Hindawi
Wireless Communications and Mobile Computing
Volume 2020, Article ID 8881616, 10 pages
https://doi.org/10.1155/2020/8881616

Research Article

Emotional Dialogue Generation Based on Conditional Variational Autoencoder and Dual Emotion Framework

Zhenrong Deng,[1] Hongquan Lin,[2] Wenming Huang,[2] Rushi Lan,[3] and Xiaonan Luo[4]

[1] *Guangxi Key Laboratory of Image and Graphic Intelligent Processing, Guilin 541004, China*
[2] *School of Computer Science and Information Security, Guilin University of Electronic Technology, Guilin 541004, China*
[3] *School of Computer Science & Engineering, South China University of Technology, Guangzhou 510006, China*
[4] *National and Local Joint Engineering Research Center of Satellite Navigation and Location Service, Guilin University of Electronic Technology, Guilin 541004, China*

Correspondence should be addressed to Wenming Huang; 995456524@qq.com

Received 19 August 2020; Revised 2 November 2020; Accepted 5 December 2020; Published 28 December 2020

Academic Editor: Yaguang Lin

An excellent dialogue system needs to not only generate rich and diverse logical responses but also meet the needs of users for emotional communication. However, despite much work, these two problems have not been solved. In this paper, we propose a model based on conditional variational autoencoder and dual emotion framework (CVAE-DE) to generate emotional responses. In our model, latent variables of the conditional variational autoencoder are adopted to promote the diversity of conversation. A dual emotion framework is adopted to control the explicit emotion of the response and prevent the conversation from generating emotion drift indicating that the emotion of the response is not related to the input sentence. A multiclass emotion classifier based on the Bidirectional Encoder Representations from Transformers (BERT) model is employed to obtain emotion labels, which promotes the accuracy of emotion recognition and emotion expression. A large number of experiments show that our model not only generates rich and diverse responses but also is emotionally coherent and controllable.

1. Introduction

With the development of privacy protection and incentive technology in the Internet of Things and mobile social networks driven by artificial intelligence, intelligent dialogue systems have entered our daily lives [1–4]. The enormous demands of privacy protection for dialogue systems have promoted the accuracy of speech recognition and semantic understanding, greatly improving the experience of human-machine dialogue. At the same time, people have put forward increasing requirements for intelligent dialogue systems to produce more human-like dialogues. As an important part of human intelligence, emotional intelligence is defined as the ability to perceive, integrate, understand, and regulate emotions [5]. Thus, machines will be able to communicate at the human level only when they have the ability to perceive and express emotions.

Currently, deep neural networks have been successfully applied in various applications [6–9]. In dialogue generation tasks, the sequence to sequence (Seq2Seq) model [10] is a commonly used model. It is mainly based on the language ability learned from a large number of corpora to conduct dialogue and on the powerful calculation ability and abstraction ability to automatically summarize and extract valuable knowledge and features from massive data. In an open-domain dialogue system, there are multiple reasonable replies to a given query from a user. This phenomenon is called "one-to-many" diversity. However, for the dialogue system based on the Seq2Seq model and the maximum likelihood estimation (MLE) objective, the characteristics of the model determine the general utterance with a greater probability of its tendency to respond, such as "I don't know" and "Yes."

To generate more informative and meaningful responses, much work has been carried out in the open-domain dialogue [11–13]. These methods focus on the consistency of the conversation content rather then on emotion. Based on the past progress of dialogue systems, Zhou et al. [14] first integrated emotional factors into large-scale dialogue generation using embedding of emotional tags, internal memory networks, and external memory networks. Subsequently, Asghar et al. [15] used emotion word embedding and emotion-based objective functions to improve performance. Zhou et al. [16] proposed to use the emoticon-rich Twitter corpus as a data set for emotional dialogue generation. However, the above work only considers the characteristics of target emotions and not the emotion of the input sentence, with the hope that the machine generates corresponding emotional responses; this will lead to the phenomenon of emotional drift, that is, the emotional response is incoherent and inconsistent with the emotion of the input sentence.

The generation of emotional dialogue needs to consider two main factors: one is the content of the generated response, and the other is the emotion of the generated response. In addition to avoiding the generation of a large number of general replies and increase the diversity of replies, it is necessary to consider the connection between the output emotion and the emotion of the user's input sentence, as well as the controllability of the output emotion. For example, if the user is sad, we can generate comforting words to make the user feel better.

The contributions of our work are summarized as follows:

(1) We propose a dual-emotional framework for emotional dialogue generation, which comprehensively considers the impact of the emotion of the input sentence and the target emotion on emotional response in order to make our emotional response consistent with the user's emotion and ensure that the emotional response is controllable

(2) We combine the conditional variational autoencoder [17] with the dual emotion framework to train an emotional generation system, and experiments prove that our model has strong performance

(3) A multiclass emotion classifier based on the BERT [18] model is employed to obtain emotion labels, which improves the accuracy of emotion recognition and emotion expression.

The rest of the paper is organized as follows. In "Related Work," we outline the related work on emotional conversational agents. Then, we describe the proposed model in "Proposed Model." "Experiment" provides the experimental results. Finally, we summarize this article and propose directions for the future work in "Conclusion."

