Implicit Objective Network for Emotion Detection*

Hao Fei[1], Yafeng Ren[2], and Donghong Ji[1]

[1] Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan, China
[2] Guangdong Collaborative Innovation Center for Language Research & Services, Guangdong University of Foreign Studies, Guangzhou, China
{hao.fei,renyafeng,dhji}@whu.edu.cn

Abstract. Emotion detection has been extensively researched in recent years. However, existing work mainly focuses on recognizing explicit emotion expressions in a piece of text. Little work is proposed for detecting implicit emotions, which are ubiquitous in people's expression. In this paper, we propose an Implicit Objective Network to improve the performance of implicit emotion detection. We first capture the implicit sentiment objective as a latent variable by using a variational autoencoder. Then we leverage the latent objective into the classifier as prior information for better make prediction. Experimental results on two benchmark datasets show that the proposed model outperforms strong baselines, achieving the state-of-the-art performance.

Keywords: sentiment analysis · variational model · neural network · implicit emotion.

1 Introduction

Emotion detection, as one heated research topic in natural language processing (NLP), aims to automatically determine the emotion or sentiment polarity in a piece of text. Emotion detection is usually modeled as a classification task, and a large number of methods are proposed in past ten years. Previous work is mainly divided into lexicon-based methods [10] and machine learning based methods [15]. Recently, various neural networks models have been proposed for this task, achieving highly competitive results on several benchmark datasets [21]. However, these work mainly focuses on explicit emotion detection. Little work is proposed for detecting implicit emotion expressions.

In real world, people are more likely to express their emotion in an implicit way, so the emotional expressions contain explicit emotion and implicit emotion. Taking the following sentences as examples:

(S1) *It is such a great thing for me to meet my best friend again today.* (happy)
(S2) *I ate a grain of sand while eating dumplings in that restaurant.* (angry)
(S3) *A friend forgot his appointment with me.* (sad)

In sentence S1, the explicit emotion *happy* can easily be identified via the keyword *great* based on sentiment lexicon. In S2, the customer is complaining about the terrible

* Corresponding authors are Yafeng Ren and Donghong Ji. This work is supported by the National Natural Science Foundation of China (No.61702121, No.61772378).

quality of food about the restaurant, and a person is disappointed to friend's missing appointment in S3. For S2 and S3, however, it is difficult to automatically detect the emotions by using sentiment lexicon or defining appropriate features, since the sentences lack of explicit sentimental clues. Obviously, the models of explicit emotion detection can not achieve satisfactory performance for implicit emotion detection.

In S2 and S3, the emotion *angry* and *sad* are expressed towards the implicit sentiment objectives 'quality of food' and'appointment and credit', respectively, which both cannot be observed explicitly. Intuitively, if these semantic-rich objective information can be learned and leveraged as prior information into the final prediction, the performances should be improved. Taking sentence S3 as example, if the implicit objective 'appointment and credit' can be captured in advance, with the cue word 'forgot' further discovered, a classifier can easily infer the emotion *sad*. This motivates us to build a method that can capture such implicit sentiment objective for better detecting implicit emotion.

In this paper, we propose a variational based model, Implicit Objective Network (named ION), for implicit emotion detection. We model the sentiment objective as a latent variable, and employ a variational autoencoder (VAE) to learn it. The overall model consists of two major components: a variational module and a classification module. First, the variational module captures the semantic-rich objective representation in the latent variable during the reconstruction of input sentence. Then, the classification module leverages such prior information, and uses a multi-head attention mechanism to effectively retrieve the clues concerning the objective within the sentence.

To our knowledge, we are the first study that employs the variational model to capture the implicit sentiment objective, and leverages such latent information for better make prediction. We conduct experiments on two datasets, which are widely used for implicit emotion detection, including ISEAR [23] and IEST [12]. Experimental results show that our model outperforms strong baselines, achieving the state-of-the-art performance.

