Relevant Emotion Ranking from Text Constrained with Emotion Relationships

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

Abstract

Text might contain or invoke multiple emotions with varying intensities. As such, emotion detection, to predict multiple emotions associated with a given text, can be cast into a multi-label classification problem. We would like to go one step further so that a ranked list of relevant emotions are generated where top ranked emotions are more intensely associated with text compared to lower ranked emotions, whereas the rankings of irrelevant emotions are not important. A novel framework of relevant emotion ranking is proposed to tackle the problem. In the framework, the objective loss function is designed elaborately so that both emotion prediction and rankings of only relevant emotions can be achieved. Moreover, we observe that some emotions co-occur more often while other emotions rarely co-exist. Such information is incorporated into the framework as constraints to improve the accuracy of emotion detection. Experimental results on two real-world corpora show that the proposed framework can effectively deal with emotion detection and performs remarkably better than the state-of-the-art emotion detection approaches and multi-label learning methods.

1 Introduction

With the growing prosperity of Web 2.0, people tend to share their feelings, attitudes and opinions through the social platforms such as online news sites, blogs. Detecting emotions from text 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 analyze and predict emotions from text accurately (Picard and Picard, 1997).

Research on emotion detection can be roughly categorized into two types: lexicon-based and learning-based approaches. Lexicon-based approaches usually rely on emotion lexicons (Lei et al., 2014; Rao et al., 2012). They cannot deal with text when words can't be found in emotion lexicons. Learning-based approaches can be furthered classified into unsupervised and supervised learning methods. Unsupervised approaches do not require annotated data for training. For example, by adding an emotion layer into traditional topic models, emotion-topic models were constructed to detect users' emotions (Bao et al., 2012, 2009). Supervised learning approaches consider each emotion category as a class label and emotion detection is cast as a classification problem. If only choosing the strongest emotion as the emotion label for a given text, emotion detection is essentially a single-label classification problem (Lin et al., 2008; Quan et al., 2015). To predict multiple emotions simultaneously, emotion detection can be solved in the multi-label classification framework (Bhowmick, 2009). Moreover, to predict both multiple emotions and their intensities, some approaches have been proposed using emotion distribution learning (Zhou et al., 2016). Some lexicon-based approaches such as (Wang and Pal, 2015) can also output multiple emotions with intensities using non-negative matrix factorization.

In this paper, we are interested in exploring emotion ranking from either readers' perspective or writers' perspective in two different real-world corpora. In both cases, a given text is associated with multiple emotions. For example, Figure 1 illustrates an online news article crawled from Sina News *Society* Channel together with readers' emotion votes. It can be observed that when reading the news article, readers expressed different emotions with the majority showed "*Sadness*" and "*Anger*". We notice that some emotions such as "*Touching*", "*Curiosity*" and "*Amusement*" only

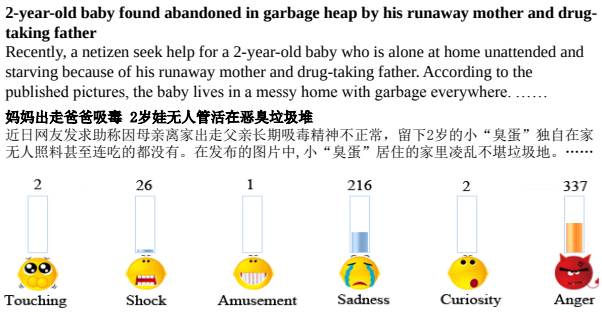

Figure 1: An example of an online news article from Sina *Society* Channel with voted emotions.

received 1 to 3 votes. In comparison to the total number of votes received, these votes could be considered as outliers or irrelevances. Also, the extremely low emotion votes might be due to readers' clicking errors. Taking into account such emotions during the learning process could introduce bias. Therefore, we aim to differentiate relevant emotions from irrelevant ones and only learn the rankings of relevant emotions while neglecting the irrelevant ones.

Our work makes the following contributions:

- We propose a novel framework based on relevant emotion ranking to identify multiple emotions and produce the rankings of relevant emotions from text. In the framework, the objective emotion loss function is designed elaborately so that both emotion prediction and rankings for only relevant emotions are achieved without being affected by irrelevant ones. To the best of our knowledge, it is the first attempt to perform emotion detection and relevant emotion ranking at the same time.

- As some emotions co-occur more often while others rarely co-exist, the prior knowledge of emotion relationships is incorporated into the framework as a constraint. Such emotion relationship can provide important cues for emotion detection.

- Experimental results on two real-world corpora show that the proposed framework can effectively deal with the emotion detection problem and performs better than the state-of-the-art emotion detection methods and multi-label learning methods.

2 Related work

Emotion detection is one of the subfields of sentiment analysis where emotions are more fine-grained and expressive. In general, emotion detection approaches can be categorized into two types: lexicon-based and learning-based approaches.

Lexicon-based approaches usually rely on emotion lexicons consisting of words and their corresponding emotion labels. For example, Aman and Szpakowicz (2007) classified emotional and non-emotional sentences with a predefined emotion lexicon. Emotional dictionaries could also be constructed from training corpora of news articles and be used to predict the readers' emotion of a new articles (Lei et al., 2014; Rao et al., 2012). Agrawal and An (2012) proposed a context-based approach to detect emotions from text at sentence level. An emotion vector for each potential affect bearing word based on the semantic relation between emotion concepts and words was generated. The emotion vector was then tuned based on the syntactic dependencies within a sentence structure. Other lexicon-based approach such as (Wang and Pal, 2015) can also output multiple emotions with intensities using non-negative matrix factorization with constraints derived based on an emotion lexicon.

