Multi-label Emotion Detection via Emotion-Specified Feature Extraction and Emotion Correlation Learning

Jiawen Deng[1], Fuji Ren[2]

Abstract—Textual emotion detection is an attractive task while previous studies mainly focused on polarity or single-emotion classification. However, human expressions are complex, and multiple emotions often co-occur with non-negligible emotion correlations. In this paper, a Multi-label Emotion Detection Architecture (MEDA) is proposed to detect all associated emotions expressed in a given piece of text. MEDA is mainly composed of two modules: Multi-Channel Emotion-Specified Feature Extractor (MC-ESFE) and Emotion Correlation Learner (ECorL). MEDA captures underlying emotion-specified features through MC-ESFE module, which is composed of multiple channel-wise ESFE networks. Each channel in MC-ESFE is devoted to the feature extraction of a specified emotion from sentence-level to context-level through a hierarchical structure. With underlying features, emotion correlation learning is implemented through an emotion sequence predictor in ECorL. Furthermore, we define a new loss function: multi-label focal loss. With this loss function, the model can focus more on misclassified positive-negative emotion pairs and improve the overall performance by balancing the prediction of positive and negative emotions. The evaluation of proposed MEDA architecture is carried out on emotional corpus: RenCECps and NLPCC2018 datasets. The experimental results indicate that the proposed method can achieve better performance than state-of-the-art methods in this task.

Index Terms—Multi-label, Emotion Detection, Emotion Correlation, Multi-label Focal Loss.

1 Introduction

As the rapid development of social media platforms, such as microblogs and twitter, it is convenient for users to share their attitudes about any topic. Essentially, understanding the latent emotions expressed in such user-generated content has gained much attention because of its vast potential applications [1], [2], such as emotional chatbots [3], emotional text-to-speech synthesizer [4], and patient emotion monitoring [5].

As a fundamental task in sentiment analysis, emotion detection has been deeply studied in the literature. Not only the basic emotional polarity classification [6], but emotion detection has also delved into more granular analysis [7], [8], such as love, hate, angry, and surprise. While many such kinds of researches have been implemented, most of them are conducted in the single-emotion environment [9]. They are based on the assumption that certain textual data is associated with only one emotion. However, in real-world conditions, people often hold multiple complex emotions simultaneously, and a textual expression is often associated with multiple emotions simultaneously. Therefore, multi-label emotion detection has gained burgeoning attention because of its vast potential applications.

Multi-label emotion detection task aims to recognize all possible emotions in a piece of textual expression [10]. In conventional emotion detection networks, textual information is often encoded together into a representation vector and then directly fed into the classifier[11], [12]. However, in a textual expression with multiple emotions, there may be some emotions with relatively weaker intensity. If information of each emotion is mixed and encoded together into a shared vector, the weaker emotions with subtle features could be covered by stronger emotions and be challenging to recognize. To accurately recognize the emotions expressed, the quality of underlying emotional feature representation has an important influence on the final prediction.

In most previous researches, multi-label emotion detection task is often narrowed down into multiple binary classifications [13], in which each emotion is detected respectively without considering their correlations. However, emotion correlation information provides non-ignorable features and is useful for improving the performance of emotion detection. The definition of emotion correlation can be illustrated based on Plutchik's work. In an emotional expression, emotion correlation mainly refers to positive or negative emotional correlation. Positively correlated emotions are similar to each other and often appearing together but with different intensities. Such as the emotion pair 'Joy' and 'Love' tend to appear simultaneously. Negatively associated emotions are often opposite to each other and rarely appear together, such as 'Love' and 'Sorrow'. Emotion correlation can be utilized to facilitate more in-depth emotion analysis in multi-label emotion recognition task.

In this paper, a Multi-label Emotion Detection Architecture (MEDA) is proposed to address the above challenges. MEDA is mainly composed of two modules: Multi-Channel Emotion-Specified Feature Extractor (MC-ESFE) and Emotion Correlation Learner (ECorL). MC-ESFE consists of multiple channels by which the features of each emotion are separately

- *Jiawen Deng is with the Institute of Technology and Science, Tokushima University, Tokushima, Japan. E-mail: c501847002@tokushima-u.ac.jp.*
- *Fuji Ren is with the Institute of Technology and Science, Tokushima University, Tokushima, Japan. E-mail: ren@is.tokushima-u.ac.jp.*

encoded. Each channel is devoted to the underlying feature representation of a specified emotion from both sentence-level and context-level. Furthermore, an external emotion lexicon is introduced as prior knowledge to integrate more detailed emotional information. ECorL module is devoted to learning emotion correlation based on extracted emotion-specified features from MC-ESFE. In ECorL, multi-label emotion detection task is transformed as an emotion sequence prediction task. Bidirectional GRU network is taken as the emotion sequence predictor, and the emotions are sequentially predicted in a fixed path. In the hidden state of each step, the emotion correlations of current emotion are learned by information interaction with the context of other emotions flowed from both forward and backward directions. Considering that the proposed MEDA network extracts emotional information from sentence level, context level, and emotion correlation level, an ensemble model called MEDA-FS is proposed to integrate emotional information from different levels. MEDA-FS can realize the maximization of information retention and avoid information loss during bottom-up learning. During the training, positive-negative emotion correlation is incorporated into the proposed multi-label focal loss function. By introducing a weighting factor, our loss will focus more on misclassified emotion pairs and balance the prediction between positive and negative emotions.

Compared with existing multi-label emotion detection methods, the proposed MEDA architecture extracts both emotion-specified features and emotion correlations. The performance of the proposed MEDA is verified in Chinese emotional corpus: RenCECps. Experimental results show that the proposed architecture achieves state-of-the-art performance on RenCECps and demonstrates the effectiveness of MEDA.

The major contributions of our paper can be summarized as follows.

1. MEDA architecture composed of MC-ESFE and ECorL modules is proposed for the textual multi-label emotion detection task. MC-ESFE can encode emotion-specified features in the corresponding channel respectively, which strengthens the underlying feature representation of each emotion. ECorL is proposed to learn emotion correlations by transforming multi-label emotion detection task into emotion sequence prediction task.
2. MEDA-FS is proposed to fuse the information at sentence-level, context-level, and emotion correlation level, which can realize the maximization of information retention during bottom-up learning.
3. Multi-label focal loss function considering emotion correlation information is proposed for multi-label learning. This loss function contributes to model training by focusing on misclassified emotion pairs and balancing the prediction of positive and negative emotions.

The rest of this paper is organized as follows: Section 2 presents a brief overview of related work on multi-label emotion detection task. Details of the proposed MEDA are given in Section 3. The experimental setting and details are presented in Section 4. The performance of MEDA is discussed in Section 5. Finally, conclusions are drawn in Section 6.

2 Related Work

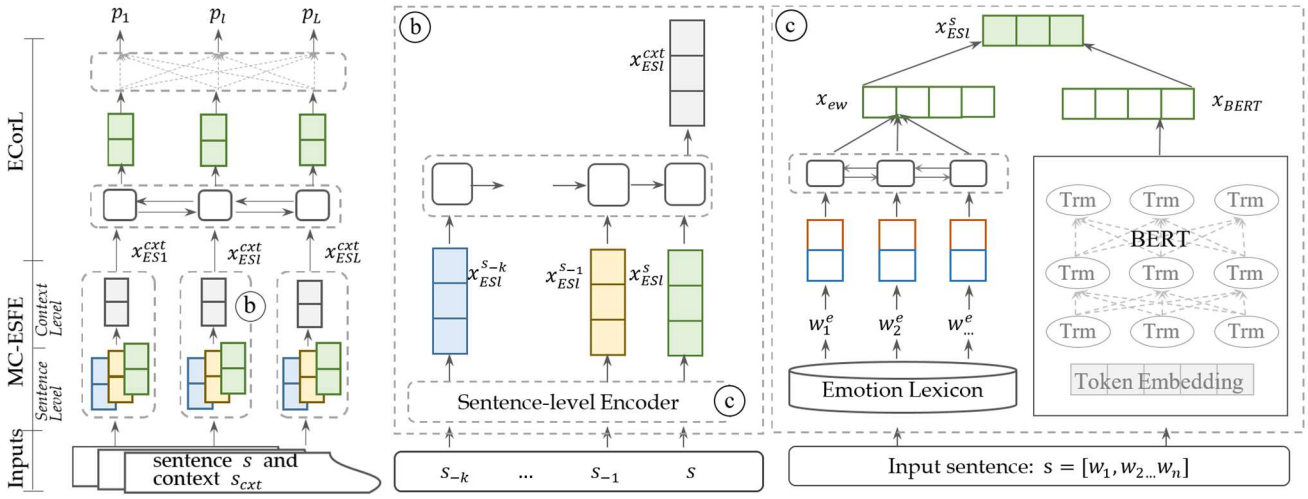
As an important task in natural language processing, textual sentiment analysis is an emerging research field. Traditionally, sentiment analysis is applied to predict a polarity label or star level of product reviews [14], [15], or film reviews [16]. With the delving of related researches, emotion classification is becoming more granular [17]. There are many different taxonomies [18], such as Paul Ekman's six basic emotions [19] and Fuji Ren's eight basic emotions [20].