2. Related Work

With the popularity of social media, massive quantities of dialogue data can be accumulated and saved, allowing researchers to solve the problems of dialogue systems in a purely data-driven manner. Vinyals et al. [19] applied the Seq2Seq model in machine translation for dialogue generation for the first time, using an encoder to encode input sentences and generating a reply through a decoder. Bahdanau et al. [20] proposed an attention mechanism and applied it to the field of machine translation to improve the accuracy of machine translation. Shang et al. [21] first built a corpus based on Sina Weibo and used a Seq2Seq model that introduced an attention mechanism to implement a single-round dialogue generation system.

Depending on the dialogue object and the dialogue scene, some work introduces latent variables, samples the distribution of latent variables, and then decodes the distribution to generate responses based on latent variables. Cao et al. [22] proposed a single-round dialogue generation model based on latent variables, including random variables z of the variational autoencoder in the decoder. Serban et al. [23] introduced the method of latent variables into a hierarchical dialogue model. The latent variables can be either topics or emotions. Zhao et al. [13] constructed a dialogue model based on a conditional variational autoencoder model using multiple semantic intentions as conditions.

Emotion perception is an indispensable part of a successful and intelligent dialogue system. Zhou et al. [14] proposed the emotional chat machine (ECM), which first focuses on how to generate a response with a specific emotion. ECM uses emotion embedding, internal memory network, and external memory network, but it considers neither the influence of the input sentence content on the decoder output nor the influence of the input sentence emotion on the emotional response. We believe that to learn higher-level dialogue skills and logic from a real corpus, a more elaborate mechanism is needed to capture the relationship between the utterance and emotional response. Therefore, we focus on the extraction and expression of the content and emotions of input sentences and produce more human-like emotional responses.

In terms of emotional dialogue research, [16] is most similar to our work, but they mainly focus on emotions in the Twitter corpus to train emotional chat robots, and their work did not further consider the emotional characteristics of the input sentences. Sun et al. [24] proposed a model that takes a sequence containing an emotion category of the input sentence and an emotion category of the output response as input. Xu et al. [25] proposed a dual-attention mechanism that pays attention to the content and emotion of input statements. Song et al. [26] proposed an emotion dialogue system that can express the desired emotion explicit or implicitly. Li et al. [27] used generative adversarial networks to generate emotional responses. Su et al. [28] proposed a stylistic dialogue generation system, which is achieved by adopting an information-guided reinforcement learning strategy.

3. Proposed Model

3.1. Task Definition and Model Overview. Our task is defined as follows: given a post $x = (x_1, x_2, \cdots, x_m)$, input emotion label E_x, and target emotion label E_y, the goal is to generate a response $y = (y_1, y_2, \cdots, y_n)$. The input emotion label E_x is obtained through the multiemotion classifier, x_i is the token of the input sentence, and y_i is the token of the output sentence. The response not only is consistent with the post in terms of both content and emotion but also corresponds to the target emotion.

An overview of CVAE-DE is given in Figure 1. E_y is the emotion label of the response, E_x is the emotion label of the post, vector v_y represents the text features of the response, vector v_x represents the text features and emotion features of the post, vector e_y represents the emotion features of the response, and vector c is obtained by concatenation of v_x and e_y. In the training process, E_x and E_y are obtained from the BERT emotion classifier, post x and E_x are encoded by the post encoder to obtain v_x, E_y obtains vector e_y through a full connection network, and v_x is concatenated with e_y to obtain vector c. Then, c and v_y are fed to the prior/recognition network, and the hidden variable z, which is sampled from the recognition network, is fed to the decoder. In the inference process, the response does not exist, E_y is directly given by the user, z is sampled from the prior probability distribution $p(z \mid c)$, and we use an attention mechanism between the encoder and the decoder. Finally, the decoder will generate an emotional response that matches the post in content, is coherent with the post emotion, and corresponds to the target emotion based on attention memory, as well as c and z.

3.2. Multiemotion Classifier Based on the BERT Model. Most existing models use word2vec or Glove to obtain pretrained word vectors. However, the word vectors trained by these models are a type of static encoding. The same word is the same expression in different contexts, and it does not solve the problem of polysemy, in which words have different meanings in different contexts. In response to this problem, this paper trains a multiemotion classifier based on the BERT model [18]. BERT is a new language representation model that can not only obtain the rich grammatical and semantic features of the corpus text but also solve the problem of traditional language feature representation ignoring word polysemy, ultimately improving the accuracy of emotion classification. The structure of the BERT model is shown in Figure 2.

The most important part of the BERT model is the bidirectional Transformer encoder [29] encoding structure, which uses the encoder structure in the Transformer model as the feature extractor. The encoder is composed of a self-attention mechanism and a feed-forward neural network, abandoning the RNN's cyclic network structure [30], and completely uses an attention-based mechanism to model a segment of text. The attention mechanism in the encoder is called self-attention, and its core idea is to calculate the relationship between each word in a sentence and other words to adjust the importance of each word in order to obtain a context-related word vector. The encoder structure is shown in Figure 3.

In the experiments in this article, the pretrained Chinese model "BERT-Base, Chinese" released by Google is used to train our classifier; it uses a 12-layer Transformer with a hidden size of 768, a multihead attention parameter of 12, and a total model size of 110 MB. First, we load the pretrained model, and then, we use the emotion classification data set to fine-tune our model. Finally, the final model will be employed in the CVAE-DE model as our multiemotion classifier.

