




# EmoChannel-SA: exploring emotional dependency towards classification task with self-attention mechanism

Zongxi Li<sup>1</sup> · Xinhong Chen<sup>2</sup> · Haoran Xie<sup>3</sup> · Qing Li<sup>4</sup> · Xiaohui Tao<sup>5</sup> · Gary Cheng<sup>6</sup> 

Received: 22 May 2021 / Revised: 1 August 2021 / Accepted: 15 September 2021 /

Published online: 6 October 2021

© The Author(s) 2021

## Abstract

Exploiting hand-crafted lexicon knowledge to enhance emotional or sentimental features at word-level has become a widely adopted method in emotion-relevant classification studies. However, few attempts have been made to explore the emotion construction in the classification task, which provides insights to how a sentence's emotion is constructed. The major challenge of exploring emotion construction is that the current studies assume the dataset labels as relatively independent emotions, which overlooks the connections among different emotions. This work aims to understand the coarse-grained emotion construction and their dependency by incorporating fine-grained emotions from domain knowledge. Incorporating domain knowledge and dimensional sentiment lexicons, our previous work proposes a novel method named *EmoChannel* to capture the intensity variation of a particular emotion in time series. We utilize the resultant knowledge of 151 available fine-grained emotions to comprise the representation of sentence-level emotion construction. Furthermore, this work explicitly employs a self-attention module to extract the dependency relationship within all emotions and propose *EmoChannel-SA* Network to enhance emotion classification performance. We conducted experiments to demonstrate that the proposed method produces competitive performances against the state-of-the-art baselines on both multi-class datasets and sentiment analysis datasets.

**Keywords** Sentiment analysis · Emotion classification · Emotion lexicon · EmoChannel

---

This article belongs to the Topical Collection: *Special Issue on Web Intelligence = Artificial Intelligence in the Connected World*

Guest Editors: Yuefeng Li, Amit Sheth, Athena Vakali, and Xiaohui Tao

This paper is an extension of a conference paper [23], which was published at 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT).

---

✉ Gary Cheng  
chengks@eduhk.hk

Extended author information available on the last page of the article.

## 1 Introduction

Text emotion classification is an important branch of Natural Language Processing (NLP) research, aiming to identify the prominent emotion from short texts by predicting the label from a set of pre-defined emotions. Upon the identified emotions from user-generated texts (e.g., comments, reviews, blogs, and news reports), user attitudes and opinions can be retrieved and analyzed, and hence the task has great potential applications in various aspects of daily lives [17, 22, 43]. Existing works commonly extract semantic representation from texts via various deep modules in order to understand the semantic meaning for neural decision making. Considering the domain-specific nature of the task, researchers have made continuous efforts to improve classifier's robustness by enriching word-level sentimental features, i.e., creating and exploiting different types of hand-crafted lexicon dictionaries [24, 27, 37].

In existing works employing lexicons in sentiment analysis [15, 39], the lexicon information is adopted as additional knowledge in the time-series context. Meanwhile, such works mainly investigate emotions from a coarse-grained perspective [1] and do not consider the interconnection between emotions. Based on such an assumption, only the most prominent emotion is recognized from a sentence. Nevertheless, due to the complexity of text expressions and the fuzzy nature of emotions, multiple types of emotions can be easily spotted in a single sentence and they are usually interconnected [9]. To address such an issue, a fine-grained emotion perspective can be helpful to bridge different fine-grained emotions with each other and profile a general emotion composition. Therefore, instead of only predicting a coarse-grained emotion label for the input text, we propose to study the intensity and distribution of various fine-grained emotions expressed by the sentence, which comprises a fuzzy-style emotion construction of the text. With such an emotion construction, we can shed light on the dependency relationship and interaction within a wide spectral of emotions towards the classification task.

To examine the emotion construction, we need to identify the implicit connections between text and emotion. To address this issue, we utilize emotion lexicons to bridge word with emotions. Li et al. [24] categorizes commonly-used lexicon dictionaries as categorical lexicons and dimensional lexicons based on the emotion theories they follow [8]. In categorical lexicons, word is labeled with at least one tag from an identified collection of emotions, e.g., the six basic emotions identified by Ekman [9]. Although such lexicons directly associate words with emotions, they can only work with a restricted scope of emotions and provide no clue about the intensity. On the other hand, dimensional lexicons, like NRC-VAD [26], following dimensional emotion theories that conceptualize emotions with measurable variables [29, 32, 33], annotate each word with intensity values on two dimensions (i.e., *valence* and *arousal*) or three dimensions (i.e., *valence*, *arousal*, and *dominance* (VAD)).

To model the relationship between text and emotions, we incorporate psychological domain knowledge together with dimensional lexicons. Russell and Mehrabian [33] recognized the sufficiency of VAD factors on the definition of emotional states and identified 151 emotions in terms of *valence*, *arousal*, and *dominance* with mean and standard deviation in the Three-Factor Theory.

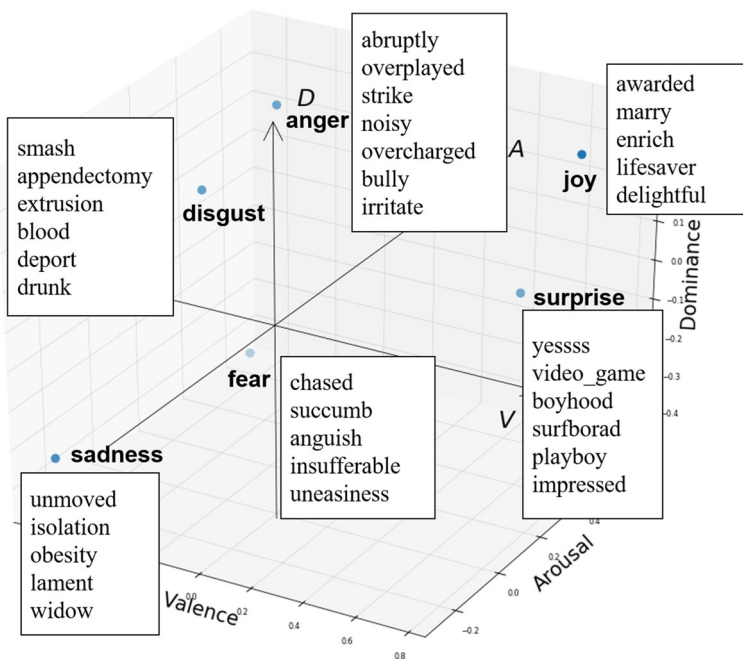
We consider a probabilistic model to combine the *Knowledge of Emotion* (KoE) and NRC-VAD lexicon as the *Knowledge of Words* (KoW), so as to quantitatively measure the relationship between a word and an emotion in the VAD space. Such a relationship can be modeled as the emotion intensity. Given a sentence, a vector of sentence length is generated for each emotion by transiting the sentence and computing the intensity expressed by each

word. The generated vector contains the intensity variation values and distributions across the sentence for a specific emotion and hence we name such a vector as *EmoChannel*. In Figure 1, the visualizations demonstrate the compatibility of KoE and KoW in VAD space, and in Figure 2, an example of how to generate *EmoChannel* is provided. So far, we obtain 151 channel vectors for 151 fine-grained emotions as the emotion construction of a sentence.

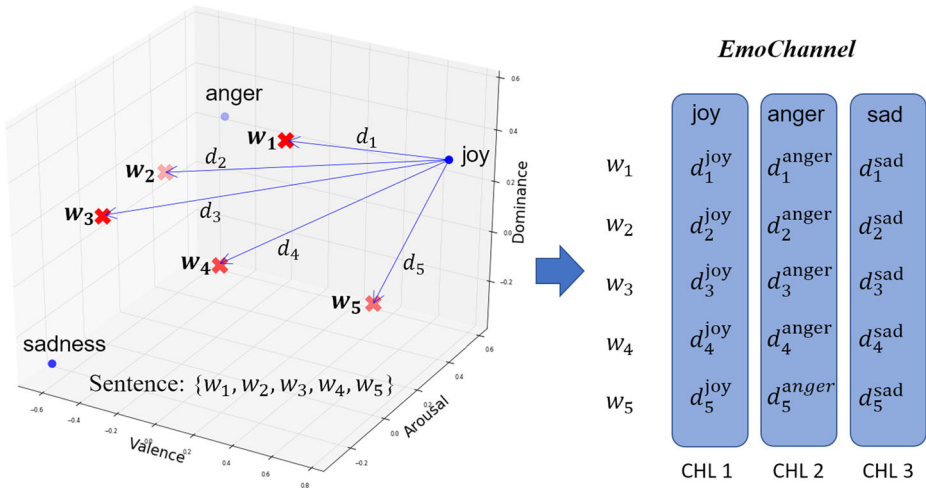
To further extract the dependency relationships among fine-grained emotions, we propose a self-attention-based model, *EmoChannel-SA* Network, which employs self-attention blocks over the constructed *EmoChannels*. The intermediate output of the self-attention block is exploited as the sentence-level emotion representation to enhance the decision-making process. Furthermore, we visualize the attention weights of multiple selected samples via heat maps and provide in-depth discussions about how the emotion dependency contributes to the learning process. Our main contributions are summarized as follows:

- We propose an self-attention-based model to enhance classification performance by exploiting dependency of fine-grained emotions;
- Extensive experiments on multi-class classification and sentiment analysis datasets of different topics show that our methodology produces competitive outperforms against the baseline models.