As one of the most obvious clues of sentiment analysis, emotional lexical resources directly encode sentimental knowledge [1], [21], and are widely used, such as WordNet-affect [22], NRC emotion lexicon [23], and Hownet [24]. These emotional lexicons are the basis for early hand-crafted features based emotional analysis [21], [25]. Given the recent success of deep learning models, various neural network models have been proposed and have achieved highly competitive performance in sentiment analysis. LSTM networks [26] have shown its superior performance in context information encoding [27], [28], and CNN networks [29], [30] are often utilized to extract local information within a sentence. Multi-task ensemble network performs well in emotional information integration and can alleviate the impact of insufficient corpus[31], [32]. To achieve robust emotion feature representation, Emo2Vec is trained in [33] to encode emotional semantics by multi-task learning six different emotion-related tasks. Different modalities, such as text, emoji, and images, can also be combined [34] to express emotion and complement each other for emotion classification. To further model the effects of sentimental relations, a modified GRNN is proposed [6] to encode the sentiment polarity and sentiment modifier context separately.

In most previous studies, the complexity of emotion detection task is often narrowed down by focusing on single emotion classification. However, human emotion is complex in reality, and the textual expression often contains multiple emotions simultaneously. To address this problem, the multi-label emotion detection task can be viewed as a special multi-label classification [10], [35], in which emotions are the multiple labels. Attention-based multi-label sentence classifier is proposed in [36] to imitate how humans comprehend and classify emotions. A dual attention-based transfer learning model is proposed in [37] to extract both general sentiment words and other emotion-specific words. Linguistic characteristics are explored in [12] to reduce the limitation of lexicon coverage and size.

However, most of the above models do not take multi-label correlations into account, and they assume that the multiple labels are independent of each other. Some studies try to explore the label correlations. To reduce the effect of irrelevant labels, prior knowledge of co-occur label relationships are incorporated [38] as a constraint for emotion prediction and ranking. Some approaches attempt to implicitly estimate label correlations by the modification of loss function. The label-correlation sensitive loss function is first proposed in [39] with the BP-MLL algorithm. Joint binary cross-entropy (JBCE) loss is proposed by Huihui He [14] in his joint binary neural network (JBNN) to capture label relations. Multi-label classification can also be transformed into a sequence generation problem [40], [41] to capture label correlations. To reduce the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3034215, IEEE Transactions on Affective Computing



(a) MEDA (b) lth ESFE channel of MC-ESFE (c) Sentence-level encoder of lth ESFE channel

Fig. 1. (a) The illustration of the proposed MEDA: Multi-label Emotion Detection Architecture. MEDA mainly composes two modules: Multi-Channel Emotion-Specified Feature Extractor (MC-ESFE) and Emotion Correlation Learner (ECorL). (b) The illustration of lth ESFE channel of emotion e_l in MC-ESFE module. (c) The illustration of sentence-level encoder of lth ESFE channel in MC-ESFE module.

computational complexity, partial label dependence can also contribute to this task, which is demonstrated in [42]. Deep canonical correlation analysis (DCCA) performs well in feature-aware label embedding and label-correlation aware prediction [43], [44]. A semisupervised multi-label method is proposed in [45] while label correlations are incorporated by modifying the loss function. Multi-Label classification and label correlations learning can also be realized in a joint learning framework [46], [47].

In contrast with most current methods, we focus on emotion-specified feature extraction and emotion correlation. There are mainly two fundamental differences:

1. The information of each emotion is encoded separately, which can concentrate more on underlying emotion-specified feature extraction. This implementation contributes to further emotion correlation learning in emotion sequence predictor.
2. The proposed multi-label focal loss function considers emotion correlation information. It can pay more attention to misclassified emotion pairs and balance the prediction of positive and negative emotions.

3 Proposed Method

To comprehensively obtain emotional information of texts, Multi-label Emotion Detection Architecture (MEDA) is proposed in this paper. It mainly composes two modules: Multi-Channel Emotion-Specified Feature Extractor (MC-ESFE) and Emotion Correlation Learner (ECorL). The framework of MEDA is shown as Fig. 1.

Multi-label emotion detection task aims to detect all possible emotions from the pre-defined emotional label set: $E = [e_1, e_2, \dots e_L]$. Considering the important influence of contextual information on this task, the previous k sentences occurred before current sentence s are taken as the context sentences: $s_{cxt} = [s_{-k}, \dots s_{-2}, s_{-1}]$. Given a sentence $s = [w_1, w_{2\dots}w_n]$ and its context s_{cxt}, our proposed multi-label emotion recognition model MEDA is trained to output the predicted probability distribution $P_{ML} = [p_1, p_2, \dots p_L]$ of each emotion, denoted as:

$$P_{ML} = f_{MEDA}(s, s_{cxt}) \tag{1}$$

MC-ESFE module is composed of L parallel channel-wise ESFE. In each channel, ESFE extracts emotion-specified features from sentence-level to context-level through a hierarchical structure. The output of each channel is combined into an emotion-specified feature matrix: $X_{ES}^{cxt} = [x_{ES1}^{cxt}, x_{ES2}^{cxt}, \dots, x_{ESL}^{cxt}]$. In ECorL module, emotion correlations are further learned from X_{ES}^{cxt} and multi-label emotions are predicted. Specifically, MEDA architecture is very flexible, and the algorithm applied in each module can be replaced by other state-of-the-art algorithms.

3.1 MC_ESFE: Multi-Channel Emotion Specified Feature Extractor

In this paper, a Multi-Channel Emotion-Specified Feature Extractor (MC-ESFE) is proposed for underlying fundamental feature extraction. MC-ESFE is composed of L channel-wise ESFE, and L is equal to the number of emotions. Each channel focuses on the feature extraction of a specified emotion, and each emotion's information is separately encoded in each channel. In this way, more details of each emotion could be summarized, and the features of weak emotions are prevented from being covered by strong emotions to some extent.

Fig. 1 shows the hierarchical structure of lth ESFE-channel corresponding to emotion e_l, $l \in [1, L]$. Each channel contains a sentence-level encoder and a context-level encoder, which focus on feature extraction of emotion e_l on both sentence-level and context-level.

3.1.1 Sentence Level Encoder

In lth ESFE channel, given a sentence s, the sentence-level encoder f_{S-En}^{l} projects input sentence s to emotion-specified feature x_{ESl}^{s}:

$$x_{ESl}^{s} = f_{S-En}^{l}(s),\ l \in [1, L] \tag{2}$$

In sentence-level encoder f^{l}_{S-En}, as shown in Fig. 1 (c), two parallel architectures with different embedding methods are employed to generate: (1) emotional feature representation x_{ew}, (2) general sentence representation x_{BERT}. They are integrated into emotion-specified sentence-level representation x^{s}_{ESl} for further context-level learning.

General sentence representation. Inspired by the pre-trained language model learning approach and transfer learning techniques, pre-trained Chinese BERT model [48] is applied to yield general sentence representation in this paper. BERT stands for Bidirectional Encoder Representations from Transformers. Chinese BERT is designed to pre-train deep bidirectional representations from unlabeled Chinese text by jointly conditioning on both left and right context in all layers. It remedies the limitation of insufficient training corpora and contributes to syntactic and semantic sentence representation. Given a sentence s, the general sentence representation x_{BERT} is generated from Chinese BERT.

Emotional feature representation. Arguably, it is accepted that general sentence representation generated by pre-trained language model does not contain specific emotional features, as no emotion-related knowledge has been included in the training process. To generate emotional sentence representation, emotional features are further extracted based on an external n-dimensional emotion lexicon.

With the input sentence $s = [w_1, w_2, \dots]$, emotional words $w^{\text{e}} = [w^{e}_1, w^{e}_2, \dots]$ occurred in s are firstly extracted by matching the emotion lexicon. The embedding of emotional words consists of two parts. The first is general word embedding, which is realized by mapping the pre-trained Word2vec word embedding matrix. Each word is embedded as $v_{\text{w2v}} \in R^{1*D}$, in which D is the embedding dimension. The second is emotional word embedding, which is realized based on n-dimensional emotion lexicon. Each word is embedded as $v_{\text{emo}} \in R^{1*n}$, in which n is the number of emotions annotated in emotion lexicon and the value means the intensity of corresponding emotion. Finally, emotional word embedding is represented as $E = [v^{e}_1, v^{e}_2, \dots]$, in which $v^{e}_i \in R^{1*(D+n)}$.