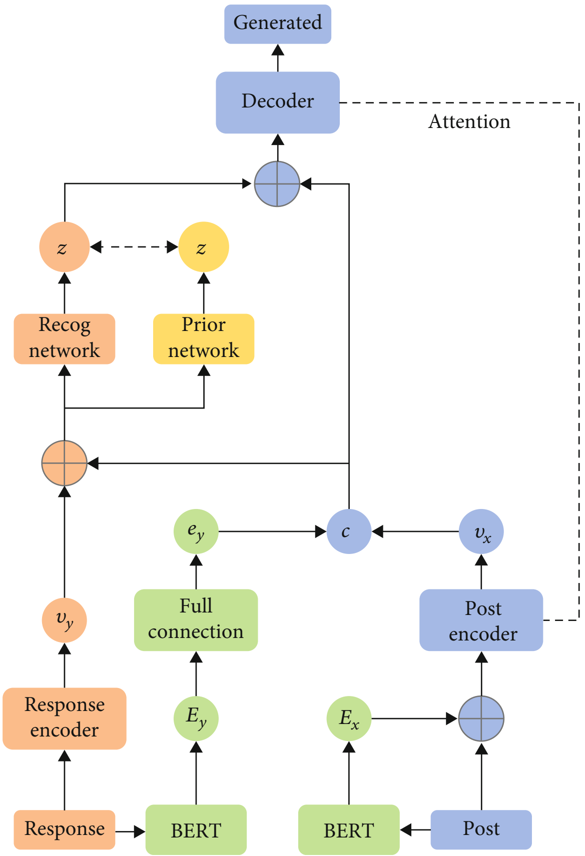

Figure 1: Overview of CVAE-DE. The blue part represents a Seq2Seq model based on the attention mechanism, which is the basis of our model. The orange and gold parts represent the conditional variational autoencoder model. The green part represents the dual emotion frame.

3.3. Sequence to Sequence Model Based on the Attention Mechanism. The basis of our model is a Seq2Seq model based on the attention mechanism [21]. The encoder and decoder of the model are implemented by GRU [31]. The role of the encoder is to map the post $x = (x_1, x_2, \cdots, x_m)$ to the hidden feature state $h = (h_1, h_2, \cdots, h_m)$. For moment t, h_t is defined as follows:

$$r_t = \delta(W_r x_t + U_r h_{t-1} + b_r), \tag{1}$$

$$z_t = \delta(W_z x_t + U_z h_{t-1} + b_z), \tag{2}$$

$$\widetilde{h}_t = \tanh\left(W_h x_t + U_h (r_t * h_{t-1}) + b_h\right), \tag{3}$$

$$h_t = (1 - z_t) h_{t-1} + z_t \widetilde{h}_t, \tag{4}$$

where the initial hidden state h_0 is zero vector, r_t represents the reset gate, z_t represents the update gate, δ is the sigmoid activation function, and W_r, W_z, W_h, U_r, U_z, U_h, b_r, b_z, b_h are training parameters. The sigmoid activation function can map the data to $[0, 1]$ to determine the gating signal. The update door has two functions: the forgetting function and memory function. It can not only selectively forget the historical information that is not related to the original

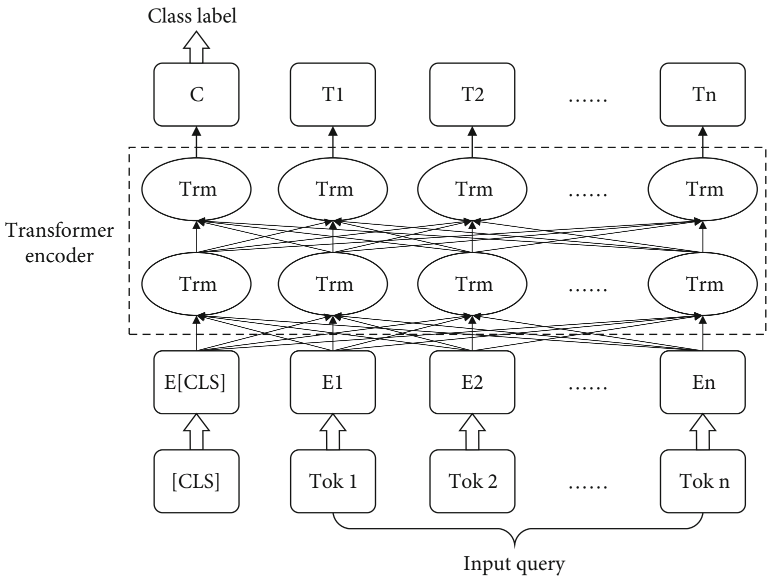

Figure 2: Fine-tuning BERT on emotion classification tasks.

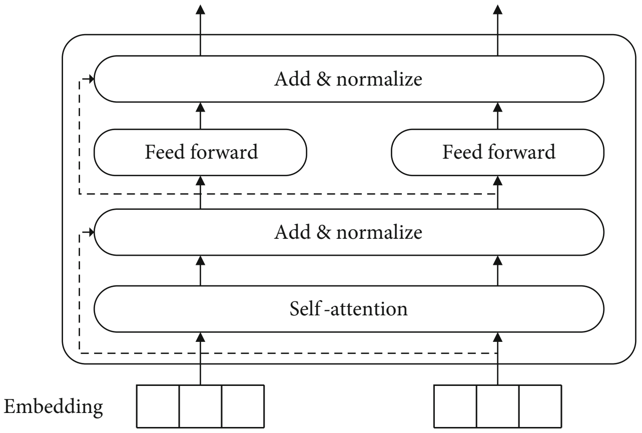

Figure 3: Transformer Encoder for feature extraction.

hidden state but also selectively remember the candidate hidden state and retain the long short-term information that is strongly dependent on the current moment. The above equations can be written as $h_t = GRU(h_{t-1}, x_t)$.