The early stage of this work has been published in [23]. We have since expanded the work on both technical and experimental contents. The main enhancements from the conference version include: (1) instead of using an additive attention, we employ self-attention module



**Figure 1** Visualization of the six basic emotions from Ekman [9] in the VAD space with the nearest words from NRC-VAD lexicon



**Figure 2** A toy example demonstrates the process of generating *EmoChannels* for three emotions given a sentence of five words. “CHL” is an abbreviation of “Channel”

to extract more informative dependencies within fine-grained emotions; (2) we experiment on both multi-class classification and sentiment analysis tasks, to test the generality of the proposed model; (3) we provide insightful discussions by visualizing the attention weights via heat maps.

## 2 Related work

In this section, we first introduce the major emotion theories in psychological domain, and then briefly review the recent development of emotion classification approaches.

The commonly adopted emotion theories are of two types [8]: categorical and dimensional. Categorical theories consider the emotions are discretely and differently constructed, and a set of basic emotions can be understood among different cultures. The most popular emotion taxonomy is Ekman’s Six Basic Emotions theory [9], which concludes a primary emotion set with six emotions: *anger*, *disgust*, *fear*, *happiness (joy)*, *sadness* and *surprise*. In contrast, dimensional models define emotion categories with measurable variables [29] that can be to conceptualize all emotion states. Most dimensional theories incorporate valence and arousal or intensity dimensions. As one of the most prominent two-dimensional model following dimensional theory, Circumplex model developed by [32] maps emotions into a circular two-dimensional space. where the vertical axis and the horizontal axis represent activation-deactivation and pleasant-unpleasant, respectively. As a representative three-dimensional model, Russell and Mehrabian [33] suggest that a three-dimensional system is sufficient to define emotional states. The system consists of three independent and bipolar dimensions, which are *pleasure* (pleasure-displeasure), *arousal* (degree of arousal), and *dominance* dominance-submissiveness”. Furthermore, their work identifies 151 terms denoting emotions in terms of the three factors with mean and standard deviation. In this work, we incorporate the Three-Factor Theory and dimensional emotion lexicon knowledge to construct the EmoChannel vector for a sentence.

Recently, exploiting deep neural networks for supervised text classification has become a mainstream approach, and it has achieved much remarkable progress. In the following, we review several widely adopted semantic feature extractors. Kim [18] proposed a classic TextCNN model with max-over-time pooling mechanism, which shows superior ability to extract local and position-invariant features. Another popular deep learning model is Recurrent Neural Networks (RNN), which can extract sequential feature from a sequence. Hochreiter and Schmidhuber [14] and Socher et al. [35] used recursive networks to explicitly exploit time-series features. Several variants based on RNN have been proposed, for example, BiLSTM [13] and GRU [5] with more complex gate mechanisms. However, these methods fail to give adequate weights to some discriminative words. To address this problem, Bahdanau et al. [3] introduced and applied attention mechanism to machine translation. Afterwards, the attention mechanism has been widely employed in various NLP tasks. Vaswani et al. [40] employed the stacked self-attention blocks to learn the global dependency of a sentence as a more robust sentence-level representation. Delvin et al. [6] revisited the language model, combining Transformer-based architecture to pre-train language model over large-scale textual resources, which has been proven effective in improving downstream tasks.

Next, we introduce recent development in emotion classification field. Researchers have made continuous efforts on emotion classification task recently. It is a common approach in deep learning framework to exploit the CNN-based, RNN-based, and self-attention-based models to extract semantic features from word embedding vectors. For example, Chen et al. [4] employed a stacked BiLSTM model to obtain both forwards and backwards sequence information to identify the emotion categories and their corresponding causes. Feng et al. [11] address the group-based emotion detection exploiting topic exploration. Lai et al. [20] adopt graph convolutional networks to enhance fine-grained emotion classification performance. However, semantic feature is not sufficient in a domain-specific task due to the data sparsity issue [19]. Several works attempt to construct dedicated word representation, which encodes sentiment information into low-dimensional embedding vectors. Several works of effective representation learning rely on training embeddings from scratch on large tweets datasets [10, 38]. Besides, numerous attempts have been made to improve classifier performance using additional knowledge, such as emotion lexicons, syntax structure, and causality relationship. Qian et al. [30] adopted a BiLSTM with linguistic-inspired regularizers considering sentiment lexicons to predict text sentiment. Teng et al. [39] proposed a context-sensitive lexicon-based method based on a weighted-sum model to calculate sentiment aggregation using RNN architecture. Li et al. [21] presented Adaptive Gate Network to incorporate statistical features into classification task.

### 3 Methodology

In this section, we will elaborate on the process of constructing the fine-grained emotion intensity vector *EmoChannel* and our proposed *EmoChannel-SA* model framework.

#### 3.1 *EmoChannel*: Emotion Distribution over Sentence

We first present the formal definition of *EmoChannel* as the intensity distribution of the emotion over a sentence.

**Definition 1** Given a sentence and an emotion, the *EmoChannel* is an emotion-based sentence-level representation, depicting the emotion intensity variation over time across the sentence. The representations of different emotions are independent [23].

We model the emotion intensity of each word as the probability of corresponding word's belongingness to the emotion. For an emotion  $E_k$ , we retrieve the VAD mean  $\mu_{E_k} = [V_{E_k}^m, A_{E_k}^m, D_{E_k}^m]$  and standard deviation  $\sigma_{E_k} = [V_{E_k}^{sd}, A_{E_k}^{sd}, D_{E_k}^{sd}]$  from the Three-Factor Theory. Given a sentence of  $M$  words,  $w_1, w_2, \dots, w_M$ , for each word  $w_i$  that is included in the NRC-VAD dictionary, we can retrieve the VAD tuple  $\mathbf{w}_i = [V^{w_i}, A^{w_i}, D^{w_i}]$ . We consider the three emotion factors following multi-variate Gaussian distribution and compute  $\mathbf{d}_{w_i}^{E_k}$  as the intensity of emotion  $E_k$ :

$$\mathbf{d}_{w_i}^{E_k} = \frac{\exp\left(-\frac{1}{2}(\mu_{E_k} - \mathbf{w}_i)\Sigma^{-1}(\mu_{E_k} - \mathbf{w}_i)\right)}{\sqrt{(2\pi)^3|\Sigma|}}, \quad (1)$$

where  $\Sigma = \text{diag}(\sigma_{E_k})$ . To this end, we obtain a sentence-level affective representation by modeling the *belongingness* of words within the sentence with the emotion. For the  $M$  words, the EmoChannel of emotion  $E_k$  over the sentence is  $\mathbf{C}_k = [d_{w_1}^{E_k}, d_{w_2}^{E_k}, \dots, d_{w_M}^{E_k}]$ . In total, 151 emotions were identified in the Three-Factor Theory, so as 151 independent EmoChannel representations are constructed for each short text. We concatenate all the channels to form the EmoChannel matrix as the emotion construction of the sentence.

We randomly initialize vectors for the OOL (out-of-lexicon) tokens following a similar way as implementing the word embedding model, because the large number of OOL tokens leads to an undesirable sparsity issue. Specifically, to differentiate the affective words included by lexicons and the OOL tokens, the values are sampled from a distribution with a much smaller standard deviation, say 0.001. Meanwhile, all weights in emotion construction matrix are trainable during the training process, making the decision-making more adaptive.