Learning-based approaches can be further categoried as unsupervised and supervised learning methods. Unsupervised learning approaches do not require labelled data for training. For example, the emotion-topic models (Bao et al., 2012, 2009) were proposed by adding an extra emotion layer into traditional topic models such as Latent Dirichlet Allocation (Blei et al., 2003), thus capturing the generation of both emotion and text at the same time.

Supervised learning approaches typically cast emotion detection 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, Yang and Chen (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 emo-

tion detection. Latent variables were introduced to model the latent structure of input text. 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. Zhou et al. (2016) proposed a novel approach based on emotion distribution learning to predict multiple emotions with different intensities in a single sentence.

3 Methodology

Assuming a set of T emotions $E = \{e_1, e_2, ...e_T\}$ and a set of n 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 $R_i \subseteq E$ and also a list of irrelevant emotions $\overline{R_i} = E - R_i$. Relevant emotion ranking aims to learn a score function $\mathbf{g}(x_i) = [g_1(x_i), ..., g_T(x_i)]$ assigning a score $g_t(x_i)$ to each emotion $e_t, (t \in \{1, ..., T\})$. 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. In order to differentiate relevant emotions from irrelevant ones, we need to define a threshold $g_\Theta(x)$ which could be simply set to 0 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 ranking function $\mathbf{g}$. Here, the predicted relevant emotions of instance x_i are denoted as $\hat{R}_i = \{e_t \in E | g_t(x_i) > g_\Theta(x_i)\}$.

3.1 Emotion Loss Function

The goal of relevant emotion ranking is to learn the parameter of the ranking function $\mathbf{g}$. Without loss of generality, we assume that $\mathbf{g}$ are linear models, i.e., $g_t(x_i) = w_t^\intercal \cdot x_i$, $t \in \{1, 2, 3, ..., T\} \cup \{\Theta\}$, where Θ denotes the threshold. Relevant emotion ranking can be regarded as a special case of multi-label learning. Several evaluation criteria typically used in multi-label learning can also be used to measure the ranking function's ability of distinguishing relevant emotions from irrelevant ones, such as hamming loss, one error, coverage, ranking loss, and average precision as suggested in (Zhang and Zhou, 2014). However, these multi-label criteria cannot meet our requirement exactly as none of them considers the ranking among emotions which are considered relevant. Therefore, by incorporating PRO loss (Xu et al., 2013), the loss function for the instance x_i is defined as follows:

$$L(x_i, R_i, \prec, \mathbf{g}) = \sum_{e_t \in R_i \cup \{\Theta\}} \sum_{e_s \in \prec(e_t)} \frac{1}{norm_{t,s}} l_{t,s} \tag{1}$$

where e_t refers to the emotion belonging to relevant emotion set R_i or the threshold Θ of instance x_i while e_s refers to the emotion which is less relevant than e_t denoted as $\prec$. Thus, (e_t, e_s) represents four types of emotion pairs: i.e., *(relevant, relevant)*, *(relevant, irrelevant)*, *(relevant, threshold)*, and *(threshold, irrelevant)*. The normalization term $norm_{t,s}$ is used to balance those four types of emotion pairs to avoid dominated terms by their respective set sizes. The set sizes of the four different types of emotion pairs mentioned above are $|R_i| \times (|R_i| - 1)/2$, $|R_i| \times |\overline{R_i}|$, $|R_i|$, and $|\overline{R_i}|$, respectively. Here, $l_{t,s}$ refers to a modified 0-1 error. Specifically,

$$l_{t,s} = \begin{cases} 1, & g_t(x_i) < g_s(x_i) \\ \frac{1}{2}, & g_t(x_i) = g_s(x_i) \\ 0, & \text{otherwise} \end{cases}$$

Note that $l_{t,s}$ is non-convex and difficult to optimize. Thus, a large margin surrogate convex loss (Vapnik and Vapnik, 1998) implemented in hinge form is used instead as follows:

$$\hat{L}(x_i, R_i, \prec, \mathbf{g}) = \sum_{e_t \in R_i \cup \{\Theta\}} \sum_{e_s \in \prec(e_t)} \frac{1}{norm_{t,s}} (1 + g_s(x_i) - g_t(x_i))_+ \tag{2}$$

where $(u)_+ = \max\{0, u\}$.

However, Eq. 2 ignores the relationships between different emotions. As mentioned in Introduction section, some emotions often co-occur such as "*joy*" and "*love*" while some rarely coexist such as "*joy*" and "*anger*". Such relationship information among emotions can provide important clues for emotion ranking. Therefore, we incorporate this information into the emotion loss function as constraints. The objective function

$\hat{L}(x_i, R_i, \prec, \mathbf{g})$ can be redefined as:

$$\hat{L_\omega}(x_i, R_i, \prec, \mathbf{g}) = \sum_{e_t \in R_i \cup \{\Theta\}} \sum_{e_s \in \prec(e_t)} \frac{1}{norm_{t,s}} \times (1 + g_s(x_i) - g_t(x_i) + \omega_{ts}(w_t - w_s))_+ \quad (3)$$

where the weight ω_{ts} models the relationship between the *t*-th emotion and the *s*-th emotion in the emotion set and can be calculated in multiple ways. Since the Pearson correlation coefficient (Nicewander, 1988) is the most familiar measure of relationship between two variables, we use it to measure the relationship of two emotions using their original emotion scores across each corpus.