Considered the polysemy of emotional words in different contexts, BiGRU (Bidirectional Gated Recurrent Neural Networks) [49] and attention network [50] are utilized to make the network pay more attention to significant emotional words. Take emotional embedding E as input, the output of the hidden state of BiGRU in each step is $h_i = [\overrightarrow{h_i}: \overleftarrow{h_i}]$, in which $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ are the output of hidden states from forward and backward directions, respectively. The attention mechanism considers the contributions of different emotional words to the prediction of specified-emotion e_l. More attention weight will be assigned to words related to emotion e_l in the current lth ESFE channel. Attention weight a_i and weighted emotional feature vector x_{ew} are defined as follows:

$$e_i = W_2^T[\sigma(W_1^T \cdot h_i + b_1)] + b_2 \tag{3}$$

$$a_i = \frac{exp(e_i)}{\sum_{k=1}^{n} exp(e_k)} \tag{4}$$

$$x_{ew} = [a_1 h_1 : a_2 h_2 : \dots a_i h_i : \dots] \tag{5}$$

in which σ indicates the sigmoid activation function, w_1, b_1, w_2, b_2 indicate the model parameters, and [:] indicates the concatenation operation.

Finally, emotional feature vector x_{ew} and general embedding x_{BERT} is integrated, and emotion-specified sentence-level representation x^{s}_{ESl} is generated as follows:

$$x^{s}_{ESl} = tanh(W_{\text{e}} \cdot x_{\text{ew}} + W_B \cdot x_{BERT} + b_s) \tag{6}$$

in which W_{e}, W_B and b_s indicate the model parameters.

3.1.2 Context Level Encoder

Context level encoding is channel-wise implemented as well. As shown in Fig. 1 (b), in lth ESFE channel corresponding to emotion e_l, given a sentence s and its context $s_{cxt} = [s_{-k}, \dots\ s_{-2}, s_{-1}]$, context-level encoder f^{l}_{C-En} gives contextual emotional feature x^{cxt}_{ESl}. GRU network is utilized to learn contextual information from previous k sentences, and the output of final step is captured as the context-level representation, which is denoted as:

$$\begin{aligned} x^{cxt}_{ESl} &= f^{l}_{C-En}\left(x^{s_{-k}}_{ESl}, \dots, x^{s_{-1}}_{ESl}, x^{s}_{ESl}\right), l \in [1, L] \\ &= f_{\text{GRU}}(x^{s_{-k}}_{ESl}, \dots, x^{s_{-1}}_{ESl}, x^{s}_{ESl}) \end{aligned} \tag{7}$$

$$x^{s_{-i}}_{ESl} = f^{l}_{S-En}(s_{-i}), i \in [1, k] \tag{8}$$

Contextual emotion-specified features x^{cxt}_{ESl} learned from each channel in MC-ESFE is output and combined as the emotional feature matrix: $X^{cxt}_{ES} = [x^{cxt}_{ES1}, x^{cxt}_{ES2}, \dots, x^{cxt}_{ESL}]$. X^{cxt}_{ES} is flowed into ECorL module for further emotion correlation learning.

3.2 ECorL: Emotion Correlation Learner

Emotion correlations are indispensable in multi-label emotion detection task. In this paper, ECorL (Emotion Correlation Learner) is proposed to give emotion prediction based on emotion correlation learning.

MC-ESFE module project inputs into a sequence of continuous emotional representations $X^{cxt}_{ES} = [x^{cxt}_{ES1}, x^{cxt}_{ES2}, \dots]$. ECorL module takes X^{cxt}_{ES} as input. In ECorL module, multi-label emotion detection task is transformed as emotion sequence prediction task, and emotions are predicted in a fixed path. Refer to the previous work [40], the order of emotion sequence is set according to its occurred cumulative number in the corpus. BiGRU is taken as the emotional sequence predictor. The operation is formulated as follows:

$$H_e = f_{\text{BiGRU}}(x^{cxt}_{ES1}, x^{cxt}_{ES2}, \dots, x^{cxt}_{ESL}) \tag{9}$$

$$P_{ML} = tanh(W_{ECor} \cdot H_e + b_{ECor}) \tag{10}$$

in which $H_e = [h_{e1}, \dots h_{el}, \dots h_{eL}]$, are the hidden states of each step, W_{ECor} and b_{ECor} are the learned weight and biases, and P_{ML} is the predicted probability of each emotion. In lth step of BiGRU, the learning of hidden state h_{el} can be viewed as the feature extraction of a specified emotion e_l. Invalid information of current input x^{cxt}_{ESl} can be filtered because of the gating mechanism. With the bidirectional network, emotional feature h_{el} is learned based on the information of other emotions flowed from both forward and backward hidden state. In this way, emotional information interaction is realized. The hidden states of BiGRU are output and fed into emotion interaction layer. This layer is a fully-connected layer and aimed to realize further emotional information interaction. In this way, the final emotion prediction is obtained: $P_{ML} = [p_1, p_2, .. p_L]$.

3.3 Network Pre-training in MC-ESFE

Each channel in MC-ESFE is dedicated to obtaining corresponding emotional information, which belongs to the underlying feature extraction in the MEDA framework. The quality of feature representation has a direct impact on the performance of upper-level emotion predictions. To improve the underlying feature representation, network-based transfer learning is employed to pre-train the sentence-level encoder in each channel. During transfer learning, a prediction layer is added to emotional sentence representation x^s_{Esl} for single-emotion prediction:

$$p^s_{Esl} = \sigma(w_l \cdot x^s_{Esl} + b_l). \tag{11}$$

in which x^s_{Esl} is the sentence-level representation of input sentence s, and p^s_{Esl} indicates the predicted probability of emotion e_l expressed in sentence s. The first-step pre-training is implemented on positive-negative annotated emotional datasets. The second-step is fine-tuning. During fine-tuning, the multi-emotion annotation $\{s, [y_1, y_2, \dots y_L]\}$ of each sentence s in dataset D is transformed into multiple single-emotion annotations: $\{s, y_1\}$, $\{s, y_2\} \dots \{s, y_L\}$. In this way, we reconstructed multiple binary-dataset: $\widehat{D} = \{\widehat{D}_1, \widehat{D}_2, \dots \widehat{D}_L\}$. For each binary-dataset $\widehat{D}_l$, $l \in [1, \mathrm{L}]$, sentence $s \in \widehat{D}_l$ is annotated as $\{s, y_l\}$. $\widehat{D}_l$ is fed into lth ESFE channel to fine-tune the sentence-level parameters. During pre-training, binary focal loss [51] is utilized:

$$E_{FL} = -\alpha_t (1 - p_t)^r log(p_t) \tag{12}$$

$$p_t = \begin{cases} p^s_{Esl} & if\ y_l = 1 \\ 1 - p^s_{Esl} & otherwise \end{cases} \tag{13}$$

in which r is a modulating factor, and it aimed to reduce the relative loss of well-classified examples. $\alpha \in [0,1]$, is a weighting factor to address the problem of class imbalance. $\alpha_t = \alpha$ for positive label and $\alpha_t = 1 - \alpha$ for negative label.

3.4 MEDA-FS: Multi-level Information Fusion

The proposed MEDA architectural learns emotional information from sentence-level to context-level, from single-emotion level in MC-ESFE to multi-emotion level in ECorL module. Each layer in MEDA network learns different levels of information. To realize the maximization of information retention and avoid information loss during bottom-up learning, MEDA-FS is proposed to fuse the information from different levels. MEDA-FS consists of three sub-models: S-MC-ESFE, C-MC-ESFE, and MEDA, which give emotion predictions on sentence-level, context-level, and emotion correlation level, respectively.

S-MC-ESFE, gives sentence-level predictions $P^s_{ES} = [p^s_{ES1}, \dots, p^s_{ESL}]$. It is obtained during the pre-training step of sentence-level encoder in MC-ESFE, which is detailed in section 3.3. P^s_{ES} represents the prediction based on the underlying information, without considering the emotion correlations and contextual information.

C-MC-ESFE, gives context-level predictions P^{cxt}_{ES} based on sentence-level predictions of current sentence s, denoted as P^s_{ES}, and sentence-level predictions of its context s_{cxt}, denoted as $[P^{s-k}_{ES}, \dots, P^{s-1}_{ES}]$. GRU network is utilized to learn contextual information and its final output is taken as the prediction:

$$P^{cxt}_{ES} = f_{\mathrm{GRU}}([P^{s-k}_{ES}, \dots P^{s-1}_{ES}, P^s_{ES}]) \tag{14}$$

MEDA, gives prediction $P_{ML} = [p_1, p_2, .. p_L]$ by considering both contextual information and emotion correlation.

MEDA-FS, gives final predictions by comprehensively fuse the information from above three level, denoted as:

$$P = w_s \cdot P^s_{ES} + w_c \cdot P^{cxt}_{ES} + w_{ML} \cdot P_{ML} \tag{15}$$

in which w_s, w_c and w_{ML} are the weight parameters of each level's information.