The current state of the decoder can be updated according to the state s_{t-1} at the previous time, the output y_{t-1} of the decoder at the previous time, and the context vector vc_t at the current time. The probability distribution of the words output by the decoder is

$$p(y_t \mid y_1, y_2, \cdots, y_{t-1}, x) = g(y_{t-1}, s_t, vc_t), \tag{5}$$

$$s_t = GRU(s_{t-1}, [y_{t-1}, vc_t]), \tag{6}$$

where g is the maxout activation function, the context vector vc_t is the result of using the attention mechanism to weight the encoder state sequence h, and typically, we use Bahdanau attention [20], which is defined as:

$$e_{tj} = v_a^T \tanh \left(W_a s_{t-1} + U_a h_j\right), \tag{7}$$

$$\alpha_{tj} = \frac{\exp\left(e_{tj}\right)}{\sum_{k=1}^{m} \exp\left(e_{tk}\right)}, \tag{8}$$

$$vc_t = \sum_{j=1}^{m} \alpha_{tj} h_j, \tag{9}$$

where v_a, W_a, and U_a are the attention parameters that need to be learned. The attention mechanism is in fact a weighted sum of the hidden states of the encoder, which can dynamically capture the dependence of the decoder on the input utterance. The objective function of the Seq2Seq model based on the attention mechanism can be expressed as

$$p(y \mid x) = \prod_{t=1}^{n} p(y_t \mid y_1, y_2, \cdots, y_{t-1}, vc_t). \tag{10}$$

3.4. Conditional Variational Autoencoder Model. The variational autoencoder (VAE) is a generative network structure based on the variational Bayesian inference proposed by Kingma et al. [32]. The VAE has been used to establish two probability density distribution models: one model, called the inference network, involves generating a variational probability distribution of hidden variables according to the variational inference of the original input data; and the other model, called the generation network, involves restoring the approximate probability distribution of the original data according to the generated variational probability distribution of hidden variables. In this model, a prior distribution $p(z)$ is added to the hidden variable z, which often follows the standard Gaussian distribution, so that the model can generate samples that are closer to the original data distribution. The goal of the variational autoencoder is to maximize the probability $p(y)$ under the premise of sampling by z, which can be expressed as

$$p(y) = \int p(y \mid z)p(z)dz. \tag{11}$$

The VAE introduces a recognition model $q_\phi(z \mid y)$ in the inference network to replace the undetermined true posterior distribution $p_\theta(z \mid y)$. To make $q_\phi(z \mid y)$ approximately equal to $p_\theta(z \mid y)$, the VAE uses the KL divergence to measure the similarity between the two distributions and minimizes the KL divergence. In this case, the objective function of the model can be expressed as

$$L(\theta, \phi ; y) = -KL\left(q_\phi(z \mid y) \| p_\theta(z)\right) + \mathbb{E}_{q_\phi(z|y)}[\log p_\theta(z \mid y)], \tag{12}$$

where ϕ is the parameter of the inferred network, θ is the parameter of the generated network, $KL(q_\phi(z \mid y) \| p_\theta(z))$ indicates the KL divergence between the prior distribution $p_\theta(z)$ of z and the posterior distribution $q_\phi(z \mid y)$ of the model encoder, and $\mathbb{E}_{q_\phi(z|y)}[\log p_\theta(y \mid z)]$ represents the reconstruction loss of the data samples by the decoder $p_\theta(y \mid z)$. The model's decoder learning goal is to restore the real data as much as possible, and the goal of the variational autoencoder becomes to maximize its objective function, which can be achieved by minimizing the first term on the right side of Equation (12), that is, making $q_\phi(z \mid y)$ of the hidden variable z approximate $p_\theta(z)$.

The traditional VAE belongs to an unsupervised model. Although it can generate similar output data based on the input, it cannot control its orientation to generate specific types of data. For this purpose, Makhzani et al. [17] proposed a conditional variational autoencoder (CVAE) model. Based on the Seq2Seq model, we introduce the latent variable z in the CVAE model. For a given input utterance, multiple appropriate responses may exist, and each response corresponds to a potential variable configuration that does not appear in the input utterance. CVAE is trained by maximizing the conditional likelihood variational lower bound of y for a given c situation.

$$p(y \mid c) = \int p(y \mid z, c)p(z \mid c)dz. \tag{13}$$

In our model, the decoder is used to approximate $p_D(y \mid z, c)$, the prior network is used to approximate $p_P(z \mid c)$, and the recognition network is used to approximate the real posterior $p_R(z \mid y, c)$. θ_D, θ_P, and θ_R are the parameters of their networks. The objective function is given by

$$\begin{aligned} L(\theta_D, \theta_P, \theta_R ; y, c) = &-KL(q_R(z \mid y, c) \| p_P(z \mid c)) \\ &+ \mathbb{E}_{q_R(z|y,c)}[\log p_D(y \mid z, c)]. \end{aligned} \tag{14}$$

