An Interpretable Neural Network with Topical Information for Relevant Emotion Ranking

Yang Yang[†] **Deyu Zhou**[*†] **Yulan He**[§]
[†]School of Computer Science and Engineering, Key Laboratory of Computer Network
and Information Integration, Ministry of Education, Southeast University, China
[§]Department of Computer Science, University of Warwick, UK
{yyang, d.zhou}@seu.edu.cn, y.he@cantab.net

Abstract

Text might express or evoke multiple emotions with varying intensities. As such, it is crucial to predict and rank multiple relevant emotions by their intensities. Moreover, as emotions might be evoked by hidden topics, it is important to unveil and incorporate such topical information to understand how the emotions are evoked. We proposed a novel interpretable neural network approach for relevant emotion ranking. Specifically, motivated by transfer learning, the neural network is initialized to make the hidden layer approximate the behavior of topic models. Moreover, a novel error function is defined to optimize the whole neural network for relevant emotion ranking. Experimental results on three real-world corpora show that the proposed approach performs remarkably better than the state-of-the-art emotion detection approaches and multi-label learning methods. Moreover, the extracted emotion-associated topic words indeed represent emotion-evoking events and are in line with our common-sense knowledge.

1 Introduction

With the growth of social web, people tend to share their opinions, feelings and attitudes on the social platforms such as online news sites and blogs. Emotion detection can enhance the understanding of users' emotional states, which is useful in many downstream applications such as human-computer interaction and personalized recommendation. Therefore, it is crucial to predict emotions from texts accurately (Picard and Picard, 1997).

Research on emotion detection can be roughly categorized into two types: generative model based and discriminative model based. Generative model based approaches (Bao et al., 2012; Rao et al., 2014a) usually build on topic models and assume texts are generated from emotions and hidden topics. While these models can extract emotion-associated topics, they perform less satisfactorily in emotion classification since they are not optimized directly to minimize the misclassification rate. Discriminative model based approaches consider each emotion category as a class label and typically cast emotion detection as a classification problem. Approaches to the prediction of both multiple emotions and their intensities include (Zhou et al., 2018, 2016; Wang and Pal, 2015). Those approaches usually assumed word-level representations and ignored the latent topical information behind words, therefore failed to effectively distinguish different emotions carried by the same word in different topical contexts.

In this paper, we focus on relevant emotion ranking (RER) by differentiating relevant emotions from irrelevant ones and only learning the rankings of relevant emotions while ignoring the irrelevant ones. A neural network with a novel loss function is proposed to tackle the RER problem. A topic representing a real-world event, an abstract entity, or an object could indicate the subject or context of the emotion. Different topics might contain or invoke different emotions (Stoyanov and Cardie, 2008). Incorporating such latent topics is essential for discovering topic-associated emotions. Motivated by transfer learning, we incorporate hidden topics and the topic distributions generated from a topic model into a neural network for RER. The main contributions of the paper are summarized below:

- A novel Interpretable Neural Network for Relevant Emotion Ranking (INN-RER) is proposed. A novel error function is employed to optimize the whole network for parameter estimation. To the best of our knowledge, it is the first neural network based approach for RER.

[*]Corresponding author

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3423–3432
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

- To understand how the emotions are evoked, the neural network is initialized to make its hidden layer approximate the behavior of topic models so that the topical information is unveiled and incorporated.

- Experimental results on three different real-world corpora show that the proposed method can effectively deal with the emotion detection problem and perform better than the state-of-the-art emotion detection methods and multi-label learning methods. Moreover, emotion-association topic words extracted by INN-RER indeed represent emotion-evoking events.

2 Related Work

In general, approaches for emotion detection can be divided into two categories: generative model based and discriminative model based. Generative model based approaches typically built on topic models. For example, the emotion-topic model (Bao et al., 2012) was proposed by adding an extra emotion layer into traditional topic models to capture the generation of both emotions and topics from text at the same time. Other topic model based approaches such as affective topic model (Rao et al., 2014a), multi-label supervised topic model and sentiment latent topic model (Rao et al., 2014b) also modeled the emotions and topics simultaneously. Contextual sentiment topic model (Rao, 2016) assumed each word is either drawn from a background theme, a contextual theme or a topic and explicitly distinguished between context-dependent and context-independent topics.

For discriminative model based methods, emotion detection is often casted as a classification problem by considering each emotion category as a class label. If only choosing the strongest emotion as the label for a given text, emotion detection is essentially a single-label classification problem. Lin et al. (2008) studied the classification of news articles into different categories based on readers' emotions with various combinations of feature sets. Strapparava and Mihalcea (2008) proposed several knowledge-based and corpus-based methods for emotion classification. Quan et al. (2015) proposed a logistic regression model with emotion dependency for emotion detection. Latent variables were introduced to model the latent structure of input text. Li et al. (2016) combined bi-term topic model and conventional neural network to detect single social emotion from short texts. To predict multiple emotions simultaneously, emotion detection can be solved using multi-label classification. Bhowmick (2009) presented a method for classifying news sentences into multiple emotion categories using an ensemble based multi-label classification technique. Wang and Pal (2015) output multiple emotions with intensities using non-negative matrix factorization with several novel constraints such as topic correlation and emotion bindings. To predict multiple emotions with different intensities in a single sentence, Zhou et al. (2016) proposed a novel approach based on emotion distribution learning. Following this way, a relevant label ranking framework for emotion detection was proposed for predict multiple relevant emotions as well as the ranking of emotions based on their intensities (Zhou et al., 2018).