3.5 Definition of Multi-label Focal Loss

Multi-label (ML) loss function [52] is one of the most commonly used loss functions in multi-label learning. Instead of concentrating on individual label discrimination like traditional cross-entropy loss function, ML-loss focused on considering the correlations between the different labels. Inspired by [51], we rewrite ML-loss and called it multi-label focal loss. Multi-label focal loss not only considers emotion correlation but also focus more on misclassified emotion pairs. Besides, it introduces a harmonic parameter to reduce the influence of the imbalance prediction of positive and negative emotions. The definition of multi-label focal loss is defined as follows:

$$E_{ML-FL} = \sum_{i=1}^{N} \frac{1}{|Y_i||\overline{Y_i}|} \sum_{(k,l) \in Y_i \times \overline{Y_i}} \alpha^i_{kl} \cdot exp(-(p^i_k - p^i_l)) \tag{16}$$

$$\alpha^i_{kl} = w \cdot (1 - p^i_k)^r + (1 - w) \cdot (p^i_l)^r \tag{17}$$

in which Y_i denotes the set of positive emotions expressed in ith instance s_i, and $\overline{Y_i}$ denotes the negative emotion set. p^i_k and p^i_l are the predicted probability of positive emotion e_k and negative emotion e_l respectively. Therefore, the training with above loss function is equivalent to maximizing the difference of negatively related emotion pair of $(p^i_k - p^i_l)$. This leads the system to output a higher probability for positive emotion while a lower probability for negative emotion. In this way, the emotion correlation of negatively related emotion pairs can be taken into consideration.

α^i_{kl} is a weighting factor and mainly affected by two parameters: $w \in (0,1)$ is a harmonic factor aimed to balance the prediction between positive and negative emotions, and $r > 0$ is a modulating factor aimed to make the loss put more focus on hard and misclassified examples during training. Significantly, while $r = 0$, the proposed multi-label focal loss is equivalent to the multi-label loss function.

For well-classified positive-negative emotion pairs (e_k, e_l), predicted probability p^i_k tends to 1 while p^i_l tends to 0. In this case, the difference of $(p^i_k - p^i_l)$ tends to the maximum, which means the minimum of $exp(-(p^i_k - p^i_l))$, and the weighting factor α^i_{kl} tends to 0. Thereby the loss of well-classified positive-negative emotion pairs is minimized. Conversely, for hard-classified emotion pairs, the difference of $(p^i_k - p^i_l)$ tends to the minimum, which could be caused by p^i_k tending to 0 or p^i_l tending to 1. In response to the above two cases, $w \cdot (1 - p^i_k)^r$ and $(1 - w) \cdot (p^i_l)^r$ are introduced to give more focus on misclassified p^i_k and p^i_l respectively.

4 Experiments Setup

4.1 Datasets

We employ two different datasets to evaluate the proposed

architecture, which are listed below:

Ren-CECps Dataset is an annotated emotional corpus with Chinese blog texts [21]. The corpus is annotated in the document, paragraph, and sentence level. Each level is annotated with eight emotional categories ('Joy', 'Hate', 'Love', 'Sorrow', 'Anxiety', 'Surprise', 'Anger', and 'Expect') and corresponding discrete emotional intensity value from 0.0 to 1.0. In our experiments, those emotions with an intensity greater than 0.0 are labeled as 1, otherwise 0. 'Neutral' is regarded as the $9th$ emotion label in case the sentence holds no emotion. After pre-processing, there is a total of 27091 sentences in training data and 7681 sentences in testing data. The average number of emotions expressed in a sentence is 1.4468.

NLPCC2018 Dataset consists of code-switching texts in Chinese, and concerns another language English on a small scale [53]. There are total 5 emotions annotated: 'Happiness', 'Sadness', 'Anger', 'Fear', and 'Surprise'. After pre-processing, there is 4611 texts in training data and 955 texts in testing data. The average number of emotions expressed in a sentence is 1.1466.

The cumulative number of each emotion e_i on Ren-CECps and NLPCC Datasets is calculated:

$$CN_i = \sum_{n=1}^{N}(y_{n,i} = 1) \quad (18)$$

in which $y_{n,i}$ is the annotation of emotion e_i in n_{th} sample. The statistical results are shown in Table 1.

TABLE 1
CUMULATIVE NUMBER OF EACH EMOTION IN REN-CECPS AND NLPCC2018 DATASETS.

Ren-CECps				NLPCC2018	
Love	11909	Hate	3533	Happiness	2534
Anxiety	10099	Anger	2236	Sadness	1502
Sorrow	8184	Surprise	1121	Surprise	811
Joy	6223	Neutral	2488	Anger	765
Expect	4633	-	-	Fear	770

4.2 Experimental details

In this section, we illustrate the experimental details during the model training.

In terms of the embedding of emotional words, it mainly consists of two parts. The first part is the general word embedding. It is initialized by 300-dimensional Word2vec word embedding, which is trained on Chinese microblog data [54]. The second part is emotional embedding by mapping an external n-dimensional emotion lexicon. Existing emotion lexicons are very rare due to the subjective and inconsistent annotation. In our experiments, a dimensional emotion lexicon is manually built based on word-level annotation on Ren-CECps. In our lexicon, each emotion word is annotated as an 8-dimensional vector v. Each dimension corresponding an emotion in ['Love', 'Anxiety', 'Sorrow', 'Joy', 'Expect', 'Hate', 'Anger', 'Surprise'], and the value represents the emotion intensity. For example, the emotional word '不幸' ('Unfortunately' in English) is represented as [0., 0.23, 0.62, 0., 0., 0., 0., 0.], which means that this word expresses stronger emotion of 'Sorrow' and weaker emotion of 'Anxiety', and the intensities are 0.62, and 0.23 respectively. Specially, an extra token named '[EMO_PAD]' is added to emotion lexicon, and its embedding vector is initialized by zeros. This token will be treated as emotional word if the current sentence does not contain any other emotional words.

For RenCECps, because of the non-coexistence of 'Neutral' label with other emotion labels, the final prediction is subject to the condition: only if the prediction of 'Neutral' label obtained the highest probability among all labels, the sentence is predicted as 'Neutral'. Otherwise, it is predicted as emotions contained.

In terms of the number of context sentences k, we set $k = 3$, which means that the previous 3 sentences are taken as the contextual information. Particularly, replication padding with the last sentence is utilized while the number of contextual sentences is less than 3.

We set the dropout as 0.2 in EcorL module to avoid overfitting. The hidden size of BiGRU in sentence-level encoder is 64 in each direction. For the binary focal loss utilized during the network pre-training, modulating factor r is set to 2, and the weighting factor α is set to 0.75. In multi-label focal loss, we set the modulating factor r to 2 and harmonic factor w to 0.4. Adam optimization method is applied to train the model by minimizing the proposed multi-label focal loss.

4.3 Metrics

In multi-label emotion detection task, the evaluation is more complicated than traditional single-label emotion classification. In this paper, some popular evaluation measures typically utilized in this task are utilized to measure the performance of proposed methods [52].

Micro F1-score and Macro F1-score are utilized as the main metrics to evaluate the global performance of each model. F1 score is the harmonic mean of precision and recall. Micro F1-score gives each sample the same importance, while Macro F1-score takes all classes as equally important. Hamming Loss (HL) is the fraction of labels that are incorrectly predicted. Average precision (AP) evaluates the average fraction of labels ranked above a particular label y: $y \in Y_i$ are actually in Y_i, in which Y_i is positive emotion set of sentence. Coverage evaluates how far it is needed to go down the ranked emotion list to cover all the relevant emotions in the instance. One Error (OE) evaluates the fraction of sentences whose top-ranked emotion is not in the relevant emotion set. Ranking Loss (RL) evaluates the average fraction of label pairs that are reversely ordered for instance.

4.4 Baseline models

To demonstrate the performance of the proposed MEDA model, some baseline methods are compared in our experiments:

BR [55], Binary Relevance, based on the label independence assumption, transforms a multi-label classification problem into multiple binary classification problems.

CC [56], Classifier Chains, a multi-label model that arranges binary classifiers into a classifier chain to capture the label correlations.

LP, LabelPowerset, creates one multi-class classifier for every label combination attested in the training set.

BP-MLL [52], is derived from the backpropagation algorithm by employing a novel error function to capture the characteristics of multi-label learning.

