A Sentiment Similarity-oriented Attention Model with Multi-task Learning for Text-based Emotion Recognition

Yahui Fu[1], Lili Guo[1], Longbiao Wang[1(✉)], Zhilei Liu[1], and Jiaxing Liu[1], Jianwu Dang[1,2]

[1] Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, China
{fuyahui,liliguo,longbiao_wang,zhileiliu,jiaxingliu}@tju.edu.cn
[2] Japan Advanced Institute of Science and Technology, Ishikawa, Japan jdang@jaist.ac.jp

Abstract. Emotion recognition based on text modality has been one of the major topics in the field of emotion recognition in conversation. How to extract efficient emotional features is still a challenge. Previous studies utilize contextual semantics and emotion lexicon for affect modeling. However, they ignore information that may be conveyed by the emotion labels themselves. To address this problem, we propose the sentiment similarity-oriented attention (SSOA) mechanism, which uses the semantics of emotion labels to guide the model's attention when encoding the input conversations. Thus to extract emotion-related information from sentences. Then we use the convolutional neural network (CNN) to extract complex informative features. In addition, as discrete emotions are highly related with the Valence, Arousal, and Dominance (VAD) in psychophysiology, we train the VAD regression and emotion classification tasks together by using multi-task learning to extract more robust features. The proposed method outperforms the benchmarks by an absolute increase of over 3.65% in terms of the average F1 for the emotion classification task, and also outperforms previous strategies for the VAD regression task on the IEMOCAP database.

Keywords: Sentiment similarity-oriented attention · Text emotion recognition · VAD regression · Multi-task learning · Convolutional neural network

1 Introduction

Text emotion recognition has emerged as a prevalent research topic that can make some valuable contributions, not only in social media applications like Facebook, Twitter and Youtube, but also in more innovative area such as human-computer interaction. It is significant to extract effective textual features for emotion recognition but still a challenging task.

In the traditional studies, distributed representations or pre-trained embeddings are playing important roles in state-of-the-art sentiment analysis systems. For example, predictive methods Word2Vec [1] and Glove [2], which can capture multi-dimensional word semantics. Beyond word-semantics, there has been a big efforts toward End-to-End neural network models [3] and achieved better performance by fine-tuning the well pre-trained models such as ELMO [4] and BERT [5]. However, these representations are based on syntactic and semantic information, which do not enclose specific affective information.

To conduct affective information into training, [6–9] introduced lexical resources to enrich previous word distributions with sentiment-informative features, as lexical values are intuitively associated with the word's sentiment polarity and strength. Especially, [8] proposed a lexicon-based supervised attention model to extract sentiment-enriched features for document-level emotion classification. Similarly, [7] introduced a kind of affect-enriched word distribution, which was trained with lexical resources on the Valence-Arousal-Dominance dimensions. These studies demonstrate the effectiveness of sentiment lexicons in emotion recognition. However, it's limited in lexicon vocabulary coverage, and the valence of one sentence is not simply the sum of the lexical polarities of its constituent words [10]. Emojis are also thought high correlated to affect, therefore, [11] proposed a model named Deepmoji that adopted a bidirectioinal long short-term memory (BLSTM) with an

* Yahui Fu and Lili Guo have contributed equally to this work.

attention mechanism. The Deepmoji predicted emojis from text on a 1246 million tweet corpus and achieved a good results. Nevertheless, it needs huge effort to collect tweets. In addition, none of these researches consider the semantics of the emotion labels themselves.

To address the above problems, we propose a sentiment similarity-oriented attention (SSOA) mechanism, which uses the label embeddings to guide the network to extract emotion-related information from input sentences. First of all, we compute the sentiment similarity between input sentences and emotion labels. Then we apply the valence value that selected from an affective lexicon as the sentiment polarity. After training the model, we can obtain SSOA, the value of which represents the weight of each emotion contributes to the final representations. Finally, we use CNN to capture complex linguistic features as it has been wildly used for text emotion recognition and shown promising performances such as [12, 13]. Furthermore, [14] indicated that emotion state can be considered as a point in a continuous space, which is described by the dimensions of valence (V, the pleasantness of the stimulus), arousal (A, the intensity of emotion produced) and dominance (D, the degree of power/control exerted by a stimulus), meanwhile, discrete emotions are highly correlated with VAD in psychophysiology. Therefore, in this work, we adopt a multi-task model for both discrete emotion classification and dimensional VAD regression to enrich robustness.