Our work is partly inspired by (Zhou et al., 2018) for relevant emotion ranking, but with the following differences: (1) our model takes into account latent topics in texts for emotion detection, which was ignored in the model proposed in (Zhou et al., 2018); (2) our model is built upon topic models and neural networks with a novel objective function defined to consider the interplay between topics and emotions, while the model in (Zhou et al., 2018) was developed based on a ranking framework with a linear objective function which was not able to describe complex relations between the input texts and their emotions.

3 The Proposed Approach

Assuming a set of T emotions $L = \{e_1, e_2, ...e_T\}$ and a set of n text instances $X = \{x_1, x_2, x_3, ..., x_n\}$, each instance $x_i \in \mathbb{R}^d$ is associated with a ranked list of its relevant emotions $R_i \subseteq L$ and also a list of irrelevant emotions $\overline{R_i} = L - R_i$. Relevant emotion ranking aims to learn a score function $\mathbf{g}(x_i) = [g_1(x_i), ..., g_T(x_i)]$ which assigns a score $g_j(x_i)$ to each emotion $e_j, (j \in \{1, ..., T\})$. In order to differentiate relevant emotions from irrelevant ones, we need to define a threshold Θ which could be simply set to a fixed value or learned from data (Fürnkranz et al., 2008). Those emotions with scores lower than the threshold will be considered as irrelevant and hence discarded. The identification of relevant emotions and their ranking can be obtained simultaneously according to their scores assigned by the learned ranking function g. As mentioned before, it is unnecessary to consider the rankings of irrelevant emotions since they might introduce errors into the model during the learning process.

We propose an Interpretable Neural Network for Relevant Emotion Ranking (INN-RER) built upon a multi-layer feed-forward neural network. Instead of using the simple sum-of-squares error function, a novel loss function is designed and employed. Accordingly, a new learning algorithm is proposed to minimize the new loss function. Furthermore, motivated by transfer learning, topical information generated from a topic model is transferred into the neural network by making its hidden layer approximate the behavior of topic models.

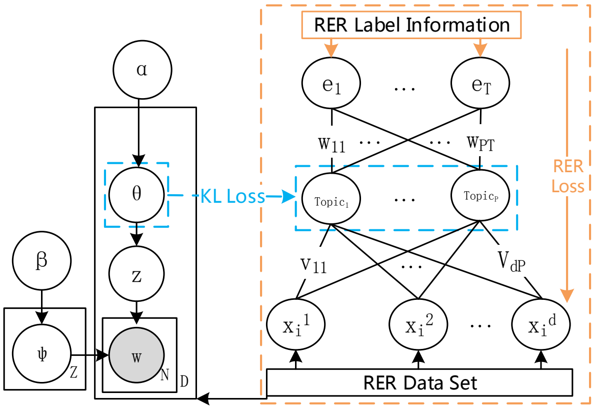

Figure 1: The overall framework of Interpretable Neural Network for Relevant Emotion Ranking (INN-RER).

The overall framework of INN-RER is shown in Figure 1. The left part is a typical topic model (Blei et al., 2003). It is designed for discovering the main topics that pervade a large unstructured collection of documents. A document is allowed to contain a mixture of topics with different weights. As such, a document d can be represented by its topic distribution θ_d. The right part is a three-layer neural network. It has d input units corresponding to the d-dimensional feature vector of each training sample x_i, T output units corresponding to all possible emotion labels, and one hidden layer with P hidden units corresponding to the hidden topics. The input layer is fully connected to the hidden layer with weights $V = [v_{qh}](1 \leq q \leq d, 1 \leq h \leq P)$ and the hidden layer is fully connected to the output layer with weights $W = [w_{hj}](1 \leq h \leq P, 1 \leq j \leq T)$. The bias parameters $\alpha_h(1 \leq h \leq P)$ of the hidden units are considered as weights from an extra input unit with a fixed value of 1. Similarly, the bias parameters $\beta_j(1 \leq j \leq T)$ of the output units are considered as weights from an extra hidden unit, with a fixed value of 1.

The learning process of INN-RER consists of two main steps. Firstly, the first two layers of the network are initialized based on the output of the topic model. The feature transformation in neural network is conducted by minimizing the Kullback-Leibler (KL) divergence between the topic distribution θ produced by the topic model and the approximated distribution $[Topic_1, Topic_2, ..., Topic_P]$ learned by the first two layers of the neural network, which is denoted by the blue rectangular dash line boxes in Figure 1. Then, the whole network is learnt and fine-tuned based on the novel loss function, which is denoted as the orange rectangular dash line boxes. Each step will be described in details in the following subsections.

3.1 INN-RER Initialization

As the number of hidden neurons and its semantic meaning is usually treated as a black box in conventional neural networks, the generated topics from the topic model are employed for guiding the construction of the hidden layer in the proposed neural network. By doing that, semantic topic information is incorporated to enhance the interpretability and accuracy of the proposed neural network. For a particular text sample x_i in training set G, the input layer takes its term-frequency representation x_i^q as the input and feeds it to the hidden layer. Assuming the total number of topics is fixed as P, then the hidden layer would contain P neurons. The topic mixture θ_{x_i} generated from the topic model is approximated by the weights connecting the input and the hidden layers. Mathematically, the initialization procedure learns a function $f(x_i^q|v_{qh}, \alpha_h)$ so that the output of $f(x_i^q|v_{qh}, \alpha_h)$ is as close to θ_{x_i} as possible, where x_i^q, v_{qh}, α_h and θ_{x_i} denote the input, weight vector, bias of the first two layers of the network and the topic distribution of text x_i generated from the topic model, respectively. A softmax function is applied to the output of the hidden layer, i.e., $f(x_i^q|v_{qh}, \alpha_h)$, and the Kullback-Leibler divergence (Leahy, 2006) is employed as follows:

$$L(\theta, f) = \theta \log \frac{\theta}{f} + (1-\theta) \log \frac{1-\theta}{1-f} \quad (1)$$