From the above, it can be observed that the goal of relevant emotion ranking can be achieved through predicting an accurate relevant emotion set as well as the ranking of relevant emotions.

3.2 Relevant Emotion Ranking

After defining an appropriate loss function, we need to define a way to minimize the empirical error measured by the appropriate loss function and at the same time to control the complexity of the resulting model. It can be done by introducing a maximum margin strategy and regularization to deal with emotion ranking data, where a set of linear classifiers are optimized to minimize the emotion loss function mentioned before while having a large margin. We could potentially use an approach based on a label ranking method (Elisseeff and Weston, 2001). It is worth mentioning that the margin of the *(relevant, relevant)* label pair needs to be dealt with carefully, which is not considered in (Elisseeff and Weston, 2001).

The learning procedure of relevant emotion ranking (RER) is illustrated in Figure 2. The big rectangular dash line boxes denoted by x_1 to x_n represent n instances in the training set. In each small box, $e_i, i \in \{1, ...T\} \cup \{\Theta\}$ represents an emotion of the instance where the shaded small boxes represent the relevant emotions while the dashed small boxes represent irrelevant ones and the last one e_Θ is the threshold. Each emotion's corresponding weight vector is w_i. We use $m_{t,s}$ to represents the margin between label e_t and e_s. There are four types of emotion pairs' margins in total, i.e., *(relevant, relevant)*, *(relevant, irrelevant)*, *(relevant, threshold)*, and *(threshold, irrelevant)*. Different types of emotion pairs' margins are denoted using different text/line colors. For each training instance x_i, $margin(x_i)$ represents the margin of instance x_i which can be obtained by taking the minimum margin of all its possible label pairs $m_{t,s}$. Similarly, the margin of the learning system $margin(learningsystem)$ can be obtained by taking the minimum margin of all the training instances. By maximizing the margin of the learning system, the weight vector of each emotion can be derived from which the predicted emotion set and the ranking of relevant emotions can be obtained.

The learning system is composed of $T + 1$ linear classifiers $[w_1; ...; w_T; w_\Theta]$ with one classifier for each emotion label and the threshold, where $w_t, t \in \{1, ...T\} \cup \{\Theta\}$ is the weight vector for the t-th classifier of emotion e_t. For a training instance x_i and its corresponding emotion label set E_i, the learning system's margin on instance x_i is defined as follows by considering its ranking ability on x_i's four types of emotion pairs, i.e., *(relevant, relevant)*, *(relevant, irrelevant)*, *(relevant, threshold)*, and *(threshold, irrelevant)*:

$$\min_{e_t \in R_i \cup \{\Theta\}, e_s \in \prec(e_t)} \frac{\langle w_t - w_s, x_i \rangle}{||w_t - w_s||} \quad (4)$$

Here, $\langle u, v \rangle$ returns the inner product $u^\top v$. For each emotion pair (e_t, e_s), its discrimination boundary corresponds to the hyperplane $\langle w_t - w_s, x_i \rangle = 0$. Therefore, Eq. 4 returns the minimum value as the margin on instance x_i. The margin on the whole training set G can be calculated as follows:

$$\min_{x_i \in G} \min_{e_t \in R_i \cup \{\Theta\}, e_s \in \prec(e_t)} \frac{\langle w_t - w_s, x_i \rangle}{||w_t - w_s||} \quad (5)$$

If the learning algorithm is capable of properly ranking the four types of label pairs for each training instance, Eq. 5 will return a positive margin. In this ideal case, the final goal is to maximize the margin in Eq. 5:

$$\max_{w_j} \min_{x_i \in G} \min_{e_t \in R_i \cup \{\Theta\}, e_s \in \prec(e_t)} \frac{1}{||w_t - w_s||} \quad s.t. \langle w_t - w_s, x_i \rangle \geq 1, 1 \leq i \leq n, 1 \leq j \leq T + 1 \quad (6)$$

