Emotion Cause Detection for Chinese Micro-Blogs Based on ECOCC Model

Kai Gao[1,2], Hua Xu[1](✉), and JiushuoWang[1,2]

[1] State Key Laboratory of Intelligent Technology and Systems, Tsinghua National Laboratory for Information Science and Technology, Department of Computer Science and Technology, Tsinghua University, Beijing, China
xuhua@tsinghua.edu.cn

[2] School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang, Hebei, China
gaokai68@139.com, wangjiushuo@126.com

Abstract. Micro-blog emotion mining and emotion cause extraction are essential in social network data mining. This paper presents a novel approach on Chinese micro-blog emotion cause detection based on the ECOCC model, focusing on mining factors for eliciting some kinds of emotions. In order to do so, the corresponding emotion causes are extracted. Moreover, the proportions of different cause components under different emotions are also calculated by means of combining the emotional lexicon with multiple characteristics (e.g., emoticon, punctuation, etc.). Experimental results show the feasibility of the approach. The proposed approaches have important scientific values on social network knowledge discovery and data mining.

Keywords: Micro-blog · Text mining · Emotion cause detection · Emotion analysis

1 Introduction

Nowadays, millions of people present and share their opinions or sentiments in micro-blogs, and as a result, to produce a large number of social network data, containing almost all kinds of opinions and sentiment. These micro-blog data usually includes the description of emergencies, incidents, disasters and some other hot events, and some of them have some kinds of emotions and sentiments. If we can mine or discover the hidden emotions behind the big data, it is essential for public opinion surveillance. Meanwhile, individual emotion generation, expression and perception are influenced by many factors, so emotion cause feature extraction and analysis are necessary.

According to the related works, the cause events are usually composed of verbs, nominalizations and nouns. They evoke the presence of the corresponding emotions by some linguistic cues, which are based on Chinese emotion cause annotated corpus [1]. For example, as for the Chinese micro-blog post *"Ta1*

T. Cao et al. (Eds.): PAKDD 2015, Part II, LNAI 9078, pp. 3–14, 2015.
DOI: 10.1007/978-3-319-18032-8_1

Men Ying2 De2 Le Bi3 Sai4 Shi3 Wo3 Hen3 Kai1 Xin1." ("They won the game makes me very happy."), the causative verb is *"Shi3"* ("makes"), and the emotion keyword is *"Kai1 Xin1"* ("happy"). With the help of the corresponding linguistic rules, we can infer that the emotion cause event is *"Ta1 Men Ying2 De2 Le Bi3 Sai4"* ("they won the game"). Meanwhile, the emotion causes also were detected by extracting cause expressions and constructions in [2]. And Li et al. [3] proposed and implemented a novel method for identifying emotions of micro-blog posts, and tried to infer and extract the emotion causes by using knowledge and theories from other fields such as sociology. On the one hand, Rao et al. [4] proposed two sentiment topic models to extract the latent topics that evoke emotions of readers, then the topics were seen as the causes of emotions as well.

In this paper, an emotion model with cause events is proposed, it describes the causes that trigger bloggers' emotions in the progress of cognitive evaluation. Meanwhile, all of the sub-events are extracted in micro-blogs. This paper detects the corresponding cause events on the basis of the proposed rule-based algorithm. Finally, the proportions of different cause components under different emotions are calculated by constructing the emotional lexicon from the corpus and combining multiple features of Chinese micro-blogs.

2 ECOCC Model Construction

According to the research on cognitive theory, this paper improves the structure about the eliciting conditions of emotions on the basis of the OCC model referred in [5], and presents an emotion model named as ECOCC (Emotion-Cause-OCC) model for micro-blog posts. It describes the combination of psychology and computer science to analyze the corresponding emotion cause events. The improved model describes a hierarchy that classifies 22 fine-grained emotions, i.e., "hope", "fear", "joy", "distress", "pride", "shame", "admiration", "reproach", "liking", "disliking", "gratification", "remorse", "gratitude", "anger", "satisfaction", "fears-confirmed", "relief", "disappointment", "happy-for", "resentment", "gloating" and "pity". This hierarchy contains three main branches (i.e., results of events, actions of agents and aspects of objects), and some branches are combined to form a group of the compound and extended emotions.