### 3.2 The model framework

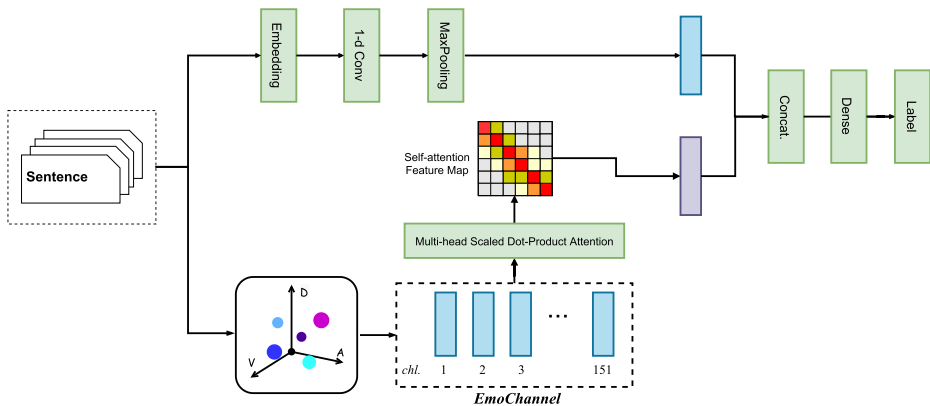
In this section, we introduce the proposed model framework according to each component. In general, we first build a classifier to extract semantic features on textual input and employ emotion construction of the sentence to enhance the decision making. A generic framework is demonstrated in Figure 3.

#### 3.2.1 Input layer

The input of our model consists of a sentence  $s$  with fixed length  $M$  and the EmoChannel vectors  $\mathbf{C} = [\mathbf{C}_1, \mathbf{C}_2, \dots, \mathbf{C}_k]$  of  $s$ . For non-Bert models, we first map each word into a  $D$ -dimensional continuous space and obtain the word embedding vector  $\mathbf{x}_M \in \mathbb{R}^D$ . Then we concatenate all word vectors to form a  $D \times M$  matrix as the model input:

$$\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_M] \quad (2)$$

Following the same way as in [18], we pad the sentences to maintain a uniform length for all sentences. For Bert-based models, we tokenize the textual input with Bert tokenizer.



**Figure 3** The generic framework of the proposed *EmoChannel-SA* Network

### 3.2.2 Semantic feature extraction layer

For a non-Bert model, we employ TextCNN [18] as the extractor to produce semantic representation from the text. We apply a filter  $\mathbf{W}^f \in \mathbb{R}^{h \times D}$  with window size  $h$  to slide through the embedding matrix. The new feature  $z_i$  is generated from a window of word vectors  $\mathbf{x}_{i:i-h+1}$ :

$$z_i = f(\mathbf{W}_f \otimes \mathbf{x}_{i:i-h+1} + b), \tag{3}$$

where,  $b \in \mathbb{R}$  is the bias term, and  $f(\cdot)$  is a non-linear function, and each filter produces a feature vector  $\mathbf{z} = [z_1, z_2, \dots, z_m]^T$  with padding. We employ  $d$  filters to produce a feature map  $\mathbf{Z} \in \mathbb{R}^{d \times M}$  in the semantic space. Afterwards, we employ a max-over-time pooling operation over each feature map and capture the maximum value  $\hat{z} = \max(\mathbf{z}_i)$ . By doing this, we obtain a latent semantic vector  $\mathbf{z}^s = [\hat{z}_1, \hat{z}_2, \dots, \hat{z}_d]$ . For the Bert model, we retrieve the [CLS] token as the sentence representation for further processing.

### 3.2.3 Self-attention block

We apply a scaled dot-product attention block [40], or self-attention, over EmoChannel vectors to extract emotional dependency,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}V\right), \tag{4}$$

where  $Q, K$ , and  $V$  are identical EmoChannel matrix, and  $\sqrt{d_k}$  is the square root of  $K$ 's dimension as a scaling factor.

The intensity variation within a channel reveals the changes of an emotion at different positions, which is helpful to understand the semantic compositionality and sentiment deviation. Therefore, we employ a multi-head self-attention module to analyze emotional dependency at different subspaces of EmoChannel matrix. We first linearly project the queries  $Q$ , keys  $K$ , and values  $V$  into  $d_k, d_k$ , and  $d_v$  dimensions for  $h$  times, respectively. Then, we compute the self-attention matrix as follows:

$$\mathbf{z}^s = \text{MultiHead}(Q, K, V) = \text{Dense}(\text{Concat}(\text{head}_1 \cdots \text{head}_h)), \tag{5}$$

where  $\text{head}_i = \text{Attention}(\text{Dense}_i(Q), \text{Dense}_i(K), \text{Dense}_i(V))$ .

The latent semantic representation  $\mathbf{z}^s$  and the output of attention layer  $\mathbf{z}^a$  are concatenated as an enhanced feature vector of the sentence and fed into the classifier.

### 3.2.4 Output layer & loss function

The obtained representation  $\mathbf{a}$  is mapped into the label space by passing through the fully-connected layers and a softmax layer for predicting the labels. To optimize the classification model, we maximize the probability of the correct label  $y$  by minimizing the cross-entropy loss  $L$ , which is defined as:

$$L(\mathbf{a}, y) = -\frac{1}{N} \cdot \sum_i^N \sum_j^c \mathbf{1}(y_i = j) \ln \mathbf{a}_j, \quad (6)$$

where  $\mathbf{1}(y_i = j) = 1$  if  $y_i = j$  and  $\mathbf{1}(y_i = j) = 0$ , otherwise.

## 4 Experiments

### 4.1 Datasets

We conduct experiments on different classification tasks to test the generality of the proposed method. The summary statistics of datasets are listed in Table 1.

#### 4.1.1 Multi-class classification task

We experimented on three multi-class datasets.

**ISEAR** contains descriptions about personal experience from personnel with and without psychological background, in which they had experienced according to seven emotions, i.e., “Anger”, “Disgust”, “Fear”, “Joy”, “Sadness”, “Shame”, “Guilt” [34].

**TEC** includes emotional tweets with prespecified hashtags. Each tweet is labeled by one of the six emotions, i.e., “Anger”, “Disgust”, “Fear”, “Joy”, “Sadness”, “Surprise” [25].

**Table 1** Statistics summary for the datasets

Dataset	Class	V	N	Test
ISEAR	7	8,741	7,444	CV
TEC	6	21,862	21,051	CV
CrowdFlower	5	16,588	40,000	CV
SST-1	5	17,089	11,855	2,210
SST-2	2	11,501	9,613	1,821

Class: Number of classes

V: Vocabulary size

N: Training set or dataset size

Test: Size of test set if train/test split is available

CV: Apply cross-validation



**CrowdFlower** (CF) includes tweets annotated by crowd-sourcing with 12 emotions. A subset of **CF** with emotions of *happiness*, *worry*, *surprise*, *love*, and *sadness* is adopted due to limited samples of the remaining emotions.

#### 4.1.2 Sentiment analysis task

The above datasets with coarse-grained emotional labels are used to test the performance on emotion classification task. We further demonstrate the generality of the proposed method on sentiment analysis task with sentiment polarity labels. We conducted experiments on **SST-1**<sup>1</sup> with five sentiment labels [36], *i.e.*, very negative/negative, very positive/positive, neutral, and **SST-2** with binary labels.

### 4.2 Baselines

For multi-class classification task, we evaluate our model with the following recent baselines on ISEAR, TEC, and CF:

**DPCNN** [16] exploits word-level pyramid convolutional model employing downsampling and shortcut connections.

**ASEDS** [31] is a sentiment representation learning model based on Facebook posts and reactions.

**DERNN** [42] is a representation learning model encoding both sentence syntactic dependency and document topical information.

**WLTM** [28] addresses the data sparsity issue of sentiment mining using topic modeling.

**Word Rep.** [2] examines multiple representation-based models applied to affective systems and investigates their effectiveness.

**TESAN** [41] is a deep topic model encoding topical information into topic embedding. A self-attention module and a fusion gate are proposed to predict the emotion label.

**DACNN** [44] applies a multi-channel convolutional network with attention to improve emotion categorization performance.

**ESTeR** [12] is a unsupervised emotion detection model incorporating word co-occurrences and word associations based on a word graph.

**WED** [24] adopts domain knowledge and existing lexicons to generate an affective representation of a word using fine-grained emotion concepts. Distribution learning is adopted to facilitate the classification task.