In addition, as described by Bowman et al. [33], it is difficult to encode useful information in hidden variables by directly combining the RNN decoder and the variational autoencoder in the field of text generation. Because the RNN-based decoder is a general function approximator, which has a strong ability to model sequence information, it can learn the representation without hidden variables information in the decoding process. The hidden variables lose their function, and VAE mathematically degenerates into a simple Seq2Seq model. Therefore, training a Seq2Seq dialogue generation model based on CVAE needs to balance the reconstruction loss and KL loss. In our experiments, we use the techniques of KL annealing, early stop, and bag loss to balance the reconstruction loss and KL loss. The bag of words loss is added to the training objective function on the previous basis, and the objective function is rewritten as

$$L' = L + L_{bow}. \tag{15}$$

3.5. Dual Emotion Framework. To make the emotional responses more coherent, we add an emotion label of the post to the input of the post encoder. The input becomes $[E_x ; x]$, enabling our model to mine the emotional information of the post and make the emotional response compatible with the post emotion. The estimated probability of the model can be rewritten as

$$p(y \mid E_x, x) = \prod_{t=1}^{n} p(y_t \mid y_1, y_2, \cdots, y_{t-1}, E_x, x). \tag{16}$$

To make our emotional responses more human-like, we stitch the target emotion vector e_y into the vector c to control the emotion replied by decoder. The vector c becomes $[e_y ; v_x]$. Thus, we can choose different emotions to reply to the users, and even affect the user's emotion. For example, when the user is unhappy, we can make the user happy by outputting a response with a happy emotion.

3.6. Summary of the CVAE-DE Model. In this section, we introduce the mathematical derivation and structural framework of the model. The goal of our model is to generate dialogue responses that are rich in content, diverse in form, and rich in emotion. To improve our model's ability to understand emotion and improve the accuracy of emotion recognition, we use the BERT model as the emotion classifier. At the same time, to prevent the Seq2Seq model from generating a large number of general responses, we introduce the hidden variables of the conditional variational autoencoder to enable our model to generate rich and diverse responses. Finally, to make the emotion contained in the responses more natural and appropriate, we design a dual emotion framework that considers not only the controllability of the output emotion but also the continuity of the emotion with the input sentence.

4. Experiment

4.1. Data Preparation and Implementation Details. We use different data sets to train the multiemotion classifier and dialogue generation model. The multiemotion classifier is trained with the Weibo corpus data with emotion labels, which are derived from the Chinese Weibo emotion recognition task in NLPCC 2013 and the Chinese Weibo text emotion analysis task in NLPCC 2014. After sorting and filtering, the data set has a total of 40133 sentences, each of which contains an emotion label, which are divided into six categories: Null, Like, Sad, Disgust, Anger, and Happiness. The dialogue generation model is trained with the data set that is derived from the emotion dialogue generation task in NLPCC 2017. The data set contains 1119207 pieces of training data, each including an original sentence and a response sentence.

In the training of the multiemotion classifier, we divide the data set into a training set, a validation set, and a test set, with a ratio of 36133 : 2000 : 2000. We train the classifier on the basis of the pretrained Chinese model "BERT-Base, Chinese" released by Google.

In the training of the dialogue generation model, the ratio of the training set, validation set, and test set is 1099239 : 9984 : 9984. Our vocabulary size is set to 40000, the word embedding vector and the emotion label embedding vector are both set to 128, the encoder and decoder use 128 hidden units of RNN layer, and the latent variable size is set to 268. We randomly initialize all of the parameters of the model and set the batch size to 128.

4.2. Baselines. In the experiments, we compare CVAE-DE with the following baselines:

Seq2Seq: A standard Seq2Seq model with attention method that is widely used as a baseline in the conversation generation task [21].

ECM: A Seq2Seq model that uses the emotion category embeddings, internal and external memory mechanisms to generate emotional responses [14].

CVAE: A conditional variational autoencoder model that takes the target emotion label as input to formulate latent variable [16].

CVAE-MTDA: A conditional variational autoencoder model with a dual-attention mechanism used to ensure that specific emotional responses are coherent with the content and the emotion of the input [25].

EDGAN: A model based on generative adversarial networks with multiple generators for generating responses with specific emotion and a multiclass discriminator [27].

4.3. Evaluation Indicators. In this paper, we introduce the evaluation metrics for the following two aspects.

4.3.1. Multiemotion Classifier. Emotion classification accuracy is used as the evaluation index of the emotion classifier. For comparison, we train a variety of emotion classifiers, including RNN [30], LSTM [34], and Bi-LSTM [35].

4.3.2. Dialogue Generation Model. The evaluation indicators of the dialogue generation model are mainly divided into the categories of automatic evaluation and manual evaluation. Since there is no correct answer in the open-domain dialogue generation, the bilingual evaluation (BLUE) algorithm [36] is not suitable for the evaluation of the dialogue generation model [37]. Therefore, according to the perplexity [38], the accuracy of emotion expression, and the Distinct-1 and Distinct-2 methods [11], the responses generated by our model are automatically evaluated.

Table 1: Accuracy of emotion classifiers.

Model	Accuracy
RNN [30]	56.2%
LSTM [34]	59.7%
Bi-LSTM [35]	62.1%
BERT [18]	*65.1%*

Table 2: Objective evaluation with perplexity and accuracy.

Model	Perplexity	Accuracy
Seq2Seq [21]	67.2	0.205
ECM [14]	66.1	0.724
CVAE [16]	37.2	0.675
CVAE-MTDA [25]	34.6	0.692
EDGAN [27]	62.8	0.716
CVAE-DE	*33.5*	*0.749*

Table 3: Diversity scores for the CVAE-DE and the baselines. Distinct-1 and Distinct-2 are the ratios of distinct unigrams and bigrams in the generated responses.