According to the three main branches, the components of the model matching the emotional rules in the model are divided into six sub-classes (i.e., event_state, event_norm, action_agent, action_norm, object_entity and object_norm). Within the components, the evaluation-schemes can be defined from event_state (i.e., the state that something will happen, the state that something has happened, the state that something didn't happen), action_agent (i.e., the emotional agent or others) and object_entity (i.e., the elements of objects). They are also divided into six classes (i.e., prospective, confirmation, disconfirmation, main-agent, other-agent and entity). Meanwhile, the corresponding evaluation-standards are defined from the aspects of event_norm (i.e., being satisfactory or unsatisfactory for the event), action_norm (i.e.,being approval or disapproval for the agent) and object_norm (i.e., being attractive or unattractive to the entity). And they are separated into six types (i.e., desirable, undesirable, praiseworthy, blameworthy, positive and negative).

As for the production process of the 22 types of emotions, the following rules (see Table 1) are used to describe them according to the components in the proposed ECOCC model. Here, "s" means the micro-blog post, "C" represents the cause components that trigger emotions, "→" represents the state that changes from one to the other, "∩" has the same meaning as "and" and "∪" has the same meaning as "or".

Table 1. The emotional rules

Classes	The emotional rules
The emotions in results of events	Hope(C)$\overset{def}{=}$Prospective(s)∩ Desirable(s)
	Fear(C)$\overset{def}{=}$Prospective(s)∩ Undesirable(s)
	Joy(C)$\overset{def}{=}$[Confirmation(s)∪Disconfirmation(s)]∩ Desirable(s)
	Distress(C)$\overset{def}{=}$[Confirmation(s)∪Disconfirmation(s)]∩ Undesirable(s)
The emotions in actions of agents	Pride(C)$\overset{def}{=}$main-agent∩Praiseworthy(s)
	Shame(C)$\overset{def}{=}$main-agent∩Blameworthy(s)
	Admiration(C)$\overset{def}{=}$other-agent∩Praiseworthy(s)
	Reproach(C)$\overset{def}{=}$other-agent∩Blameworthy(s)
The emotions in aspects of objects	Liking (C)$\overset{def}{=}$ entity ∩Positive(s)
	Disliking(C)$\overset{def}{=}$entity∩Negative(s)
Compound emotions	Gratification(C)$\overset{def}{=}$Pride(C)∩Joy(C)
	Remorse(C)$\overset{def}{=}$Shame(C)∩Distress(C)
	Gratitude(C)$\overset{def}{=}$ Admiration (C)∩Joy (C)
	Anger(C)$\overset{def}{=}$Reproach(C)∩Distress(C)
Extended emotions	Satisfaction(C)$\overset{def}{=}$Joy(C)∩[Prospective(s)→Confirmation(s)]∩ Hope(C)
	Fears-confirmed(C)$\overset{def}{=}$Distress(C)∩[Prospective(s)→Confirmation(s)] ∩ Fear(C)
	Relief(C)$\overset{def}{=}$Joy(C)∩[Prospective(s)→Disconfirmation(s)] ∩ Fear(C)
	Disappointment(C)$\overset{def}{=}$Distress(C)∩[Prospective(s)→ Disconfirmation(s)]∩ Hope(C)
	Happy-for(C)$\overset{def}{=}$Joy(C)∩[other-agent∩Praiseworthy(s)∩Desirable(s)]
	Resentment(C)$\overset{def}{=}$Distress(C)∩[other-agent ∩ Blameworthy(s) ∩Desirable(s)]
	Gloating(C)$\overset{def}{=}$Joy(C)∩[other-agent∩Blameworthy(s)∩Undesirable(s)]
	Pity(C)$\overset{def}{=}$Distress(C)∩[other-agent∩Praiseworthy(s)∩Undesirable(s)]

3 Emotion Causes Extraction

In this section, the internal events can be taken into account in the process of detecting emotion cause components, which are usually the direct reasons for triggering the change of individual emotion. They can be extracted from the domain of results of events, actions of agents and aspects of objects based on the ECOCC model.