TABLE 2
COMPARISON RESULTS OF PROPOSED MODEL AND BASELINES ON RENCECPS DATASET

	Micro F1: % (↑)	Macro F1: % (↑)	AP: % (↑)	HL (↓)	Coverage (↓)	OE (↓)	RL (↓)
BR	46.40	34.79	63.69	0.2464	2.8313	0.5221	0.1789
CC	46.97	33.62	63.16	0.2282	2.9721	0.5234	0.1965
LP	45.15	42.51	62.62	0.2069	2.9117	0.5275	0.1861
BP-MLL	48.89	38.13	55.45	0.2241	3.1272	0.4625	0.3234
DPCNN	49.99	35.47	65.43	0.1583	3.0555	0.4834	0.1993
HANs	54.54	41.36	70.65	0.1504	2.4631	0.4520	0.1362
SGM	55.60	-	-	0.1758	-	-	-
DATN	-	45.70	73.20	-	-	0.4150	-
SGM-IFC	58.60	-	-	0.1613	-	-	-
S-MC-ESFE	59.24	***47.73***	75.19	0.1367	2.3170	***0.3760***	0.1163
C-MC-ESFE	55.30	34.34	74.76	**0.1213**	2.2765	0.3915	0.1134
MEDA	***59.71***	47.25	***75.76***	0.1378	***2.2369***	0.3763	***0.1084***
MEDA-FS	**60.76**	**48.31**	**76.51**	***0.1249***	**2.2226**	**0.3618**	**0.1062**

TABLE 3
COMPARISON RESULTS OF PROPOSED MODEL AND BASELINES ON NLPCC2018 DATASET

	Micro F1: % (↑)	Macro F: % (↑)	AP: % (↑)	HL (↓)	Coverage (↓)	OE (↓)	RL (↓)
BR	48.92	41.07	67.74	0.2975	2.1645	0.4958	0.2771
CC	49.92	40.51	68.63	0.2790	2.1221	0.4883	0.2668
LP	47.67	36.81	67.04	0.2456	2.1592	0.5159	0.2758
BP-MLL	55.66	41.65	74.78	0.2584	1.8896	0.4002	0.2066
DPCNN	46.07	34.25	64.22	0.2420	2.3482	0.5414	0.3231
HANs	55.69	42.78	76.92	0.2805	1.7930	***0.3758***	0.1835
SGM	57.11	36.28	64.24	0.1843	2.7813	0.4395	0.4267
S-MC-ESFE	**63.32**	***49.23***	**77.19**	0.1849	***1.7340***	**0.3780**	***0.1694***
C-MC-ESFE	60.59	46.90	76.43	***0.1719***	1.7592	0.3895	0.1749
MEDA	61.21	47.70	75.90	***0.1696***	1.7665	0.4021	0.1775
MEDA-FS	***63.02***	**49.42**	***77.12***	0.1728	**1.7288**	***0.3812***	**0.1681**

DPCNN [57], a low-complexity word-level deep pyramid CNN network that can efficiently capture global representations of text.

HANs [50], hierarchical attention networks that mirror the hierarchical structure of documents. HANs can find the essential words and sentences in a document while taking the contextual information into consideration.

SGM [40], transfers multi-label classification task to a sequence generation problem and can capture the correlations between labels.

In previous studies, several emotion classification methods have been implemented in RenCECps datasets and achieved the previous state-of-the-art performances. Therefore, we take them as baselines to verify the performance of our method in RenCECps, which includes:

DATN [58], divides the sentence representation into two different feature spaces, which aims to capture the general sentiment words and the other critical emotion-specific words via a dual attention mechanism.

SGM-IFC [59], utilizes the attention-based Seq2Seq model to solve the multi-label problem. An initialized fully connection layer is employed to capture the correlation between any two different labels.

For the baselines of BR, CC and LP, we take pre-trained BERT model as sentence encoder and Gaussian Naive Bayes as the classifier, and all experiments are implemented based on Scikit-multilearn library. The results of baselines BP-MLL, SGM, DATN, and SGM-IFC on RenCECps dataset are adopted from the published papers [38], [58], [59]. For others, the comparison experiments are implemented based on the open-source codes shared on GitHub.

5 EXPERIMENTAL RESULTS AND DISCUSSION

Experimental results of the proposed method and baseline models are reported in section 5.1. The discussions are organized into two sections. In section 5.2, we analyzed the contribution of multi-level information from each sub-model. In section 5.3, we evaluate the effectiveness of emotional features by ablation experiments. In section 5.4, we explore the effectiveness of proposed multi-label focal loss on this task.

5.1 Experimental Results

Experimental results of the proposed methods against baselines are shown in Table 2 and Table 3, the best two results on each metric are in bold and in bold italics, respectively.

As the results shown in Table 2, the proposed model significantly outperforms baseline models and achieves state-of-

the-art performance on RenCECps. Compared with SGM-IFC [59], which has previously achieved the state-of-the-art performances, proposed MEDA-FS has improved micro-F1 score from 58.60% to 60.76% and reduced hamming loss from 0.1613 to 0.1249. Compared with DATN, the proposed MEDA-FS has improved macro-F1 score from 45.70% to 48.31%, improved average precision from 73.20% to 76.51%, and reduced one error from 0.4150 to 0. 3618. Besides, our model outperforms other deep learning methods and commonly used machine learning methods to a great extent, such as BR algorithm and SGM model.

Table 3 shows the experimental results of proposed model and baselines on NLPCC2018 dataset. Our proposed model achieved excellent results on almost all metrics except hamming loss. The hamming loss of proposed MEDA-FS is 0.1728, while the best is 0.1617 (achieved by LP). HL is the fraction of wrong labels to the total number of labels and penalizes only the individual labels. There are mainly two reasons for the higher hamming loss. One reason is that weak emotions are difficult to predict accurately. MC-ESFE module can prevent the features of weak emotions from being covered by strong emotions to some extent, but not completely. Their emotional features are not noticeable and are difficult to recognize. The classifier tends to conservatively predict them as negative emotions to ensure the whole performance among all emotion labels. Another reason is that the data distribution is imbalanced. It is hard to guarantee the performance of low-source emotion categories. In future work, more attention will be paid to the detection of weak and low- source emotions. In addition to hamming loss, the global performance of proposed method can also be reflected by other multi-label metrics, such as micro-F1, macro-F1, and average precision, on which the proposed method has achieved satisfying performance.

5.2 Discussion of Sub-models

MEDA-FS is composed of 3 sub-models: MEDA, S-MC-ESFE and C-MC-ESFE. These sub-models are devoted to learning information from different levels and contributing to a more comprehensive ensemble model. To further explore the contribution of each sub-model, we further analyze their performance on RenCECps in this section. The comparison results are shown in Table 4.

TABLE 4
COMPARISON RESULTS OF SUB-MODELS ON RENCECPS.

	Micro			Macro		
	P	R	F1	P	R	F1
S-MC-ESFE	52.16,	68.55,	59.24	42.48	56.72	47.73
C-MC-ESFE	**59.34**,	51.77,	55.30	43.15	32.01	34.34
MEDA	51.81,	**70.46**,	59.71	41.44	**57.54**	47.25
MEDA-FS	55.77,	66.72,	**60.76**	**46.10**	52.21	**48.31**

MEDA: As the global performance shown in Table 2. MEDA (micro-F1 = 59.71%, HL = 0.1378) outperforms the previous state-of-the-art model SGM-IFC (micro-F1 = 58.60%, HL = 0.1613), and outperforms another two sub-models on micro-F1, AP and ranking loss. MEDA network consists of two modules. The first is MC-ESFE, which is a hierarchical network and extracted emotion-specified features from both sentence-level and context-level in each channel. This feature matrix is extracted from the under-layer and each dimension focused on a certain emotion, which could conclude more detailed emotion-specified information. Another is ECorL module, which learns more global semantic information and emotion correlations based on above emotion-specified features. These two modules enable MEDA to give emotion predictions based on context and emotion correlation information.

To verify whether the emotion correlation information is learned in MEDA, we visualize the emotional correlation coefficients matrix. It is calculated with Pearson product-moment correlation coefficients, which indicates the level to which two emotions vary together:

$$R_{ij} = cov(E_i, E_j)/\sigma E_i \cdot \sigma E_j \quad (19)$$

Where $E_i = [E_{1i}, E_{2i}, \dots E_{Ni}]$ and E_{ni} is the emotional intensity of emotion e_i in the nth sample. $cov(E_i, E_j)$ is the covariance of e_i and e_j, and σ is the standard deviation. Fig. 2 and Fig. 3 show the comparison of the actual correlation coefficients matrix on Ren-CECps and the predicted correlation coefficients matrix in MEDA model. We can observe that the distribution of positively/negatively related emotion pairs predicted in MEDA is similar to the real distribution on Ren-CECps. Taking 'Love' as an example. Fig. 3 shows that in actual distribution, the most positively related emotion with 'Love' is 'Joy' (+0.20) while the most negatively related emotion is 'Anxiety' (-0.38). This means that emotions 'Love' and 'Joy' often occur together while 'Love' and 'Anxiety' rarely appear together. The above emotion correlation information can also be learned by MEDA: correlation coefficient of 'Love' and 'Joy' is +0.51 while 'Love' and 'Anxiety' is -0.65. Besides, there are some emotion pairs with emotion correlation that have been learned, such as 'Love' and 'Sorrow' (-0.39), 'Anxiety' and 'Joy' (-0.43), 'Hate' and 'Anger' (+0.40), etc. The results demonstrate the ability of emotion correlation learning in proposed MEDA.