Model	Distinct-1	Distinct-2
Seq2Seq [21]	0.0045	0.0353
ECM [14]	0.0062	0.0396
CVAE [16]	0.0256	0.2635
CVAE-MTDA [25]	0.0287	0.2712
EDGAN [27]	0.0273	0.2658
CVAE-DE	*0.0308*	*0.2836*
Target responses	0.0952	0.5897

Table 4: Manual evaluation of the generated responses in terms of content and emotion.

Model	Content	Emotion
Seq2Seq [21]	1.258	0.154
ECM [14]	1.286	0.422
CVAE [16]	1.302	0.392
CVAE-MTDA [25]	1.325	0.405
EDGAN [27]	1.314	0.416
CVAE-DE	*1.337*	*0.437*

Perplexity: Defined by Eq. (17) where S is the generated sentence, L is the length of the sentence, and $p(w_i)$ is the probability of the i-th word. A lower PPL score corresponds to a better model, more natural response, and smoother sentence.

$$PPL(S) = \sqrt[L]{\prod_{i=1}^{L} \frac{1}{p(w_i \mid w_1, w_2, \cdots, w_{i-1})}}. \tag{17}$$

Table 5: Sample responses generated by CVAE-DE and other baseline models. For each input sequence, a generated example response is listed for each emotion category.

Post	Model		Response (original)	Response (translated)
港澳游很好 玩啊!	Seq2Seq [21]		哈哈, 是啊, 我也想去。	Haha, yeah, I want to go too.
	ECM [14]	Like	嗯, 我喜欢旅游。	Well, I like traveling.
		Sad	唉, 可惜我去不了	Oh, it's a pity I cannot go.
		Disgust	我觉得太贵了。	I think it is too expensive.
		Angry	难道你们觉得好玩?	Do you guys find it fun?
		Happy	哈哈, 太好玩了。	Haha, it's so fun.
Traveling in Hong Kong and Macau is fun!	CVAE-DE	Like	不错哦, 我也很想去, 我的兴趣就是去各地旅游!	Not bad, I also want to go, my hobby is to travel everywhere!
		Sad	很郁闷, 我还要加班, 只能等下次去马尔代夫了。	Very depressed, I have to work overtime and can only wait for the next time I go to the Maldives.
		Disgust	这个旅游实在是很无聊的, 还需要港澳通行证。	This tour is really boring and requires a Hong Kong and Macau pass.
		Angry	你们都不告诉我, 难道我是空气?	You do not tell me, am I the air?
		Happy	哈哈, 收获很多, 太开心了。	Haha, I have gained a lot, so happy.
你换发型了?	Seq2Seq [21]		哈哈, 是啊, 我也觉得。	Haha, yeah, I think too.
	ECM [14]	Like	我喜欢这个发型。	I like this hairstyle.
		Sad	很郁闷, 我也觉得不好。	Very depressed, I also feel bad.
		Disgust	换发型很无聊啊!	Changing hair style is boring!
		Angry	我不喜欢理发。	I do not like haircuts.
		Happy	哈哈, 剪了短发。	Haha, I cut my hair short.
Have you changed your hair style?	CVAE-DE	Like	新发型是蛮爽的哦! 我就是喜欢这种风格!	The new hairstyle is pretty cool! I just like this style!
		Sad	太恐怖了, 我只是想换, 但是没有勇气。	It's horrible, I just want to change, but I do not have the courage.
		Disgust	无聊啊, 换了一个不喜欢的发型。	Bored, changed to a hairstyle that I do not like.
		Angry	难道你不喜欢我的发型?	Do not you like my hairstyle?
		Happy	哈哈, 这个发型, 你会爱上我的。	Haha, with this hairstyle, you will fall in love with me.
上海下雨了吗?	Seq2Seq [21]		是啊, 下雨了。	Yes, it's raining
	ECM [14]	Like	我喜欢这种天气。	I like this weather.
		Sad	估计会被淋湿的。	It is estimated that it will get wet.
		Disgust	下雨太难受了。	The rain is too uncomfortable.
		Angry	我还不能出门。	I cannot go out yet.
		Happy	哈哈，已经凉快了。	Haha, it's already cool.
Is it raining in Shanghai?	CVAE-DE	Like	在上海, 天气预报不错, 下雨很凉快耶~	In Shanghai, the weather forecast is good, and the rain is very cool ~
		Sad	可惜最近完全没有下雨, 是一个干旱的季节。	It is a pity that it has not rained at all recently, it is a dry season.
		Disgust	是的, 上海的天气太热了!	Yes, the weather in Shanghai is too hot!
		Angry	没啊!我说过不会的, 你们不相信。	No! I said no, you do not believe it.
		Happy	哈哈哈, 估计大暴雨就要来了。	Hahaha, it is estimated that the heavy rain is coming.