2 Related Work

Emotion Detection Emotion detection is a heated research topic in NLP. A large number of methods are proposed for the task in past ten years. Existing work are mainly divided into three classes: lexicon-based methods, machine learning based methods and deep learning methods. Pang et al., (2002) first explored this task by using bag-of-word features [18]. Later, many machine learning models are explored for the task [6]. More recently, various neural networks models have been proposed for this task [3, 20, 22], achieving highly competitive results. However, existing work largely focuses on explicit emotion detection. For implicit emotion detection, some preliminary work also is proposed [2, 19]. Meanwhile, a shared task is proposed for implicit emotion detection [3, 12]. However, these models fail to effectively use the information of implicit sentiment objective for the final prediction, limiting the performance of the task.

Variational Models Our proposed method is also related to variational models in NLP applications. Bowman et al. (2015) introduced a RNN-based VAE model for generating diverse and coherent sentences [4]. Miao et al. (2016) proposed a neural variational

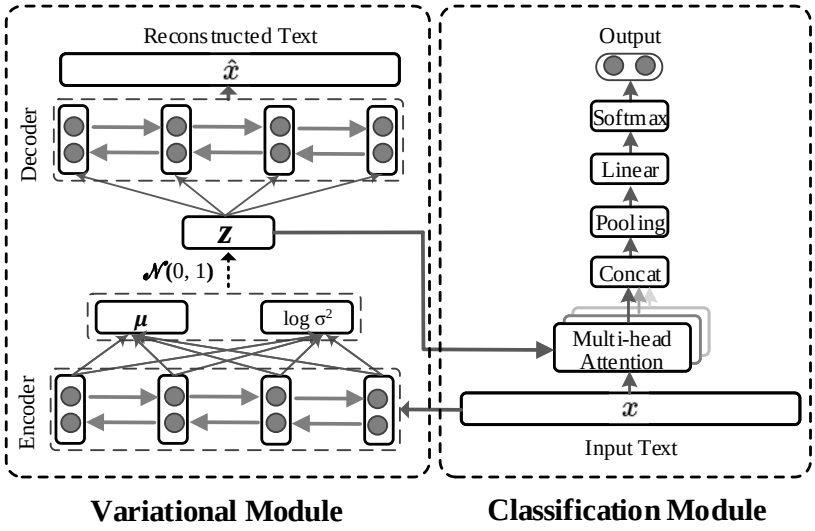

Fig. 1. The overall framework of the proposed model.

framework incorporating multilayer perceptrons (MLP), CNN and RNN for generative models of text [17]. Bahuleyan et al. (2017) proposed a variational attention-based seq2seq model for alleviating the attention bypassing effect [1]. Different from the above work, we employ the VAE model to make reconstruction for original sentence, during which we make use of the intermediate latent representation as prior information for facilitating downstream prediction.

3 Implicit Objective Network

The proposed model is shown in Figure 1, which consists of two main components: a variational module and a classification module.

3.1 Variational Module

For an input sentence, the variational module can capture the semantic-rich representation in the latent space. The main motivation is that we model the implicit sentiment objective as latent variables, since it cannot be observed explicitly. Previous studies have already proved that variational autoencoder (VAE) had strong ability to learn latent representation that was semantic-rich [4, 17]. Therefore, we use VAE to learn such objective representation, during the reconstruction of the input sentence.

The model takes a sequence of words as input. For each word $\boldsymbol{w}_i$, we use a look-up table $\boldsymbol{E} \in \mathbb{R}^{L \times V}$ (L represents the dimension of embedding vector and V is the vocabulary size) to obtain its embedding $\boldsymbol{e}_i \in \mathbb{R}^L$. We employ a bi-directional Long Short-Term Memory (BiLSTM) network as encoder, transforming the inputs $\boldsymbol{x} =$

$\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_T\}$ ($\boldsymbol{x} \in \mathbb{R}^{T \times L}$, T is the sequence length) into prior parameters: $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$.

$$\boldsymbol{\mu} = BiLSTM(\boldsymbol{x}), \tag{1}$$

$$log\boldsymbol{\sigma} = BiLSTM(\boldsymbol{x}). \tag{2}$$

We define a latent variable $\boldsymbol{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \cdot \epsilon$, where ϵ is Gaussian noise variable sampled from $\mathcal{N}(0, 1)$, $\boldsymbol{z} \in \mathbb{R}^D$ (D denotes the dimension of latent variable).