As for the domain of results of events, the LTP (Language Technology Platform) is used to set up the model of extracting sub-events based on the named entity recognition, dependency parsing, semantic role labeling, and so on [6]. Firstly, we label the parts of speech of Chinese (e.g., nouns, verbs, adjectives) within the micro-blog posts by using ICTCLAS, and identify the person names, place names and institutions by using the named entity recognition. Then the core relation of subject-verbs and verb-objects can be identified by using dependency parsing. Finally, we identify the phrases labeled with the semantic role (i.e., A0) for the actions of the agent, the semantic role (i.e., A1) for actions of the receiver, or other four different core semantic roles (i.e., A2-A5) for different predicates by using the semantic role labeling, respectively.

In detail, this paper first selects the phrases which are labeled as A0, A1, A2-A5 (if it exists) and then combines the above components as the basis of event recognition. Otherwise, the triple $U = (nouns, verbs, nouns)$ will be used as another basis of event recognition. By describing the results of events, it is easy to decide the corresponding emotion and its causes according to the evaluation-schemes and the evaluation-standards.

As for the domain of actions of agents, the feature words can be used to describe agents' actions. Firstly, this paper extracts the class of ACT in HowNet[1] which contains a large number of words describing different kinds of people's actions. If the predicate verb belongs to the class of ACT and there exists an initiative relationship between itself and the agent, then this structure can be as a kind of agent's action. On the other hand, as for main-agent, it contains the explicit-agent and the implicit-agent. The former is highlighted in the text and has the subject-predicate relationship with the predicate verb. The latter does not appear in the text, but it is usually expressed by the context and the act of the verb. Finally, on the basis of the description of the actions of agents, it is easy to decide the corresponding emotion and its causes according to the evaluation-schemes and the evaluation-standards.

As for the domain of aspects of objects, the corresponding emotions "Liking" and "Disliking" describe the reactions of the agent to the corresponding object. For recognizing the characteristic information of aspects of objects, it needs to extract the entities with the help of the HowNet corpus and find the subject-predicate relationship by using dependency parsing. The features of objects are extracted by using the semantic role labeling to confirm the evaluation-standards of object_norm, and then we get the final emotion and its cause components.

[1] http://www.keenage.com/

4 Emotion Cause Components Analysis

4.1 Emotional Lexicon Construction

Generally, the emotional lexicon can be constructed manually and automatically from the corpus. Firstly, the standard lexicon can be constructed manually. In this process, the 22 fine-grained emotions based on the ECOCC model are chosen as the basal emotions. Then the intensity scores of the emotional words can be divided into five level ranges (i.e., 0-1.0, 1.0-2.0, 2.0-3.0, 3.0-4.0 and 4.0-5.0). Among them, 0-1.0 represents the word with the weakest emotion intensity; while 4.0-5.0 represents the corresponding word with the strongest emotion intensity. Meanwhile, the standard emotional words which belong to 22 different types of emotions are selected by three different annotators, and those words are from HowNet Dictionary, National Taiwan University Sentiment Dictionary[2], and the Affective Lexicon Ontology [7]. And then we give them the corresponding emotion intensities by the setting of the emotion intensity.

The lexicon will be expanded by acquiring automatically from the corpus for getting the larger capacity. It can be completed by the word2vec[3]. The word2vec provides an efficient implementation for computing vector representation of words by using the continuous bag-of-words and skip-gram architectures [8]. Firstly, a large number of posts are randomly crawled from Sina Micro-blog website (weibo.com) to constitute a 1.5G micro-blog dataset and can be transformed to vector representation of words, and then the synonyms of the standard emotional words are chosen as the candidate words. Secondly, it needs to compute the similarity and choose the maximum between the candidate words and the standard words. Meanwhile, the emotion intensity E_i of the i^{th} selected word can be defined as the formula (1) below, where WD_i represents the i^{th} word in the candidate word list, ST_j represents the j^{th} word in the standard word list, $SIM(WD_i, ST_j)$ represents the maximum similarity between the candidate word and the standard word, and $I(ST_j)$ represents the emotion intensity of the standard word.