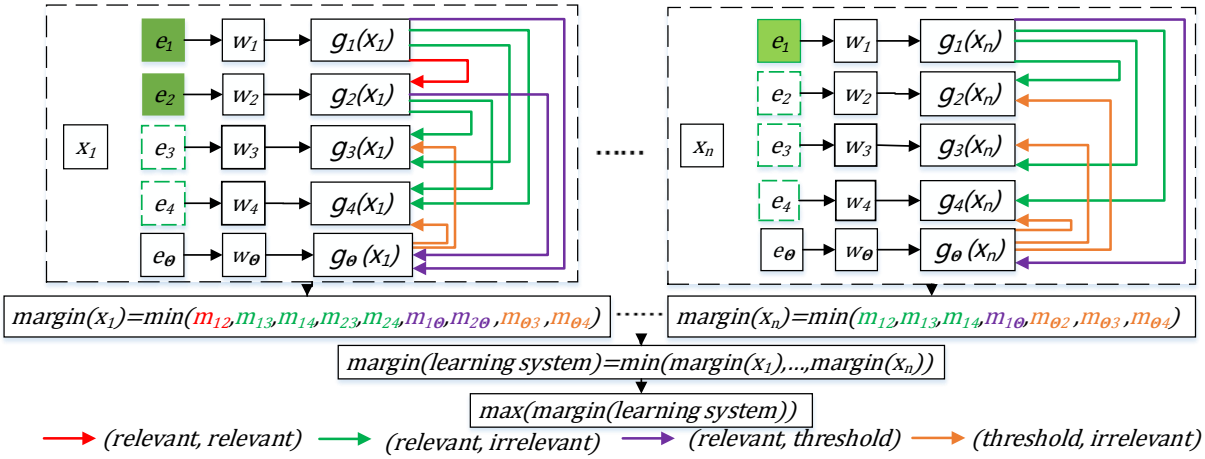

Figure 2: The overall framework of our proposed Relevant Emotion Ranking (RER) method.

Suppose we have sufficient training examples such that for each label pair (e_t, e_s), there exists $x_i \in G$satisfying $e_t \in R_i \cup \{\Theta\}, e_s \in \prec (e_t)$. Thus, the objective in Eq.6 becomes equivalent to $\max_{w_j} \min_{1\le s<t\le T+1} \frac{1}{||w_t - w_s||}$ and can be rewritten as $\min_{w_j} \max_{1\le s<t\le T+1} ||w_t - w_s||$.

Moreover, to overcome the complexity brought in by the max operator, the objective of the optimization problem can be re-written by approximating the max operator with the sum operator. Thus, the objective of Eq. 6 can be transformed as:

$$\begin{aligned} &\min_{w_j} \sum_{t=1}^{T+1} ||w_t||^2 \\ &s.t.\ \langle w_t - w_s, x_i \rangle \ge 1, 1 \le i \le n, \\ &1 \le j \le T+1, e_t \in R_i \cup \{\Theta\}, e_s \in \prec (e_t) \end{aligned} \tag{7}$$

To accommodate real-world scenarios where constraints in Eq. 7 can not be fully satisfied, slack variables can be incorporated into the objective function:

$$\begin{aligned} &\min_{w_j, \xi_{its}} \sum_{t=1}^{T+1} ||w_t||^2 + \lambda \sum_{i=1}^{n} \sum_{e_t \in R_i \cup \{\Theta\}} \sum_{e_s \in \prec (e_t)} \frac{1}{norm_{t,s}} \xi_{its} \\ &s.t. \langle w_t - w_s, x_i \rangle \ge 1 - \xi_{its}, 1 \le j \le T+1, \xi_{its} \ge 0 \end{aligned} \tag{8}$$

Since ξ_{its} does not need to be optimized since it can be easily determined by w_t, w_s. The final objective function can be reformulated as:

$$\min_{w_t, \hat{L}} \sum_{t=1}^{T+1} ||w_t||^2 + \lambda \sum_{i=1}^{n} \hat{L}(x_i, R_i, \prec, \mathbf{g}) \tag{9}$$

As can be seen, Eq.9 consists of two parts balanced by the trade-off parameter λ. Specifically, the first part corresponds to the maximum margin of the learning system and it can also represent the complexity of the learning system, while the second part corresponds to the emotion loss function of the learning system implemented in hinge form.

3.3 Parameter Estimation

Let $\mathbf{w} = [w_1; ...; w_T; w_\Theta]$, Eq. 9 is cast into a general form in SVM-type:

$$\begin{aligned} &\min_{\mathbf{w}, \boldsymbol{\xi}} \frac{1}{2} ||\mathbf{w}||^2 + \lambda C^\top \boldsymbol{\xi} \\ &s.t.\ A\mathbf{w} \ge \mathbf{1}_p - \boldsymbol{\xi}, \boldsymbol{\xi} \ge \mathbf{0}_p \end{aligned} \tag{10}$$

where p is the total number of label pairs, calculated by $\sum_{i=1}^{n} \sum_{e_t \in R_i \cup \{\Theta\}} \sum_{e_s \in \prec (e_t)} norm_{t,s}$ and $\mathbf{1}_p (\mathbf{0}_p)$ is the $p \times 1$ all one (zero) vector. The entries in vector C correspond to the weights of hinge losses, i.e., the normalization term to balance the four kinds of label pairs. The matrix A corresponds to the constraints for instances which reflects the emotion relationships and the margin of the label pairs.

$\boldsymbol{\xi}$ does not need to be optimized since it can be easily determined by $\mathbf{w}$. Hence the objective function can be reformulated into the following form without $\boldsymbol{\xi}$:

$$\min_{\mathbf{w}} F(\mathbf{w}, G) = \frac{1}{2} ||\mathbf{w}||^2 + \lambda C^\top (1_p - A\mathbf{w})_+ \tag{11}$$

Through minimizing the objective function $F(\mathbf{w}, G)$, we can finally obtain parameter $\mathbf{w}$ and the ranking function $\mathbf{g}$. Eq. 11 involves a large scale optimization. To address Eq. 11, we consider an efficient Alternating Direction Method of Multipliers (ADMM) solution (Bertsekas and Tsitsiklis, 1989). The basic idea of ADMM is to take the decomposition-coordinate procedure such that the solution of subproblems can be coordinated to find the solution to the original problem. We decompose G into M disjoint subsets, i.e., $\{G_1, G_2, ..., G_M\}$ and then Eq. 11 is converted into the following form:

$$\min_{\mathbf{w}^0, \mathbf{w}^1, \mathbf{w}^m} \sum_{m=1}^{M} F(\mathbf{w}^m, G^m), \quad s.t.\ \mathbf{w}^m = \mathbf{w}^0, \forall m = 1, ..., M \tag{12}$$