Distinct-1 and Distinct-2: Used to judge whether the model will generate a large number of universal and repetitive responses, which can reflect the diversity of responses. The definition is given in Equation (18) where Count(unique ngram) is the number of unigrams/bigrams that are not repeated in the responses and Count(word) is the total number of unigrams/bigrams in the responses. A larger value of Distinct-1 and Distinct-2 indicates a higher diversity of the generated responses.

$$\mathrm{Distinct}(n)=\frac{\mathrm{Count}(\text{unique ngram})}{\mathrm{Count}(\text{word})}. \tag{18}$$

Manual Evaluation: To better understand the quality of the generated response in terms of content and emotion, we invite 4 volunteers to evaluate the results of our generation models. The reviewer scores of the generated response are based on content and emotion. The content scores are mainly based on whether the response is appropriate and natural or whether it may be generated by people; it is a widely accepted measurement standard by researchers and was proposed by Shang et al. [21]. The emotion scores are mainly based on whether the emotion of response meets the given target emotion. The content scores are divided into 0 point, 1 point, and 2 points. The emotion scores are divided into 0 point and 1 point.

4.4. Experimental Results and Analysis. (1) Classification Accuracy of the Multiemotion Classifier: As shown in Table 1, the classifier based on the BERT model has the highest accuracy, reaching 65.1%. The higher the accuracy of the emotion classifier is, the more accurate the emotion label is, and the higher the accuracy of the emotion expression. Therefore, we will use a classifier based on the BERT model to generate the emotion labels.

(2) Perplexity and Accuracy of Emotion Expression: As shown in Table 2, CVAE-DE obtains better score than all of the other models in perplexity and emotion expression accuracy. The best score in emotion expression accuracy indicates that the dual emotion framework can generate a response that is closer to the emotional response in the real human conversation corpus than the other models. The emotional responses of CVAE-DE model are not only controlled by the target emotion but also affected by the emotion of the input sentence. As communicated in real life, the responding party is not only controlled by their own emotion but also affected by the emotion expressed by the other party. The emotion accuracy of Seq2Seq is quite low because it generates the same response with different emotion types.

(3) Distinct-1 and Distinct-2: It is observed from Table 3 that the CVAE-DE model is far superior to the pure Seq2Seq model in response diversity. The CVAE-based models can enjoy the superiority of stochastic sampling from probabilistic latent variable, enabling them to generate various and meaningful responses, while pure Seq2Seq models tend to generate monotonous responses.

(4) Manual Evaluation: From the results shown in Table 4, CVAE-DE outperforms other models in content and emotion. This result indicates that CVAE-DE can generate high-quality emotional responses without sacrificing the grammatical correctness and logic of the content.

(5) Case Study: In Table 5, we show some example responses generated by CVAE-DE and other baselines. For a given post, there are a variety of proper responses with different emotion types. Intuitively, different people generate different emotion types for the same post.

It is observed that the CVAE-DE model generates emotional responses on every emotion type, while Seq2Seq with attention chooses a random emotion type for response. Compared with ECM, our model can produce responses with richer content, more diverse forms, and greater emotional accuracy. When fed with different target emotions, the CVAE-DE model uses different emotional words to control expressions.

Although our model has achieved a relatively satisfactory performance compared to that of other models, there are still some limitations. Our model is mainly limited to some coarse-grained emotional labels, including like, sadness, and anger. Such coarse-grained classification labels make it difficult to capture the nuances of human emotion. Therefore, our future work direction may be to train our model to make it easier to capture the nuances of human emotions by building a corpus with fine-grained emotional labels.

5. Conclusion

In this paper, we propose an emotional dialogue generation model, CVAE-DE, to produce high-quality responses with multiple emotion types. An emotion classifier based on the BERT model is used to classify a variety of emotions, which to a certain extent improves the problem of previous methods obtaining a low classification accuracy of emotion categories. To enable the model to produce more rich and diverse responses, we introduce a conditional variable autoencoder on the basis of the Seq2Seq model based on the attention mechanism. At the same time, to enable the model to generate coherent and controllable emotional responses, we propose a dual-emotional framework. The experimental results show that the model proposed in this paper can produce high-quality responses with specific emotions.

In future work, we will use more complex generation models to further improve the quality of generated responses and use a corpus with fine-grained emotion classification labels to enrich the emotion of responses. At the same time, we will also explore the application of the method in this article to multiple rounds of dialogue, using contextual information to infer the user's emotional information, rather than the emotional information specified by the user. This will be a challenging task because it depends on the topic, contextual information, and the user's emotions.

Data Availability

We use different data sets to train the multiemotion classifier and dialogue generation model. The multiemotion classifier is trained with the Weibo corpus data with emotion labels, which are derived from the Chinese Weibo emotion recognition task in NLPCC 2013 and the Chinese Weibo text emotion analysis task in NLPCC 2014. After sorting and filtering, the data set has a total of 40133 sentences, each of which contains an emotion label, which are divided into six categories: Null, Like, Sad, Disgust, Anger, and Happiness. The dialogue generation model is trained with the data set that is derived from the emotion dialogue generation task in NLPCC 2017. The data set contains 1119207 pieces of training data, each including an original sentence and a response sentence. All researchers can access the data at the following site: https://www.biendata.xyz/ccf_tcci2018/datasets.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Nos. 6202780103, 61772149, 61762028, and U1701267), Guangxi Science and Technology Project (Nos. AB20238013, ZY20198016, 2019GXNSFFA245014), and Guangxi Key Laboratory of Image and Graphic Intelligent Processing Project (No. GIIP2003).

References

[1] Z. Cai, Z. He, X. Guan, and Y. Li, "Collective data-sanitization for preventing sensitive information inference attacks in social networks," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 4, pp. 577–590, 2018.