where θ denotes a topic distribution derived from the topic model, and f denotes the output of the hidden layer. The KL divergence is a measure of the difference between two distributions. It is always non-negative and equals to zero when the two distributions are the same. As shown in Equation 1, the KL divergence can describe the difference between the topic distribution generated from topic models and the approximate distributions learned in the initialization procedure. Note that the topic distribution for a document generated by the topic model is used as the supervision information for initializing INN-RER. Thus, maximizing the log-likelihood is equivalent to minimizing the KL divergence according to Equation 1, and its gradients are as follows:

$$\frac{\partial L(\theta, f)}{\partial v_{qh}} = -(\theta_{x_i,h} - f_h(x_i^q | v_{qh}, \alpha_h)) \cdot x_i^q \quad (2)$$

$$\frac{\partial L(\theta, f)}{\partial \alpha_h} = -(\theta_{x_i,h} - f_h(x_i^q | v_{qh}, \alpha_h)) \quad (3)$$

According to the gradient descent method, the first two layers can be initialized iteratively by Equation 2 and 3. The initialization procedure for INN-RER is shown in Algorithm 1. η_{init} with the subscript $init$ represents the learning rate during the initialization procedure and λ is the penalty term. Note that the first two layer should be learnt from topic model as much as possible in order to incorporate topic information, thus the learning rate term η_{init} should be larger than the learning rate during training procedure.

3.2 INN-RER Learning

This step aims to optimize the three-layer neural network to tackle the relevant emotion ranking problem. It can adjust the neural network initialized at previous step at the same time. An intuitive way is to define the global error function for the network on the training set. However, some important characteristics of relevant emotion ranking, such as ranking, not considering irrelevant emotions, are not considered in the classical back propagation algorithm (Rumelhart et al., 1988). The error function defined in traditional neural network such as mean-square error only focuses on individual label discrimination, i.e. whether a predicted label is correct or not. It does not consider the correlations between different labels of a training instance, e.g., relevant emotions should be ranked higher than irrelevant ones and there is a ranking for relevant emotions according to their intensities. Therefore, to fulfil the requirements of relevant emotion ranking, a novel global error function is defined as follows:

Algorithm 1 Algorithm of INN-RER Initialization.

Input: x_i^q: Term frequency of text x_i; θ_{x_i}: Topic distribution of text x_i
Output:$\Delta v, \Delta\alpha$: gradient approximation of initialization procedure

```
Initialize $\Delta v, \Delta\alpha$ as random values
for each iteration do
    for each text $x_i \in G$ do
        for $q = 1, ..., d, h = 1, ..., P$ do
            $\Delta v_{qh} \leftarrow \Delta v_{qh} + \eta_{init} \cdot (\theta_{x_i,h} - f_h(x_i^q|v_{qh}, \alpha_h)) \cdot x_i^q + \lambda \cdot \Delta v_{qh}$
        end for
        for $h = 1, ..., P$ do
            $\Delta\alpha_h \leftarrow \Delta\alpha_h + \eta_{init} \cdot (\theta_{x_i,h} - f_h(x_i^q|v_{qh}, \alpha_h)) + \lambda \cdot \Delta\alpha_h$
        end for
    end for
end for
```

$$E = \sum_{i=1}^{n} \sum_{e_t \in R_i} \sum_{e_s \in \prec(e_t)} \frac{1}{norm_{t,s}} \Big[\exp(-(g_t(x_i) - g_s(x_i))) + \omega_{ts}(g_t(x_i) - g_s(x_i))^2 \Big] \quad (4)$$

Here, emotion e_t and emotion e_s are two emotion labels and e_s is less relevant than emotion e_t, represented by $e_s \in \prec (e_t)$. The normalization term $norm_{t,s}$ is used to balance emotion pairs (e_t, e_s) to avoid dominated terms by their set sizes. The term $g_t(x_i) - g_s(x_i)$ measures the difference between two emotion outputs, e_t and e_s, of a given input text x_i. We want the difference as larger as possible. Furthermore, the negation of this difference is fed to the exponential function in order to severely penalize the i-th error term if emotion e_t is much smaller than e_s. As the relationships among different emotions can provide important clues for emotion detection, we further incorporate the information into the loss function as constraints. Here, ω_{ts} is the relationship between emotion e_t and e_s which is calculated by Pearson correlation coefficient (Nicewander, 1988).

The minimization of the global relevant emotion ranking loss function defined in Equation 4 is carried out by gradient descent combined with the back propagation (Rumelhart et al., 1988). For training instance x_i and its label set L_i, the actual output of the j-th output neuron is(omitting the superscript i without loss of generality):

$$g_j = f(netg_j + \beta_j) \tag{5}$$

where β_j is the bias of the j-th output neuron which is a ”tanh” function:

$netg_j$ is the input to the j-th output neuron:

$$netg_j = \sum_{h=1}^{P} b_h w_{hj} \tag{6}$$

where w_{hj} is the weight which connects the h-th hidden neuron and the j-th output neuron, and P is the number of hidden neurons, i.e., the topics. b_h is the output of the h-th hidden neuron:

$$b_h = f(netb_h + \alpha_h) \tag{7}$$

where α_h is the bias of the h-th hidden neuron, $f()$ is also a “tanh” function. $netb_h$ is the input to the h-th hidden neuron:

$$netb_h = \sum_{q=1}^{d} x^q v_{qh} \tag{8}$$

where x^q is the q-th dimension of instance x. v_{qh} is the weight which connects the q-th input neuron and the h-th hidden neuron.

