

# Emotionally-Aware Chatbots: A Survey

Endang Wahyu Pamungkas  
pamungka@di.unito.it  
University of Turin  
Turin, Italy

Applications of  
NLP.

## ABSTRACT

Textual conversational agent or chatbots' development gather tremendous traction from both academia and industries in recent years. Nowadays, chatbots are widely used as an agent to communicate with a human in some services such as booking assistant, customer service, and also a personal partner. The biggest challenge in building chatbot is to build a humanizing machine to improve user engagement. Some studies show that emotion is an important aspect to humanize machine, including chatbot. In this paper, we will provide a systematic review of approaches in building an emotionally-aware chatbot (EAC). As far as our knowledge, there is still no work focusing on this area. We propose three research question regarding EAC studies. We start with the history and evolution of EAC, then several approaches to build EAC by previous studies, and some available resources in building EAC. Based on our investigation, we found that in the early development, EAC exploits a simple rule-based approach while now most of EAC use neural-based approach. We also notice that most of EAC contain emotion classifier in their architecture, which utilize several available affective resources. We also predict that the development of EAC will continue to gain more and more attention from scholars, noted by some recent studies propose new datasets for building EAC in various languages.

## CCS CONCEPTS

- **Human-centered computing** → **Natural language interfaces**;
- **Computing methodologies** → **Natural language generation**.

## KEYWORDS

chatbot, conversational agent, affective computing, natural language generation

## ACM Reference Format:

Endang Wahyu Pamungkas. 2018. Emotionally-Aware Chatbots: A Survey. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/1122445.1122456>

## 1 INTRODUCTION

Conversational agents or dialogue systems development are gaining more attention from both industry and academia [15, 17] in the

latest years. Some works tried to model them into domain-specific tasks such as customer service [10, 18], and shopping assistance [22]. Other works design a multi-purpose agents such as SIRI<sup>1</sup>, Amazon Alexa<sup>2</sup>, and Google Assistance<sup>3</sup>. This domain is a well-researched area in Human-Computer Interaction research community but still become a hot topic now. The main development focus right now is to have an intelligent and humanizing machine to have a better engagement when communicating with human [16]. Having a better engagement will lead to higher user satisfaction, which becomes the main objective from the industry perspective.

In this study, we will only focus on textual conversational agent or chatbot, a conversational artificial intelligence which can conduct a textual communication with a human by exploiting several natural language processing techniques. There are several approaches used to build a chatbot, start by using a simple rule-based approach [1, 46] until more sophisticated one by using neural-based technique [44, 47]. Nowadays, chatbots are mostly used as customer service such as booking systems [38, 59], shopping assistance [10] or just as conversational partner such as Endurance<sup>4</sup> and Insomnobot<sup>5</sup>. Therefore, there is a significant urgency to humanize chatbot for having a better user-engagement. Some works were already proposed several approaches to improve chatbot's user-engagement, such as building a context-aware chatbot [52] and injecting personality into the machine [63]. Other works also try to incorporate affective computing to build emotionally-aware chatbots [18, 45, 65].

Some existing studies shows that adding emotion information into dialogue systems is able to improve user-satisfaction [43, 61]. Emotion information contribute to a more positive interaction between machine and human, which lead to reduce miscommunication [28]. Some previous studies also found that using affect information can help chatbot to understand users' emotional state, in order to generate better response [41]. Not only emotion, another study also introduce the use of tones to improve satisfactory service. For instance, using empathetic tone is able to reduces user stress and results in more engagement. [18] found that tones is an important aspect in building customer care chatbot. They discover eight different tones including anxious, frustrated, impolite, passionate, polite, sad, satisfied, and empathetic.

In this paper, we will try to summarize some previous studies which focus on injecting emotion information into chatbots, on discovering recent issues and barriers in building engaging emotionally-aware chatbots. Therefore, we propose some research questions to have a better problem definition:

<sup>1</sup><https://www.apple.com/es/siri/>

<sup>2</sup><https://developer.amazon.com/alexa>

<sup>3</sup><https://assistant.google.com/>

<sup>4</sup><https://bit.ly/2RGM40S>

<sup>5</sup><http://insomnobot3000.com/>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

Conference'17, July 2017, Washington, DC, USA

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-9999-9/18/06...\$15.00

<https://doi.org/10.1145/1122445.1122456>

- RQ1 How to incorporate emotion information in building an emotionally-aware chatbot?
- RQ2 What are available resources that can be used in building emotionally-aware chatbots?
- RQ3 How to evaluate the performance of emotionally-aware chatbots?

This paper will be organized as follows: Section 2 introduces the history of the relation between affective information with chatbots. Section 3 outline some works which try to inject affective information into chatbots. Section 4 summarizes some affective resources which can be utilized to provide affective information. Then, Section 5 describes some evaluation metric that already applied in some previous works related to emotionally-aware chatbots. Last Section 6 will conclude the rest of the paper and provide a prediction of future development in this research direction based on our analysis.