**AGN** [21] employs a variational autoencoder to encode the corpus-level word-to-label frequency and proposes an Adaptive Gate Network to consolidate semantic representation with statistical features selectively.

We also compared with the widely-adopted text classification baselines, *i.e.*, **TextCNN** [18], **Bi-LSTM** [13], **C-LSTM** [45], **Transformer** [40], and **Bert** [7], on multi-class datasets and sentiment analysis datasets.

### 4.3 Evaluation metrics

We adopt the **accuracy** and **Macro F1 score** to evaluate the model performance. Furthermore, we conducted **T-test** and report *p*-value on datasets without a standard train/test split to reveal how significant the improvements are against the baseline models.

<sup>1</sup><http://nlp.stanford.edu/sentiment/>

#### 4.4 Word embedding and hyperparameter setting

To focus on the effect of the proposed framework, we randomly initialize word embedding vectors (with a size of 300, except for Bert) to eliminate the influence of using different pre-trained language models. The preprocessing of all datasets follows the procedures reported in [18].

The hyperparameters involved are set as follows. The CNN-based models have a filter size of [3, 4, 5] with 100 filters of each, and the RNN-based models have hidden dimension of 128. For the Transformer, we use an encoder with 8 heads and 3 blocks. The employed Bert model is the Bert-base Uncased, including 12 layers, 768 hidden units, and 110M parameters. We adopt Adam optimizer with a batch size of 64 for non-Bert models and 16 for Bert models. The model dropout rate is set to 0.5. For the leaky dropout layer, the rate is 0.1, and the leaky parameter  $c$  is set to 200 according to our results on a validation trial. We select three fine-grained concepts for each emotion label because we can only identify three connections for some coarse-grained labels.

#### 4.5 Experiment results

The results of our proposed EmoChannel-SA model against other baseline methods were reported in Tables 2 and 3. In general, our proposed method can achieve the best results in most of the comparisons (except Accuracy on TEC and F1 score on SST-1), which means that the proposed method can produce comparable results against the state-of-the-art baselines on both multi-class datasets and sentiment analysis datasets. For the Bert-based method, our proposed method also yields substantial improvement to the Bert baseline.

Besides the aforementioned baseline models, we conducted experiments with two variants on ISEAR dataset for ablation study. We compare different attention mechanisms, i.e., additive attention (AA) from [23] and self-attention (SA) of the proposed model, to investigate the effects of incorporating different dependency relations. We also tested the performance of the models with self-attention blocks only or additive attention only to evaluate how informative the EmoChannel vectors are. The ablation study's results were reported in Table 4.

### 5 Discussion

In what follows, we first elaborate on how our proposed approach promotes the prediction performance by learning the dependencies among fine-grained emotions via self-attention module, and then we illustrate the specific effect of the self-attention modules through case study from multiple perspectives.

#### 5.1 Effect of adopting self-attention module

According to the results in Tables 2 and 3, we observed that EmoChannel with self-attention module yields impressive improvements against TextCNN on all datasets. Especially, compared with TextCNN, the proposed method using CNN as the feature extractor increases accuracy and F1 score by 1.48% and 1.61% on ISEAR, and 2.44% and 4.63% on TEC, respectively, indicating that the proposed method is effective. Moreover, using self-attention and EmoChannel matrix only can achieve 51.31% on accuracy and 52.72% on F1 according

**Table 2** Results on multi-class classification task

Model	ISEAR		TEC		CF		
	Accu.	F1	Accu.	F1	Accu.	F1	
(results from references)							
ASEDS	( <sup>18</sup> )	–	42.2	–	47.3	–	–
DERNN	( <sup>19</sup> )	–	60.44	–	–	–	–
WLTm	( <sup>19</sup> )	36.50	–	–	–	–	–
Word Rep	( <sup>20</sup> )	–	59.48	–	–	–	–
TESAN	( <sup>20</sup> )	61.14	–	–	–	–	–
DACNN	( <sup>20</sup> )	–	–	62.73	–	–	–
ESTeR	( <sup>20</sup> )	–	–	–	39.8	–	42.2
(reproduced results)							
TextCNN	( <sup>14</sup> )	61.55 ±0.94	61.19 ±0.98	59.08 ±1.07	47.70 ±1.52	45.08 ±0.95	39.18 ±0.86
BiLSTM	( <sup>13</sup> )	59.25 ±1.75	59.06 ±1.93	59.57 ±1.04	49.18 ±1.45	42.84 ±1.20	39.60 ±1.36
CLSTM	( <sup>15</sup> )	58.66 ±1.32	58.47 ±1.36	58.79 ±1.13	48.25 ±1.37	43.23 ±1.19	39.49 ±1.18
DPCNN	( <sup>17</sup> )	60.36 ±1.59	60.10 ±1.80	58.93 ±0.87	47.42 ±1.53	44.71 ±0.72	39.54 ±0.81
Transformer	( <sup>18</sup> )	61.07 ±1.20	60.60 ±1.31	60.17 ±1.29	49.28 ±1.59	45.29 ±1.44	40.47 ±0.88
AGN	( <sup>20</sup> )	62.87 ±1.48	61.41 ±1.26	<b>62.16</b> ±1.33	52.13 ±1.62	45.34 ±1.38	40.53 ±1.18
WED	( <sup>21</sup> )	62.19 ±0.87	61.24 ±0.76	61.58 ±0.97	51.18 ±1.62	45.12 ±1.54	40.24 ±1.18
EmoChannel-SA							
Ours		<b>63.03</b> <sup>§</sup> ±1.17	<b>62.80</b> <sup>§</sup> ±1.28	61.52 ±0.63	<b>52.33</b> <sup>†</sup> ±1.42	<b>45.59</b> ±1.01	<b>40.79</b> ±1.38
Bert	( <sup>19</sup> )	65.33	65.35	64.37	55.58	47.98	44.12
Bert+Ours		<b>65.88</b> <sup>†</sup>	<b>65.52</b> <sup>†</sup>	<b>64.89</b> <sup>†</sup>	<b>55.72</b> <sup>†</sup>	<b>48.36</b>	<b>45.62</b>

<sup>†</sup>  $p < .05$ , <sup>‡</sup>  $p < .01$ , <sup>§</sup>  $p < .001$ 

to the ablation results in Table 4. Therefore, we can conclude that it is beneficial to exploit the dependency within fine-grained emotions to understand of sentence polarity.

Moreover, we noticed that the proposed model performs differently on dataset aspect. Particularly, the improvements against baseline models on tweets datasets, such as TEC and CrowdFlower, were less impressive compared with those on ISEAR. We speculate that such a difference is caused by the special language profile of tweets data. The informal language usage of tweets causes trouble on machine understanding due to the increasing difficulty when identifying affective words, which makes it hard for the self-attention module to extract meaningful emotion dependencies. Furthermore, a large quantity of typos and

**Table 3** Results on sentiment analysis task.

Model		SST-1		SST-2	
		Accu.	F1	Accu.	F1
TextCNN	( <sup>†</sup> 14)	41.72	37.81	81.37	81.35
BiLSTM	( <sup>†</sup> 13)	41.85	38.82	81.13	81.10
CLSTM	( <sup>†</sup> 15)	42.13	<b>39.15</b>	80.76	80.74
Transformer	( <sup>†</sup> 18)	42.31	38.83	80.28	80.16
Ours		<b>42.49<sup>†</sup></b>	39.00	<b>81.84</b>	<b>81.84<sup>†</sup></b>
Bert	( <sup>†</sup> 19)	53.20	51.00	91.20	91.20
Bert+Ours		<b>53.52<sup>†</sup></b>	<b>51.37<sup>†</sup></b>	<b>91.78<sup>†</sup></b>	<b>91.77<sup>†</sup></b>

<sup>†</sup> $p < .05$ , <sup>‡</sup> $p < .01$ , <sup>§</sup> $p < .001$

slang in twitter posts leads to a great amount of vacancies in the EmoChannel vector, which can bias the emotion construction of a sentence. In contrast, the AGN baseline, which is a self-attention-based model with statistical features, produced competitive results on tweets datasets. Therefore, statistical information has potential to benefit EmoChannel framework on such challenging datasets, where emotion dependency is difficult to extract.