“tanh” function is differentiable, the error of the j-th output neuron can be defined as:

$$d_j = \begin{cases} \left[\frac{1}{norm}exp(-(g_j - g_s)) + 2\omega_{js}(g_j - g_s)\right] \\ (1 + g_j)(1 - g_j), if e_j \in R_i \& e_s \in \prec (e_j) \\ \left[-\frac{1}{norm}exp(-(g_t - g_j)) - 2\omega_{tj}(g_t - g_j)\right] \\ (1 + g_j)(1 - g_j), \\ if(e_j \in R_i \& e_j \in \prec (e_t)) or \\ (e_j \in \overline{R_i} \& e_j \in \prec (e_t)) \end{cases} \tag{9}$$

Similarly, the error of the h-th hidden neuron can be defined as:

$$e_h = \left(\sum_{j=1}^{T} g_j w_{hj}\right)(1 + b_h)(1 - b_h) \tag{10}$$

In order to reduce the error of the neural network INN-RER, we can use gradient descent strategy:

$$\begin{aligned} \Delta w_{hj} &= -\eta\frac{\partial E_i}{\partial w_{hj}} = -\eta\frac{\partial E_i}{\partial netg_j}\frac{\partial netg_j}{\partial w_{hj}} \\ &= \eta d_j \left[\frac{\partial[\sum_{h=1}^{P} b_h w_{hj}]}{\partial w_{hj}}\right] = \eta d_j b_h \end{aligned} \tag{11}$$

$$\begin{aligned} \Delta v_{qh} &= -\eta\frac{\partial E_i}{\partial v_{qh}} = -\eta\frac{\partial E_i}{\partial netb_h}\frac{\partial netb_h}{\partial v_{qh}} \\ &= \eta e_h \left[\frac{\partial[\sum_{q=1}^{d} x^q v_{qh}]}{\partial v_{qh}}\right] = \eta e_h x^q \end{aligned} \tag{12}$$

the biases are updated as follows:

$$\Delta\beta_j = \eta d_j; \Delta\alpha_h = \eta e_h \tag{13}$$

where η is the learning rate.

The training process of the neural network is presented in Algorithm 2.

Algorithm 2 Algorithm of INN-RER Learning.

Input: x_i^q: Term frequency of text x_i; $\Delta v, \Delta\alpha$: Parameters after initialization; L: emotion labels

Output: A predictable neural network INN-RER.

```
Initialize INN-RER network parameters from Algorithm 1
for each iteration do
    for each text $x_i \in G$ do
        Forward compute output of INN-RER's score function g given $x_i$.
        Backward compute the gradient according to g and $L$ based on the relevant emotion ranking loss function with learning rate of $\eta_{learn}$ and penalty term $\lambda$.
    end for
end for
```

4 Experiments

We evaluate the proposed approach on the following three corpora:

Sina Social News (News) was collected from the Sina news *Society* channel where readers can choose one of the six emotions such as *Amusement*, *Touching*, *Anger*, *Sadness*, *Curiosity*, and *Shock* after reading a news article. As Sina is one of the largest online news sites in China, it is sensible to carry out experiments to explore the readers' emotion (social emotion). News articles with less than 20 votes were discarded since few votes can not be considered as proper representation of social emotion. In total, 5,586 news articles published from January 2014 to July 2016 were kept, together with the readers' emotion votes.

Ren-CECps corpus (Blogs) (Quan and Ren, 2010) contains 1,487 blogs in Chinese. Each document is annotated with eight basic emotions from writer's perspective, including *anger*, *anxiety*, *expect*, *hate*, *joy*, *love*, *sorrow* and *surprise*, together with their emotion scores indicating the level of emotion intensity in the range of $[0, 1]$. Higher scores represent higher emotion intensity.

SemEval (Strapparava and Mihalcea, 2007) is an English data set containing 1,250 news headlines extracted from Google news, CNN, and many other portals. The news headlines are typically short. Each headline was manually scored in a fine-grained valence scale of 0 to 100 across 6 emotions (i.e., *anger*, *disgust*, *fear*, *joy*, *sad* and *surprise*). After pruning 4 items with the total scores equal to 0, 1246 headlines are got for the experiments.

News		Blogs		SemEval	
Category	#Votes	Category	#Scores	Category	#Scores
Touching	694,006	Joy	349.2	anger	12042
Shock	572,651	Hate	174.2	disgust	7634
Amusement	869,464	Love	610.6	fear	20306
Sadness	837,431	Sorrow	408.4	joy	23613
Curiosity	212,559	Anxiety	422.6	sad	24039
Anger	1,109,315	Surprise	59.2	surprise	21495
		Anger	116.4		
		Expect	385.5		
All	4,295,426	All	2526.1	All	109,129

Table 1: Statistics for the three corpora used in our experiments.

The statistics of the three corpora are shown in Table 1. The first two corpora were preprocessed by using the python jieba segmenter[1] for word segmentation and filtering. The third corpus SemEval is in English and can be tokenized by white spaces. Stop words and words appeared only once or in less than two documents were removed to alleviate data sparsity. Next, TF-IDF (term frequency-inverse document frequency) was used to extract the features from text. TF-IDF is a numerical statistic method that is designed to reflect how important a word is to a document in a corpus. In our experiments, we set the dimension of each text representation to 2,000 according to the ranking of the TF-IDF weights with each dimension of term-frequency(TF) features. After that, the text representations are fed into the proposed INN-RER method.

η_{init}, η_{learn}, λ, the number of iterations and the number of topics are set to 0.9, 0.1, 0.001, 100 and 60 respectively. The parameters were chosen by 10-fold cross-validation. The topic distribution used in INN-RER are derived in different ways. For long text such as News and Blogs, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is employed for generating topic distributions. For short texts in Semeval, bi-term topic model (BTM) (Cheng et al., 2014) was used, since short text typically contains a few words which results in sparse word co-occurrence patterns. BTM is a variant of LDA which effectively infers the latent topic distribution of short text by modeling the generation of bi-terms in the whole corpus and it alleviates the problem of sparsity at the document level. For each method, 10-fold cross validation is conducted using the same feature construction method to get the final performance.