## 2 HISTORY OF EMOTIONALLY-AWARE CHATBOT

The early development of chatbot was inspired by Turing test in 1950 [57]. Eliza was the first publicly known chatbot, built by using simple hand-crafted script [58]. Parry [8] was another chatbot which successfully passed the Turing test. Similar to Eliza, Parry still uses a rule-based approach but with a better understanding, including the mental model that can stimulate emotion. Therefore, Parry is the first chatbot which involving emotion in its development. Also, worth to be mentioned is ALICE (Artificial Linguistic Internet Computer Entity), a customizable chatbot by using Artificial Intelligence Markup Language (AIML). Therefore, ALICE still also use a rule-based approach by executing a pattern-matcher recursively to obtain the response. Then in May 2014, Microsoft introduced Xiaolce [66], an empathetic social chatbot which is able to recognize users' emotional needs. XiaoIce can provide an engaging interpersonal communication by giving encouragement or other affective messages, so that can hold human attention during communication.

Nowadays, most of chatbots technologies were built by using neural-based approach. Emotional Chatting Machine (ECM) [65] was the first works which exploiting deep learning approach in building a large-scale emotionally-aware conversational bot. Then several studies were proposed to deal with this research area by introducing emotion embedding representation [2, 9, 50] or modeling as reinforcement learning problem [23, 55]. Most of these studies used encoder-decoder architecture, specifically sequence to sequence (seq2seq) learning. Some works also tried to introduce a new dataset in order to have a better gold standard and improve system performance. [45] introduce EMPATHETICDIALOGUES dataset, a novel dataset containing 25k conversations include emotional contexts information to facilitate training and evaluating the textual conversational system. Then, work from [18] produce a dataset containing 1.5 million Twitter conversation, gathered by using Twitter API from customer care account of 62 brands across several industries. This dataset was used to build tone-aware customer care chatbot. Finally, [27] tried to enhance SEMAINE corpus [29] by using crowdsourcing scenario to obtain a human

judgement for deciding which response that elicits positive emotion. Their dataset was used to develop a chatbot which captures human emotional states and elicits positive emotion during the conversation.

## 3 BUILDING EMOTIONALLY-AWARE CHATBOT (EAC)

As we mentioned before that emotion is an essential aspect of building humanize chatbot. The rise of the emotionally-aware chatbot is started by Parry [8] in early 1975. Now, most of EAC development exploits neural-based model. In this section, we will try to review previous works which focus on EAC development. Table 1 summarizes this information includes the objective and exploited approach of each work. In early development, EAC is designed by using a rule-based approach. However, in recent years mostly EAC exploit neural-based approach. Studies in EAC development become a hot topic start from 2017, noted by the first shared task in Emotion Generation Challenge on NLPCC 2017 [20]. Based on Table 1 this research line continues to gain massive attention from scholars in the latest years.

Based on Table 1, we can see that most of all recent EAC was built by using encoder-decoder architecture with sequence-to-sequence learning. These seq2seq learning models maximize the likelihood of response and are prepared to incorporate rich data to generate an appropriate answer. Basic seq2seq architecture structured of two recurrent neural networks (RNNs), one as an encoder processing the input and one as a decoder generating the response. long short term memory (LSTM) or gated recurrent unit (GRU) was the most dominant variant of RNNs which used to learn the conversational dataset in these models. Some studies also tried to model this task as a reinforcement learning task, in order to get more generic responses and let the chatbot able to achieve successful long-term conversation. Attention mechanism was also introduced in this report<sup>6</sup>. This mechanism will allow the decoder to focus only on some important parts in the input at every decoding step.

Another vital part of building EAC is emotion classifier to detect emotion contained in the text to produce a more meaningful response. Emotion detection is a well-established task in natural language processing research area. This task was promoted in two latest series of SemEval-2018 (Task 1) and SemEval-2019 (Task 3). Some tasks were focusing on classifying utterance into several categories of emotion [34]. However, there is also a task which trying to predict the emotion intensities contained in the text [33]. In the early development of emotion classifier, most of the studies proposed to use traditional machine-learning approach. However, the neural-based approach is able to gain better performance, which leads more scholars to exploit it to deal with this task. In chatbot, the system will generate several responses based on several emotion categories. Then the system will respond with the most appropriate emotion based on emotion detected on posted utterance by emotion classifier. Based on Table 1, studies have different emotion categories based on their focus and objective in building chatbots.