The surrogate augmented Lagrangian Function (LF) was introduced into Eq. 12 and it was cast into the following form:

$$LF(\{\mathbf{w}^0, \mathbf{w}^1, ..., \mathbf{w}^m\}, \{\alpha^m\}_{m=1}^M, \beta) = \sum_{m=1}^{M} F(\mathbf{w}^m, G^m) + \sum_{m=1}^{M} (\alpha^m)^\top (\mathbf{w}^m - \mathbf{w}^0) + \frac{\beta}{2} \sum_{m=1}^{M} ||\mathbf{w}^m - \mathbf{w}^0||^2 \tag{13}$$

where α, β are the Lagrange multiplies. The updating process of Eq. 13 is shown in Algorithm 1.

Algorithm 1 Parameter updating process.

1: Decompose data set G into M disjoint subsets i.e., $\{G_1, G_2, ..., G_M\}$. Set iteration $i = 0$.
2: Initialize $\{\mathbf{w}_0^0, \mathbf{w}_0^1, ..., \mathbf{w}_0^M, \alpha_0^1, ..., \alpha_0^M\}$ as zeros.
3: **while** not converged **do**
4: Set $i = i + 1$
5: Update $\mathbf{w}_i^0, \{\mathbf{w}_i^m, \alpha_i^m\}_{m=1}^M$ as:
$\{\mathbf{w}_i^m\}_{m=1}^M = \underset{\mathbf{w}^1..\mathbf{w}^m}{\operatorname{argmin}} LF(\mathbf{w}_{i-1}^0, \{\mathbf{w}_{i-1}^m, \alpha_{i-1}^m\}_{m=1}^M, \beta)$
$\mathbf{w}_i^0 = \underset{\mathbf{w}^0}{\operatorname{argmin}} LF(\mathbf{w}^0, \{\mathbf{w}_{i-1}^m, \alpha_{i-1}^m\}_{m=1}^M, \beta$
$\alpha_i^m = \alpha_{i-1}^m + \beta(\mathbf{w}_i^m - \mathbf{w}_i^0)^\top, \forall m = 1, 2, ..., M$
6: **end while**
Output: Final $\mathbf{w}^0$

4 Experiments

4.1 Setup

We evaluate the proposed approach on two real-world corpora, one is document level and the other is sentence level:

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 34,719 sentences selected from blogs in Chinese. Each sentence 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 which range from 0 to 1. Higher scores represents higher emotion intensity.

The statistics of the two corpora are shown in Table 1.

Sina Social News		**Ren-CECps Corpus**	
Category	#Votes	Category	#Scores
Touching	694,006	Joy	1,349.6
Shock	572,651	Hate	6,103.9
Amusement	869,464	Love	2,911.1
Sadness	837,431	Sorrow	2,042.5
Curiosity	212,559	Surprise	3,873.9
Anger	1,109,315	Anger	7,832.1
		Anxiety	5,006.4
		Expect	610.4
All	4,295,426	All	29,729.9

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

The two corpora were preprocessed by using word segmentation and filtering. The python jieba segmenter is used for the segmentation and a removal of stop words is performed based on a stop word thesaurus. Words appeared only once or appeared in less than two documents were re-

moved to alleviate data sparsity. We used the single layer long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) to extract the features of each text. LSTM is one kind of recurrent neural networks, which can capture sequence information from text and can represent meanings of inputs in the reduced dimensional space. It treats text as a sequence of word embeddings and outputs a state vector over each word, which contains the information of the previous words. The final state vector can be used as the representation of the text. In our experiments, we set the dimension of each text representation to 100. During LSTM model training, we optimized the hyper parameters using a development dataset which is built using external data. We train LSTM using a learning rate of 0.001, a dropout rate of 0.3 and categorical cross-entropy as the loss function. The mini batch (Cotter et al., 2011) size is set to 32. After that, the learned text representations are fed into the proposed system for relevant emotion ranking as has been previously presented in the Methodology section.

4.2 Comparison with Baselines

There are only few baselines which address multiple emotions learning from text. We first compare the proposed framework with two baselines which have previously achieved the state-of-the-art performances on multi-emotion detection.

- **Emotion Distribution Learning (EDL)** (Zhou et al., 2016): It learns a mapping function from texts to their emotion distributions describing multiple emotions and their respective intensities based on label distribution learning. Moreover, the relationships of emotions are captured based on the Plutchik's wheel of emotions which are subsequently incorporated into the learning algorithm in order to improve the accuracy of emotion detection.