To summarize, our main contributions are as follows: 1) we propose a sentiment similarity-oriented attention mechanism to encode sentiment-informative representations by incorporating label semantics. 2) we propose to leverage the inter-dependence of two related tasks (i.e. discrete emotion recognition and dimensional VAD recognition) in improving each other's performance. The rest of this paper is organized as follows. Section 2 introduces the proposed method, sentiment similarity-oriented attention mechanism with multi-task learning. We then conduct a series of comparative experiments and validation studies in Section 3. Section 4 gives the conclusions.

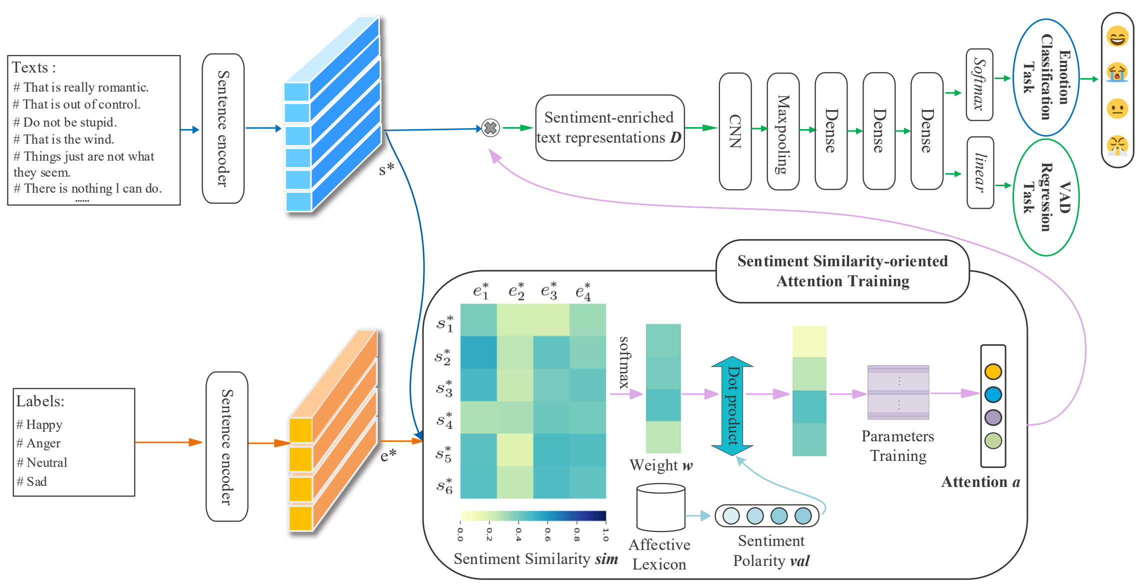


Fig. 1. This is the overall framework: sentiment similarity-oriented attention model with multi-task learning for text-based emotion recognition. We introduce sentiment similarity and sentiment polarity to compute affective attention. Then, we use this attention to construct sentiment-enriched text representations for both emotion classification and VAD regression task with multi-task learning.

2 Sentiment similarity-oriented attention model with multi-task learning

Figure 1 gives the overall framework. First, the sentence encoder approach is used to generate representations for all the input texts and emotional labels. Then we adopt the proposed sentiment similarity-oriented attention mechanism to obtain the sentiment-enriched text representations, followed by a CNN to extract deep informative features. In addition, we introduce multi-task learning for both emotion classification and VAD regression to extract more robust representations.

2.1 Sentence encoder

[15] has published two kinds of universal sentence encoder for sentence embedding, one is trained with Transformer encoder [16], while the other is based on deep averaging network (DAN) architecture [17], and all of them can be obtained from the TF Hub website. We use the first one (USE_T) for our sentence encoder part to encode texts and emotion labels into sentence embeddings. Rather than learning label embeddings from radome, we also explore using contextual embeddings from transformer-based models. This allow us to use richer semantics derived from pre-training. The reason that we use sentence embeddings not conventional pre-trained word embeddings as when computing emotion of one sentence based on word level may cause sentiment inconsistency. For example, in a sentence sample *'You are not stupid.'* word *not* and *stupid* are both represent negative emotion, if just concatenate them to represent the emotion of this sentence, it is negative, which should be positive.

