-disclosure
17	[68]	To develop a model that selects an appropriate reaction by learning the context and underlying emotion
18	[36]	Used an affective lexicon to embed sentiments into the word vectors and used a CVAE-based dialog model to generate diverse and emotional responses
19	[62]	To develop an AI-driven chat-oriented dialog system that dynamically imitates human emotions in the conversation
20	[63]	To elicit a more positive emotional valence throughout a chat-based interaction in order to promote positive emotional states
21	[72]	To develop three weakly supervised models that can generate diverse, polite (or rude) dialog responses using data from separate style and dialog domains
22	[51]	To propose a generative model that fuses word- and sentence-level emotions to model the dialog text and learn emotional expression in order to control the emotional feature of the generated response
23	[57]	To present a topic-enhanced emotional conversation generation model that incorporates emotional factors and topic information into the conversation system
24	[17]	To use a custom-built empathetic conversational dataset and explore different ways of combining information from related tasks that can lead to more empathetic responses
25	[76]	To generate meaningful responses embedded with explicit or implicit emotion
26	[54]	To develop a new approach of context-relevant emotional responses using the bidirectional Seq2Seq model
27	[7]	To create an emotionally intelligent chatbot using emotional tags on the posts and recognize the emotional dimension
28	[55]	To use reinforcement learning with emotional editing constraints to generate more meaningful and customizable emotional responses
29	[79]	To create a bilingual-aided interactive approach that can simultaneously and interactively generate bilingual emotional replies to monolingual posts
30	[80]	To provide social support for community members in an online health community using a Seq2Seq model-based chatbot that recognizes emotion and produces diverse responses
31	[59]	To build a unified neural architecture in order to encode the semantics and affect for generating more intelligent responses with expressed emotions
32	[58]	To extract the emotional and semantic information of the interlocutor to generate logical responses embedded with emotion
33	[83]	To develop an anthropomorphic model and present its ability to understand the human interlocutor using both the subjective and objective measures

Table 8: Continued.

#	Study	Purpose
34	[84]	To create an empathetic conversation system that incorporates emotional factors added to semantics and to enhance the context-response through a multitask learning framework
35	[52]	To design an artificial conversational chatting machine that generates nondeterministic responses providing the same input with different emotional contexts that are empathetically coherent
36	[73]	To propose a multiemotional conversation system (MECS) and evaluate the model at both the context level and the emotion level
37	[53]	To develop a dual-factor generation model that fits the conversation data and actively controls the generation of the response with respect to sentiment or topic specificity
38	[65]	To develop an intelligent open-domain neural conversational model that produces responses that are syntactically and semantically appropriate and rich in emotion
39	[74]	To propose a neural conversation generation with auxiliary emotional supervised models where the dialog generation system is characterized by emotional intelligence
40	[61]	To propose a model to generate emotional responses using internal and external memory
41	[48]	To present the design and implementation of XiaoIce, a multimodal chatbot that recognizes and responds with emotion in an open-domain conversation
42	[75]	To apply emotion detection with emojis using a reinforced CVAE model to generate affective responses that contain emojis

domains. Consequently, there are now a growing number of chatbots available for public use. Today, more attention is being paid to the development of emotionally intelligent chatbots. Developing chatbots that can generate emotional responses to user requests is challenging yet crucial to its successful adoption.

In this study, we conducted a systematic literature review exploring a spectrum of topics regarding the development of emotionally intelligent chatbots, exploring the technique of embedding and generating emotional responses, the challenges, the datasets used, and the evaluation processes used to measure the chatbot's performance. This study was based on available publications from 2011 to 2022 using six digital databases: Scopus, IEEE Xplore, ProQuest, ScienceDirect, ACM Digital Library, and EBSCO. We use a systematic approach to gather and assimilate our findings. This study is aimed at generating evidence-based guidelines for researchers and developers to gain insights into emotionally intelligent chatbot development research. Thus, researchers and practitioners in the related fields will gain a deeper understanding of emotionally intelligent chatbots based on the findings of this study presented in the discussion section.