<sup>6</sup>[http://web.stanford.edu/class/cs224s/reports/Honghao\\_Wei.pdf](http://web.stanford.edu/class/cs224s/reports/Honghao_Wei.pdf)

Authors	Year	Focus	Approach
Kenneth Mark Colby [8]	1975	Designing chatbot behave like paranoid person	Rule-based approach with capability to simulate emotion.
Zhou et. al. [66]	2014	Building empathetic chatbot for social interaction	Seq2seq learning with GRU-RNN model which takes into account intelligent quotient (IQ) and emotion quotient (EQ).
Zhou et. al. [65]	2018	Building emotion chatting machine (ECM) for large scale conversation generation	Seq2seq learning with GRU which incorporates emotion detection to capture implicit internal emotion states.
Colombo et. al. [9]	2019	Developing affect-driven dialogue system which generates emotional responses in a controlled manner	Seq2seq with GRU which incorporates emotion classifier, also affective re-ranking in the last step to produce the response.
Li et. al. [23]	2019	Building conversational system emotional by using editing constraints to generate more meaningful and customizable emotional replies	<u>Encoder-decoder architecture</u> with reinforcement learning approach by using asynchronous decoder which uses keyword predictor to predict the topic and emotion editor to produce emotional-embedded response.
Zhang et. al. [62]	2017	Building emotional conversation systems which produces several responses for every emotion category	Multi-task seq2seq model with GRU which utilize bidirectional long short term memory (Bi-LSTMs) as emotion classifier.
Lubis et. al. [27]	2019	Building a fully data driven chat-oriented dialogue system that can dynamically mimic affective human interactions	Proposing a seq2seq response generator, includes emotion encoder which is trained jointly with the entire network to encode and maintain the emotional context throughout the dialogue, but focusing only on positive emotion
Ashgar et. al. [2]	2018	Building open-domain neural dialogue models by augmenting them with affective intelligence.	<u>Encoder-decoder architecture</u> with LSTM model using cognitively engineered affective word embedding, and also affectively diverse beam search for decoding
Sun et. al. [55]	2018	Developing conversation generation which addresses the emotional factor by changing the model's input.	<u>Encoder-decoder framework</u> based on LSTM using three inputs, a sequence without an emotional category, a sequence with an emotional category for the input sentence, and a sequence with an emotional category for the output responses.
Hu et. al. [18]	2018	Building a tone-aware chatbot for customer care on social media.	<u>Encoder-decoder architecture</u> with LSTM, which the decoder is modified to handle meta information by add an embedding vector as tone indicator to produce several tone responses including passionate, empathetic, and neutral.
Sun et. al. [55]	2018	Building emotional human machine conversation chatbot that be able to understands the user's emotions, and consider the user's emotions then give a satisfactory response.	Model this problem as reinforcement learning task by proposing a new neural model based on Seq2Seq model which uses generative adversarial network (GAN) called EM-SeqGAN.
Shantala et. al. [50]	2018	Building neural dialogue system which addresses emotional aspect.	Seq2Seq model with LSTM by using a novel emotion embedding.
Zhong et. al. [64]	2018	Building affect-rich open domain human-machine conversation system.	Extends Seq2Seq model and adopts VAD (Valence, Arousal and Dominance) affective notations. Proposed model also considers the effect of negators and intensifiers via a novel affective attention mechanism.
Catania et. al. [6]	2019	Building a modular framework to facilitate and accelerate the realization and the maintenance of intelligent Conversational Agents with both rational and emotional capabilities.	The model consists of several modules including intent predictor, sentiment analysis, emotion analysis, topic analysis, and profiling module.

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

## 4 RESOURCE FOR BUILDING EAC

In this section, we try to investigate the available resources in building EAC. As other artificial intelligent agents, building chatbot also needs a dataset to be learned, to be able to produce a meaningful conversation as a human-like agent. Therefore, some studies propose dataset which contains textual conversation annotated by different emotion categories. Table 2 summarizes the available dataset that found in recent years. We categorize the dataset based on the language, source of the data, and further description which contains some information such as annotation approach, size of instances, and emotion labels. All datasets were proposed during 2017 and 2018, started by dataset provided by NLPCC 2017 Shared Task on Emotion Generation Challenge organizers. This dataset gathered from Sina Weibo social media <sup>7</sup>, so it consists of social conversation in Chinese. Based on our study, all of the datasets that we discover are only available in two languages, English and Chinese. However, the source of these datasets is very diverse such as social media (Twitter, Sina Weibo, and Facebook Message), online content, and human writing through crowdsourcing scenario. Our investigation found that every dataset use different set of emotion label depends on its focus and objective in building the chatbot.

As we mentioned in the previous section, that emotion classifier is also an integral part of the emotionally-aware chatbot. In building emotion classifier, there are also available several affective resources, which already widely used in the emotion classification task. Table 3 shows the available affective resources we discovered, range from old resources such as LIWC, ANEW, and DAL, until the modern version such as DepecheMood and EmoWordNet. Based on some prior studies, emotion can be classified into two fundamental viewpoints, including discrete categories and dimensional models. Meanwhile, this feature view emotion as a discrete category which differentiates emotion into several primary emotion. The most popular was proposed by [14] that differentiate emotion into six categories including anger, disgust, fear, happiness, sadness, and surprise. Three different resources will be used to get emotion contained in the tweet, including EmoLex, EmoSentNet, LIWC, DepecheMood, and EmoWordNet. Dimensional model is another viewpoint of emotion that define emotion according to one or more dimension. Based on the dimensional approach, emotion is a coincidence value on some different strategic dimensions. We will use two various lexical resources to get the information about the emotional dimension of each tweet, including Dictionary of Affect in Language (DAL), ANEW, and NRC VAD.