Evaluation metrics typically used in multi-label learning and label ranking are employed which are different from those of classical single-label learning systems (Sebastiani, 2001). The detailed explanation of evaluation metrics are presented in Table 2. Note that metrics from PRO Loss to $F1_{exam}$ work by evaluating performance on each test example separately and returning the mean value across test set. MicroF1 and MacroF1 work by evaluating performance on each emotion category separately and returning the macro/micro-averaged value across all emotion categories.

4.1 Experimental Results

There are several approaches addressing multiple emotions detection from texts. Three generative model based baselines and three discriminative model based baselines are chosen.

- **Emotion Distribution Learning (EDL)** (Zhou et al., 2016) learns a mapping

[1] https://github.com/fxsjy/jieba

Corpus	Category	Method	Criteria								
			PL(↓)	HL(↓)	RL(↓)	OE(↓)	AP(↑)	Cov(↓)	F1(↑)	MiF1(↑)	MaF1(↑)
News	Generative	MSTM	0.3343	0.4065	0.3097	0.2123	0.6677	3.3202	0.5666	0.5853	0.5044
		SLTM	0.3205	0.3639	0.2753	0.2008	0.7326	2.9863	0.6095	0.6429	0.4899
		ATM	0.3192	0.3743	0.2507	0.1947	0.7490	2.9369	0.6127	0.6412	0.4885
	Discriminative	EDL	0.2348	0.2510	0.1616	0.2243	0.8372	2.1940	0.6260	0.6454	0.5703
		EmoDetect	0.2157	0.2575	0.1538	0.1627	0.8605	2.1761	0.6697	0.6739	0.5359
		RER	0.2142	0.2498	0.1491	0.1513	0.8633	2.1989	0.6820	0.6919	0.6198
	Our model	**INN-RER(-t)**	0.1998	0.2420	0.1393	0.1456	0.8715	2.1377	**0.7116**	0.7137	0.6242
		INN-RER	**0.1973**	**0.2312**	**0.1353**	**0.1331**	**0.8764**	**2.1339**	0.7108	**0.7161**	**0.6282**
Blogs	Generative	MSTM	0.3567	0.4171	0.3030	0.4761	0.6046	3.7005	0.5236	0.4978	0.4758
		SLTM	0.3148	0.3769	0.2397	0.4598	0.6547	**3.2513**	0.5757	0.5865	0.5283
		ATM	0.3493	0.3890	0.2885	0.4385	0.6278	3.4278	0.5105	0.5260	0.5026
	Discriminative	EDL	0.3385	0.3916	0.2550	0.4206	0.6962	4.2491	0.5060	0.5396	0.4131
		EmoDetect	0.3115	0.3848	0.2123	0.2880	0.7617	4.1650	0.5340	0.5492	0.4387
		RER	0.3007	0.3657	0.2043	0.2728	0.7746	4.1638	0.5957	0.6084	**0.5342**
	Our model	**INN-RER(-t)**	0.2868	0.3268	0.1993	0.2695	0.7751	3.9653	0.6132	0.6165	0.5069
		INN-RER	**0.2829**	**0.3209**	**0.1924**	**0.2626**	**0.7784**	3.6418	**0.6187**	**0.6225**	0.5133
SemEval	Generative	MSTM	0.3524	0.3835	0.2796	0.3698	0.7653	3.1986	0.6902	0.7133	0.5854
		SLTM	0.3155	0.3253	0.2370	0.3150	0.8052	2.9589	0.7016	0.7278	0.5889
		ATM	0.3138	0.3276	0.2389	0.3767	0.8302	2.8976	0.7039	**0.7292**	0.5244
	Discriminative	EDL	0.4130	0.4291	0.3401	0.3875	0.7345	3.3433	0.4002	0.4136	0.3813
		EmoDetect	0.3176	0.3167	0.2411	0.2308	0.8241	3.0439	0.6275	0.6245	0.5385
		RER	**0.2907**	0.3128	0.2389	**0.2220**	0.8302	2.9963	0.6839	0.6898	**0.6283**
	Our model	**INN-RER(-t)**	0.3213	0.3026	0.2331	0.2388	0.8364	2.8773	0.7019	0.7118	0.5973
		INN-RER	0.3194	**0.3005**	**0.2302**	0.2261	**0.8379**	**2.8632**	**0.7081**	0.7156	0.6093

Table 3: Experimental results of the proposed approach and the baselines. 'PL' represent Pro Loss, 'HL' represents Hamming Loss, 'RL' represents ranking loss, 'OE' represents one error, 'AP' represent average precision, 'Cov' represent coverage, 'F1' represents $F1_{exam}$, MiF1' represents MicroF1, 'MaF1' represents MacroF1. "↓" indicates "the smaller the better", while "↑" indicates "the larger the better". The best performance on each evaluation measure is highlighted by boldface.