Our study shows that Chinese is the most commonly used interface language in developing emotionally intelligent chatbots. Weibo and Twitter datasets are the most popular datasets used to develop open-domain AI-powered chatbots. Additionally, most chatbots are developed for the open domain due to the availability of conversational datasets. However, these datasets are not labeled; therefore, a common preprocessing step is to label the dataset using a classifier, lexicon-based, or hybrid approach. Furthermore, we identified that the lexicon-based approach, such as the VAD vector, provides fine-grained emotion detection. Classifiers are also used to detect emotion and generate diverse responses, which is the ultimate objective of the evaluation. Most studies use automatic and human evaluation measures. BLEU and perplexity are the most commonly used metrics in automatic evaluation. Human evaluations are essential to test the quality of the responses. Several studies sourced participants from MTurk or used other human judges to evaluate the response diversity and emotional relevance. Statistical measures such as Fleiss' kappa are used to determine the validity of human responses.

This study may have had limitations due to several factors. First, there was a limited amount of time to conduct the study. Moreover, although six bibliographic databases were used to retrieve relevant studies, the lack of research due to the relatively new and emerging topic resulted in the possibility of specific unexplored areas that the readers may notice. Furthermore, due to limited resources, some retrieval studies may not be thorough and may compromise the effectiveness of the study.

Data Availability

The search keywords and databases used in the systematic review are provided in the paper. Table 8 list of all the papers included in the systematic literature review. Furthermore, the data encoding of the papers analyzed during the current study is available from the corresponding author upon reasonable request.

Conflicts of Interest

All authors declare that they have no conflicts of interest.

References

[1] M. Allouch, A. Azaria, and R. Azoulay, "Conversational agents: goals, technologies, vision and challenges," *Sensors*, vol. 21, no. 24, p. 8448, 2021.

[2] M. Adam, M. Wessel, and A. Benlian, "AI-based chatbots in customer service and their effects on user compliance," *Electronic Markets*, vol. 31, no. 2, pp. 427–445, 2021.

[3] M. Milne-ives, C. CockDe, E. Lim et al., "The effectiveness of artificial intelligence conversational agents in health care:

systematic review," *Journal of Medical Internet Research*, vol. 22, no. 10, article e20346, 2020.

[4] M. Moran, "25+ top chatbot statistics for 2022: usage, demographics, trends," Startup Bonsai, 2022, September 2022, https://startupbonsai.com/chatbot-statistics/.

[5] A. Rapp, L. Curti, and A. Boldi, "The human side of human-chatbot interaction: a systematic literature review of ten years of research on text-based chatbots," *International Journal of Human Computer Studies*, vol. 151, article 102630, 2021.

[6] J. S. Chen, T. T. Y. Le, and D. Florence, "Usability and responsiveness of artificial intelligence chatbot on online customer experience in e-retailing," *International Journal of Retail and Distribution Management*, vol. 49, no. 11, pp. 1512–1531, 2021.

[7] X. Sun, X. Chen, Z. Pei, and F. Ren, "Emotional human machine conversation generation based on SeqGAN," in *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, Beijing, China, May 2018.

[8] T. Hu, A. Xu, Z. Liu et al., "Touch your heart: a tone-aware chatbot for customer care on social media," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, Montréal, Canada, 2018.

[9] E. Adamopoulou and L. Moussiades, "An overview of chatbot technology," in *BT - Artificial Intelligence Applications and Innovations*, I. Maglogiannis, L. Iliadis, and E. Pimenidis, Eds., p. 373, Springer International Publishing, 2020.

[10] S.-M. Tan and T. W. Liew, "Multi-chatbot or single-chatbot? The effects of m-commerce chatbot interface on source credibility, social presence, trust, and purchase intention," *Human Behavior and Emerging Technologies*, vol. 2022, article 2501538, 14 pages, 2022.