## 5 EVALUATING EAC

We characterize the evaluation of Emotionally-Aware Chatbot into two different parts, qualitative and quantitative assessment. Qualitative assessment will focus on assessing the functionality of the software, while quantitative more focus on measure the chatbots' performance with a number.

### 5.1 Qualitative Assessment

Based on our investigation of several previous studies, we found that most of the works utilized ISO 9241 to assess chatbots' quality by focusing on the usability aspect. This aspect can be grouped into

three focuses, including efficiency, effectiveness, and satisfaction, concerning systems' performance to achieve the specified goals. Here we will explain every focus based on several categories and quality attributes.

**5.1.1 Efficiency.** Efficiency aspect focuses on several categories, including robustness to manipulation and unexpected input [56]. Another study tries to assess the chatbots' ability to control damage and inappropriate utterance [36].

**5.1.2 Effectiveness.** Effectiveness aspect covers two different categories, functionality and humanity. In functionality point of view, a study by [13] propose to assess how a chatbot can interpret command accurately and provide its status report. Other functionalities, such as chatbots' ability to execute the task as requested, the output linguistic accuracy, and ease of use suggested being assessed [7]. Meanwhile, from the human aspect, most of the studies suggest each conversational machine should pass Turing test [58]. Other prominent abilities that chatbot needs to be mastered can respond to specific questions and able to maintain themed discussion.

**5.1.3 Satisfaction.** Satisfaction aspect has three categories, including affect, ethics and behaviour, and accessibility. Affect is the most suitable assessment categories for EAC. This category assess several quality aspects such as, chatbots' ability to convey personality, give conversational cues, provide emotional information through tone, inflexion, and expressivity, entertain and/or enable the participant to enjoy the interaction and also read and respond to moods of human participant [30]. Ethic and behaviour category focuses on how a chatbot can protect and respect privacy [13]. Other quality aspects, including sensitivity to safety and social concerns and trustworthiness [31]. The last categories are accessibility, which the main quality aspect focus to assess the chatbot ability to detect meaning or intent and, also responds to social cues <sup>12</sup>.

## 5.2 Quantitative Assessment

**5.2.1 Automatic Evaluation.** In automatic evaluation, some studies focus on evaluating the system at emotion level [23, 65]. Therefore, some common metrics such as precision, recall, and accuracy are used to measure system performance, compared to the gold label. This evaluation is similar to emotion classification tasks such as previous SemEval 2018 [34] and SemEval 2019 <sup>13</sup>. Other studies also proposed to use perplexity to evaluate the model at the content level (to determine whether the content is relevant and grammatical) [23, 26, 45]. This evaluation metric is widely used to evaluate dialogue-based systems which rely on probabilistic approach [48]. Another work by [45] used BLEU to evaluate the machine response and compare against the gold response (the actual response), although using BLEU to measure conversation generation task is not recommended by [25] due to its low correlation with human judgment.

**5.2.2 Manual Evaluation.** This evaluation involves human judgement to measure the chatbots' performance, based on several criteria. [65] used three annotators to rate chatbots' response in two

<sup>7</sup><https://www.weibo.com/login.php>

<sup>12</sup><https://sloanreview.mit.edu/article/will-ai-create-as-many-jobs-as-it-eliminates/>

<sup>13</sup><https://www.humanizing-ai.com/emocontext.html>



Authors	Year	Language	Source	Description
Zhou et. al. [65]	2018	Chinese	Weibo	This dataset used STC dataset [49], and performed automatic annotation by using the best performing classifier trained on NLPCC 2013 <sup>8</sup> and NLPCC 2014 <sup>9</sup> datasets. This dataset contains 217,905 conversations where each response annotated by 6 emotion labels.
Rashkin et. al. [45]	2018	English	ParlAI Platform	This dataset is built by using crowd-sourcing scenario using ParlAI which involves 810 different participants. This dataset contains 24,850 conversations/prompts where each conversation annotated by <u>32 emotion labels</u> .
Huang et. al. [20]	2017	Chinese	Weibo	Similar to [65], the training data is automatically annotated by using trained emotion classifier. This dataset contains more than 1 million conversations as training and 200 conversation which was manually annotated as test set, where each response labeled by 5 emotion labels.
Hu et. al. [18]	2018	English	Twitter	This dataset gathered from Twitter based on customer care conversation in 62 brands across different industries. 500 conversation was chosen randomly and annotated by <u>8 major tones</u> using CrowdFlower <sup>10</sup> platform
Lubis et. al. [26]	2018	English	Crowd source	This dataset was built by enhancing SEMAINE dataset [29], adding a label on instance which elicit positive emotions (8 labels including alert, excited, elated, happy, content, serene, relaxed, calm). The final dataset contains 2,349 conversations.
Li et. al. [24]	2017	English	Online Website	This dataset (called "DailyDialog") is crawled from various websites which serve for English learner to practice English dialog in daily life. DailyDialog contains 13,118 dialogs, and was manually annotated with communication intention and emotion information. They use <u>six primary basic emotion</u> (Anger, Disgust, Fear, Happiness, Sadness, Surprise) to label the conversations.
Huang et. al. [19]	2018	English	Facebook Message	This dataset was collected from private conversation on Facebook messenger application. 8,818 messages were manually annotated by using <u>seven emotion labels</u> including joy, anger, sadness, anticipation, tired, fear and neutral.