[2] Y. Lin, Z. Cai, X. Wang, and F. Hao, "Incentive mechanisms for crowdblocking rumors in mobile social networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 9, pp. 9220–9232, 2019.

[3] Y. Lin, X. Wang, F. Hao et al., "Dynamic control of fraud information spreading in mobile social networks," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, pp. 1–14, 2019.

[4] Z. Cai and X. Zheng, "A private and efficient mechanism for data uploading in smart cybe physical systems," *IEEE Transactions on Network Science and Engineering*, vol. 7, no. 2, pp. 766–775, 2020.

[5] P. Salovey and J. Mayer, "What is emotional intelligence? Emotional development and emotional intelligence: implications for educators," *New York: Basic Books. Senge, PM (1998). Sharing Knowledge. Executive Excellence*, vol. 15, no. 6, pp. 11-12, 1997.

[6] R. Lan, L. Sun, Z. Liu, H. Lu, C. Pang, and X. Luo, "Madnet: a fast and lightweight network for single-image super resolution," *IEEE Transactions on Cybernetics*, pp. 1–11, 2020.

[7] B. Li, R. Liu, J. Cao, J. Zhang, Y.-K. Lai, and X. Liu, "Online low-rank representation learning for joint multi-subspace recovery and clustering," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 335–348, 2017.

[8] Y. Wang, Y. Gao, Y. Li, and X. Tong, "A worker-selection incentive mechanism for optimizing platform-centric mobile crowdsourcing systems," *Computer Networks*, vol. 171, pp. 107–144, 2020.

[9] R. Lan, Y. Zhou, Z. Liu, and X. Luo, "Prior knowledge-based probabilistic collaborative representation for visual recognition," *IEEE Transactions on Cybernetics*, vol. 50, no. 4, pp. 1498–1508, 2020.

[10] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of the 27th International Conference on Neural Information Processing Systems*, vol. 2, pp. 3104–3112, Cambridge, MA, United States, 2014.

[11] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 110–119, San Diego, California, 2016.

[12] C. Xing, W. Wu, Y. Wu et al., "Topic aware neural response generation," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 3351–3357, San Francisco, California, USA, 2017.

[13] T. Zhao, R. Zhao, and M. Eskenazi, "Learning discourse-level diversity for neural dialog models using conditional variational autoencoders," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 654–664, Vancouver, Canada, 2017.

[14] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: emotional conversation generation with internal and external memory," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 730–739, New Orleans, Louisiana, USA, 2018.

[15] N. Asghar, P. Poupart, J. Hoey, X. Jiang, and L. Mou, "Affective neural response generation," in *European Conference on Information Retrieval*, pp. 154–166, Springer, 2018.

[16] 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. 1128–1137, Melbourne, Australia, 2018.

[17] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," 2015, https://arxiv.org/abs/1511.05644.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, 2019.

[19] O. Vinyals and Q. Le, *A neural conversational model*, ICML Deep Learning Workshop, 2015.

[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR*, San Diego, USA, 2015.

[21] L. Shang, Z. Lu, and H. Li, "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)*, pp. 1577–1586, Beijing, China, 2015.

[22] K. Cao and S. Clark, "Latent variable dialogue models and their diversity," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 182–187, Valencia, Spain, 2017.

[23] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 3776–3783, Phoenix, Arizona, USA, 2016.

[24] X. Sun, X. Peng, and S. Ding, "Emotional human-machine conversation generation based on long short-term memory," *Cognitive Computation*, vol. 10, no. 3, pp. 389–397, 2018.

[25] W. Xu, X. Gu, and G. Chen, "Generating emotional controllable response based on multi-task and dual attention framework," *IEEE Access*, vol. 7, pp. 93734–93741, 2019.

[26] Z. Song, X. Zheng, L. Liu, M. Xu, and X.-J. 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, 2019.

[27] Y. Li and B. Wu, “Emotional dialogue generation with generative adversarial networks,” in *2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC)*, vol. 1, pp. 868–873, Chongqing, China, 2020.

[28] Y. Su, D. Cai, Y. Wang et al., “Stylistic dialogue generation via information-guided reinforcement learning strategy,” 2020, https://arxiv.org/abs/2004.02202.

[29] A. Vaswani, N. Shazeer, N. Parmar et al., “Attention is all you need,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010, Red Hook, NY, United States, 2017.

[30] T. Mikolov, M. Karafiát, and L. Burget, “Recurrent neural network based language model,” in *Eleventh Annual Conference of the International Speech Communication Association*, pp. 1045–1048, Makuhari, Chiba, Japan, 2010.

[31] K. Cho, B. van Merriënboer, C. Gulcehre et al., “Learning phrase representations using rnn encoder–decoder for statistical machine translation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, Doha, Qatar, 2014.

[32] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” in *International Conference on Learning Representations*, Banff, Canada, 2014.

[33] S. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio, “Generating sentences from a continuous space,” in *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 10–21, Berlin, Germany, 2016.

[34] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[35] A. Graves, S. Fernández, and J. Schmidhuber, “Bidirectional lstm networks for improved phoneme classification and recognition,” in *International Conference on Artificial Neural Networks*, pp. 799–804, Springer, 2005.

[36] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, USA, 2002.

[37] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau, “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*, pp. 2122–2132, Austin, Texas, 2016.

[38] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, “A neural probabilistic language model,” *Journal of Machine Learning Research*, vol. 3, pp. 1137–1155, 2003.