S-MC-ESFE: Results in Table 2 indicate that the prediction of S-MC-ESFE is better than baselines on most metrics. Compared with MEDA, S-MC-ESFE achieves a higher macro-F1 value (47.73% while 47.25%). Although this gap is small, it can reflect the average level of emotion detection of each emotion category in S-MC-ESFE. The higher macro-F1 of S-MC-ESFE suggests that for some sparse-resources emotion categories, it could give more accurate prediction than MEDA model. S-MC-ESFE is an intermediate model derived from MC-ESFE during sentence-level pre-training. In S-MC-ESFE, each channel is trained channel-wise and can be considered as multiple binary emotion classifier. During the training of each classifier, only the parameters of the corresponding channel are updated, which could make the model focus more on the feature extraction of a specified emotion. Take a sentence as an example: 'For a long time, I write about funny things in my blog, but this time, my heart is heavy.' In the channel of 'Joy', feature extraction will pay more attention to the words 'funny things', while 'Sorrow' channel focused more on 'heart is heavy'. Therefore, in each channel, the prediction of whether the sentence contains the corresponding emotion will be more accurate.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2020.3034215, IEEE Transactions on Affective Computing

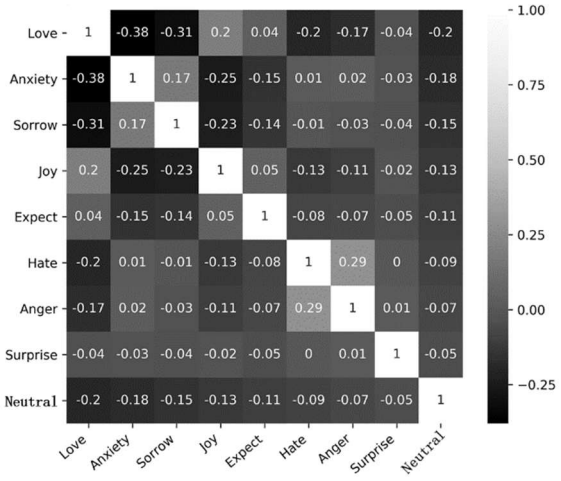


Fig. 2. Emotional correlation coefficients matrix in RenCECp

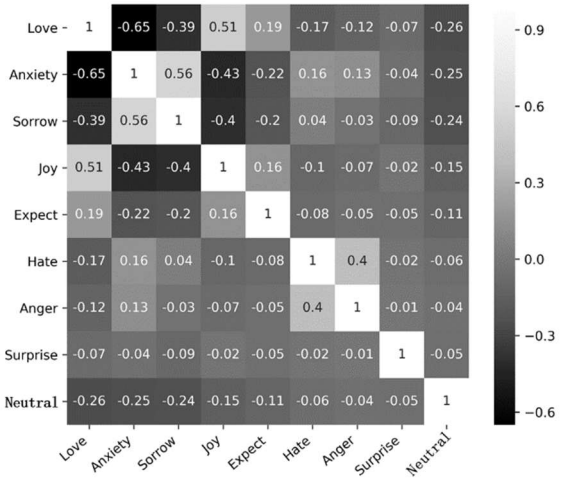


Fig. 3. Emotional correlation coefficients matrix learned by MEDA

C-MC-ESFE: In C-MC-ESFE model, contextual information is further considered compared with S-MC-ESFE model. The results in Table 2 shows that both macro-F1 and micro-F1 are inferior to S-MC-ESFE. However, C-MC-ESFE achieves a better hamming loss (HL = 0.1213) than MEDA (HL = 0.1378). To further explore the role of C-MC-ESFE in MEDA-FS model, we further compared the micro/macro precision and recall of each sub-model. The results are shown in Table 4.

From Table 4, we can see that the lower F1-score of C-MC-ESFE mainly because of the lower recall during prediction. Its micro recall is 51.77% while S-MC-ESFE is 68.55%. Although its recall is lower, it can ensure that the prediction is more accurate: micro-precision of C-ESFE is 59.34% while S-MC-ESFE is 52.16%. This means that the prediction given by C-MC-ESFE model is more rigorous. Therefore, with higher precision, the C-MC-ESFE model improves the confidence for the final prediction of the ensemble model.

S-MC-ESFE, C-MC-ESFE, and MEDA mean different levels of information from sentence-level, context-level, and emotion correlation level. They are integrated into MEDA-FS and contribute to more accurate and stable predictions in emotion detection task.

5.3 Ablation experiments

In the sentence-level embedding, we extract the emotional features based on the external emotional lexicon. To evaluate the effect of emotional features on experimental results, we train the model without this feature on RenCECps dataset. The experimental ablation results are shown in Table 5.

From Table 5, both MEDA model and MEDA-FS model with emotional features outperform the models without emotional features on almost all metrics. It is revealed that considering emotional features can make contributions to the classification improvement. In deep emotion recognition models, low-resource emotional datasets have been challenging, and effectively incorporating existing emotional resources is the key to improving performance. In proposed MEDA, external emotional lexicon works as prior knowledge and is directly incorporated in sentence-level encoding. This method implements external knowledge supplementation in the simplest way and contributes to the effective extraction of emotional features.

TABLE 5
ABLATION STUDY ON RENCECPS DATASET

	MEDA		MEDA-FS	
	With	Without	With	Without
Micro F1: %	**59.71**	56.22	**60.76**	57.33
Macro F1: %	**47.25**	42.91	**48.31**	44.16
AP: %	**75.76**	73.22	**76.51**	74.11
Hamming Loss	0.1378	**0.1342**	**0.1249**	0.1313
Coverage	**2.2369**	2.3703	**2.2226**	2.3322
One Error	**0.3763**	0.4101	**0.3618**	0.3959
Ranking Loss	**0.1084**	0.1239	**0.1062**	0.1189

'With' and 'without' denote with and without emotional features.

5.4 Discussion of multi-label focal-loss

In this section, we discuss the effectiveness of proposed multi-label focal loss (ML-FL) on emotion detection results. In the definition of multi-label focal loss, w is a harmonic factor aimed to balance the prediction of positive and negative labels. In this way, it has an effect on balancing the results of precision and recall, thus obtains an optimal F1 value. To verify the influence of w in emotion detection, we vary the value of w from 0. to 1. and compared it with two other commonly used loss functions: binary cross-entropy loss function(CE-loss) and multi-label loss function (ML-loss). The comparison experiments are implemented on RenCECps dataset, and the results are shown in Fig. 4, Table 6.

The results of CE-loss and ML-loss both show a higher recall (81.06% and 80.60% in micro-recall) while lower precision (41.95% and 44.39% in micro-precision). Precision is the average probability of relevant retrieval, while recall is the average probability of complete retrieval. They are two metrics restrain mutually [60]. In this emotion detection task, we hope to recognize as many emotions as possible, based on the premise of ensuring precision. It can be seen from the trend of the curve in Fig. 4: a proper w can modulate the value between recall and precision, thus achieve both higher precision and F1-score to alleviate the above problems.

TABLE 6
RESULTS COMPARISON OF MEDA MODEL WITH DIFFERENT LOSS FUNCTIONS

	Micro % (↑)			Macro: % (↑)			AP: %	HL	Coverage	OE	RL
	P	P	F1	P	P	F1	(↑)	(↓)	(↓)	(↓)	(↓)
CE-loss	41.95	**81.06**	55.29	37.26	64.71	42.79	70.64	0.1900	2.4496	0.4598	0.1349
ML-loss	44.39	80.60	57.25	35.49	**69.84**	46.07	72.27	0.1745	2.3652	0.4421	0.1251
ML-FL (w = 0.4)	**51.81**	70.46	**59.71**	**41.44**	57.54	**47.25**	**75.76**	**0.1378**	**2.2369**	**0.3763**	**0.1084**

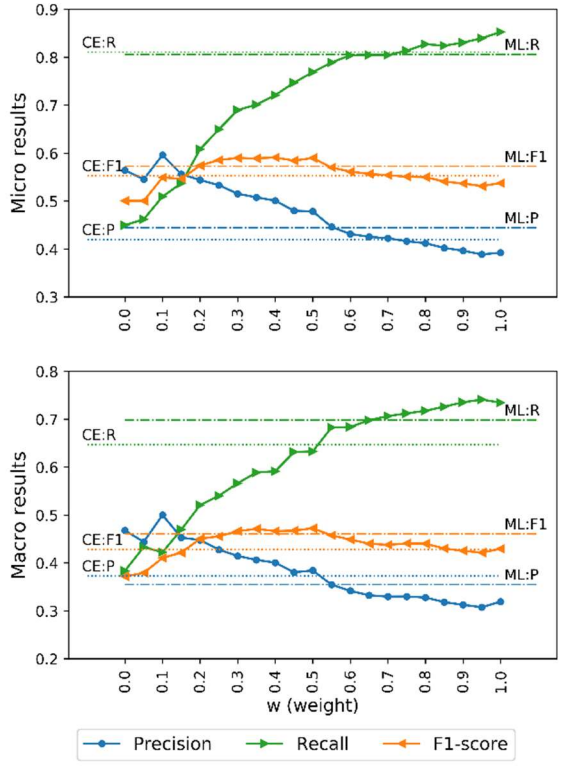


Fig. 4. The comparison results for cross-entropy(CE), multi-label loss function(ML), and proposed multi-label focal loss for different values of weight.