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

criteria, content (scale 0,1,2) and emotion (scale 0,1). Content is focused on measuring whether the response is natural acceptable and could plausible produced by a human. This metric measurement is already adopted and recommended by researchers and conversation challenging tasks, as proposed in [49]. Meanwhile, emotion is defined as whether the emotion expression contained in the response agrees with the given gold emotion category. Similarly, [23] used four annotators to score the response based on consistency, logic and emotion. Consistency measures the fluency and grammatical aspect of the response. Logic measures the degree whether the post and response logically match. Emotion measures the response, whether it contains the appropriate emotion. All of these aspects were measured by three scales 0, 1, and 2. Meanwhile, [26] proposed naturalness and emotion impact as criteria to evaluate the chatbots' response. Naturalness evaluates whether the response is intelligible, logically follows the context of the conversation, and acceptable as a human response, while emotion impact measures whether the response elicits a positive emotional or triggers an emotionally-positive dialogue, since their study focus only on positive emotion. Another study by [45] uses crowdsourcing to

gather human judgement based on three aspects of performance including empathy/sympathy - did the responses show understanding of the feelings of the person talking about their experience?; relevance - did the responses seem appropriate to the conversation? Were they on-topic?; and fluency - could you understand the responses? Did the language seem accurate?. All of these aspects recorded with three different response, i.e., (1: not at all, 3: somewhat, 5: very much) from around 100 different annotators. After getting all of the human judgement with different criteria, some of these studies used a t-test to get the statistical significance [23, 26], while some other used inter-annotator agreement measurement such as Fleiss Kappa [45, 65]. Based on these evaluations, they can compare their system performance with baseline or any other state of the art systems.

## 6 RELATED WORK

There is some work which tried to provide a full story of chatbot development both in industries and academic environment. However, as far as my knowledge, there is still no study focused on summarizing the development of chatbot which taking into account the emotional aspect, that getting more attention in recent years.

Authors	Name	Description
Pennebaker et. al. [39]	Linguistic Inquiry and Word Count (LIWC)	LIWC is a computer application that offers an efficient instrument for word-by-word study of the emotional, cognitive, and structural elements in English. LIWC also provide a dictionary which has 4500 words distributed into 64 different emotional categories including positive and negative.
Mohammad et. al. [35]	Emolex	Emolex was built by crowdsourcing scenario. Emolex contains 14.182 words associated with eight primary emotion based on [40] : joy, sadness, anger, fear, trust, surprise, disgust, and anticipation.
Poria et. al. [42]	EmoSenticNet	EmoSenticNet(EmoSN) is an enriched version of SenticNet[5]. Work by [42] added emotion label by mapping WordNet-Affect label to the SenticNet concepts. As a WordNet-Affect label, EmoSenticNet uses Six Ekman's basic emotion : anger, disgust, fear, happiness, sadness, and surprise. The whole list of this lexica contains 13.189 emotion labeled words.
Whissell et. al. [60]	Dictionary of Affect in Language (DAL)	DAL was developed by [60] and composed of 8742 English words. These words were labeled by three scores represent three emotion dimension : Pleasantness (the degree of pleasure), Activation (the degree of human response under some emotional condition), and Imagery (the degree of how easy the given words to be formed a mental picture).
Bradley et. al. [4]	Affective Norms for English Words (ANEW)	This lexicon consists of 1,034 English words rated with ratings based on the Valence-Arousal-Dominance (VAD) model [37].
de Albornoz et. al. [11]	SentiSense	SentiSense attaches emotional meanings to concepts from WordNet database. It is composed by a list of 5,496 words words labeled with emotional labels consisted of 14 emotional categories, which refer to a merge of Arnold, Plutchik and Parrot models.
Strapparava et. al. [54]	WordNet-Affect	WordNet-Affect is a lexical resource created based on WordNet which contains information about the emotions that the words convey. This lexicon is organized in six basic emotions: anger, disgust, fear, joy, sadness, surprise and, contains 2,874 synsets and 4,787 words
Saif M. Mohammad [32]	NRC VAD	The NRC VAD Lexicon is a list of English words and their valence, arousal, and dominance scores ranged between 0 (low) and 1 (high). This lexicon contains 20,000 terms and was built by crowdsourcing scenario.
Staiano and Guerini [53]	DepecheMood	DepecheMood is an emotion lexicon that crowdsource emotions through rapper website <sup>11</sup> . This lexicon contains 37,000 entries rated by eight emotion categories (happy, sad, angry, afraid, annoyed, inspired, amused, and don't care).
Badaro et.al. [3]	EmoWordNet	EmoWordNet is created by expanding an existing emotion lexicon, DepecheMood [53], by leveraging semantic knowledge from English WordNet. By aligning DepecheMood and WordNet, EmoWordNet consists of 67K terms, almost 1.8 times the size of DepecheMood. This lexicon has same emotion label as DepecheMood i.e., happy, sad, angry, afraid, annoyed, inspired, amused, and don't care.