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

For each method, 10-fold cross validation is conducted using the same feature construction

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(t,s)
Hamming Loss	$\frac{1}{nT}\sum_{i=1}^{n}\lvert\hat{R}_i\triangle R_i\rvert$
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	$\frac{1}{n}\sum_{i=1}^{n}\delta[\operatorname{argmax}_{e_t} g_t(x_i)\notin R_i]$
Average Precision	$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\lvert R_i\rvert}\times$ $(\sum_{t:e_t\in R_i}\lvert\{e_s\in R_i\lvert g_s(x_i)>g_t(x_i)\}\rvert)/(\lvert\{e_s\lvert g_s(x_i)>g_t(x_i)\}\rvert)$
Coverage	$\frac{1}{n}\sum_{i=1}^{n}\max_{t:e_t\in R_i}\lvert\{e_s\lvert g_s(x_i)>g_t(x_i)\}\rvert$
Subset Accuracy	$\frac{1}{n}\sum_{i=1}^{n}\delta[\hat{R}_i=R_i]$
$F1_{exam}$	$\frac{1}{n}\sum_{i=1}^{n}2\lvert R_i\cap\hat{R}_i\rvert/(\lvert R_i\rvert+\lvert\hat{R}_i\rvert)$
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).

method to get the final performance. Supposing n test instances and T emotion categories, several evaluation criteria as presented in Table 2 typically used in multi-label learning can be used to measure the efficiency of the proposed framework and the baseline approaches. PRO Loss concerning the prediction on all labels as well as the rankings of only relevant labels. Hamming loss evaluates how many times an emotion label is misclassified. Ranking loss evaluates the fraction of reversely ordered emotion pairs. One-error evaluates the fraction of sentences whose top-ranked emotion is not in the relevant emotion set. Average precision evaluates the average fraction of the relevant emotions ranked higher than a particular emotion. Coverage evaluates how many steps are needed to move down the ranked emotion list so as to cover all the relevant emotions of the example. Subset Accuracy evaluates the fraction of correctly classified examples, i.e. the predicted label set is identical to the ground-truth label set. $F1_{exam}$ evaluates the averaged F1 (Manning et al., 2008) over instances. MicroF1 pools each document decisions across categories, and then computes an effectiveness measure on the pooled contingency table. MacroF1 take the average of F1 for all categories. Note that the threshold Θ is removed before evaluation. It should be pointed out that metrics from PRO Loss to $F1_{exam}$ work by evaluating the learning systems performance on each test exam-

ple separately, and then returning the mean value across the test set. MicroF1 and MacroF1 work by evaluating the learning systems performance on each emotion category separately, and then returning the macro/micro-averaged value across all emotion categories.

The evaluation results using 10 different evaluation criteria are shown in Table 3. It can be observed that our proposed method Relevant Emotion Ranking(RER) outperforms baseline approaches on all 10 evaluation metrics on both datasets.

Datasets	Evaluation Criterion	Methods			
		RER	RERc	EDL	EmoDetect
News	PRO loss(↓)	0.1992	**0.1913**	0.2596	0.2465
	Hamming Loss(↓)	0.2318	**0.2277**	0.2671	0.2696
	Ranking Loss(↓)	0.1477	**0.1405**	0.1689	0.1769
	One-error(↓)	0.1579	**0.1562**	0.2115	0.1903
	Average Precision(↑)	0.8775	**0.8789**	0.8028	0.7865
	Coverage(↓)	2.1398	**2.1316**	2.1595	2.2348
	Subset Accuracy(↓)	0.1899	**0.1822**	0.2026	0.2243
	$F1_{exam}$(↑)	0.7062	**0.7143**	0.6503	0.6469
	MicroF1(↑)	0.7086	**0.7171**	0.6346	0.6375
	MacroF1(↑)	0.6244	**0.6291**	0.5641	0.5767
Blogs	PRO loss(↓)	0.2354	**0.2321**	0.2739	0.2912
	Hamming Loss(↓)	0.2054	**0.2014**	0.2102	0.2202
	Ranking Loss(↓)	0.2137	**0.2102**	0.2589	0.2781
	One-error(↓)	0.4556	**0.4550**	0.5227	0.5352
	Average Precision(↑)	0.6749	**0.6803**	0.6411	0.5663
	Coverage(↓)	2.1269	**2.1268**	2.1699	2.8956
	Subset Accuracy(↓)	**0.1663**	**0.1663**	0.2116	0.2321
	$F1_{exam}$(↑)	0.5080	**0.5114**	0.4606	0.4650
	MicroF1(↑)	0.5093	**0.5116**	0.4620	0.4552
	MacroF1(↑)	0.4102	**0.4161**	0.3923	0.3622

Table 3: Comparison with emotion detection baselines. "↓" indicates "the smaller the better", while "↑" indicates "the larger the better". The best performance on each evaluation measure is highlighted by boldface.

We have also extended RER by incorporating emotion relationships as constraints into the learning framework, denoted as RERc in Table 3. The correlation of every pair of emotions is calculated based on their respective votes from news articles or scores from blogs. It can be observed from Table 3 that in almost all cases, incorporating the constraints gives better performance. It should be pointed out that the results of the baseline approach EDL are slightly different from those reported in (Zhou et al., 2016) since we employ LSTM for feature construction instead of recursive autoencoders.