2.2 Sentiment similarity-oriented attention

In this section, we introduce our proposed SSOA mechanism more explicitly. The main idea behind the SSOA mechanism is to compute affective attention scores between the labels and the representations of the input sentences that is to be classified. Formally, let $S = \{s_1 ... s_i ... s_N\}$ be the set of the sentences in the database, where N is the total number of training data set. $E = \{e_1, e_2, e_3, e_4\}$ be the set of four emotion labels (Happy, Angry, Neutral, Sad) same as in [18], $Val = \{val_1, val_2, val_3, val_4\}$ be the set of valence scores of the emotions, which selected from ANEW lexicon [19]. We define val_i as the sentiment polarity of each emotion e_j, which is a real number and indicates the strength of each emotion.

For each s_i in S, $1 \leq i \leq l$, where l is batch size. And each e_j in E, $1 \leq j \leq 4$, we directly assess their sentence embedding s_i^* and e_j^* respectively, produced by the sentence encoder. For the pairwise sentiment similarity $sim\left(s_i^*, e_j^*\right)$, we compute it based on the method proposed in [15], that first compute the cosine similarity of the sentence embedding and emotion embedding, then use arccos to convert the cosine similarity into an angular distance, which had experimented to have better performance on sentiment similarity computing, that is,

$$sim\left(s_i^*, e_j^*\right) = \left(1 - \arccos\left(\frac{s_i^{*\top} e_j^*}{\parallel s_i^* \parallel \parallel e_j^* \parallel} / \pi\right)\right) \tag{1}$$

where $s_i^{*\top}$ represents the transpose of s_i^*. For each $sim\left(s_i^*, e_j^*\right)$, we use the softmax function to compute the weight probability $w_{i,j}$ as:

$$w_{i,j} = \frac{\exp\left(sim\left(s_i^*, e_j^*\right)\right)}{\sum_{j=1}^{4} \exp\left(sim\left(s_i^*, e_j^*\right)\right)} \tag{2}$$

Then the affective attention $a_{i,j}$ that sentence s_i oriented on each emotion is computed as below:

$$a_{i,j} = \alpha * \left(val_j w_{i,j}\right) \tag{3}$$

We add a scaling hyper-parameter α to increase the range of possible probability values for each conditional probability term. The sentiment-enriched text representations D can be induced as follows:

$$D = \sum_{i=1}^{l} \sum_{j=1}^{4} W_s s_i^* a_{i,j} \tag{4}$$

where W_s denotes sentence-level weight matrices, $D \in R^{l \times 4d^s}$, and d^s is the size of sentence embedding.

2.3 Multi-task learning

In this subsection, we introduce multi-task learning for both emotion classification and VAD regression task, as the knowledge learned in one task can usually improve the performance of another related task and enrich robustness of different type tasks [20, 21]. Each sentence s_i in the training corpus has the following feature and label set $[s_i^*, (y_{emo,i}, y_{val,i}, y_{aro,i}, y_{dom,i})]$, where s_i^* represents the sentence embedding of s_i, and $(y_{emo,i}, y_{val,i}, y_{aro,i}, y_{dom,i})$ represent the associated categorical emotion, dimensional valence, arousal and dominance label separately. We apply CNN and three dense layers as informative feature extractor, then H^* is the final document vector. The probability of emotion classification task is computed by a *softmax* function:

$$P\left(y_{emo}\right) = softmax\left(W_e H^* + b_e\right) \tag{5}$$

where W_e and b_e are the parameters of the *softmax* layer. We use categorical cross entropy loss function for the first task, the objective function of this system is as follows:

$$J_e = -\frac{1}{l} \sum_{i=1}^{l} logP\left(y_{emo,i}\right)\left[y_{emo,i}\right] \tag{6}$$

where $y_{emo,i}$ is the expected class label of sentence s_i and $P\left(y_{emo,i}\right)$ is the probability distribution of emotion labels for s_i. However, for the continuous labels, the *softmax* layer is not applicable, we use the *linear* function to predict the values for the VAD regression task. Then the predict value $y^p_{val|aro|dom,i}$ for sentence s_i is calculated using the following formula:

$$y^p_{val|aro|dom,i} = linear\left(W_s h_i^* + b_s\right) \tag{7}$$

where h_i^* represents the final vector of sentence s_i, W_e and b_e represent weights and bias respectively. Given l training sentences, we use the mean squared error loss function for VAD analysis, the loss between predicted dimensional values $y^p_{val|aro|dom,i}$ and original continuous labels $y^o_{val|aro|dom,i}$ is calculated as below:

$$L_{s,val|aro|dom} = \frac{1}{3l} \sum_{i=1}^{l} \left(y^p_{val|aro|dom,i} - y^o_{val|aro|dom,i}\right)^2 \tag{8}$$