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

[51] provides a long history of chatbots technology development. They also described several uses of chatbots' in some practical domains such as tools entertainment, tools to learn and practice language, information retrieval tools, and assistance for e-commerce of other business activities. Then, [12] reviewed the development of chatbots from rudimentary model to more advanced intelligent

system. They summarized several techniques used to develop chatbots from early development until recent years. Recently, [21] provide a more systematic shape to review some previous works on chatbots' development. They classify chatbots into two main categories based on goals, including task-oriented chatbots and non-task oriented chatbot. They also classify chatbot based on its development technique into three main categories, including rule-based, retrieval-based, and generative-based approach. Furthermore, they

check.

also summarized the detailed technique on these three main approaches.

## 7 DISCUSSION AND CONCLUSION

In this work, a systematic review of emotionally-aware chatbots is proposed. We focus on three main issues, including, how to incorporate affective information into chatbots, what are resources that available and can be used to build EAC, and how to evaluate EAC performance. The rise of EAC was started by Parry, which uses a simple rule-based approach. Now, most of EAC are built by using a neural-based approach, by exploiting emotion classifier to detect emotion contained in the text. In the modern era, the development of EAC gains more attention since Emotion Generation Challenge shared task on NLPCC 2017. In this era, most EAC is developed by adopting encoder-decoder architecture with sequence-to-sequence learning. Some variant of the recurrent neural network is used in the learning process, including long-short-term memory (LSTM) and gated recurrent unit (GRU). There are also some datasets available for developing EAC now. However, the datasets are only available in English and Chinese. These datasets are gathered from various sources, including social media, online website and manual construction by crowdsourcing. Overall, the difference between these datasets and the common datasets for building chatbot is the presence of an emotion label. In addition, we also investigate the available affective resources which usually use in the emotion classification task. In this part, we only focus on English resources and found several resources from the old one such as LIWC and Emolex to the new one, including DepecheMood and EmoWordNet. In the final part, we gather information about how to evaluate the performance of EAC, and we can classify the approach into two techniques, including qualitative and quantitative assessment. For qualitative assessment, most studies used ISO 9241, which covers several aspects such as efficiency, effectiveness, and satisfaction. While in quantitative analysis, two techniques can be used, including automatic evaluation (by using perplexity) and manual evaluation (involving human judgement). Overall, we can see that effort to humanize chatbots by incorporation affective aspect is becoming the hot topic now. We also predict that this development will continue by going into multilingual perspective since up to now every chatbot only focusing on one language. Also, we think that in the future the studies of humanizing chatbot are not only utilized emotion information but will also focus on a contextual-aware chatbot.