We also observe that the proposed method does not outperform baseline model regarding F1 score on SST-1. We speculate that the reason is that the dependency within different EmoChannels cannot explicitly present sentiment orientation of the sentence. In contrast, Li et al. [24] indicate that incorporating the sequential feature of the emotional word embedding, which exploits the emotion distribution from the other perspective, can yield significant improvements on sentiment analysis task. Therefore, we think it does not mean that the EmoChannel method is not informative. Instead, it would be beneficial to refine the EmoChannel information to fit the sentiment analysis task like [24].

## 5.2 Comparison between additive attention and self-attention

We compare the models with different attention mechanisms and report the results in Table 4. The model with self-attention shows substantial improvements to the model with additive attention, which indicates self-attention can better extract and exploit dependency relationship within EmoChannels than additive attention for the classification task. In [23], additive attention calculates a weighted arithmetic mean of aligned EmoChannels, and

**Table 4** Results of ablation study

Model	ISEAR	
	Accu.	F1
TextCNN	61.55	61.19
EmoChannel		
AA w/o WE	32.12	29.58
AA	62.13	61.83
SA w/o WE	51.31	52.72
Ours (SA)	<b>63.03</b>	<b>62.80</b>

WE stands for word embedding, AA stands for additive attention [23], and SA stands for self-attention

the weights are chosen according to the relevance of each EmoChannel towards the final emotional latent feature vector. The additive attention module is expected to highlight the fine-grained emotions that have more contributions to the label emotion. However, such an approach cannot utilize the dependency within fine-grained emotions and reveal how different emotion constructions interact. Furthermore, the EmoChannel vector cannot reflect a global sentiment deviation at the sentence-level, thus the incremented EmoChannel vector can be less informative, which can explain why the model employing additive attention without word embedding presents low performance. In contrast, the model employing self-attention only achieves satisfactory results, indicating the self-attention module is superior in extracting valid features from EmoChannel vectors.

### 5.3 Case study

To better illustrate how the self-attention contributes to the classification task and how it can be affected by the other components, we visualize the resultant self-attention weights among the fine-grained emotions of several testing data samples from the **ISEAR** dataset, and analyze them in the following section. Specifically, the self-attention weights are visualized via a heat map, where a brighter entry denotes a larger weight and a darker entry denotes a smaller weight.

As described in Section 3.1, we have in total 151 fine-grained emotions, and a  $151 \times 151$  heat map may not be straightforward for us to analyze the dependencies among these fine-grained emotions. Therefore, we select several representatives for each emotion type contained in the **ISEAR** dataset to construct the heat map. Since there are relatively more fine-grained emotions related to *joy* than the other labeled emotion types of the **ISEAR** dataset, we select 6 fine-grained emotions for *joy* and 3 fine-grained emotions for the other emotion types, the details of which can be found in Table 5. With these selected fine-grained emotions, we construct a heat map with size  $24 \times 24$  for each selected testing sample, and show them together with their corresponding sentence content and emotion labels.

Next, we will analyze the selected testing samples from multiple perspectives. To be more specific, we focus on the following three questions concerning the effect of the proposed self-attention module: (1) what self-attention learns during training; (2) the effect of semantic features on the self-attention module; (3) the effect of pre-trained word embedding vectors on the self-attention module. For the first question, we aim to illustrate that the proposed self-attention module has learned the dependencies among the fine-grained emotions and hence improved the performance of the emotion classification tasks. As for the other two questions, we focus on studying the working mechanism of the proposed self-attention

**Table 5** The selected fine-grained emotion tags for each coarse-grained emotion label in Dataset ISEAR

Labels	Selected tags
<i>anger</i>	[1] Angry [2] Irritated [3] Cold-anger
<i>disgust</i>	[4] Disgusted [5] Nauseated [6] Disdainful
<i>fear</i>	[7] Fearful [8] Terrified [9] Anguished
<i>joy</i>	[10] Joy [11] Leisurely [12] Enjoyment [13] Free [14] Lucky [15] Love
<i>sadness</i>	[16] Sad [17] Upset [18] Depressed
<i>shame</i>	[19] Shame [20] Humiliated [21] Embarrassed
<i>guilt</i>	[22] Guilty [23] Regretful [24] Sinful

module, that is, whether the propose self-attention module can still perform as expected if without semantic features or pre-trained word embedding. The above mentioned questions will be addressed one-by-one in the following.

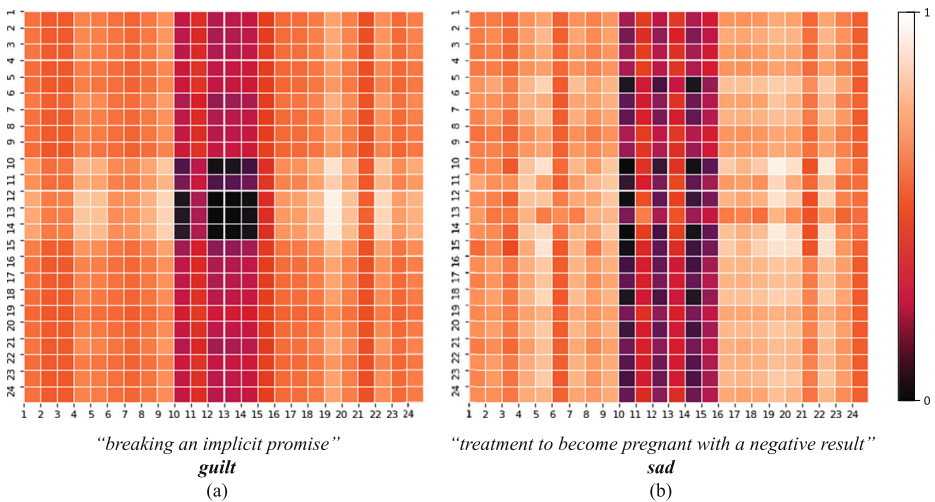
### 5.3.1 What self-attention learns

To begin with, we utilize 2 examples to visualize what our SA module has learnt during prediction. Figure 4 (a) contains a case about breaking a promise and is labeled to have a *guilt* emotion. In the heat map, for each fine-grained emotion, the attention weights of the fine-grained emotions related to *joy* (i.e., from the 10<sup>th</sup> one to the 15<sup>th</sup> one) are relatively smaller, indicated by their darker entries. This naturally makes sense since *joy* is the least likely emotion type that this sentence can convey.

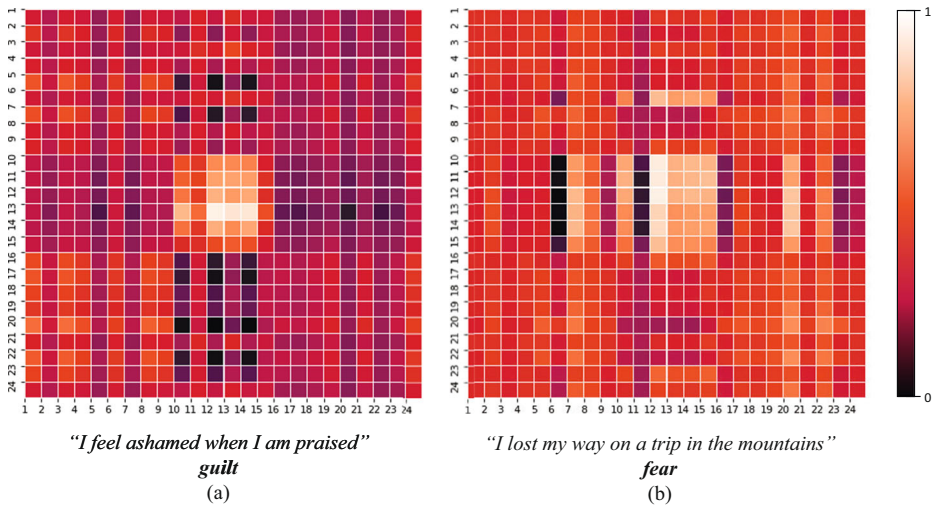
Similar phenomena can also be observed in Figure 4 (b), which describes a situation with *sad* emotion that someone finds out that her pregnancy test turns out to be negative. Under such a situation, the fine-grained emotions that are the least related are still the ones of *joy*. Hence the entries from the 10<sup>th</sup> column to the 15<sup>th</sup> one are relatively darker than those of the others.

Such observations reveal that our self-attention module can suppress the contributions from the fine-grained emotions of the emotion type that is the least likely to occur, and promotes the contributions of the fine-grained emotions related to those emotion types that are similar to the labeled one.