The implicit objective is reflected in $\boldsymbol{z}$. Besides, we believe the sentimental objective should be compound of several semantic-rich meaning, instead of a single simple topic. Therefore, the dimension D of the latent variable, which controls the capacity on carrying the sentimental objective, should be decided empirically based on development experiments.

We use variational inference to approximate a posterior distribution $p(\boldsymbol{x}|\boldsymbol{z})$ over $\boldsymbol{z}$, and use BiLSTM with a same length of time step in encoder, to decode $\boldsymbol{z}$. The final terminals of BiLSTM at each time step will be followed with softmax function to make a prediction for possible words. Thus, the decoding is formulated as:

$$\widehat{\boldsymbol{x}}_i = softmax_i(BiLSTM(\boldsymbol{z})), \tag{3}$$

where $i = (0, \cdots, T)$, $\widehat{\boldsymbol{x}}_i$ are the reconstructed words.

Following previous work [14], parameters in VAE are learned by maximizing the variational lower bound on the marginal log likelihood of features:

$$logp_\theta(\boldsymbol{x}) \geq \mathbb{E}_{\boldsymbol{z} \sim q_\phi(\boldsymbol{z}|\boldsymbol{x})}[logp_\theta(\boldsymbol{x}|\boldsymbol{z})] - KL(q_\phi(\boldsymbol{z}|\boldsymbol{x})||p(\boldsymbol{z})), \tag{4}$$

where ϕ and θ are the parameters of encoder and decoder respectively, and KL-divergence term ensures that the distributions $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ is near to prior probability $p(\boldsymbol{z})$, $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ describes the decoding process.

Since the loss objective of decoder is to reconstruct the input, it has a direct access to the source features. Thus, when the decoder is trained, it will be $q(\boldsymbol{z}|\boldsymbol{x}) = q(\boldsymbol{z}) = p(\boldsymbol{z})$, which means that the KL loss is zero. It makes the latent variables $\boldsymbol{z}$ fail to capture information. We thus employ KL cost annealing and word dropout of encoder [5].

3.2 Classification Module

After capturing the sentiment objective $\boldsymbol{z}$, we leverage it into the classification module. We employ a multi-head attention mechanism to attend the objective representation to the input sentence representation.

In practice, we first repeat the implicit sentiment objective T times to obtain the query representation:

$$\boldsymbol{q} = \underbrace{[\boldsymbol{z}, \cdots, \boldsymbol{z}]}_{T}. \tag{5}$$

Following previous work [25], we first compute the match between the sentence and query representation via Scaled Dot-Product[3] alignment function. Then one single at-

[3] The reason we scale the dot products by $\sqrt{D}$ is to counteract the effect that, if D is large enough, the sum of the dot products will grow large, pushing softmax into regions 0 or 1 [25].

tention matrix is computed by a softmax function after the dot-product of $\boldsymbol{x}$:

$$\boldsymbol{v} = tanh(\frac{\boldsymbol{q} \cdot \boldsymbol{x}^T}{\sqrt{D}}), \tag{6}$$

$$\boldsymbol{\alpha} = softmax(\boldsymbol{v}) \cdot \boldsymbol{x} \quad . \tag{7}$$

We maintain K attention heads by repeating the computation K times via above formulation. We then concatenate context representation from K attention heads into an overall context matrix $\boldsymbol{H}$.

$$\boldsymbol{H} = [\boldsymbol{\alpha}_1; \cdots; \boldsymbol{\alpha}_K] \cdot \boldsymbol{W} \tag{8}$$

$$= \{\boldsymbol{h}_1, \cdots, \boldsymbol{h}_T\}, \tag{9}$$

where $\boldsymbol{W} \in \mathbb{R}^{(K \cdot L) \times L}$ is a parameter matrix, transforming the shape of context representation as $\mathbb{R}^{T \times L}$, and each $\boldsymbol{h}_i$ is the high level context representation of i-th token within a sentence.