## REFERENCES

- [1] Hadeel Al-Zubaide and Ayman A Issa. 2011. Ontbot: Ontology based chatbot. In *International Symposium on Innovations in Information and Communications Technology*. IEEE, 7–12.
- [2] Nabihha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. 2018. Affective neural response generation. In *European Conference on Information Retrieval*. Springer, 154–166.
- [3] Gilbert Badaro, Hussein Jundi, Hazem Hajj, and Wassim El-Hajj. 2018. EmoWordNet: Automatic Expansion of Emotion Lexicon Using English WordNet. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. 86–93.
- [4] Margaret M Bradley and Peter J Lang. 1999. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical Report. Technical Report C-1, The Center for Research in Psychophysiology, University of Florida.
- [5] Erik Cambria, Catherine Havasi, and Amir Hussain. 2012. Senticnet 2: A semantic and affective resource for opinion mining and sentiment analysis. In *Twenty-Fifth International FLAIRS Conference*.
- [6] Fabio Catania, Micol Spitale, Davide Fiscaro, and Franca Garzotto. 2019. CORK: A CONversational agent framewORK exploiting both rational and emotional intelligence. (2019).
- [7] David Cohen and Ian Lane. 2016. An Oral Exam for Measuring a Dialog System's Capabilities. In *Thirtieth AAAI Conference on Artificial Intelligence*.
- [8] Kenneth Mark Colby. 2013. *Artificial paranoia: A computer simulation of paranoid processes*. Vol. 49. Elsevier.
- [9] Pierre Colombo, Wojciech Witon, Ashutosh Modi, James Kennedy, and Mubbasir Kapadia. 2019. Affect-Driven Dialog Generation. *arXiv preprint arXiv:1904.02793* (2019).
- [10] Lei Cui, Shaohan Huang, Furu Wei, Chuanqi Tan, Chaoqun Duan, and Ming Zhou. 2017. Superagent: A customer service chatbot for e-commerce websites. *Proceedings of ACL 2017, System Demonstrations* (2017), 97–102.
- [11] Jorge Carrillo de Albornoz, Laura Plaza, and Pablo Gervás. 2012. SentiSense: An easily scalable concept-based affective lexicon for sentiment analysis.. In *LREC*. 3562–3567.
- [12] Aditya Deshpande, Alisha Shahane, Darshana Gadre, Mrunmayi Deshpande, and Prachi M Joshi. 2017. A Survey of Various Chatbot Implementation Techniques. *International Journal of Computer Engineering and Applications* 11 (2017).
- [13] M v Eeuwen. 2017. *Mobile conversational commerce: messenger chatbots as the next interface between businesses and consumers*. Master's thesis. University of Twente.
- [14] Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion* 6, 3-4 (1992), 169–200.
- [15] Hao Fang, Hao Cheng, Maarten Sap, Elizabeth Clark, Ari Holtzman, Yejin Choi, Noah A Smith, and Mari Ostendorf. 2018. Sounding Board: A User-Centric and Content-Driven Social Chatbot. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*. 96–100.
- [16] Eun Go and S Shyam Sundar. 2019. Humanizing Chatbots: The effects of visual, identity and conversational cues on humanness perceptions. *Computers in Human Behavior* (2019).
- [17] Braden Hancock, Antoine Bordes, Pierre-Emmanuel Mazare, and Jason Weston. 2019. Learning from Dialogue after Deployment: Feed Yourself, Chatbot! *arXiv preprint arXiv:1901.05415* (2019).
- [18] Tianran Hu, Anbang Xu, Zhe Liu, Quanzeng You, Yufan Guo, Vibha Sinha, Jiebo Luo, and Rama Akkiraju. 2018. 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*. ACM, 415.
- [19] Chieh-Yang Huang and Lun-Wei Ku. 2018. EmotionPush: Emotion and Response Time Prediction Towards Human-Like Chatbots. In *2018 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 206–212.
- [20] Minlie Huang, Zuoxian Ye, and Hao Zhou. 2017. Overview of the NLPCC 2017 Shared Task: Emotion Generation Challenge. In *National CCF Conference on Natural Language Processing and Chinese Computing*. Springer, 926–936.
- [21] Shaifuat Hussain, Omid Ameri Sianaki, and Nedal Ababneh. 2019. A Survey on Conversational Agents/Chatbots Classification and Design Techniques. In *Workshops of the International Conference on Advanced Information Networking and Applications*. Springer, 946–956.
- [22] A. Kothari, J. Hoover, and R. Zyane. 2017. *Chatbots for ECommerce: Learn how to Build a Virtual Shopping Assistant*. Bleeding Edge Press. <https://books.google.it/books?id=DZ7cswEACAAJ>
- [23] Jia Li, Xiao Sun, Xing Wei, Changliang Li, and Jianhua Tao. 2019. Reinforcement Learning Based Emotional Editing Constraint Conversation Generation. *arXiv preprint arXiv:1904.08061* (2019).
- [24] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 986–995.
- [25] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2122–2132.
- [26] Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2018. Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- [27] Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. 2019. Positive Emotion Elicitation in Chat-Based Dialogue Systems. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27, 4 (2019), 866–877.
- [28] Bilyana Martinovski and David Traum. 2003. Breakdown in human-machine interaction: the error is the clue. In *Proceedings of the ISCA tutorial and research workshop on Error handling in dialogue systems*. 11–16.