Interestingly, under some specific cases, we can have completely different observations. Let us take a look at Figure 5. Figure 5 (a) presents a case about being ashamed when praised, which is labeled to have a *guilt* emotion. However, the heat map shows that the largest attention weights lie on the fine-grained emotions related to *joy* instead of those related to *guilt* or *shame*. We think the reason behind is as follows. The input of the



**Figure 4** Example sentences with their heat maps, content, and their emotion labels. (a) “breaking an implicit promise” is likely to convey a guilty emotion, and hence the fine-grained emotions related to *guilt* and also other similar coarse-grained emotions all receive high attention values; (b) “treatment to become pregnant with a negative result” can be deduced to express a sad emotion, and therefore only the fine-grained emotions related to *joy* are suppressed



**Figure 5** Example sentences with their heat maps, content, and emotion labels. (a) “I feel ashamed when I am praised” is an example with two explicit emotive words, “ashamed” and “praised”, and hence the model gets confused by such a conflict and fails to learn informative dependencies among fine-grained emotions; (b) “I lost my way on a trip in the mountains” is another example expressing two explicit types of emotions conveyed by “lost” and “trip”, and hence the model gets confused, incorrectly highlighting the fine-grained emotions of *joy* instead of *fear*

self-attention module is the *EmoChannel* vectors aggregating the word-level emotion information. We notice that there are two explicit emotive words in the sentence, “ashamed” and “praised”, which conveys *shame* and *joy* emotions, respectively, different from the labeled *guilt* type. Hence the self-attention module gets confused by the conflict between the labeled emotion label and the word-level emotion information, and fail to learn informative dependencies among fine-grained emotions.

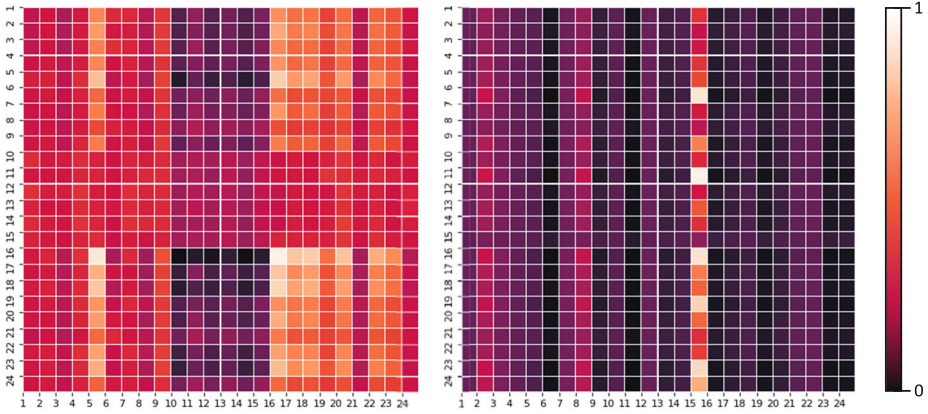
Similar conclusion can be drawn from Figure 5 (b). The example of Figure 5 (a) is about getting lost in a trip and the label emotion is *fear*, but the heat map shows that the fine-grained emotions of *joy* receive relatively larger attention weights than the others. We find that there are also two explicit emotive words in the sentence, “lost” and “trip”, expressing *sad* and *joy*, respectively. Therefore, the self-attention module is confused by such inconsistency and cannot learn the expected dependencies among emotions.

To sum up, in general the self-attention module can learn to assign more weights to the fine-grained emotions that are similar to the labeled emotion type, and suppress the attentions from those of the least possible emotion type. However, when the input sentence itself contains explicit emotive emotions expressing different emotion types from the labeled one, the self-attention module may not learn informative dependencies among fine-grained emotions, under which case the final prediction shall rely more on the semantic feature vectors.

### 5.3.2 Effect of semantic features on self-attention

To validate whether the self-attention module can still learn meaningful dependencies among fine-grained emotions without the semantic feature vectors (Section 3.2.2), we select 2 testing samples to compare their heat maps achieved by the proposed approach with the ones achieved by removing the word embedding vectors.

Figure 6 (a) is about a case that someone has to walk through a field with wild bulls since his/her car breaks down, which shall lead to *fear* or *sad*. The labeled emotion is *fear* and the heat map with semantic feature taken into consideration shows that for each fine-grained emotions, the contributions from the fine-grained emotions related to *joy* are all suppressed,



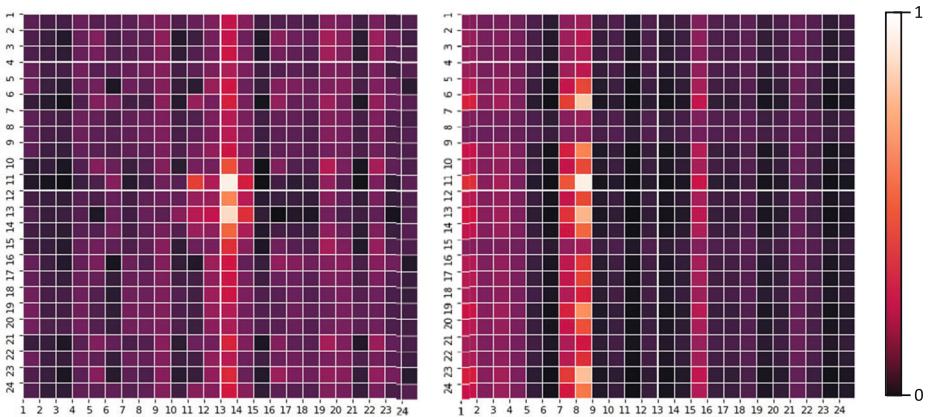
with semantic feature

without semantic feature

“When I had to walk over three kms through a field which had wild bulls, because my car had broken down”

*fear*

(a)



with semantic feature

without semantic feature

“I received a letter from a friend whom I was missing”

*joy*

(b)

**Figure 6** Example sentences of their content, emotion labels, and the heat maps resulted from including or removing semantic features. (a) is an example about walking over a field with wild bulls because of car breaking down, which naturally conveys a *fear* emotion. (b) is another example about receiving a letter from a friend, which is happy. The heatmaps obtained by using semantic features of both examples illustrate the correctly captured fine-grained emotions, while the ones without semantic features highlight the incorrect fine-grained emotions



while the other fine-grained emotions are assigned with larger attention weights. However, if we relies only on the self-attention module without concatenating the semantic feature vectors, the learnt attention weights among fine-grained emotions can be biased towards a single fine-grained emotion and not credible, such as the dramatically large attention of the 15<sup>th</sup> fine-grained emotion in the right sub-figure of Figure 6 (a).

Another selected example is Figure 6 (b), which describes that someone receives a letter from a friend. The labeled emotion type is *joy*, and the heat map with semantic features involved also reveals the same phenomenon that the largest attention weights are assigned to the 13<sup>th</sup> fine-grained emotion related to *joy*. However, by the self-attention module alone, the learnt self-attention weights indicate that the 7<sup>th</sup> and the 8<sup>th</sup> fine-grained emotions contribute the most to the final prediction, which does not make sense since these two fine-grained emotions are related to *fear*.

Therefore, we conclude that our self-attention module needs to work together with the semantic feature vectors to learn informative dependencies among fine-grained emotions for promoting the emotion classification performance. Without the word embedding vectors, the self-attention module may be biased towards some unrelated fine-grained emotions and fail to benefit the whole model.

### 5.3.3 Effect of pre-trained word embedding vectors

After showing that semantic feature vectors are indispensable for our self-attention module, we further explore the effect of using pre-trained word embedding vectors when computing semantic feature vectors. Specifically, we select 2 testing samples to compare their heat maps resulted from using pre-trained word embedding vectors and those of using random word embedding vectors.