Afterwards, we employ the pooling techniques to merge the varying number of features from the context representation $\boldsymbol{H}$. We use three types of pooling methods: *max*, *min* and *average*, and concatenate them into a new representation $\boldsymbol{H}_{pooled}$. Specifically,

$$\boldsymbol{h}_{i,pooled} = \left[\begin{bmatrix} max(\boldsymbol{h}_i^1) \\ \cdots \\ max(\boldsymbol{h}_i^L) \end{bmatrix}; \begin{bmatrix} min(\boldsymbol{h}_i^1) \\ \cdots \\ min(\boldsymbol{h}_i^L) \end{bmatrix}; \begin{bmatrix} avg(\boldsymbol{h}_i^1) \\ \cdots \\ avg(\boldsymbol{h}_i^L) \end{bmatrix} \right], \tag{10}$$

$$\boldsymbol{H}_{pooled} = \{\boldsymbol{h}_{1,pooled}, \cdots, \boldsymbol{h}_{T,pooled}\}, \tag{11}$$

where $\boldsymbol{h}_i^j$ denotes the j-th dimension of $\boldsymbol{h}_i$.

To fully utilize these sources of information, we use a linear hidden layer to automatically integrate *max*, *min* and *average* pooling features. Finally, a softmax classifier is applied to score possible labels according to the final representation of hidden layer. Specifically,

$$y_k = softmax(f_{MLP}(\boldsymbol{H}_{pooled})), \tag{12}$$

where y_k is the predicted label, $f_{MLP}(\cdot)$ is the linear hidden layer.

3.3 Training

Our training target is to minimize the following cross-entropy loss function:

$$L = -\sum_i \sum_k \hat{y}_k log y_k + \frac{\lambda}{2} \|\theta\|^2, \tag{13}$$

where the $\frac{\lambda}{2}\|\theta\|^2$ is a l_2 regularization term, $\hat{y}_k$ is ground truth label.

The variational module and the classification module can be jointly trained. However, directly training the whole framework with cold-start can be difficult and causes high variance. The training of the variational module can be particularly difficult. Thus

Table 1. Statistics of two datasets. **Avg.Len.** denotes the average length of the sentence.

Dataset	Sent.	Avg.Len.	Train	Dev	Test
ISEAR	7,666	21.20	5,366	767	1,533
IEST	191,731	18.28	153,383	9,591	28,757

we first pre-train the variational module until it is close to the convergence via Eq. (4). Afterwards, we jointly train all the components via Eq. (4) and (13). Once the classification loss is close to convergence, we again train the variational module alone, until it is close to the convergence. We then co-train the overall ION. We keep such training iterations until the overall performance reaches its plateau.

4 Experiments

4.1 Datasets

We conduct experiments on two widely used datasets for implicit emotion detection, including ISEAR dataset [23] and IEST dataset [12]. ISEAR contains seven emotions, including *joy*, *fear*, *anger*, *sadness*, *disgust*, *shame* and *guilt*. IEST contains six emotions, including *sad*, *joy*, *disgust*, *surprise*, *anger* and *fear*. Statistics of the datasets is shown in Table 1.

4.2 Experimental Settings

We employ the GloVe[4] 300-dimensional embeddings as the pre-trained embedding, which are trained on 6 billion words from Wikipedia and web text. The dimension of latent variable and the number of attention heads are fine-tuned according to each dataset. We follow Glorot et al. (2010) and initialize all the matrix and vector parameters with Xavier methods. In the learning process, we set 300 epochs for pretraining variational module, and 1000 total training epochs with early-stop strategy. To mitigate overfitting, we apply the dropout method to regularize our model, with 0.01 dropout rate. We use Adam to schedule the learning, with the initial learning rate of 0.001. Experimental results are reported by precision, recall and F1 score.