$$E_i = SIM(WD_i, ST_j) * I(ST_j) \tag{1}$$

4.2 Multiple Features Recognition of Chinese Micro-Blogs

Emoticons. In this paper, we will combine the emoticon with the corresponding intensity to assist in calculating the proportion of the emotion cause. Firstly, the micro-blogs can be formulated as a triple $U = (C, R, T)$, where C means the emoticon list, R is the emotional keyword list, and T represents a list of micro-blog posts. As for one post, it can be regarded as a triple $u_x = (c_i, r_{kj}, t_n)$, where c_i means the i^{th} emoticon in C, r_{kj} is the j^{th} emotion keyword in R of the k^{th} (1≤k≤22) emotion, t_n represents the n^{th} post in T. If c_i and r_{kj} appear

[2] http://nlg18.csie.ntu.edu.tw:8080/opinion/index.html
[3] http://word2vec.googlecode.com/svn/trunk/

within t_n at the same time, the corresponding co-occurrence frequency is called $|CO(c_i, r_{kj})|$.

As for the co-occurrence intensity between the corresponding emoticon and the emotion keyword, it can be represented as $\delta_{ij}(c_i, r_{kj})$, see the formula(2), where $|c_i|$ is the number of c_i appearing in t_n, and $|r_{kj}|$ is the number of r_{kj} appearing in t_n.

$$\delta_{ij}(c_i, r_{kj}) = \frac{|CO(c_i, r_{kj})|}{(|c_i| + |r_{kj}|) - |CO(c_i, r_{kj})|} \tag{2}$$

According to the above definitions, it is easy to construct the co-occurrence graph (see Figure 1). Within the figure, E is the set of center nodes containing emoticons (e.g., c_1, c_2, c_3, etc.), and D is the set of leaf nodes containing emotion keywords (e.g., r_{11}, r_{12}, r_{13}, etc.). P is the side set between E and D, which represents the degree of closeness between the emotion intensities of c_i and r_{kj}. The longer the side is, the closer the emotion intensities of both are [9].

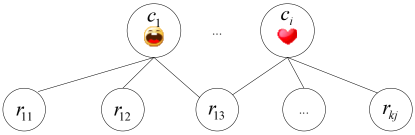

Fig. 1. The co-occurrence graph

As for the weight of the side in the co-occurrence graph (it is defined as $W_{ij}(c_i, r_{kj})$), it can be set to a value equal to the co-occurrence intensity, which is shown in the formula (3):

$$W_{ij}(c_i, r_{kj}) = \delta_{ij}(c_i, r_{kj}) \tag{3}$$

According to the above equations and definitions, the emotion intensity of the i^{th} emoticon (which is called as I_{ICON_i}) can be inferred as follows, see the formula (4), and E_j is the emotion intensity of the j^{th} emotion keyword.

$$I_{ICON_i} = E_j * \max_{1 \leq k \leq 22} W_{ij}(c_i, r_{kj}) \tag{4}$$

Degree Adverbs. In addition, the modification of degree adverbs is helpful in computing the emotion intensities of cause events. This paper uses the intensifier lexicon including 219 degree adverbs, which are divided into five levels: “Ji2 Qi2 | extreme”; “Hen3 | very”; “Jiao4 | more”; “Shao1 | -ish” and “Qian4 | insufficiently” [10]. Then the influence coefficient is set to x, and the value of x is +0.5, +0.3, -0.1, -0.3 and -0.5 in order. The “-” has the function of weakening the emotion intensities of the corresponding words; and the “+” has the function of strengthening the emotion intensities of the corresponding words. The exponential function e^x is applied to adjust the emotion intensity. Finally, the emotion intensity of the i^{th} emotion keyword (called I_{DA_i}) can be calculated by the following formula (5). Here, γ is the adjustable parameter.

$$I_{DA_i} = \gamma e^x E_i (\gamma \geq 1) \tag{5}$$

Negation Words. As for negation words, they can also impact the negative transformation of emotions or impact the emotion intensity scores. With the location of the negation word changing, the emotion and the corresponding emotion intensity will also change. The following Table 2 describes the influence of the locations of negation words in four situations. "NE" means the negation word, "DA" means the degree adverb, "EWD" means the emotion keyword. The parameter α, η and β stand for the adjustable parameters respectively, and "$-$" is the sign of the transformation of emotion. Here, I_{NEGA_i} is used to express the modified result of the emotion intensity of the i^{th} emotion keyword.