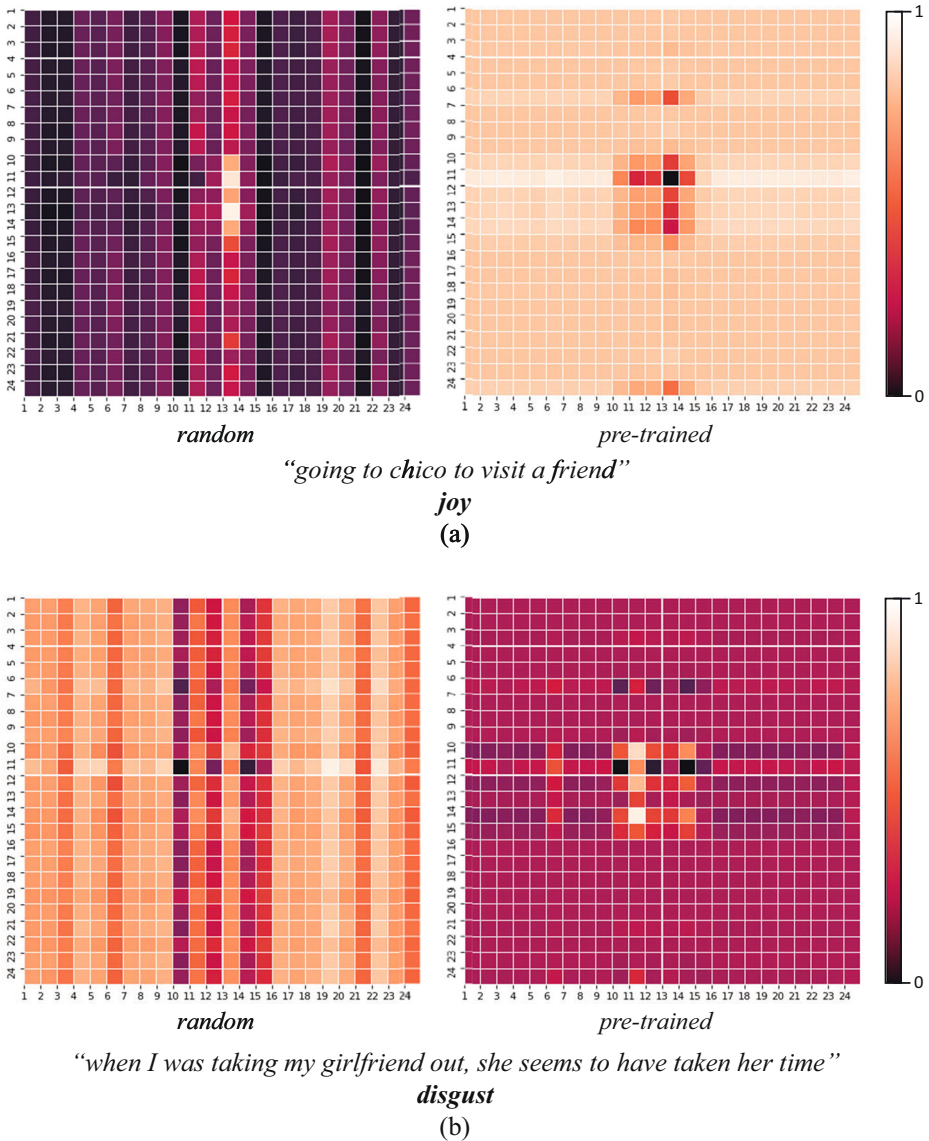
The first example is Figure 7 (a), which is about visiting a friend and is labeled to have *joy* emotion. The heat map of using random word embedding vectors shows that the fine-grained emotions related to *joy* contribute the most to all the fine-grained emotions, especially the 13<sup>th</sup> fine-grained emotion. However, the heat map of using pre-trained word embedding vectors looks completely different, where the contributions of the fine-grained emotion related to *joy* are suppressed.

Similarly, Figure 7 (b) describes a case that someone's girlfriend keeps him waiting when he tries to take her out. The labeled emotion is *disgust*, and the heat map of random embedding vectors shows that the fine-grained emotions related to *joy* are assigned with smaller attention weights. On the contrary, when using pre-trained word embedding vectors, the fine-grained emotions related to *joy* are assigned with larger weights than the others, which conflicts with the observation of using random word embedding vectors.

We think such interesting conflicts are caused by that the pre-trained word embedding vectors can already provide strong and informative semantic features. With these pre-trained word embedding vectors, the model learns to rely more on the semantic feature vectors. Hence the self-attention module is not trained enough and fails to perform as expected. Therefore, to guarantee that the self-attention module can get enough training, using random word embedding vectors is a better choice.

### 5.3.4 Emotion distribution

Since the self-attention block is directly employed over EmoChannel matrix, the attention weights can reflect an integration of all the affective words in a sentence, thanks to the inner product between queries and keys, as shown in Figure 4. As analyzed in Section 5.3.1,



**Figure 7** Example sentences of their content, emotion labels, and the heat maps using random or pre-trained word embedding vectors. (a) is an example about going to visit a friend, which conveys a *joy* emotion, while (b) is another example about feeling unhappy about waiting his girlfriend. The heatmaps obtained by pre-trained word embeddings of both examples seem to focus on conflicting fine-grained emotions against the labeled coarse-grained emotions, while the ones using random word embeddings show more reasonable attention weights

the self-attention module can learn to assign less weights to the fine-grained emotions that are the least likely to occur, and increase the weights of the other fine-grained emotions. Moreover, according to our discussion in Section 5.3.3, training without randomly initialized word embedding vectors encourages the self-attention module to encode semantic information, so that the attention map is able to present the real sentiment composition to

some extent. Therefore, the learnt self-attention weights can be viewed as a modification or adaptation towards the inherent word-level emotion information with the sentence's emotion label. The result of modified emotion information could be a natural fit as an emotion distribution. In the next step of our research, we plan to investigate the potential connection between attention map and sentence-level emotion distribution by implementing dimension deduction method on the attention map, which could be a feasible approach to construct a distribution label that facilitates the classification task.

## 6 Conclusion and future work

In this work, we proposed an *EmoChannel-SA* Network to enhance emotion classification performance by exploiting the dependency relationship with the emotion construction of the text. We examined 151 fine-grained emotions incorporating domain knowledge and a dimensional lexicon dictionary. Our experimental results show that the proposed method leads to a more robust classifier on datasets of various topics. To examine the generality of our proposed model, we conduct experiments on multi-class classification and sentiment analysis tasks. The experiment results indicate our model can produce competitive performance against state-of-the-art baselines. Furthermore, we provide in-depth discussions about how the interaction within fine-grained emotions affects decision-making by visualizing self-attention weights. We conclude that the self-attention module can learn to assign more weights to the fine-grained emotions that are similar to the labeled emotion type, and suppress the attentions from those of the least possible emotion type. Meanwhile, the learnt self-attention weights can be regarded as a modification towards the inherent word-level emotion information with the sentence's emotion label. Therefore, the attention map has potential to shed lights on the exploration of a new approach towards emotion distribution learning. In future work, we plan to investigate the possible connection between self-attention weights and emotion distribution, which would be beneficial to enhance the information of the one-hot label.

**Acknowledgements** This journal paper is an extension based on a conference paper, entitled “Emochannelattn: Exploring emotional construction towards multi-class emotion classification”, which has been published in the 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) [23]. For a concise expression and the consistency with the original paper, we reuse a part of the content from the conference version. We have significantly revised the conference version by proposing a novel model, adding extensive experiments, and discussing parameter issues. Compared to the conference version, there are at least 70% new content in this extended journal version. Some notations, equations, algorithms and so on are reused for smooth presentation. The research has been supported by the One-off Special Fund from Central and Faculty Fund in Support of Research from 2019/20 to 2021/22 (MIT02/19-20), the Research Cluster Fund (RG 78/2019-2020R), and the Interdisciplinary Research Scheme of the Dean's Research Fund 2019-20 (FLASS/DRF/IDS-2) of The Education University of Hong Kong, the Hong Kong Research Grants Council under the Collaborative Research Fund (project number: C1031-18G), and the Direct Grant (DR21A5) and the Faculty Research Grant (DB21A9) of Lingnan University, Hong Kong.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References