4.3 Baselines

To show the effectiveness of the proposed model, we compare the proposed model with the following baselines:

- **TextCNN:** A classical convolutional neural network for text modeling [11].
- **BiLSTM:** Bi-directional LSTM makes prediction by considering the information from both forward and backward directions [24].

[4] `http://nlp.stanford.edu/projects/glove/`

Table 2. Results on two datasets. **ION-l.o.** denotes **ION** model without latent objective representation z. **ION+lt** denotes **ION** does not take pre-trained embedding. **ION-att** denotes that we remove the multi-head attention. **ION-pool** denotes **ION** removes the pooling operation.

System	ISEAR			IEST		
	Precision	**Recall**	**F1 score**	**Precision**	**Recall**	**F1 score**
TextCNN	0.645	0.643	0.642	0.594	0.594	0.594
BiLSTM	0.636	0.643	0.638	0.589	0.572	0.578
RCNN	0.643	0.647	0.644	0.602	0.617	0.608
AttLSTM	0.661	0.666	0.663	0.615	0.623	0.618
FastText	0.620	0.634	0.629	0.605	0.614	0.612
BLSTM-2DCNN	0.716	0.701	0.712	0.617	0.627	0.620
ohLSTMp	0.708	0.699	0.701	0.637	0.601	0.612
DPCNN	0.749	0.739	0.743	0.644	0.647	0.646
ION	**0.755**	**0.746**	**0.752**	**0.664**	**0.647**	**0.658**
ION-l.o.	0.702	0.719	0.707	0.617	0.625	0.620
ION+lt	0.736	0.740	0.739	0.656	0.638	0.648
ION-att	0.667	0.679	0.671	0.609	0.615	0.612
ION-pool	0.723	0.719	0.729	0.656	0.635	0.640

- **RCNN:** Lat et al., (2015) build a model for text modeling by concatenating the output of Recurrent network to the following Convolutional Neural network [13].
- **AttLSTM:** Attention based LSTM model can learn high level context representation by assigning different weights for different elements in sequence [27].
- **FastText:** Joulin et al., (2016) propose a model for text classification which employs the global average pooling technique to extract high level features [9].
- **BLSTM-2DCNN:** Zhou et al., (2016) improve the sequence modeling by integrating bi-directional LSTM with two-dimensional max pooling operation [26].
- **ohLSTMp:** Johnson et al., (2016) explore a region embedding with one hot LSTM [7].
- **DPCNN:** Johnson et al., (2017) improve the word-level CNN to 15 weight layers to better capture global representations [8].

5 Results and Analysis

5.1 Main Results

Experimental results are shown in Table 2. We can see that our proposed model achieves 0.752 and 0.658 F1 score on ISEAR and IEST, respectively. Compared with the best baseline model (DPCNN), our model achieves better performance on both datasets. This demonstrates the effectiveness of the proposed method for implicit emotion detection.

We further compare our model with several baselines on each emotion. Results on ISEAR dataset are shown in Table 3. We can see that ION achieves the best performance on most of the emotions. For the emotions *joy*, *sadness* and *guilt*, our model also achieves comparative performance. Note that the emotion *anger* gives 0.850 F1

Table 3. Results of each emotions on ISEAR.

Model	joy	fear	anger	sadness	disgust	shame	guilt	Avg
FastText	0.621	0.736	0.552	0.645	0.528	0.755	0.565	0.629
AttLSTM	0.675	0.725	0.620	0.643	0.618	0.777	0.584	0.663
ohLSTMp	0.700	0.767	0.683	0.691	0.644	0.796	0.627	0.701
DPCNN	**0.767**	0.802	0.741	**0.705**	0.661	**0.843**	**0.680**	0.743
ION	0.738	**0.835**	**0.850**	0.573	**0.822**	0.845	0.604	**0.752**

Table 4. Results of each emotion on IEST.