Table 2. The detail of the locations of negation words

The four situations	The formula
NE+NE+EWD	$I_{NEGA_i} = \eta E_i (\eta > 1)$
NE+DA+EWD	$I_{(NEGA_i)} = \beta E_i (0 < \beta < 1)$
NE+EWD	$I_{(NEGA_i)} = -E_i$ or $I_{(NEGA_i)} = 0$
DA+NE+EWD	$I_{(NEGA_i)} = -\alpha e^x E_i (\alpha \geq 1)$

Punctuations. In micro-blogs, the emotional punctuation marks, such as the exclamation mark, the interrogation mark etc., usually strengthen the emotion intensities of sentences in some manners. For example, "You can't be so crazy!!!!!", which expresses the emotion of "anger", and the serial exclamation marks can strengthen the emotional intensities. On the other hand, as the repetitive punctuations strengthen the emotion intensities more obviously than the single punctuation, the former is emphasized. The formula (6) is used to compute the emotion intensity which is influenced by punctuations (called I_{PUNC_i}). ε is the adjustable parameter that can be confirmed by the degree of repeatability which has a direct relationship with the number of punctuations.

$$I_{PUNC_i} = \varepsilon * E_i (\varepsilon \geq 1) \tag{6}$$

Furthermore, with regard to the interrogative sentences, they have different means in different circumstances. Sometimes, the interrogative sentence describes a positive attitude literally, but its actual mean is negative. In another case, it can also strengthen this emotion.

Conjunctions. As we all know, there are different kinds of conjunctions such as coordinating conjunctions, adversative conjunctions, causal conjunctions and so on. Under certain circumstance, conjunctions can play an important role in the emotional expression. For example, the conjunction *"Dan4 Shi4"* ("but") mainly emphasizes the event with a strong emotion which is behind it. And the influencing parameters are divided into two classes. One is that the conjunction impacts on the emotion intensity of its front event, and which is set to F_{before}; the other is that the conjunction impacts on the emotion intensity of its back event, and which is set to F_{after}. If $F_{before} = F_{after}$, it means that the conjunction has the same effect on the front event and the back event, so the values of the two parameters are set to 1; if $F_{before} < F_{after}$, it means that the conjunction has no effect on the front event but strengthens the emotion intensity of the back event, so we set that F_{before}=1 and $F_{after} > 1$; and if $F_{before} > F_{after}$, it presents the opposite case with $F_{before} < F_{after}$, so we set that $F_{before} > 1$ and $F_{after} = 1$. Therefore, this paper proposes the formula (7) to express the modified result.

$$I_{CONJ_i} = \begin{cases} F_{before} * E_i & F_{before} > F_{after} \\ E_i & F_{before} = F_{after} \\ F_{after} * E_i & F_{before} < F_{after} \end{cases} \tag{7}$$

4.3 Emotion Cause Components Proportions Calculation

In this section, we combine the characteristics of Bayesian probability model to describe the proportions of emotion causes from the perspective of the prior probability and the conditional probability.

Firstly, this paper constructs an emotion cause component matrix $\rho(s)$ for micro-blog posts, $E(C_m)$ represents the emotion vector with cause components, m is the serial number of 22 types of emotions, E_{nm} represents the n^{th} emotion cause intensity score of the m^{th} emotion, see the formula (8).

$$\rho(s) = (E(C_1), E(C_2), \cdots, E(C_m))^T = \begin{pmatrix} E_{11} & \cdots & E_{1m} \\ \vdots & \ddots & \vdots \\ E_{n1} & \cdots & E_{nm} \end{pmatrix} \tag{8}$$

The proportion of the n^{th} cause component under the m^{th} emotion (which is defined as $P(Emo_m|Cau_n)$) can be computed based on the Bayesian probability, see the formula (9).

$$P(Emo_m|Cau_n) = \frac{P(Cau_n|Emo_m)P(Emo_m)}{\sum_{m=1}^{22} P(Emo_m)P(Cau_n|Emo_m)} \tag{9}$$

In the above formula (9), the parameter Emo_m is the m^{th} emotion and Cau_n is the n^{th} cause component under Emo_m. The prior probability $P(Emo_m)$ is the probability distribution of Emo_m. It can be calculated by the formula (10) and (11), Where $SCORE(Emo_m)$ is the m^{th} emotion intensity score which can be modified by the multiple features in micro-blogs. And I_{ICON_i} is the result modified by emoticons in micro-blogs; I_{DA_i} is the result modified by degree adverbs; I_{NEGA_i} is the result modified by negation words; $I_{PUNC_i}$is the result modified by punctuations; and I_{CONJ_i} is the result modified by conjunctions which are described in the above sections. If there is no any linguistic feature in micro-blogs, the modified result will be ignored and set to 0.