Since relevant emotion ranking can be seen as an extension of multi-label learning, the proposed framework is also compared with 8 widely used multi-label learning methods using the threshold Θ which is initialized as 0.15 after normalization, such as ML-KNN (Zhang and Zhou, 2007), LIFT (Zhang, 2011), CLR (Fürnkranz et al., 2008), Rank-SVM (Zhang and Zhou, 2014) , MLLOC (Huang and Zhou, 2012), BP-MLL (Zhang and Zhou, 2006), ECC (Read et al., 2009) and ML-RBF (Zhang, 2009). ML-KNN is based on the traditional k-nearest neighbor (KNN) algorithm. LIFT constructs features specific to each emotion by conducting clustering analysis on its positive or negative instances. CLR transforms MLL into a label ranking problem by pairwise comparison which considers each label pairs and rank all the labels without recognizing that only the rankings of relevant ones are crucial. Rank-SVM focuses on distinguishing relevant from irrelevant while neglecting the rankings of relevant ones. MLLOC tries to exploit emotion correlations in the expression data locally. The global discrimination fitting and local correlation sensitivity are incorporated into a unified framework. BP-MLL is derived from the back propagation algorithm through employing a novel error function capturing the characteristics of multi-label learning. ECC applies classifier chains in an ensemble framework. ML-RBF gets the multi-label neural networks adopted from the traditional Radial Basis Function (RBF) neural networks.

Anger	Anxiety	Expect	Hate
生气(angry)	害怕(fear)	祝福(blessing)	讨厌(hate)
愤怒(rage)	失去(lose)	幸福(happy)	虚伪(hypocrisy)
抱怨(complain)	孤独(lonely)	美好(fine)	炒作(hype)
批评(criticize)	压力(pressure)	梦想(dream)	无耻(shameless)
利益(interest)	现实(reality)	自由(freedom)	手段(means)
歧视(discriminate)	陌生(strange)	渴望(long for)	愚蠢(silly)
制止(stop)	心灵(heart)	希望(hope)	浪费(waste)
指责(accuse)	痛苦(pain)	学习(learn)	背后(behind)
懊恼(annoy)	想象(imagine)	信念(faith)	肮脏(dirty)
无耻(shameless)	伤害(hurt)	家里(home)	欺骗(lie)
Joy	**Love**	**Sorrow**	**Surprised**
快乐(happy)	美丽(beautiful)	孤独(lonely)	好奇(curious)
高兴(joyful)	爱情(love)	眼泪(tears)	惊讶(surprise)
朋友(friend)	朋友(friend)	爱情(love)	震惊(shock)
感动(touching)	幸福(happiness)	寂寞(solitude)	惊奇(wonder)
心情(mood)	孩子(child)	痛苦(pain)	惊人(amazing)
温暖(warm)	生命(life)	感情(feeling)	意外(accident)
享受(enjoy)	阳光(sunshine)	伤害(hurt)	惊吓(fright)
兴奋(excited)	温暖(warmth)	失去(lose)	惊呼(scream)
收获(harvest)	思念(miss)	思念(miss)	不经意(accidently)
微笑(smile)	可爱(lovely)	生活(life)	诧异(amazed)

Figure 3: The top 10 words for each emotion label from Blogs dataset.

For the MLL methods, the value of k is set to 8 in ML-KNN, ratio is 0.02 and μ is 2 in ML-RBF. Linear kernel is used in LIFT. Rank-SVM uses the RBF kernel with the width σ equals to 1. The CR4.5 is used as the base classifier for CLR and ECC. The evaluation results of the proposed

Datasets	Evaluation Criterion	Methods								
		RERc	ML-KNN	LIFT	CLR	Rank-SVM	MLLOC	BP-MLL	ECC	ML-RBF
News	PRO loss(↓)	**0.1913**	0.2551	0.2426	0.2487	0.2670	0.3429	0.2603	0.2823	0.2658
	Hamming Loss(↓)	**0.2277**	0.2876	0.3118	0.3023	0.3127	0.3241	0.3040	0.3079	0.3599
	Ranking Loss(↓)	**0.1405**	0.1898	0.1987	0.2142	0.2271	0.3234	0.1897	0.2563	0.1949
	One-error(↓)	**0.1562**	0.2366	0.1881	0.2242	0.2258	0.2025	0.2043	0.2151	0.2240
	Average Precision(↑)	**0.8789**	0.8095	0.7945	0.7916	0.8001	0.7545	0.8044	0.6245	0.8106
	Coverage(↓)	**2.1316**	2.3602	2.4641	2.3453	2.6093	3.1272	2.4032	2.4122	2.4390
	Subset Accuracy(↓)	**0.1822**	0.1916	0.1857	0.2386	0.1839	0.2107	0.2765	0.2222	0.2609
	$F1_{exam}$(↑)	**0.7143**	0.6215	0.6262	0.6032	0.6244	0.5193	0.5879	0.5108	0.6147
	MicroF1(↑)	**0.7171**	0.6280	0.6131	0.6177	0.6268	0.5389	0.6231	0.5699	0.6160
	MacroF1(↑)	**0.6291**	0.5587	0.5593	0.5658	0.5613	0.4913	0.5563	0.4573	0.5543
Blogs	PRO loss(↓)	**0.2321**	0.3036	0.2912	0.3041	0.2869	0.3523	0.3429	0.2867	0.2922
	Hamming Loss(↓)	**0.2014**	0.2409	0.2242	0.2162	0.2585	0.2156	0.2241	0.2301	0.2204
	Ranking Loss(↓)	**0.2102**	0.2928	0.2881	0.2947	0.3024	0.4532	0.3234	0.3345	0.2364
	One-error(↓)	**0.4550**	0.5543	0.5152	0.5229	0.5606	0.6143	0.4625	0.6635	0.4679
	Average Precision(↑)	**0.6803**	0.5897	0.5963	0.6370	0.5832	0.4532	0.5545	0.5256	0.6412
	Coverage(↓)	**2.1268**	2.4448	2.4356	2.2671	2.5962	3.5634	3.1272	2.7756	2.5067
	Subset Accuracy(↓)	**0.1663**	0.1978	0.2116	0.1938	0.2321	0.2251	0.2107	0.2236	0.1803
	$F1_{exam}$(↑)	**0.5114**	0.4616	0.4620	0.4509	0.4832	0.4931	0.5093	0.4986	0.4997
	MicroF1(↑)	**0.5116**	0.4720	0.4552	0.4859	0.4962	0.4902	0.4889	0.5003	0.5051
	MacroF1(↑)	**0.4161**	0.3632	0.3656	0.4056	0.3965	0.3853	0.3813	0.3957	0.4086