Then the objective function for the whole system is:

$$J = J_e + \beta * \left(L_{s,act} + L_{s,aro} + L_{s,dom}\right) \tag{9}$$

where β is the hyper-parameter to control the influence of the loss of the regression function to balance the preference between classification and regression disagreements.

3 Experiments and analysis

3.1 Database and lexicon

The IEMOCAP emotion database The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [22] contains videos of ten unique speakers acting in two different scenarios: scripted and improvised dialog with dyadic interactions. We only use the transcript data. To compared with state-of-the-art approaches, we use four emotion categories and three sentiment dimensions with 5531 utterances in this study. The four-class emotion distribution is: 29.6% happy, 30.9% neutral, 19.9% anger and 19.6% sad. Note that happy and excited category in the original annotation are included into happy class to balance data distribution between classes. For valence, arousal and dominance labels, self-assessment are used for annotation, in which the scale is from 1 to 5. In this paper, we focus on speaker-independent emotion recognition. We use the first eight speakers from session one to four as the training set, and session five as the test set.

The ANEW affective lexicon The emotional values of the English words in Affective Norms for English Words (ANEW) [19] were calculated by means of measuring the psychological reaction of a person to the specific word. It contains real-valued scores for valence, arousal and dominance (VAD) on a scale of 1-9 each, corresponding to the degree from low to high for each dimension respectively. We select the *Valence* rating as the sentiment polarity which can distinguish different emotions of distinct words with the scale ranging from unpleasant to pleasant.

3.2 Experimental setup

Following [15], we set the dimension of the sentence embedding to 512. We use a convolutinoal layer with 16 filters each for kernel size of (4,4) and a max-pooling layer with the size of (2,2). As for dense layers, we use three hidden dense layers with 1024, 512 and 256 units and ReLU activation [23] separately. For regularization, we employ Dropout operation [24] with dropout rate of 0.5 for each layer. We set the mini-batch size as 50 and epoch number as 120, Adam [25] optimizer with a learning rate 0.0002, clipnorm as 5. And we set the parameter β to 1.0 to control the strength of the cost function for the VAD regression task.

We evaluate the experimental results of both single-task learning (STL) and multi-task learning (MTL) architecture. In the single-task architecture, we build seperate systems for emotion classification and VAD regression, whereas in multi-task architecture a join-model is learned for both of these problems.

3.3 Experimental results and analysis

Comparison to state-of-the-art approaches: To quantitatively evaluate the performance of the proposed model, we compare our method with currently advanced approaches. The following are the commonly used benchmarks:

Tf-idf+Lexicon+DNN [9]: Introducing affective *ANEW* [19] lexicon and the term frequency-inverse document frequency (*tf-idf*) to construct the text features with DNN for emotion classification on IEMOCAP.

CNN [26]: A efficient architecture which achieves excellent results on multiple benchmarks including sentiment analysis.

LSTMs [27]: Two conventional stacked LSTM layers for emotion detection using the text transcripts of IEMOCAP.

Deepmoji [11]: Using the millions of texts on social media with emojis to pre-train the model to learn representations of emotional contents.

BiGRU+ATT [28]: A BiGRU network with the classical attention (ATT) mechanism.

BiLSTM+CNN [29]: Incorporating convolution with BiLSTM layer to sample more meaningful information.

BERT$_{BASE}$ [5]: Bidirectional encoder with 12-layer Transformer blocks, which obtains new state-of-the-art results on sentence-level sentiment analysis.

In order to evaluate the performance, we present accuracy and F1-score for emotion classification task. As for VAD regression work, we use the mean squared error (MSE) and pearson correlation coefficient (r) to evaluation the performance, in which the lower MSE value and higher r correlation, the better performance. Experimental results of different methods in single task framework are shown in Table 1 and Table 2.