[11] E. Adamopoulou and L. Moussiades, "An Overview of Chatbot Technology," in *Artificial Intelligence Applications and Innovations. AIAI 2020. IFIP Advances in Information and Communication Technology*, I. Maglogiannis, L. Iliadis, and E. Pimenidis, Eds., vol. 584, p. 373, Springer, Cham, 2020.

[12] P. Salovey and J. D. Mayer, "Emotional intelligence," *Imagination, Cognition and Personality*, vol. 9, no. 3, pp. 185–211, 1990.

[13] X. Wang and R. Nakatsu, "How do people talk with a virtual philosopher: log analysis of a real-world application," in *Entertainment Computing – ICEC 2013. ICEC 2013*, J. C. Anacleto, E. W. G. Clua, F. S. C. Silva, S. Fels, and H. S. Yang, Eds., vol. 8215 of Lecture Notes in Computer Science, pp. 132–137, Springer, Berlin, Heidelberg, 2013.

[14] A. Ghandeharioun, D. McDuff, M. Czerwinski, and K. Rowan, "Towards understanding emotional intelligence for behavior change chatbots," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 8–14, Cambridge, UK, September 2019.

[15] S. C. Paul, N. Bartmann, and J. L. Clark, "Customizability in conversational agents and their impact on health engagement," *Human Behavior and Emerging Technologies*, vol. 3, no. 5, pp. 1141–1152, 2021.

[16] J. C. Giger, N. Piçarra, P. Alves-Oliveira, R. Oliveira, and P. Arriaga, "Humanization of robots: is it really such a good idea?," *Human Behavior and Emerging Technologies*, vol. 1, no. 2, pp. 111–123, 2019.

[17] H. Rashkin, E. M. Smith, M. Li, and Y. L. Boureau, "Towards empathetic open-domain conversation models: a new benchmark and dataset," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381, Florence, Italy, 2020.

[18] M. R. Pacheco-Lorenzo, S. M. Valladares-Rodríguez, L. E. Anido-Rifón, and M. J. Fernández-Iglesias, "Smart conversational agents for the detection of neuropsychiatric disorders: a systematic review," *Journal of Biomedical Informatics*, vol. 113, article 103632, 2021.

[19] C. W. Okonkwo and A. Ade-Ibijola, "Chatbots applications in education: a systematic review," *Computers and Education: Artificial Intelligence*, vol. 2, article 100033, 2021.

[20] A. Miklosik, N. Evans, A. Mahmood, and A. Qureshi, "The use of chatbots in digital business transformation: a systematic literature review," *IEEE Access*, vol. 9, pp. 106530–106539, 2021.

[21] A. de Barcelos Silva, M. M. Gomes, C. A. da Costa et al., "Intelligent personal assistants: a systematic literature review," *Expert Systems with Applications*, vol. 147, article 113193, 2020.

[22] S. Mohamad Suhaili, N. Salim, and M. N. Jambli, "Service chatbots: a systematic review," *Expert Systems with Applications*, vol. 184, p. 115461, 2021.

[23] A. K. Wardhana, R. Ferdiana, and I. Hidayah, "Empathetic chatbot enhancement and development: a literature review," in *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*, Bandung, Indonesia, April 2021.

[24] E. W. Pamungkas, *Emotionally-aware chatbots: a survey*, Cornell University Library, 2019, http://ezproxy.hct.ac.ae/login?url=https://www.proquest.com/working-papers/emotionally-aware-chatbots-survey/docview/2246534988/se-2.

[25] Y. Ma, K. L. Nguyen, F. Z. Xing, and E. Cambria, "A survey on empathetic dialogue systems," *Information Fusion*, vol. 64, pp. 50–70, 2020.

[26] J. Grudin and R. Jacques, "Chatbots, humbots, and the quest for artificial general intelligence," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–11, Glasgow, UK, May 2019.