Name	**Definition**
PRO Loss	$\frac{1}{n}\sum_{i=1}^{n}\sum_{e_t\in R_i\cup\{\Theta\}}\sum_{e_s\in\prec(e_t)}\frac{1}{norm_{t,s}}l_{t,s}$ $l_{t,s}$ is a modified 0-1 error;$norm_{t,s}$is the set size of label pair(e_t, e_s)
Hamming Loss	$\frac{1}{nT}\sum_{i=1}^{n}\lvert\hat{R}_i\triangle R_i\rvert$ The predicted relevant emotions: $\hat{R}_i$.
Ranking Loss	$\frac{1}{n}\sum_{i=1}^{n}(\sum_{(e_t,e_s)\in R_i\times\overline{R_i}}\delta[g_t(x_i)<g_s(x_i)])/(\lvert R_i\rvert\times\lvert\overline{R_i}\rvert)$ where δ is the indicator function.
One Error	$\frac{1}{n}\sum_{i=1}^{n}\delta[\underset{\mathbf{e_t}}{\text{argmax}}\, g_t(x_i)\notin R_i]$
Average Precision	$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\lvert R_i\rvert}\times$ $(\sum_{t:e_t\in R_i}\lvert\{e_s\in R_i\mid g_s(x_i)>g_t(x_i)\}\rvert)/(\lvert\{e_s\mid g_s(x_i)>g_t(x_i)\}\rvert)$
Coverage	$\frac{1}{n}\sum_{i=1}^{n}\max_{t:e_t\in R_i}\lvert\{e_s\mid g_s(x_i)>g_t(x_i)\}\rvert$
$F1_{exam}$	$\frac{1}{n}\sum_{i=1}^{n}2\lvert R_i\cap\hat{R}_i\rvert/(\lvert R_i\rvert+\lvert\hat{R}_i\rvert)$
MicroF1	$F1(\sum_{t=1}^{T}TP_t,\sum_{t=1}^{T}FP_t,\sum_{t=1}^{T}TN_t,\sum_{t=1}^{T}FN_t)$
MacroF1	$\frac{1}{T}\sum_{t=1}^{T}F1(TP_t,FP_t,TN_t,FN_t)$

Table 2: Evaluation criteria for the Multi-Label Learning (MLL) methods. TP_t, FP_t, TN_t, FN_t represent the number of true positive, false positive, true negative, and false negative test examples with respect to emotion t respectively. $F1(TP_t, FP_t, TN_t, FN_t)$ represent specific binary classification metric F1 (Manning et al., 2008).

function from texts to their emotion distributions based on label distribution learning.

- **EmoDetect** (Wang and Pal, 2015) outputs the emotion distribution based on a dimensionality reduction method using non-negative matrix factorization which combines several constraints such as emotions bindings and topic correlations.

- **RER** (Zhou et al., 2018) predicts multiple emotions and their rankings from text based on relevant emotion ranking using support vector machines.

- **Multi-label supervised topic model (MSTM) and Sentiment latent topic model (SLTM)** (Rao et al., 2014b): As the variants of supervised topic models, MSTM and SLTM connect latent topics with evoked emotions of

Touching		Anger		Amusement	
Topic 1	Topic 2	Topic 1	Topic 2	Topic 1	Topic 2
救人(save)	教师(teacher)	歹徒(ruffian)	犯罪(sin)	男女(men and women)	网上(network)
照顾(take care of)	辛苦(hard)	强行(force)	嫌疑人(suspect)	宾馆(hotel)	醉酒(drunkenness)
身亡(sacrifice)	落水(fall into water)	猥亵(obscenity)	徒刑(imprisonment)	服务(service)	检察院(procuratorate)
治疗(cure)	年轻(youth)	女童(girl)	打人(beat)	照片(photo)	违法(illegal)
生命(life)	病情(state of an illness)	杀害(murder)	殴打(hit)	报警(call the police)	罚款(penalty)
老人(older)	坚持(persist)	造成(cause)	工地(construction site)	鉴定(authenticate)	调查(investigate)
感谢(grateful)	群众(public)	派出所(police station)	交警(traffic police)	诈骗(defraud)	违规(get out of line)
医院(hospital)	车祸(traffic accident)	作案(commit a crime)	采访(interview)	网络(internet)	现金(cash)
感动(moved)	感动(touching)	死亡(death)	曝光(exposure)	离婚(divorce)	警官(police officer)
Sadness		Curiosity		Shock	
Topic 1	Topic 2	Topic 1	Topic 2	Topic 1	Topic 2
失踪(disappear)	车祸(car accident)	家长(parents)	监控(monitoring)	抢劫(rob)	菜刀(kitchen knife)
不幸(misfortune)	小偷(thief)	中国(China)	妇女(women)	尸体(corpse)	脖子(neck)
去世(pass away)	公安(public security)	婚姻(marriage)	春节(spring festival)	紧急(emergency)	重症(sever illness)
杀害(murder)	围观(watch)	健康(health)	医院(hospital)	现场(scene)	地铁(subway）
犯罪(crime)	鉴定(identify)	女子(women)	怀孕(pregnancy)	安全(security)	新闻(news)
遭到(suffer)	道歉(apologize)	年轻(young)	早上(morning)	治疗(cure)	竟然(unexpectedly)
公安局(Public Security Bureau)	激动(excite)	结婚(marry up)	抢救(rescue)	生命(life)	银行(bank)
出事(have an accident)	执法(enforce the law)	男性(men)	无效(in vain)	检查(examine)	赔偿(compensate)
媒体(media)	派出所(police station)	现金(money)	喜欢(like)	家属(family member)	消费(consume)

Figure 2: The top topic words under each emotion category from the News corpus.

texts. MSTM first generates a set of topics from words, and then samples emotions from each topic. SLTM generates topics directly from emotions.

- **Affective topic model (ATM)** (Rao et al., 2014a) employs the exponential distribution to generate ratings for each emotion.

We also evaluated INN-RER with random initialization instead of the proposed initialization procedure, which is denoted as INN-RER(-t).

Experimental results on the three corpora are summarized in Table 3. It can be observed from the table that: (1) INN-RER outperforms the baselines on almost all evaluation metrics across all the data sets; (2) INN-RER achieves better performance on almost all the evaluation metrics than INN-RER(-t), which further verifies the effectiveness of incorporating the topic information; (3) Both INN-RER and INN-RER(-t) perform remarkably better than RER which is based on linear models. It verifies the effectiveness of using the neural networks for RER task, which are able to learn dynamic and complex functions.