$$P(Emo_m) = \frac{SCORE(Emo_m)}{\sum_{m=1}^{22} SCORE(Emo_m)} \tag{10}$$

$$SCORE(Emo_m) = \sum_{i=1}(E_i + I_{DA_i} + I_{NEGA_i} + I_{ICON_i} + I_{PUNC_i} + I_{CONJ_i}) \tag{11}$$

Within the above formula (9), $P(Cau_n|Emo_m)$ is the probability density function of the n^{th} cause component in a known condition of emotion. It can be calculated by the formula (12) and (13), where $SCORE(Cau_n)$ is the emotion intensity score of the n^{th} cause component under the m^{th} emotion. It is also influenced by the multiple features in micro-blogs.

$$P(Cau_n|Emo_m) = \frac{SCORE(Cau_n)}{\sum_{n=1} SCORE(Cau_n)} \tag{12}$$

$$SCORE(Cau_n) = \sum_{i=1}(E_{im} + I_{DA_{im}} + I_{NEGA_{im}} + I_{ICON_{im}} + I_{PUNC_{im}} + I_{CONJ_{im}}) \tag{13}$$

5 Experiments and Analysis

5.1 The Experimental Dataset

In this section, some strategies based on simulating browsers' behaviors are used to obtain the micro-blog dataset from Chinese micro-blog website (weibo.com) [11]. In the dataset, the micro-blog posts are short, and most of them are less than 140 characters in each post. After the pre-processing (i.e., removing duplicates, filtering irrelevant results and doing some conversion), 16371 posts are remained in our dataset. Meanwhile, every data is used to label the emotion type and the corresponding cause by some annotators manually. In the process of annotating, the micro-blog posts with obvious emotions are chosen to annotate. If the micro-blog does not belong to any category, it will not be labeled; and if the micro-blog contains both the emotion type and the corresponding cause component, it will be labeled both. If the micro-blog does not contain any cause component, it will be only labeled the emotion type.

5.2 Evaluation Metrics

This paper uses the following three metrics to evaluate the performance: precision (U_P), recall (U_R) and F-score (U_F), see the formula (14), (15) and (16), respectively. S is the set of all posts in the collection; NEC is the number of posts with cause components which are identified through our algorithm; NEM is the total number of posts with cause components.

$$U_P = \frac{s_i \in S|Proportion(s_i) \quad is \quad correct}{NEC} \tag{14}$$

$$U_R = \frac{s_i \in S|Proportion(s_i) \quad is \quad correct}{NEM} \tag{15}$$

$$U_F = \frac{2 * U_P * U_R}{U_P + U_R} \tag{16}$$

As for the correctness of the cause components proportions under different emotions, the following method is used to analyze. Firstly, the proportional range of each cause component is divided into ten levels, and each level is expressed as τ_i(1≤i≤10): τ_1 ∈(0-10%), τ_2 ∈(10%-20%), τ_3 ∈(20%-30%), τ_4 ∈(30%-40%), τ_5 ∈(40%-50%), τ_6 ∈(50%-60%), τ_7 ∈(60%-70%), τ_8 ∈(70%-80%), τ_9 ∈ (80%-90%) and τ_{10} ∈ (90%-100%), respectively. The sampled posts with cause components are labeled manually according to the above proportional levels. Secondly, the proportion of the cause component is obtained according to our algorithm and set to x. If the absolute error between x and τ_i is less than φ (φ is a calibration parameter), then the result is correct; otherwise, it is incorrect.

5.3 Experimental Analysis

For examining the effects of emotion cause extraction, this paper conducts the experiments from two aspects. One is to verify the feasibility of our algorithm in the aspect of emotion cause detection; the other is to verify whether using the multiple features can improve the accuracy of calculation of cause component proportion effectively.