Model	sad	joy	disgust	surprise	anger	fear	Avg
FastText	0.588	0.705	0.637	0.598	0.555	0.588	0.612
AttLSTM	0.622	0.738	0.550	0.589	0.588	0.620	0.618
ohLSTMp	0.533	0.682	0.611	0.599	0.592	0.656	0.612
DPCNN	0.620	0.706	**0.659**	0.633	**0.601**	0.658	0.646
ION	**0.636**	**0.746**	0.647	**0.650**	0.587	**0.691**	**0.658**

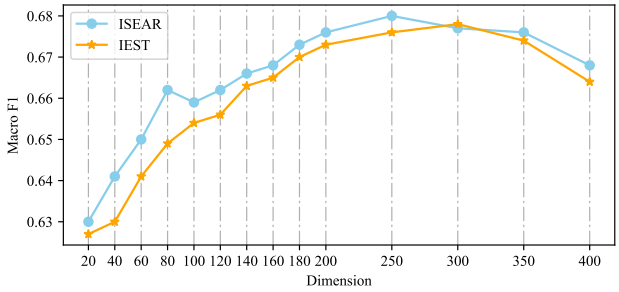

Fig. 2. Results with different dimensions of latent variable $\boldsymbol{z}$.

score and the emotion *disgust* gives 0.822 F1 score, significantly outperforming other baseline systems. This shows the effectiveness of the proposed model for the task. Meanwhile, we also find that all models perform worse on the emotion *guilt*. Results on IEST dataset are shown in Table 4, the same trend can be found.

5.2 Impact of Dimensionality of Latent Variable

The latent variable $\boldsymbol{z}$ carries the semantic sentiment objective information. Intuitively, the larger the dimension of $\boldsymbol{z}$ is, the stronger capability the ION has. We analyze the ability of our model in capturing the sentiment objective with different dimensionality of the latent variable. Figure 2 shows the results. We have the following observations. Taking the ISEAR dataset as example. First, when the dimension is below 80, the performance drops dramatically. When dimension is above 100, F1 score increases gradually. Second, too large or too small size of dimension equally hurts the performance, and ION with 250-dim $\boldsymbol{z}$ obtains the best performance on ISEAR dataset. For IEST

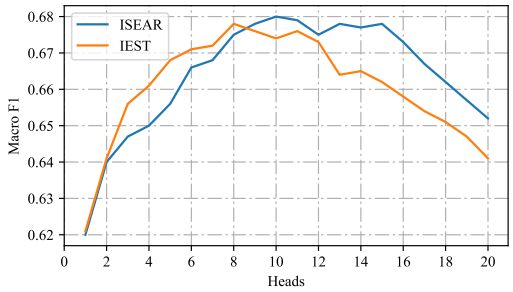

Fig. 3. Results with different head numbers of attention.

dataset, we can see the 300-dim z is the best for ION. The possible reason is that too large dimensions may produce redundant representation and cause overfitting.

5.3 Impact of Head Numbers of Attention

Multi-head attention allows the model to retrieve task-relevant information from different representation subspaces at different positions. We study how the head numbers of multi-head attention influence the performance of ION. Figure 3 shows the results. We can see that ION has strong ability when the head number is from 6 to 16. Otherwise, the accuracy drops significantly. Specifically, for the ISEAR dataset, ION achieves the best F1 score with 10 heads. For the IEST dataset, ION achieves the best F1 score with 8 heads. This demonstrates that ION needs more heads on ISEAR than IEST. The main reason is that the average length of sentences in ISEAR is larger than in IEST, so that the features should intuitively be more scattered in ISEAR, and correspondingly requires more number of heads for ION to capture these clues.

5.4 Ablation Study

To investigate the contributions of different parts in our model, we make the ablation study, which is shown in Table 2. We first remove the implicit objective representation from Eq. 5. Without such prior information, we find that the performance decreases significantly, with a striking drop of 0.045 and 0.038 F1 score on ISEAR and IEST, respectively. This verifies our idea that the implicit objective information is important for implicit emotion detection. The above analysis also shows the usefulness of the variational module, which capture the implicit sentiment objective, in our ION model.