[27] M. Jovanovic, M. Baez, and F. Casati, "Chatbots as conversational healthcare services," *IEEE Internet Computing*, vol. 25, no. 3, pp. 44–51, 2021.

[28] H. Y. Shum, X. D. He, and D. Li, "From Eliza to XiaoIce: challenges and opportunities with social chatbots," *Frontiers of Information Technology and Electronic Engineering*, vol. 19, no. 1, pp. 10–26, 2018.

[29] S. Hussain, O. Ameri Sianaki, and N. Ababneh, "A survey on conversational agents/chatbots classification and design techniques," in *Advances in Intelligent Systems and Computing*, vol. 927, Springer International Publishing, 2019.

[30] Z. Safi, A. Abd-alrazaq, M. Khalifa, and M. Househ, "Technical aspects of developing chatbots for medical applications: scoping review," *Journal of Medical Internet Research*, vol. 22, no. 12, article e19127, 2020.

[31] Z. Xiao, M. X. Zhou, W. Chen, H. Yang, and C. Chi, "If I hear you correctly: building and evaluating interview chatbots with active listening skills," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, vol. 1–14, Hawai'i, USA, April 2020.

[32] R. S. Wallace, "The anatomy of ALICE," in *Parsing the Turing Test*, pp. 181–210, Springer, 2009.

[33] A. Xu, Z. Liu, Y. Guo, V. Sinha, and R. Akkiraju, "A new chatbot for customer service on social media," in *Proceedings of the*

2017 CHI Conference on Human Factors in Computing Systems, pp. 3506–3510, Denver, Colarado, May 2017.

[34] E. Svikhnushina and P. Pu, "Social and emotional etiquette of chatbots: a qualitative approach to understanding user needs and expectations," 2020, https://arxiv.org/abs/2006.13883.

[35] A. Hutapea, "Chatbot: architecture , design , & development," University of Pennsylvania School of Engineering and Applied Science Department of Computer and Information Science, 2017.

[36] M. Liu, X. Bao, J. Liu, P. Zhao, and Y. Shen, "Generating emotional response by conditional variational auto-encoder in open- domain dialogue system," *Neurocomputing*, vol. 460, pp. 106–116, 2021.

[37] M. Aleedy, H. Shaiba, and M. Bezbradica, "Generating and analyzing chatbot responses using natural language processing," *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 9, pp. 60–68, 2019.

[38] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas," *Behavior Research Methods*, vol. 45, no. 4, pp. 1191–1207, 2013.

[39] N. Asghar, P. Poupart, J. Hoey, X. Jiang, and L. Mou, "Affective neural response generation," in *Advances in Information Retrieval. ECIR 2018*, G. Pasi, B. Piwowarski, L. Azzopardi, and A. Hanbury, Eds., vol. 10772 of Lecture Notes in Computer Science(), pp. 154–166, Springer, Cham, 2018.

[40] B. Kitchenham and S. Charters, *Guidelines for performing systematic literature reviews in software engineering*, 2007.

[41] D. Tranfield, D. Denyer, and P. Smart, "Towards a methodology for developing evidence-informed management knowledge by means of systematic review," *British Journal of Management*, vol. 14, no. 3, pp. 207–222, 2003.

[42] M. Petticrew and H. Roberts, *Systematic Reviews in the Social Sciences: A Practical Guide*, John Wiley & Sons, 2008.

[43] L. Yang, H. Zhang, H. Shen et al., "Quality assessment in systematic literature reviews: a software engineering perspective," *Information and Software Technology*, vol. 130, article 106397, 2021.

[44] N. J. Van Eck and L. Waltman, *VOSviewer Manual*, vol. 1, no. 1, 2013, Univeristeit Leiden, Leiden, 2013.

[45] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement," *International Journal of Surgery*, vol. 8, no. 5, pp. 336–341, 2010.