Table 1. F1, Accuracy for the comparative experiments in emotion classifiation framework. Acc.=Accuracy(%), Average(w)=Weighted average(%). The best results are in bold.

ID	Model	IEMOCAP									
		Happy		Anger		Neutral		Sad		Average(W)	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
1	Tf-idf+Lexicon+DNN [9]	63.80	69.29	68.24	67.64	60.68	58.84	62.86	57.69	63.89	63.39
2	CNN [26]	64.71	69.00	72.35	64.23	60.16	59.08	62.45	62.70	64.92	63.75
3	LSTMs [27]	60.41	69.08	71.18	66.30	61.72	59.18	68.98	62.25	65.57	64.20
4	Deepmoji [11]	58.37	66.15	61.18	63.03	72.14	61.56	63.67	66.10	63.84	64.21
5	BiGRU+ATT [28]	60.18	68.73	76.47	67.01	59.64	58.79	71.02	64.33	66.83	64.72
6	BiLSTM+CNN [29]	63.57	70.60	71.76	67.59	63.80	61.17	66.53	62.21	66.42	65.40
7	BERT$_{BASE}$ [5]	59.05	69.23	72.35	65.78	67.19	63.70	73.88	66.54	68.12	66.31
Proposed	USE_T+SSOA+CNN	**69.91**	**72.88**	71.18	**70.14**	**67.71**	**65.74**	72.24	**71.08**	**70.26**	**69.96**

Table 2. MSE and r for the comparative experiments in VAD regression framework

ID	Model	IEMOCAP					
		Valence		Arousal		Dominance	
		MSE	r	MSE	r	MSE	r
1	Tf-idf+Lexicon+DNN [9]	0.755	0.435	0.536	0.277	0.638	0.318
2	CNN [26]	0.731	0.471	0.544	0.345	0.619	0.359
3	LSTMs [27]	0.626	0.575	0.413	0.425	0.536	0.447
4	Deepmoji [11]	0.655	0.499	0.417	0.421	0.514	0.458
5	BiGRU+ATT [28]	0.674	0.478	0.439	0.378	0.561	0.416
6	BiLSTM+CNN [29]	0.685	0.466	0.433	0.400	0.531	0.442
7	BERT$_{BASE}$ [5]	0.566	0.587	0.416	0.464	0.564	0.460
Proposed	USE_T+SSOA+CNN	**0.523**	**0.603**	**0.402**	0.446	**0.511**	**0.486**

As shown in Table 1, our proposed model outperforms the state-of-the-art approaches with the absolute increase of more than 3.65%, 2.14% on average weighted F1, accuracy in the emotion classification task. As for VAD regression task, we can see from Table 2 that the proposed model *USE_T+SSOA+CNN* has better performance of consistently lower MAE and higher r. The results of the comparative experiments demonstrate the effectiveness of our proposed model. In order to illustrate the performance of our proposed SSOA mechanism and multi-task training, we do further researches in the following part.

Validation studies of proposed model: We apply Universal Sentence Encoder which is trained with Transformer [15] (USE_T) to encode input texts into sentence embeddings and use CNN as the feature extractor. Therefore **USE_T+CNN** is the basic architecture and we control it as invarient.

USE_T+ATT+CNN: In order to validate our proposed SSOA mechanism, we also consider the most useful self-attention mechanism [16], which decide the importance of features for the prediction task by weighing them when constructing the representation of text.

USE_T+SSOA+CNN (STL): It is our work in single task framework, which uses SSOA mechanism to compute attention scores between the label and the representations of the sentences in the input that is to be classified. This can then be used to appropriately weight the contributions of each sentence to the final representations.

USE_T+SSOA+CNN (MTL): To demonstrates the effectiveness of incorporating VAD regression with emotion classification, we experiment this model in the multi-task framework which trained with both categorical emotion labels and dimensional valence, arousal, dominance labels.