Table 4: Comparison with Multi-Label Learning (MLL) Methods."↓" indicates "the smaller the better", while "↑" indicates "the larger the better". The best performance on each evaluation measure is highlighted by boldface.

approach in comparison to all MLL baselines are presented in Table 4. RERc performs the best on all evaluation measures. It verifies the advantage of RERc due to its consideration of varying intensities of the emotion labels and the ignorance of irrelevant ones during the learning of the relevant emotion ranking model. We also observe that, in most cases, the performance on the News dataset is better than that in the Blogs dataset. This may due to different types of text observed in these two platforms. News articles are more formal while bogs typically contain informal language and are more colloquial.

4.3 Result Analysis

To fully understand the emotion detection results, we generate the top 10 most frequent words in the test set for each emotion label from Blogs dataset presented in Figure 3. Intuitively, we can find that most top words are quite expressive of their associated emotions. For example, the word "happy" delivers the emotion of *Joy* and the word "tears" tells *Sorrow*, etc. Moreover, we also observe that there are some common words across multiple emotion categories. For instance, "friend" appears in both *Joy* and *Love*. The results demonstrate that the proposed framework can learn emotions from text precisely.

5 Conclusions

In this paper, we have proposed a novel framework to detect multiple emotions from text based on relevant emotion ranking. Moreover, the relationships between emotions are incorporated into the learning framework as constraints. Experimental results on both News and Blogs datasets show that the proposed framework is able to detect multiple emotions as well as generating rankings of relevant emotions. It performs remarkably better than the state-of-the-art baselines on multi-emotion detection and also outperforms several different methods used for multi-label learning. In the future, we will explore the possibility of extending the current framework by detecting emotions at more fine-grained level, for example, emotions associated with specific events reported in text.

Acknowledgments

The work was supported by the National Key R&D Program of China (No. 2017YFB1002801), the National Natural Science Foundation of China (61772132), the Natural Science Foundation of Jiangsu Province of China (BK20161430) and Innovate UK (103652).

References

Ameeta Agrawal and Aijun An. 2012. Unsupervised emotion detection from text using semantic and syntactic relations. In *Ieee/wic/acm International Conferences on Web Intelligence and Intelligent Agent Technology*. pages 346–353.

Saima Aman and Stan Szpakowicz. 2007. Identifying expressions of emotion in text. In *Text, Speech and Dialogue, International Conference, Tsd 2007, Pilsen, Czech Republic, September 3-7, 2007, Proceedings*. pages 196–205.

Shenghua Bao, Shengliang Xu, Li Zhang, Rong Yan, Zhong Su, Dingyi Han, and Yong Yu. 2009. Joint emotion-topic modeling for social affective text mining. In *2009 Ninth IEEE International Conference on Data Mining*. IEEE, pages 699–704.

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.

Dimitri P Bertsekas and John N Tsitsiklis. 1989. *Parallel and distributed computation: numerical methods. Reprint of the 1989 edition published by Prentice-Hall.*. Prentice Hall.

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.

Andrew Cotter, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. 2011. Better mini-batch algorithms via accelerated gradient methods. *Advances in Neural Information Processing Systems* pages 1647–1655.

Andr Elisseeff and Jason Weston. 2001. A kernel method for multi-labelled classification. In *International Conference on Neural Information Processing Systems: Natural and Synthetic*. pages 681–687.

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

Sepp Hochreiter and Jrgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9(8):1735–1780.

Sheng Jun Huang and Zhi Hua Zhou. 2012. Multi-label learning by exploiting label correlations locally. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*. pages 949–955.

Jingsheng Lei, Yanghui Rao, Qing Li, Xiaojun Quan, and Liu Wenyin. 2014. Towards building a social emotion detection system for online news. *Future Generation Computer Systems* 37:438–448.

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*. IEEE Computer Society, pages 220–226.

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, Xiaojun Quan, Liu Wenyin, Qing Li, and Mingliang Chen. 2012. Building word-emotion mapping dictionary for online news. In *SDAD 2012 The 1st International Workshop on Sentiment Discovery from Affective Data*. page 28.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2009. Classifier chains for multi-label classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. pages 254–269.

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

Vladimir Naumovich Vapnik and Vlamimir Vapnik. 1998. *Statistical learning theory*, volume 1. Wiley New York.

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.

Miao Xu, Yu Feng Li, and Zhi Hua Zhou. 2013. Multi-label learning with pro loss. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Min Ling Zhang. 2009. M l-rbf : Rbf neural networks for multi-label learning. *Neural Processing Letters* 29(2):61–74.

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. 2007. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition* 40(7):2038–2048.

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, 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.