[46] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in Neural Information Processing Systems*, vol. 4, pp. 3104–3112, 2014.

[47] Y.-C. Chang and Y.-C. Hsing, "Emotion-infused deep neural network for emotionally resonant conversation," *Applied Soft Computing*, vol. 113, p. 107861, 2021.

[48] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, "The design and implementation of XiaoIce, an empathetic social chatbot," *Computational Linguistics*, vol. 46, no. 1, pp. 53–93, 2020.

[49] P. Huo, Y. Yang, J. Zhou, C. Chen, and L. He, "TERG: topic-aware emotional response generation for chatbot," in *2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, UK, July 2020.

[50] S. Li, S. Feng, D. Wang, K. Song, Y. Zhang, and W. Wang, "EmoElicitor: an open domain response generation model with user emotional reaction awareness," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 3637–3643, Yokohama, Japan, July 2020.

[51] D. Peng, M. Zhou, C. Liu, and J. Ai, "Human-machine dialogue modelling with the fusion of word- and sentence-level emotions," *Knowledge-Based Systems*, vol. 192, article 105319, 2020.

[52] K. Yao, L. Zhang, T. Luo, D. Du, and Y. Wu, "Non-deterministic and emotional chatting machine: learning emotional conversation generation using conditional variational autoencoders," *Neural Computing and Applications*, vol. 33, no. 11, pp. 5581–5589, 2021.

[53] R. Zhang, J. Guo, Y. Fan, Y. Lan, and X. Cheng, "Dual-factor generation model for conversation," *ACM Transactions on Information Systems*, vol. 38, no. 3, pp. 1–31, 2020.

[54] V. Srinivasan, S. Santhanam, and S. Shaikh, "Using reinforcement learning with external rewards for open-domain natural language generation," *Journal of Intelligent Information Systems*, vol. 56, no. 1, pp. 189–206, 2021.

[55] X. Sun, J. Li, X. Wei, C. Li, and J. Tao, "Emotional editing constraint conversation content generation based on reinforcement learning," *Information Fusion*, vol. 56, pp. 70–80, 2020.

[56] J. Li and X. Sun, "A syntactically constrained bidirectional-asynchronous approach for emotional conversation generation," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 678–683, Brussels, Belgium, 2020.

[57] Y. Peng, Y. Fang, Z. Xie, and G. Zhou, "Topic-enhanced emotional conversation generation with attention mechanism," *Knowledge-Based Systems*, vol. 163, pp. 429–437, 2019.

[58] W. Wei, J. Liu, X. Mao et al., "Target-guided emotion-aware chat machine," *ACM Transactions on Information Systems*, vol. 39, no. 4, pp. 1–24, 2021.

[59] W. Wei, J. Liu, X. Mao et al., "Emotion-aware chat machine: automatic emotional response generation for human-like emotional interaction," in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 1401–1410, Beijing, China, November 2019.

[60] J. Casas, T. Spring, K. Daher, E. Mugellini, O. A. Khaled, and P. Cudré-Mauroux, "Enhancing conversational agents with empathic abilities," in *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pp. 41–47, Japan, 2021.

[61] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: emotional conversation generation with internal and external memory," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, pp. 730–738, 2018.

[62] N. Lubis, S. Sakti, K. Yoshino, and S. Nakamura, "Eliciting positive emotion through affect-sensitive dialogue response generation: a neural network approach," in *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pp. 5293–5300, Louisiana, USA, 2018.

[63] N. Lubis, S. Sakti, K. Yoshino, and S. Nakamura, "Positive emotion elicitation in chat-based dialogue systems," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 27, no. 4, pp. 866–877, 2019.

[64] P. Colombo, W. Witon, A. Modi, J. Kennedy, and M. Kapadia, "Affect-driven dialog generation," in *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*

- *Proceedings of the Conference, , Volume 1*, pp. 3734–3743, Minneapolis, Minnesota, 2019.