From Table 3 and Table 4, some conclusions can be drawn as following: (1) Both *USE_T+ATT+CNN* with self-attention and *USE_T+SSOA+CNN* with our SSOA have a better performance than with no attention mechanism as expected. (2) Compared with *USE_T+ATT+CNN*, our *USE_T+SSOA+CNN* model achieves a relatively better result, especially achieves improvement about 2.5% in Happy,

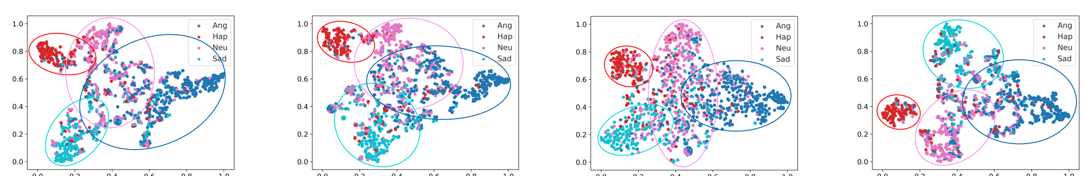
Table 3. Results (%) of Validation studies on emotion classification task

Model	IEMOCAP									
	Happy		Anger		Neutral		Sad		Average(W)	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
USE_T+CNN	60.63	69.61	70.59	67.61	73.44	66.04	69.80	67.99	68.61	67.81
USE_T+ATT+CNN	69.00	70.77	68.82	68.82	69.01	65.11	66.12	69.53	68.24	68.56
USE_T+SSOA+CNN (STL)	66.97	73.27	70.00	71.47	71.61	64.94	69.80	68.95	69.60	69.66
USE_T+SSOA+CNN (SML)	**69.91**	72.88	**71.18**	70.14	67.71	**65.74**	**72.24**	**71.08**	**70.26**	**69.96**

Table 4. Results of validation studies on VAD regression task

Model	IEMOCAP					
	Valence		Arousal		Dominance	
	MSE	r	MSE	r	MSE	r
USE_T+CNN	0.595	0.570	0.431	0.418	0.563	0.464
USE_T+ATT+CNN	0.571	0.582	0.463	0.415	0.554	0.459
USE_T+SSOA+CNN(STL)	0.546	0.591	0.405	0.441	0.526	0.470
USE_T+SSOA+CNN(MTL)	**0.523**	**0.603**	**0.402**	**0.446**	**0.511**	**0.486**

2.65% in Anger on F1-score, and have accuracy improvement about 2.6% in Neutral, 3.68% in Sad separately. The results demonstrate that semantics of emotion labels can guide a model's attention when representing the input conversation and our proposed SSOA mechanism is able to capture sentiment-aware features, meanwhile, self-attention mechanism usually weights features based on semantic and context information which is not effective enough for emotion recognition. (3) Comparatively, as is shown in the last row, when both the problems are learned and evaluated in a multi-task learning framework, we observe performance enhancement for both tasks as well, which illustrates the effectiveness of multitask framework. And as we assume there are two reasons that VAD regression and emotion classification can assist each other task. On the one hand, emotions are high correlated with valence-arousal-dominance space. On the other hand, we take emotion labels into attention computing, which can help to capture more valence and arousal features.



(a) (b) (c) (d)

Fig. 2. t-SNE visualization of validation studies on emotion classification. (a):USE_T+CNN, (b):USE_T+ATT+CNN, (c):USE_T+SSOA+CNN(STL) (d):USE_T+SSOA+CNN(MTL)

Furthermore, in order to validate the effectiveness of our proposed method on different emotions, we introduce the t-Distributed Stochastic Neighbor Embedding (t-SNE) [30] for visualizing the deep representations as shown in Figure 2. We can see that compared with Figure 2 (a), the points which represent Anger in Figure 2 (b) can be distinguished more easily. The points which represent Happy and Sad have similar performance. Compared with Figure 2 (b), all the four emotion points have better discrimination in Figure 2(c) which means the deep representations extracted by our model are more sentiment-aware. However, we can observe from Figure 2(c) that most confusions are concentrated between Anger, Sad with Neutral. We assume the reason is that Anger and Sad hold the lowest percentage in IEMOCAP, which would not trained enough in our SSOA training process. Besides, the dataset we use is multimodal, a few utterances such as *"Yeah"*, *"l know"* carrying non-neutral emotions were misclassified as we do not utilize audio and visual modality in our experiments. In Figure 2(d), Sad can be distinguished better, we assume it's because Sad is one kind of negative valence and arousal values emotion according to Valence-Arousal representation [18], whose prediction would be more easy with the help of VAD.

Overall, the proposed *USE_T+SSOA+CNN* with multi-task learning model outperforms the other comparative and ablation studies. It is reasonable to assume that the proposed model is good at capturing both semantic and emotion features not only in emotion classification but also the VAD regression task.