Joy	Anger	Sad	Disgust	Fear	Surprise
home	kill	flu	sex	kill	sue
heart	attack	cancer	immigr	danger	korea
game	violenc	terror	scandal	iran	blast
youtub	terror	danger	porn	dead	north
movie	stop	health	charg	state	fight
friend	fire	kill	insist	fear	war
sleep	blast	flood	women	terror	nuclear
miss	death	crash	held	global	shoot
award	condemn	end	girl	attack	protest

Figure 3: The top topic words under each emotion category from the Semeval corpus.

Joy	Hate	Love	Sorrow
花儿(flower)	孤独(lonely)	学习(study)	角落(corner)
新年(ney year)	面对(face with)	比赛(competition)	希望(hope)
宝贝(baby)	无情(heartless)	开心(happy)	天堂(heaven)
享受(enjoy)	重新(again)	感觉(feeling)	寂寞(lonely)
快乐(happy)	情绪(emotion)	心境(mood)	地震(earthquake)
祝福(wish)	失去(lose)	充满(full of)	使命(mission)
宝宝(baby)	脾气(temper)	文化(culture)	男朋友(boyfriend)
开心(joyful)	痛苦(pain)	作品(production)	离开(leave)
微笑(smile)	完全(entirely)	丰富(rich)	无奈(helpless)
Anxiety	Surprise	Anger	Expect
房子(house)	彩虹(rainbow)	离开(leave)	希望(hope)
婚姻(marriage)	北海道(Hokkaido)	离婚(divorce)	责任(responsible)
老公(husband)	忽然(sudden)	无奈(helpless)	女性(women)
错误(error)	记忆(memory)	法律(law)	奥运会(Olympic)
心情(mood)	礼物(gift)	银行(bank)	幸福(happiness)
陌生(strange)	奇迹(miracle)	道德(morality)	行为(action)
家里(family)	据说(reputedly)	情感(emotion)	努力(strive)
上班(on duty)	好奇(curious)	悲伤(sorrow)	以后(later)
城市(city)	季节(season)	自己(self)	精彩(splendid)

Figure 4: The top topic words for each emotion category from the Blogs corpus.

4.2 INN-RER Interpretation

In addition to comparing the performance of the proposed model with several baselines, we also present the experimental results from the perspective of result interpretation to fully understand INN-RER. The topic words of each emotion in three corpora are extracted according to the ranking of weights learned by INN-RER, i.e., the probabilities of topics conditioned on emotions (weights between the hidden layer and the output layer) and words conditioned on topics (weights between the input layer and the hidden layer). Results are shown in Figure 2, 3, 4 respectively. It can be observed that the extracted topics words under each emotion category correspond to a certain event, which evokes the emotion. It is in ac-

Corpus	Method	Criteria								
		PL(↓)	HL(↓)	RL(↓)	OE(↓)	AP(↑)	Cov(↓)	F1(↑)	MiF1(↑)	MaF1(↑)
News	RANK-SVM	0.2842	0.2872	0.2114	0.2034	0.7967	2.5358	0.5066	0.5656	0.5298
	BP-MLL	0.2118	0.2399	0.1443	0.1544	0.8677	2.1738	0.6881	0.6915	0.6013
	LIFT	0.2224	0.3363	0.1382	0.1411	0.8234	2.1394	0.6646	0.6801	0.6151
	INN-RER	**0.1973**	**0.2312**	**0.1353**	**0.1331**	**0.8764**	**2.1339**	**0.7108**	**0.7161**	**0.6282**
Blogs	RANK-SVM	0.3888	0.3786	0.3356	0.3219	0.7030	4.0801	0.3489	0.3686	0.3210
	BP-MLL	0.2987	0.3281	0.2141	0.2727	0.7267	3.9802	0.5844	0.6065	0.4833
	LIFT	0.3452	0.3817	0.3089	0.3306	0.7557	**3.1290**	0.6053	0.6113	**0.5155**
	INN-RER	**0.2829**	**0.3209**	**0.1924**	**0.2626**	**0.7784**	3.6418	**0.6187**	**0.6225**	0.5133
SemEval	RANK-SVM	0.3452	0.3617	0.3083	0.3006	0.7557	3.1290	0.6253	0.6472	0.5955
	BP-MLL	0.3790	0.3656	0.3605	0.3790	0.7495	3.2097	0.5868	0.6101	0.5402
	LIFT	0.4279	0.4651	0.3627	0.4113	0.7344	3.2823	0.6299	0.6469	**0.6112**
	INN-RER	**0.3194**	**0.3005**	**0.2302**	**0.2261**	**0.8379**	**2.8632**	**0.7081**	**0.7156**	0.6093

Table 4: Comparison with Multi-Label Learning (MLL) Methods. The evaluation criteria are same as in Table 3.

cord with what has been observed in social psychology (Stoyanov and Cardie, 2008). For example, in the Sina corpus, Topic 1 under the emotion *touching* is about **"heroic rescue"**; Topic 1 under the emotion *anger* is about **"sexual molestation of a child"** and Topic 2 under the emotion *sadness* is about an **"car accident"**. In the SemEval and the Blog corpora, we can also find that topic words listed under each emotion category are related to some social events. For example, in the SemEval corpus, the *Joy* topic is about **"home entertainment"** and the *Anger* topic is about **"terrorist attack"**. In the Blog corpus, the *sorrow* topic is about **"earthquake and the lost of their loved ones"**. The extracted emotion-associated topic words unveil how the corresponding emotion is evoked. By incorporating topical information into neural network learning, we are able to obtain more interpretable results from INN-RER.