We also compare the randomly initialized embedding with the pre-trained embeddings, to show the effect of the embedding for the task. By using the randomly initialized embedding, ION gives 0.739 and 0.648 F1 score on two datasets, respectively. We can see that the proposed model can achieve better performance with the pre-trained word embedding.

We also explore how multi-head attention module contributes to the overall performance. We remove the multi-head attention, and use BiLSTM to encode the text. From

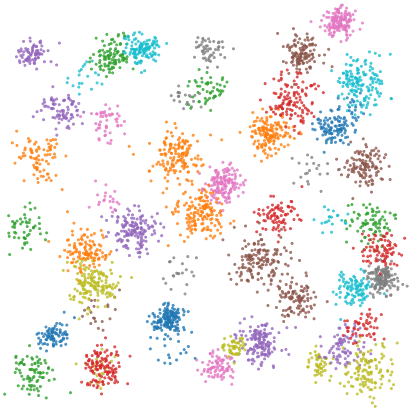

Fig. 4. Visualization of implicit sentiment objective on 1,000 sentences.

Table 2, we can find that without the multi-head attention, the performance of the model decreases a lot, only giving 0.671 and 0.612 F1 score on two datasets, respectively. The above analysis shows the usefulness of multi-head attention for the task.

We further remove the pooling operation from ION model, to show how the performance of the model changes. We can see that F1 score on two datasets slightly decreases to 0.729 and 0.640, respectively. This demonstrates the necessity of merging feature representation via pooling techniques.

5.5 Visualization of Implicit Sentiment Objective

To investigate the learned latent representations z, we visualize it and examine whether the latent space groups together the sentence share the similar implicit objective. Specifically, we select 1,000 sentences from the test set of ISEAR that are correctly predicted. We project the z representation of the sentences into two-dimensional space by using t-SNE algorithm [16],which is shown in Figure 4.

The visualization of these objective groups provides some interesting insights. First, latent representations of different sentences are clustered into different groups. As we expected, these sentences with similar sentiment objective (in the same color) are projected into adjacent position. Second, some groups are overlapped with others. We consider this as the semantic compound instead of outliers. This verifies the idea that an implicit sentiment objective is compound of several semantic-rich meaning, instead of a single or simple topic.

5.6 Case Study

We analyze how the variational module work together with multi-head attention to better make prediction. We show some cases based on examples used in previous subsection. We present 3 representative objective categories that correspond to the certain

Table 5. Representatives of objective category and the visualization of attention weights.

Objective Category	Attention Visualization
argument, attitude	When I argue with my mother about the way she treats her two children differently . (sadness)
	Once my father slapped my mother for a small quarrel . (anger)
	I had a quarrel with my parents; I was convinced to be right . (anger)
work, achievement	I felt this when I was copying homework for one of my classes . (shame)
	When I finished the work that I had planned to do - my homework . (joy)
appointment, credit	I failed to show up at an agreed date . (guilt)
	A friend forgot his appointment with me . (anger)

group in Figure 4, and assign some labels for describing the categories. For each category, we present some sentences with their visualization of the attention weights. Table 5 shows the results.

We can see that the first objective category is about 'argument' and 'attitude', and the token within the sentences that strongly support the prediction, such as 'argue', 'quarrel', 'slapped' etc, are highlighted under this category. The highly weighted tokens are closely related to the corresponding objective category. The sentences in another two objective categories have the same observation. Besides, ION can sufficiently mine task-relevant clues and assign proper weights for them, instead of merely highlighting the most relevant word. This owes to the advantage of using multi-head attention.

6 Conclusion

In this paper, we propose an implicit objective network for implicit emotion detection. The variational module in our model captures the implicit sentiment objective during the reconstructing of the input sentence. The classification module leverages such prior information, and uses a multi-head attention mechanism to effectively capture the clues for the final prediction. Experimental results on two datasets show that our model outperforms strong baselines, achieving the state-of-the-art performance.