4 Conclusion

In this paper, we proposed a sentiment similarity-oriented attention mechanism, which can be used to guide the network to extract emotion-related information from input sentences to improve classification and regression accuracy. In addition, to extract more robust features, we jointed dimensional emotion recognition using multi-task learning. The effectiveness of our proposed method has been verified under a series of comparative experiments and validation studies on IEMOCAP. The results show that the proposed method outperforms previous text-based emotion recognition by 6.57% from 63.39% to 69.96%, and show better robustness. In the future work, we will make improvements of the proposed model by introducing speech information into SSOA computation.

Acknowledgements This work was supported in part by the National Key R&D Program of China under Grant 2018YFB1305200, the National Natural Science Foundation of China under Grant 61771333 and the Tianjin Municipal Science and Technology Project under Grant 18ZXZNGX00330.

References

1. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
2. Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
3. Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Jamin Shin, Yan Xu, Peng Xu, and Pascale Fung. Caire_hkust at semeval-2019 task 3: Hierarchical attention for dialogue emotion classification. *arXiv preprint arXiv:1906.04041*, 2019.
4. Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.
5. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
6. Oscar Araque, Ganggao Zhu, and Carlos A Iglesias. A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems*, 165:346–359, 2019.
7. Sopan Khosla, Niyati Chhaya, and Kushal Chawla. Aff2vec: Affect–enriched distributional word representations. *arXiv preprint arXiv:1805.07966*, 2018.
8. Yicheng Zou, Tao Gui, Qi Zhang, and Xuan-Jing Huang. A lexicon-based supervised attention model for neural sentiment analysis. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 868–877, 2018.
9. Eesung Kim and Jong Won Shin. Dnn-based emotion recognition based on bottleneck acoustic features and lexical features. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6720–6724. IEEE, 2019.
10. Saif M Mohammad. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In *Emotion measurement*, pages 201–237. Elsevier, 2016.
11. Bjarke Felbo, Alan Mislove, Anders Søgaard, Iyad Rahwan, and Sune Lehmann. Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. *arXiv preprint arXiv:1708.00524*, 2017.

12. Jiachen Du, Lin Gui, Yulan He, and Ruifeng Xu. A convolutional attentional neural network for sentiment classification. In *2017 International Conference on Security, Pattern Analysis, and Cybernetics (SPAC)*, pages 445–450. IEEE, 2017.
13. Xiao Yang, Craig Macdonald, and Iadh Ounis. Using word embeddings in twitter election classification. *Information Retrieval Journal*, 21(2-3):183–207, 2018.
14. Stacy Marsella and Jonathan Gratch. Computationally modeling human emotion. *Communications of the ACM*, 57(12):56–67, 2014.
15. Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. Universal sentence encoder for english. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, 2018.
16. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
17. Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. Deep unordered composition rivals syntactic methods for text classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1681–1691, 2015.
18. Theodoros Giannakopoulos, Aggelos Pikrakis, and Sergios Theodoridis. A dimensional approach to emotion recognition of speech from movies. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 65–68. IEEE, 2009.
19. Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior research methods*, 45(4):1191–1207, 2013.
20. Shabnam Tafreshi and Mona Diab. Emotion detection and classification in a multigenre corpus with joint multi-task deep learning. In *Proceedings of the 27th international conference on computational linguistics*, pages 2905–2913, 2018.
21. Md Shad Akhtar, Dushyant Singh Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. Multi-task learning for multi-modal emotion recognition and sentiment analysis. *arXiv preprint arXiv:1905.05812*, 2019.
22. Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.
23. Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.
24. Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.
25. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
26. Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.
27. Samarth Tripathi and Homayoon Beigi. Multi-modal emotion recognition on iemocap dataset using deep learning. *arXiv preprint arXiv:1804.05788*, 2018.
28. Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.
29. Peng Zhou, Zhenyu Qi, Suncong Zheng, Jiaming Xu, Hongyun Bao, and Bo Xu. Text classification improved by integrating bidirectional lstm with two-dimensional max pooling. *arXiv preprint arXiv:1611.06639*, 2016.
30. Laurens Van Der Maaten. Accelerating t-sne using tree-based algorithms. *The Journal of Machine Learning Research*, 15(1):3221–3245, 2014.