1. Agrawal, A., An, A., Papagelis, M.: Learning emotion-enriched word representations. In: Proceedings of the 27th International Conference on Computational Linguistics, pp. 950–961. Association for Computational Linguistics (2018)
2. Babanejad, N., Agrawal, A., An, A., Papagelis, M.: A comprehensive analysis of preprocessing for word representation learning in affective tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5799–5810. Association for Computational Linguistics (2020)
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference on Learning Representations (2015)
4. Chen, X., Li, Q., Wang, J.: A unified sequence labeling model for emotion cause pair extraction. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 208–218. International Committee on Computational Linguistics, Barcelona, Spain (Online). <https://www.aclweb.org/anthology/2020.coling-main.18> (2020)
5. Chung, J., Gülçehre, Ç., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv:1412.3555 (2014)
6. Devlin, J., Chang, M., Lee, K., Toutanova, K.: 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, pp. 4171–4186 (2019)
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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. Association for Computational Linguistics (2019)
8. Ekkekakis, P., Russell, J.A.: The Measurement of Affect, Mood, and Emotion: A Guide for Health-Behavioral Research. Cambridge University Press, Cambridge (2013). <https://doi.org/10.1017/CBO9780511820724>
9. Ekman, P.: An argument for basic emotions. *Cognition & Emotion* **6**(3–4), 169–200 (1992)
10. Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., Lehmann, S.: Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 1615–1625. Association for Computational Linguistics (2017)
11. FENG, J., RAO, Y., XIE, H., WANG, F., LI, Q.: User group based emotion detection and topic discovery over short text. *World Wide Web* **23**(3), 1553–1587 (2020). <https://doi.org/10.1007/s11280-019-00760-3>
12. Gollapalli, S.D., Rozenshtein, P., Ng, S.K.: ESTeR: Combining word co-occurrences and word associations for unsupervised emotion detection. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 1043–1056. Association for Computational Linguistics (2020)
13. Graves, A., Jaitly, N., Mohamed, A.: Hybrid speech recognition with deep bidirectional lstm. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 273–278 (2013). <https://doi.org/10.1109/ASRU.2013.6707742>
14. Hochreiter, S., Schmidhuber, J.: Long short-term memory. *Neural Computing* **9**(8), 1735–1780 (1997). <https://doi.org/10.1162/neco.1997.9.8.1735>
15. Hu, X., Tang, J., Gao, H., Liu, H.: Unsupervised sentiment analysis with emotional signals. In: Proceedings of the 22nd International Conference on World Wide Web, WWW '13, pp. 607–618. Association for Computing Machinery (2013)
16. Johnson, R., Zhang, T.: Deep pyramid convolutional neural networks for text categorization. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 562–570. Association for Computational Linguistics (2017)
17. Khatua, A., Khatua, A., Cambria, E.: Predicting political sentiments of voters from twitter in multi-party contexts. *Appl. Soft Comput.* **97**, 106743 (2020). <https://doi.org/10.1016/j.asoc.2020.106743>
18. Kim, Y.: Convolutional neural networks for sentence classification. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1746–1751. Association for Computational Linguistics (2014)
19. Labetov, I., Lipson, H.: Re-embedding words. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 489–493. Association for Computational Linguistics (2013)
20. Lai, Y., Zhang, L., Han, D., Zhou, R., Wang, G.: Fine-grained emotion classification of chinese microblogs based on graph convolution networks. *World Wide Web* **23**(5), 2771–2787 (2020)
21. Li, X., Li, Z., Xie, H., Li, Q.: Merging statistical feature via adaptive gate for improved text classification. Proceedings of the AAAI Conference on Artificial Intelligence **35**(15), 13288–13296 (2021). <https://ojs.aaai.org/index.php/AAAI/article/view/17569>

22. Li, X., Xie, H., Lau, R.Y.K., Wong, T., Wang, F.L.: Stock prediction via sentimental transfer learning. *IEEE Access* **6**, 73110–73118 (2018)
23. Li, Z., Chen, X., Xie, H., Li, Q., Tao, X.: Emochannelattn: Exploring emotional construction towards multi-class emotion classification. In: 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), pp. 242–249 (2020). <https://doi.org/10.1109/WIIAT50758.2020.00036>
24. Li, Z., Xie, H., Cheng, G., Li, Q.: Word-level emotion distribution with two schemas for short text emotion classification. *Knowledge-Based Systems* p. 107163. <https://doi.org/10.1016/j.knsys.2021.107163>. <https://www.sciencedirect.com/science/article/pii/S0950705121004263> (2021)
25. Mohammad, S.M.: Emotional tweets. In: Proceedings of the First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation, pp. 246–255. Association for Computational Linguistics (2012)
26. Mohammad, S.M.: Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 174–184. Association for Computational Linguistics (2018)
27. Mudinas, A., Zhang, D., Levene, M.: Combining lexicon and learning based approaches for concept-level sentiment analysis. In: Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining, WISDOM '12. Association for Computing Machinery (2012). <https://doi.org/10.1145/2346676.2346681>
28. Pang, J., Rao, Y., Xie, H., Wang, X., Wang, F.L., Wong, T.L., Li, Q.: Fast supervised topic models for short text emotion detection. *IEEE Trans. Cybern.*: 1–14 (2019)
29. Posner, J., Russell, J.A., Peterson, B.S.: The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and Psychopathology* **17**(3), 715–734 (2005)
30. Qian, Q., Huang, M., Lei, J., Zhu, X.: Linguistically regularized LSTM for sentiment classification. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1679–1689. Association for Computational Linguistics, Vancouver, Canada (2017). <https://doi.org/10.18653/v1/P17-1154>
31. Raad, B., Philipp, B., Patrick, H., Christoph, M.: Aseds: Towards automatic social emotion detection system using facebook reactions. In: 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), pp. 860–866. IEEE Computer Society (2018)
32. Russell, J.A.: A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**(6), 1161 (1980)
33. Russell, J.A., Mehrabian, A.: Evidence for a three-factor theory of emotions. *J. Res. Pers.* **11**(3), 273–294 (1977)
34. Scherer, K.R., Wallbott, H.G.: Evidence for universality and cultural variation of differential emotion response patterning. *J. Pers. Soc. Psychol.* **66**(2), 310–328 (1994)
35. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1631–1642 (2013)
36. Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C.D., Ng, A., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1631–1642 (2013)
37. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M.: Lexicon-based methods for sentiment analysis. *Comput. Linguist.* **37**(2), 267–307 (2011)
38. Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., Qin, B.: Sentiment embeddings with applications to sentiment analysis. *IEEE Trans. Knowl. Data Eng.* **28**(2), 496–509 (2016)
39. Teng, Z., Vo, D.T., Zhang, Y.: Context-sensitive lexicon features for neural sentiment analysis. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pp. 1629–1638. Association for Computational Linguistics (2016)
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, vol. 30, pp. 5998–6008. Curran Associates Inc. (2017)
41. Wang, C., Wang, B.: An end-to-end topic-enhanced self-attention network for social emotion classification. In: Proceedings of The Web Conference 2020, WWW '20, p. 2210–2219. Association for Computing Machinery (2020)

42. Wang, C., Wang, B., Xiang, W., Xu, M.: Encoding syntactic dependency and topical information for social emotion classification. In: Proceedings of the 42nd International ACM SIGIR Conference, SIGIR'19, pp. 881–884. Association for Computing Machinery (2019)
43. Xie, H., Li, X., Wang, T., Lau, R.Y., Wong, T.L., Chen, L., Wang, F.L., Li, Q.: Incorporating sentiment into tag-based user profiles and resource profiles for personalized search in folksonomy. *Inf. Process. Manag.* **52**(1), 61–72 (2016). <https://doi.org/10.1016/j.ipm.2015.03.001>
44. Yang, C.T., Chen, Y.L.: Dacnn: Dynamic weighted attention with multi-channel convolutional neural network for emotion recognition. In: 2020 21st IEEE International Conference on Mobile Data Management (MDM), pp. 316–321 (2020)
45. Zhou, C., Sun, C., Liu, Z., Lau, F.C.M.: A c-lstm neural network for text classification. arXiv:1511.08630 (2015)

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Affiliations

Zongxi Li<sup>1</sup> · Xinhong Chen<sup>2</sup> · Haoran Xie<sup>3</sup> · Qing Li<sup>4</sup> · Xiaohui Tao<sup>5</sup> · Gary Cheng<sup>6</sup> 

Zongxi Li  
zongxili2@gmail.com

Xinhong Chen  
xinhchen2-c@my.cityu.edu.hk

Haoran Xie  
hrxie2@gmail.com

Qing Li  
csqli@comp.polyu.edu.hk

Xiaohui Tao  
xiaohui.tao@usq.edu.au

<sup>1</sup> School of Science and Technology, Hong Kong Metropolitan University, Ho Man Tin, Hong Kong

<sup>2</sup> Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong

<sup>3</sup> Department of Computing and Decision Sciences, Lingnan University, Tuen Mun, Hong Kong

<sup>4</sup> Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Hong Kong

<sup>5</sup> School of Sciences, University of Southern Queensland, Toowoomba, Australia

<sup>6</sup> Department of Mathematics and Information Technology, The Education University of Hong Kong, Tai Po, Hong Kong