References

1. Bahuleyan, H., Mou, L., Vechtomova, O., Poupart, P.: Variational attention for sequence-to-sequence models. arXiv preprint arXiv:1712.08207 (2017)
2. Balahur, A., Hermida, J.M., Montoyo, A.: Detecting implicit expressions of emotion in text: A comparative analysis. Decision Support Systems **53**(4), 742–753 (2012)
3. Balazs, J.A., Marrese-Taylor, E., Matsuo, Y.: Iiidyt at iest 2018: Implicit emotion classification with deep contextualized word representations. arXiv preprint arXiv:1808.08672 (2018)
4. Bowman, S.R., Vilnis, L., Vinyals, O., Dai, A.M., Jozefowicz, R., Bengio, S.: Generating sentences from a continuous space. arXiv preprint arXiv:1511.06349 (2015)
5. Goyal, A.G.A.P., Sordoni, A., Côté, M.A., Ke, N.R., Bengio, Y.: Z-forcing: Training stochastic recurrent networks. In: Proceedings of Advances in Neural Information Processing Systems. pp. 6713–6723 (2017)

6. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 168–177. ACM (2004)
7. Johnson, R., Zhang, T.: Supervised and semi-supervised text categorization using lstm for region embeddings. arXiv preprint arXiv:1602.02373 (2016)
8. 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. pp. 562–570 (2017)
9. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759 (2016)
10. Kamal, R., Shah, M.A., Maple, C., Masood, M., Wahid, A., Mehmood, A.: Emotion classification and crowd source sensing; a lexicon based approach. IEEE Access (2019)
11. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
12. Klinger, R., De Clercq, O., Mohammad, S.M., Balahur, A.: Iest: Wassa-2018 implicit emotions shared task. arXiv preprint arXiv:1809.01083 (2018)
13. Lai, S., Xu, L., Liu, K., Zhao, J.: Recurrent convolutional neural networks for text classification. In: Proceedings of the Twenty-ninth AAAI conference on Artificial Intelligence (2015)
14. Le, H., Tran, T., Nguyen, T., Venkatesh, S.: Variational memory encoder-decoder. In: Proceedings of Advances in Neural Information Processing Systems. pp. 1508–1518 (2018)
15. Liu, B.: Sentiment analysis and opinion mining. Synthesis lectures on human language technologies **5**(1), 1–167 (2012)
16. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9**(Nov), 2579–2605 (2008)
17. Miao, Y., Yu, L., Blunsom, P.: Neural variational inference for text processing. In: Proceedings of the International Conference on Machine Learning. pp. 1727–1736 (2016)
18. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up?: sentiment classification using machine learning techniques. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing. pp. 79–86. Association for Computational Linguistics (2002)
19. Ren, H., Ren, Y., Li, X., Feng, W., Liu, M.: Natural logic inference for emotion detection. In: Proceedings of the Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. pp. 424–436 (2017)
20. Ren, Y., Wang, R., Ji, D.: A topic-enhanced word embedding for twitter sentiment classification. Information Sciences **369**, 188–198 (2016)
21. Ren, Y., Zhang, Y., Zhang, M., Ji, D.: Context-sensitive twitter sentiment classification using neural network. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
22. Rozental, A., Fleischer, D., Kelrich, Z.: Amobee at iest 2018: Transfer learning from language models. arXiv preprint arXiv:1808.08782 (2018)
23. Scherer, K.R.: What are emotions? and how can they be measured? Social science information **44**(4), 695–729 (2005)
24. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing **45**(11), 2673–2681 (1997)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proceedings of Advances in Neural Information Processing Systems. pp. 5998–6008 (2017)
26. Zhou, P., Qi, Z., Zheng, S., Xu, J., Bao, H., Xu, B.: Text classification improved by integrating bidirectional lstm with two-dimensional max pooling. arXiv preprint arXiv:1611.06639 (2016)
27. Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., Xu, B.: Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. pp. 207–212 (2016)