next steps

- [29] Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2012. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing* 3, 1 (2012), 5–17.
- [30] MO Meira and AMP Canuto. 2015. Evaluation of Emotional Agents' Architectures: an Approach Based on Quality Metrics and the Influence of Emotions on Users. In *Proceedings of the World Congress on Engineering*, Vol. 1.
- [31] Adam S Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. 2016. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA internal medicine* 176, 5 (2016), 619–625.
- [32] Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 174–184.
- [33] Saif Mohammad and Felipe Bravo-Marquez. 2017. WASSA-2017 Shared Task on Emotion Intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 34–49.
- [34] Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, 1–17.
- [35] Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence* 29, 3 (2013), 436–465.
- [36] Kellie Morrissey and Jurek Kirakowski. 2013. "Reality" in Chatbots: Establishing Quantifiable Criteria. In *International Conference on Human-Computer Interaction*. Springer, 87–96.
- [37] C.E. Osgood, G.J. Suci, and P.H. Tenenbaum. 1957. *The Measurement of meaning*. University of Illinois Press, Urbana..
- [38] Baolin Peng, Xiujun Li, Lihong Li, Jianfeng Gao, Asli Celikyilmaz, Sungjin Lee, and Kam-Fai Wong. 2017. Composite Task-Completion Dialogue Policy Learning via Hierarchical Deep Reinforcement Learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2231–2240.
- [39] James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates* 71, 2001 (2001), 2001.
- [40] Robert Plutchik. 2001. The Nature of Emotions Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American scientist* 89, 4 (2001), 344–350.
- [41] Thomas S Polzin and Alexander Waibel. 2000. Emotion-sensitive human-computer interfaces. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*.
- [42] Soujanya Poria, Alexander Gelbukh, Amir Hussain, Newton Howard, Dipankar Das, and Sivaji Bandyopadhyay. 2013. Enhanced SenticNet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems* 28, 2 (2013), 31–38.
- [43] Helmut Prendinger and Mitsuru Ishizuka. 2005. The empathic companion: A character-based interface that addresses users' affective states. *Applied Artificial Intelligence* 19, 3-4 (2005), 267–285.
- [44] Minghui Qiu, Feng-Lin Li, Siyu Wang, Xing Gao, Yan Chen, Weipeng Zhao, Haiqing Chen, Jun Huang, and Wei Chu. 2017. Alime chat: A sequence to sequence and rerank based chatbot engine. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 498–503.
- [45] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. I know the feeling: Learning to converse with empathy. *arXiv preprint arXiv:1811.00207* (2018).
- [46] Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 583–593.
- [47] Iulian V Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, et al. 2017. A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349* (2017).
- [48] Iulian V Serban, Alessandro Sordani, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*.
- [49] Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Vol. 1, 1577–1586.
- [50] Roman Shantala, Gennadiy Kyselov, and Anna Kyselova. 2018. Neural Dialogue System with Emotion Embeddings. In *2018 IEEE First International Conference on System Analysis & Intelligent Computing (SAIC)*. IEEE, 1–4.
- [51] Bayan Abu Shawar and Eric Atwell. 2007. Chatbots: are they really useful?. In *Ldv forum*, Vol. 22, 29–49.
- [52] Alessandro Sordani, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A Neural Network Approach to Context-Sensitive Generation of Conversational Responses. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, 196–205. <https://doi.org/10.3115/v1/N15-1020>
- [53] Jacopo Staiano and Marco Guerini. 2014. Depeche Mood: a Lexicon for Emotion Analysis from Crowd Annotated News. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2, 427–433.
- [54] Carlo Strapparava, Alessandro Valitutti, et al. 2004. Wordnet affect: an affective extension of wordnet. In *LREC*, Vol. 4. Citeseer, 40.
- [55] Xiao Sun, Xinmiao Chen, Zhengmeng Pei, and Fuji Ren. 2018. Emotional Human Machine Conversation Generation Based on SeqGAN. In *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*. IEEE, 1–6.
- [56] Andree Thielges, Florian Schmidt, and Simon Hegelich. 2016. The devil's triangle: Ethical considerations on developing bot detection methods. In *2016 AAAI Spring Symposium Series*.
- [57] Alan M Turing. 2009. Computing machinery and intelligence. In *Parsing the Turing Test*. Springer, 23–65.
- [58] Joseph Weizenbaum. 1966. ELIZA—a computer program for the study of natural language communication between man and machine. *Commun. ACM* 9, 1 (1966), 36–45.
- [59] Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gasic, Lina M Rojas Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A Network-based End-to-End Trainable Task-oriented Dialogue System. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 438–449.
- [60] Cynthia Whissell. 2009. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *Psychological reports* 105, 2 (2009), 509–521.
- [61] Zhou Yu, Alexandros Papangelis, and Alexander Rudnicky. 2015. TickTock: A non-goal-oriented multimodal dialog system with engagement awareness. In *2015 AAAI Spring symposium series*.
- [62] Rui Zhang, Zhenyu Wang, and Dongcheng Mai. 2017. Building emotional conversation systems using multi-task Seq2Seq learning. In *National CCF Conference on Natural Language Processing and Chinese Computing*. Springer, 612–621.
- [63] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2204–2213.
- [64] Peixiang Zhong, Di Wang, and Chunyan Miao. 2018. An Affect-Rich Neural Conversational Model with Biased Attention and Weighted Cross-Entropy Loss. *arXiv preprint arXiv:1811.07078* (2018).
- [65] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- [66] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2018. The Design and Implementation of XiaoIce, an Empathetic Social Chatbot. *arXiv preprint arXiv:1812.08989* (2018).



This figure "sample-franklin.png" is available in "png" format from:

<http://arxiv.org/ps/1906.09774v1>