We will analyze the role of the parameter w in the curve change. From the tendency of the curve in Fig. 4, we can see that as the weight w increases, precision shows a downward trend, recall shows an upward trend while the overall trend of F1-score is to rise first and then fall. The loss function proposed in this paper is committed to maximizing the prediction difference between positive-negative emotion pairs. The weighting factor α is dedicated to balancing the prediction of positive and negative labels, which aimed to recognize as many emotions as possible while ensuring the accuracy of prediction. The weight α is consists of two parts to control the prediction loss of positive and negative emotions, respectively:

$$\alpha_{pos}^{i} = w \cdot \left(1 - p_k^i\right)^r, \alpha_{neg}^{i} = (1 - w) \cdot (p_l^i)^r \quad (20)$$

Considering the limit case, if harmonic factor w gradually increases to the maximum $w = 1$:

$$\alpha_{pos}^{i} \approx \left(1 - p_k^i\right)^r, \quad \alpha_{neg}^{i} \approx 0 \quad (21)$$

In this case, as long as the model predicts all the emotion as $p^i = 1$, it is possible to minimize $\alpha_{pos}^{i} = 0$, thereby minimizing the loss. In this way, the prediction gap between positive and negative emotion pairs $exp(-(p_k^i - p_l^i)$ can only play a weak role. Therefore, the results show a higher recall while precision is difficult to be guaranteed: while $w = 1.0$, the micro-precision, recall, and F1-score are 39.22%, 85.29%, and 53.74%, respectively.

Conversely, as w gradually decreases, α_{neg}^{i} gradually increases. In this way, the prediction error of the negative label will bring greater losses. To reduce the loss, the model predicts the positive label more conservatively, and thus the recall decreased and precision could be guaranteed to some extent. Therefore, it can be assumed that by choosing an appropriate value of w, it is possible to reach a balance between precision and recall, and then achieve satisfactory results. As the results in Fig. 4, take F1-score to measure the overall performance, while $w \in [0.3,0.5]$, proposed multi-label focal loss outperforms cross-entropy and multi-label loss function. To be specific, while $w = 0.4$, its micro-precision is 51.81%, micro- recall is 70.46%, micro-F1-score is 59.71%. Compared with ML-loss, although the recall drops, its micro-precision improved 7.42% and micro-F1 improved 2.46%. Table 6 shows the comparison results of different loss functions on multi-label metrics, which demonstrate that multi-label loss function outperforms others.

6 CONCLUSIONS

In this paper, a Multiple-label Emotion Detection Architecture (MEDA) was proposed for the textual multi-label emotion detection task. MEDA was composed of two modules, its key idea was to capture emotion-specified features by MC-ESFE module in advance, and then learn emotion correlations based on above features in ECorL module. In MC-ESFE module, information of each emotion reflected in the text was separately encoded from sentence-level to context-level, which contributed a lot to underlying fundamental feature extraction. In ECorL module, bidirectional-GRU network was utilized as emotion sequence predictor and emotion correlation learning was implemented among emotion-specified features. MEDA-FS integrated three sub-models derived from MEDA, and realized information fusion from sentence-level, context-level, and emotion correlation level. Furthermore, to incorporate emotion correlation information into model training, multi-label focal loss was proposed for multi-label learning. The proposed model achieved satisfactory performance and outperformed state-of-the-art models on both RenCECps and NLPCC2018 datasets, which demonstrated the effectiveness

of the proposed method for multi-label emotion detection.

There is still much space for improvements in our works. Discernible feature representation of the weak emotion category is a critical problem in multi-label emotion detection task. Our proposed MC-ESFE module can prevent the features of weak emotions from being covered by strong emotions to some extent, but not completely. In future work, we will try to explore more effective methods to recognize weak emotions more accurately. During the emotional feature extraction in our model, an external emotion lexicon was severed as prior knowledge to enhance emotional feature representation. Abundant resources are the basis of neural network training. In future work, more emotional data will be incorporated for better emotion understanding.

ACKNOWLEDGMENT

This work was partially supported by the Research Clusters program of Tokushima University (No. 2003002).

REFERENCES

[1] W. Liang, H. Xie, Y. Rao, R.Y.K. Lau, and F.L. Wang, "Universal affective model for Readers' emotion classification over short texts," *Expert Syst. Appl.*, vol. 114, pp. 322-333, Dec. 2018.

[2] F. Ren and K. Matsumoto, "Semi-Automatic Creation of Youth Slang Corpus and Its Application to Affective Computing," *IEEE Trans. Affect. Comput.*, vol.7, no.2, pp. 176-189, 2016.

[3] H.Y. Shum, X. He, and D. Li , "From Eliza to XiaoIce: challenges and opportunities with social chatbots," *Front. Inf. Technol. Electron. Eng.*, vol.19, no.1, pp. 10-26, Jan. 2018.

[4] R. Jayakrishnan, G.N. Gopal, and M.S. Santhikrishna, "Multi-class emotion detection and annotation in malayalam novels," *2018 Int. Conf. on Comput. Commun. and Inform. (ICCCI)*, Jan. 2018.

[5] T. Kowatsch, M. Nißen, C.H.I Shih, and D. Rüegger, "Text-based healthcare chatbots supporting patient and health professional teams: Preliminary results of a randomized controlled trial on childhood obesity," *Persuasive Embodied Agents for Behav. Chang. (PEACH2017)*, Aug. 2017.

[6] C. Chen, R. Zhuo, and J. Ren, "Gated recurrent neural network with sentimental relations for sentiment classification," *Inf. Sciences*, vol. 502, pp. 268–278, Oct. 2019.

[7] D. Sznycer, and A.W. Lukaszewski, "The emotion–valuation constellation: Multiple emotions are governed by a common grammar of social valuation," *Evol. Hum. Behav.*, vol. 40, no. 4, pp. 395-404, Jul. 2019.

[8] X. Kang, F. Ren and Y. Wu, "Exploring Latent Semantic Information for Textual Emotion Recognition in Blog Articles," *IEEE/CAA J. Autom. Sin.*, vol.5, no.1, pp. 204-216, 2018.

[9] A. Bandhakavi, N. Wiratunga, and D. Padmanabhan, "Lexicon based feature extraction for emotion text classification," *Pattern Recognit. Lett.*, vol. 93, no. 1, pp. 133-142, Jul. 2017.

[10] S.M. Liu, and J.H. Chen, "A multi-label classification based approach for sentiment classification," *Expert Syst. Appl.*, vol. 42, no. 3, pp. 1083-1093, Feb. 2015.

[11] N. Colneriĉ, and J. Demsar, "Emotion Recognition on Twitter: Comparative Study and Training a Unison Model," *IEEE Trans. Affect. Comput.*, pp. 1949-3045, Feb. 2018.

[12] D. Pan, and J. Nie, "Mutux at SemEval-2018 Task 1: Exploring Impacts of Context Information On Emotion Detection," in *Proc. 12th Int. Workshop on Semant. Eval.*, pp. 345-349, Jun. 2018.

[13] H. Huihui, and R. Xia, "Joint Binary Neural Network for Multi-label Learning with Applications to Emotion Classification," *Nat. Lang. Process. and Chin. Compu. (NLPCC)*, pp. 250-259, Aug. 2018.

[14] P.D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," *in Proc. Assoc. for Comput. Linguist. (ACL)*, pp. 417–424, Jul. 2002.

[15] F. Ren and Y. Wu, "Predicting User-Topic Opinions in Twitter with Social and Topical Context," *IEEE Trans. Affect. Comput.*, vol.4, no.4, pp. 412-424, 2013.

[16] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques," *in Proc. Conf. on Empirical Methods in Nat. Lang. Process. (EMNLP)*, pp. 79–86, 2012, doi: 10.3115/1118693.1118704.

[17] S. Mohammad, and S. Kiritchenko. "Understanding emotions: A dataset of tweets to study interactions between affect categories," *in Proc. 11th Int. Conf. on Lang. Resour. and Eval. (LREC-2018)*, pp. 198-209, 2018.

[18] N.H. Frijda, "The laws of emotion," *Am. Psychol.*, vol. 43, no. 5, pp. 349-358, May. 1988.

[19] P. Ekman, "An Argument for Basic Emotions," *Cogn. Emot.*, vol. 6, no. 3-4, pp. 169–200, 1992, doi:10.1080/02699939208411068.

[20] Q. Changqin, and F. Ren. "Sentence emotion analysis and recognition based on emotion words using Ren-CECps," *Int. J. Adv. Intell.*, vol. 2, no. 1, pp. 105-117, Jul. 2010.

[21] J. Li, and F. Ren, "Creating a Chinese emotion lexicon based on corpus Ren-CECps," *2011 IEEE Int. Conf. on Cloud Comp. and Intell. Syst.*, Sep. 2011, doi: 10.1109/CCIS.2011.6045036.

[22] C. Strapparava, and A. Valitutti, "Wordnet affect: an affective extension of wordnet," *in Proc. 4th Int. Conf. on Lang. Resour. and Eval. (LREC'04)*, vol. 4, pp. 1083-1086, May. 2004.

[23] S. Mohammad, "Portable features for classifying emotional text," *in Proc. 2012 Conf. of the North Am. Chapter of the Asso. for Comput. Linguist.: Hum. Lang. Technol.*, pp.587–591, 2012.

[24] Q. Liu, and S. Li, "Word Semantic Similarity Computation based on Hownet," *in Proc. 3rd Symp. on Chinese Words Semantics.* vol. 5, 2002.