[65] P. Zhong, D. Wang, and C. Miao, "An affect-rich neural conversational model with biased attention and weighted cross-entropy loss," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 7492–7500, 2019.

[66] Q. Li, H. Chen, Z. Ren, P. Ren, Z. Tu, and Z. Chen, "EmpDG: multi-resolution interactive empathetic dialogue generation," in *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4454–4466, Barcelona, Spain, 2021.

[67] Y. Li, K. Li, H. Ning et al., "Towards an online empathetic chatbot with emotion causes," in *44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2041–2045, Association for Computing Machinery, 2021.

[68] Z. Lin, A. Madotto, J. Shin, P. Xu, and P. Fung, "MOEL: mixture of empathetic listeners," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 121–132, Hong Kong, China, 2020.

[69] T. Hasegawa, N. Kaji, N. Yoshinaga, and M. Toyoda, "Predicting and eliciting addressee's emotion in online dialogue," *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, vol. 29, no. 1, pp. 90–99, 2013.

[70] L. Qiu, Y. Shiu, P. Lin et al., "What if bots feel moods?," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1161–1170, China, July 2020.

[71] C. Huang, O. R. Zaïane, A. Trabelsi, and N. Dziri, "Automatic dialogue generation with expressed emotions," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 49–54, New Orleans, Louisiana, 2018.

[72] T. Niu and M. Bansal, "Polite dialogue generation without parallel data," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 373–389, 2018.

[73] R. Zhang, Z. Wang, and D. Mai, "Building emotional conversation systems using multi-task Seq2Seq learning," in *Natural Language Processing and Chinese Computing. NLPCC 2017*, X. Huang, J. Jiang, D. Zhao, Y. Feng, and Y. Hong, Eds., vol. 10619 of Lecture Notes in Computer Science(), pp. 612–621, Springer, Cham, 2018.

[74] G. Zhou, Y. Fang, Y. Peng, and J. Lu, "Neural conversation generation with auxiliary emotional supervised models," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 19, no. 2, pp. 1–17, 2019.

[75] X. Zhou and W. Y. Wang, "MojiTalk: generating emotional responses at scale," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1-2, Melbourne, Australia, 2018.

[76] Z. Song, X. Zheng, L. Liu, M. Xu, and X. Huang, "Generating responses with a specific emotion in dialog," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3685–3695, Florence, Italy, 2020.

[77] S. Ghosh, M. Chollet, E. Laksana, L. P. Morency, and S. Scherer, "Affect-LM: a neural language model for customizable affective text generation," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 634–642, Vancouver, Canada, 2017.

[78] A. Adikari, D. de Silva, H. Moraliyage et al., "Empathic conversational agents for real-time monitoring and co-facilitation of patient-centered healthcare," *Future Generation Computer Systems*, vol. 126, pp. 318–329, 2022.

[79] J. Wang, X. Sun, and M. Wang, "Emotional conversation generation with bilingual interactive decoding," *IEEE Transactions on Computational Social Systems*, vol. 9, no. 3, pp. 818–829, 2021.

[80] L. Wang, D. Wang, F. Tian et al., "CASS: towards building a social-support chatbot for online health community," in *Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1)*, 2021.

[81] D. Griol, A. Sanchis, J. M. Molina, and Z. Callejas, "Developing enhanced conversational agents for social virtual worlds," *Neurocomputing*, vol. 354, pp. 27–40, 2019.

[82] J. Hu, Y. Huang, X. Hu, and Y. Xu, "Enhancing the perceived emotional intelligence of conversational agents through acoustic cues," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, May 2021.

[83] J. Wu, S. Ghosh, M. Chollet, S. Ly, S. Mozgai, and S. Scherer, "NADiA - towards neural network driven virtual human conversation agents," in *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pp. 2262–2264, Sydney, Australia, November 2018.

[84] R. Yan, "What if bots feel moods?," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1161–1170, China, July 2020.