4.3 Comparison with Multi-Label Methods

Since relevant emotion ranking can be seen as an extension of multi-label learning, the proposed INN-RER is also compared with three widely used well-established multi-label learning methods such as LIFT (Zhang, 2011), Rank-SVM (Zhang and Zhou, 2014) and BP-MLL (Zhang and Zhou, 2006). In our experiments, LIFT used linear kernel and Rank-SVM uses the RBF kernel with the width σ set to 1 using the threshold Θ which is initialized as 0.15 after normalization.

The results of INN-RER in comparison with MLL baselines are presented in Table 4. It can be observed that INN-RER outperforms all the baselines across all the datasets on all evaluation measures most of the time. This further verifies the effectiveness of our proposed INN-RER for multi-label emotion detection due to its consideration of rankings of the relevant emotions and the incorporation of topic models.

5 Conclusion

In this paper, we have proposed a novel interpretable neural network for relevant emotion ranking. Specifically, motivated by transfer learning, the neural network is initialized to make its hidden layer approximate the behavior of a topic model. Moreover, a novel error function is defined to optimize the whole neural network for relevant emotion ranking. Experimental results on three real-world corpora show that the proposed approach performs remarkably better than the state-of-the-art emotion detection approaches and multi-label learning methods. Moreover, the extracted emotion-associated topic words indeed represent emotion-evoking events which are in line with our common-sense knowledge. In the future, we will explore the possibility of learning a topic model and an emotion ranking function simultaneously in a unified framework.

Acknowledgments

The work was supported by the National Key R&D Program of China (No. 2017YFB1002801), the National Natural Science Foundation of China (61772132,61528302) and the Natural Science Foundation of Jiangsu Province of China (BK20161430).

References

Shenghua Bao, Shengliang Xu, Li Zhang, Rong Yan, Zhong Su, Dingyi Han, and Yong Yu. 2012. Mining social emotions from affective text. *IEEE transactions on knowledge and data engineering*, 24(9):1658–1670.

Plaban Kumar Bhowmick. 2009. Reader perspective emotion analysis in text through ensemble based multi-label classification framework. *Computer and Information Science*, 2(4):64.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Xueqi Cheng, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. 2014. Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):2928–2941.

Johannes Fürnkranz, Eyke Hüllermeier, Eneldo Loza Mencía, and Klaus Brinker. 2008. Multilabel classification via calibrated label ranking. *Machine learning*, 73(2):133–153.

D. E. Leahy, D. P. Searson. 2006. Gaussian processes for machine learning. *International Journal of Neural Systems*, 14(2):69.

Xiangsheng Li, Jianhui Pang, Biyun Mo, and Yanghui Rao. 2016. Hybrid neural networks for social emotion detection over short text. In *International Joint Conference on Neural Networks*, pages 537–544.

Kevin Hsin-Yih Lin, Changhua Yang, and Hsin-Hsi Chen. 2008. Emotion classification of online news articles from the reader's perspective. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 220–226. IEEE Computer Society.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schtze. 2008. An introduction to information retrieval. *Journal of the American Society for Information Science and Technology*, 43(3):824–825.

W. Alan Nicewander. 1988. Thirteen ways to look at the correlation coefficient. *American Statistician*, 42(1):59–66.

Rosalind W Picard and Roalind Picard. 1997. *Affective computing*, volume 252. MIT press Cambridge.

Changqin Quan and Fuji Ren. 2010. Sentence emotion analysis and recognition based on emotion words using ren-cecps. *International Journal of Advanced Intelligence Paradigms*, 2(1):105–117.

Xiaojun Quan, Qifan Wang, Ying Zhang, Luo Si, and Liu Wenyin. 2015. Latent discriminative models for social emotion detection with emotional dependency. *ACM Transactions on Information Systems (TOIS)*, 34(1):2.

Yanghui Rao. 2016. Contextual sentiment topic model for adaptive social emotion classification. *IEEE Intelligent Systems*, 31(1):41–47.

Yanghui Rao, Qing Li, Wenyin Liu, Qingyuan Wu, and Quan Xiaojun. 2014a. Affective topic model for social emotion detection. *Neural Netw*, 58(5):29–37.

Yanghui Rao, Qing Li, Xudong Mao, and Wenyin Liu. 2014b. Sentiment topic models for social emotion mining. *Information Sciences*, 266(5):90–100.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. 1988. *Learning internal representations by error propagation*. MIT Press.

Fabrizio Sebastiani. 2001. Machine learning in automated text categorization. *Acm Computing Surveys*, 34(1):1–47.

Veselin Stoyanov and Claire Cardie. 2008. Annotating topics of opinions. In *International Conference on Language Resources and Evaluation, Lrec 2008, 26 May - 1 June 2008, Marrakech, Morocco*, pages 3213–3217.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing*, pages 1556–1560. ACM.

Yichen Wang and Aditya Pal. 2015. Detecting emotions in social media: A constrained optimization approach. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, pages 996–1002.

Min Ling Zhang. 2011. Lift: multi-label learning with label-specific features. In *International Joint Conference on Artificial Intelligence*, pages 1609–1614.

Min Ling Zhang and Zhi Hua Zhou. 2006. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge Data Engineering*, 18(10):1338–1351.

Min-Ling Zhang and Zhi-Hua Zhou. 2014. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837.

Deyu Zhou, Yang Yang, and Yulan He. 2018. Relevant emotion ranking from text constrained with emotion relationships. In *Meeting of the North American Chapter of the Association for Computation Linguistics*.

Deyu Zhou, Xuan Zhang, Yin Zhou, Quan Zhao, and Xin Geng. 2016. Emotion distribution learning from texts. In *Conference on Empirical Methods in Natural Language Processing*, pages 638–647.