[25] S.M. Mohammad, and P.D. Turney, "Crowdsourcing a word–emotion association lexicon," *Comput. Intell.*, vol. 29, no. 3, pp. 436-465, Sep. 2012.

[26] F.A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *in Proc. 9th Int. Conf. on Artif. Neural Netw. (ICANN '99)*, pp. 850-855, 1999.

[27] S. Poria, E. Cambria, D. Hazarika, and N. Majumder, "Context-dependent sentiment analysis in user-generated videos," *in Proc. 55th Annu. Meet. of the Assoc. for Comput. Linguist.*, vol. 1, pp. 873-883, 2017, doi: 10.18653/v1/P17-1081.

[28] D. Tang, B. Qin, T. Liu, "Aspect level sentiment classification with deep memory network," *in Proc. 2016 Conf. on Empirical Methods in Nat. Lang. Process.*, pp. 214–224, Nov. 2016.

[29] Y. Kim, "Convolutional neural networks for sentence classification," *in Proc. 2014 Conf. on Empir. Methods in Nat. Lang. Process. (EMNLP)*, 2014.

[30] T. Rao, X. Li, H. Zhang, and M. Xu, "Multi-level region-based Convolutional Neural Network for image emotion classification," *Neurocomputing*, vol. 333, no. 14, pp. 429-439, Mar. 2019.

[31] M.S. Akhtar, D. Ghosal, and A. Ekbal, "A Multi-task Ensemble Framework for Emotion, Sentiment and Intensity Prediction," *arXiv preprint arXiv:1808.01216*, 2018.

[32] M. Chae, T.H. Kim, Y.H. Shin, J.W. Kim, and S.Y. Lee, "End-to-

end Multimodal Emotion and Gender Recognition with Dynamic Weights of Joint Loss," *arXiv preprint arXiv: 1809.00758*, 2018.

[33] P. Xu, A. Madotto, C.S. Wu, J.H. Park, and P. Fung, "Emo2vec: Learning generalized emotion representation by multi-task training," *arXiv preprint arXiv:1809.04505*, 2018.

[34] A. Illendula, and A. Sheth, "Multimodal Emotion Classification," *Companion Proc. of The 2019 World Wide Web Conf,* ACM, May, 2019.

[35] L. Buitinck, J.V. Amerongen, and E. Tan, "Multi-emotion detection in user-generated reviews," *Eur. Conf. on Inf. Retr. (ECIR 2015)*, pp. 43-48, 2015.

[36] Y. Kim, H. Lee, and K. Jung, "AttnConvnet at SemEval-2018 Task 1: attention-based convolutional neural networks for multi-label emotion classification," *in Proc. 12th Int. Workshop on Semantic Eval.*, pp. 141–145, Jun. 2018.

[37] J. Yu, L. Marujo, J. Jiang, and P. Karuturi, "Improving multi-label emotion classification via sentiment classification with dual attention transfer network," *in Proc. 2018 Conf. on Empir. Methods in Nat. Lang. Process.*, pp. 1097–1102, Oct. 2018.

[38] D. Zhou, Y. Yang, and Y. He, "Relevant emotion ranking from text constrained with emotion relationships," *in Proc. 2018 Conf. of the North Am. Chapter of the Assoc. for Comput. Linguist.: Hum. Lang. Technol.*, vol. 1, pp. 561-571, Jun. 2018.

[39] M.L. Zhang, and Z.H. Zhou, "Multi-label neural networks with applications to functional genomics and text categorization," *IEEE Trans. Knowl. Data Eng.*, vol. 18, pp. 1338–1351, 2006.

[40] P. Yang, X. Sun, W. Li, S. Ma, W. Wu, and H. Wang, "SGM: sequence generation model for multi-label classification," *in Proc. 27th Int. Conf. on Comput. Linguist.*, pp. 3915-3926, Aug. 2018.

[41] B. Zhao, X. Li, X. Lu, and Z. Wang, "A CNN–RNN architecture for multi-label weather recognition," *Neurocomputing*, vol. 322, no. 17, pp. 47-57, Dec. 2018.

[42] S. Lian, J. Liu, R. Lu, and X. Luo, "Captured multi-label relations via joint deep supervised autoencoder," *Appl. Soft Comput.*, vol. 74, pp. 709-728, Jan. 2019.

[43] C.K. Yeh, W.C. Wu, W.J. Ko, and Y.C.F. Wang, "Learning deep latent space for multi-label classification," *in Proc. 31th AAAI Conf. on Artif. Intell.*, pp. 2838-2844, Feb. 2017.

[44] K. Wang, M. Yang, and W. Yang, "Deep Correlation Structure Preserved Label Space Embedding for Multi-label Classification," *Asian Conf. on Mach. Learn.*, vol. 95, pp.1-16, 2018.

[45] D.A. Phan, Y. Matsumoto, and H. Shindo, "Autoencoder for Semisupervised Multiple Emotion Detection of Conversation Transcripts," *IEEE Trans. Affect. Comput.*, Dec. 2018.

[46] Z.F. He, M. Yang, Y. Gao, H.D. Liu, and Y. Yin, "Joint multi-label classification and label correlations with missing labels and feature selection," *Knowl. Based Syst.*, vol. 163, pp. 145-158, Jan. 2019.

[47] M. Rei, and A. Søgaard, "Jointly Learning to Label Sentences and Tokens," *in Proc. 33th AAAI Conf. on Artif. Intell.*, pp. 6916-6923, 2019.

[48] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *in Proc. NAACL-HLT*, pp. 4171-4186, Jun. 2019.

[49] J. Chung, C. Gulcehre, K.H. Cho, and Y Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *arXiv preprint arXiv:1412.3555*, Dec. 2014.

[50] Z. Yang, D. Yang, C. Dyer, X. He, and A. Smola, "Hierarchical attention networks for document classification," *in Proc. 2016 Conf. of the North Am. Chapter of the Assoc. for Comput. Linguist.: Hum. Lang. Technol.*, pp. 1480-1489, Jun. 2016.

[51] T.Y Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *in Proc. IEEE international Conf. on Comput. Vis.*, pp. 2980-2988 2017.

[52] M.L. Zhang, and Z.H. Zhou, "Multilabel neural networks with applications to functional genomics and text categorization," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 10, pp. 1338-1351, 2006.

[53] Z. Wang, S. Li, F. Wu, Q. Sun, and G. Zhou, "Overview of NLPCC 2018 Shared Task 1: Emotion Detection in Code-Switching Text," *CCF International Conf. on Nat. Lang. Process. and Chinese Comput.*, pp. 429–433, 2018.

[54] S. Li, Z. Zhao, R. Hu, W. Li, T. Liu, and X. Du, "Analogical reasoning on chinese morphological and semantic relations", *in Proc. 56th Annu. Meeting of the Assoc. for Comput. Linguist.*, pp. 138–143, 2018.

[55] O. Luaces, J. Díez, J. Barranquero, and J.D. Coz, "Binary relevance efficacy for multilabel classification," *Prog. Artif. Intell.*, vol. 1, no. 4, pp. 303-313, 2012.

[56] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Mach. Learn.*, vol. 85, no. 3, pp. 333-359, 2011.

[57] R. Johnson, and T. Zhang, "Deep pyramid convolutional neural networks for text categorization", *in Proc. 55th Annu. Meeting of the Assoc. for Comput. Linguist.*, vol. 1, pp.562-570, 2017.

[58] J. Yu, L. Marujo, J. Jiang, P. Karuturi, and W. Brendel, "Improving multi-label emotion classification via sentiment classification with dual attention transfer network," *in Proc. of the Assoc. for Comput. Linguist. (ACL)*, 2018.

[59] W. Liao, Y. Wang, Y. Yin, X. Zhang, and P. Ma, "Improved sequence generation model for multi-label classification via CNN and initialized fully connection," *Neurocomputing*, vol. 382, no.21, pp. 188-195, Mar. 2020.

[60] D.M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, no.1, pp. 37-63, Dec. 2011.

Jiawen Deng received the double degree of master in Advanced Technology and Science form Tokushima University, Japan, and Mechanical Engineering from Nantong University, China. She currently is a Ph.D student at Tokushima University. Her research interests include text mining and sentiment analysis.

Fuji Ren received his Ph.D. degree in 1991 from the Faculty of Engineering, Hokkaido University, Japan. From 1991 to1994, he worked at CSK as a chief researcher. In 1994, he joined the Faculty of Information Sciences, Hiroshima City University, as an Associate Professor. Since 2001, he has been a Professor of the Faculty of Engineering, Tokushima University. His current research interests include Natural Language Processing, Artificial Intelligence, Affective Computing, Emotional Robot. He is the Academician of The Engineering Academy of Japan and EU Academy of Sciences. He is a senior member of IEEE, Editor-in-Chief of International Journal of Advanced Intelligence, a vice president of CAAI, and a Fellow of The Japan Federation of Engineering Societies, a Fellow of IEICE, a Fellow of CAAI. He is the President of International Advanced Information Institute, Japan.