To begin with, this paper compares our method with two other methods, Method I is designed on the rule-based system proposed by Lee et al. [1], and Method II is proposed in [3]. The method of ours in extracting emotion causes is based on an emotion model, while the other two methods use some rules and linguistic cues to extract emotion causes. Meanwhile, we use the same metric which is referred in the literature [3], then the results in terms of F-score are shown in Table 3. Obviously, the performance of ours is superior to others, and the F-score improves by 12.95% and 4.21% than Method I and Method II respectively. The experimental result demonstrates the feasibility of this approach and lays a foundation for the calculation of cause component proportion.

Besides, seven experiments are organized for gaining another goal. One is the baseline experiment, we can calculate the proportions only from the emotion intensities of keywords (which is called "EIK"). "EIK_DA" is the experiment which

Table 3. Comparison among the methods

Methods	F-score (%)
Our method	65.51
Method I	52.56
Method II	61.30

calculates the proportions combining "EIK" with degree adverbs. "EIK_ICON" is the experiment which calculates the proportions combining "EIK" with emoticons in micro-blogs. "EIK_NEGA" is the experiment which calculates the proportions combining "EIK" with negation words. "EIK_PUNC" is the experiment which calculates the proportions combining "EIK" with punctuations. "EIK_CONJ" is the experiment which calculates the proportions combining "EIK" with conjunctions. "EIK_ALL" is the experiment which calculates the proportions combining "EIK" with the five features. And the results are shown in Table 4.

Table 4 shows the results of the corresponding baseline experiment and the experiments of the proposed algorithm with multiple features. The precision of "EIK_ALL" is 82.50% which is higher than the other experiments. And the baseline without any language feature has the lowest F-score which is 69.99%. If the experiment is conducted by combining with emoticons, then the F-score increases to 73.09%. By the same token, if we add the other features, the F-score also increases. Obviously, recognizing the linguistic features of emoticons, negation words, punctuations, conjunctions and degree adverbs can be helpful in extracting emotion causes and calculating the proportions of the cause components. We also find that emoticons and negation words have a greater impact on calculating the proportion. And the precisions of the two experiments are 79.91% and 79.55% respectively. That is, people tend to use the two features to express strong emotions in micro-blogs.

Table 4. The results of the experiments

Experiments	Precision (%)	Recall (%)	F-score (%)
Baseline	76.52	64.48	69.99
EIK_DA	77.05	64.94	70.48
EIK_ICON	79.91	67.34	73.09
EIK_NEGA	79.55	67.04	72.76
EIK_PUNC	77.50	65.31	70.89
EIK_CONJ	77.32	65.16	70.72
EIK_ALL	82.50	69.53	75.46

Obviously, by using the method of emotion cause detection based on the ECOCC model, we can extract the cause events that trigger different emotions effectively. Meanwhile, it is possible to find the main cause component on the basis of the proportions of causes under public emotions. Moreover it can also help researchers to study the psychological activity of bloggers, and it has a profound influence on data mining.

6 Conclusions

In this paper, according to the emotional database, we present an ECOCC model which describes the eliciting conditions of emotions. The corresponding cause components under fine-grained emotions are also extracted. Latter, the proportions of cause components in the influence of the multiple features are calculated based on Bayesian probability. The experiment results demonstrate the effectiveness of the approach.

Acknowledgments. This work is sponsored by National Natural Science Foundation of China (Grant No.: 61175110, 61272362) and National Basic Research Program of China (973 Program, Grant No.: 2012CB316301). It is also sponsored by National Science Foundation of Hebei Province and its Academic Exchange Foundation (Grant No.: F2013208105, S2014208001), Key Research Project for University of Hebei Province (Grant No.: ZD2014029) and National Banking Information Technology Risk Management Projects of China.

References

1. Lee, S.Y.M., Chen, Y., Huang, C.-R., Li, S.: Detecting emotion causes with a linguistic rule-based approach. Computational Intelligence **39**(3), 390–416 (2013)
2. Chen, Y., Lee, S.Y.M., Li, S., Huang, C.-R.: Emotion cause detection with linguistic constructions. In: 23rd COLING, pp. 179–187 (2010)
3. Li, W., Xu, H.: Text-based emotion classification using emotion cause extraction. Expert Systems with Applications **41**(4), 1742–1749 (2014)
4. Rao, Y., Li, Q., Mao, X., Liu, W.: Sentiment topic models for social emotion mining. Information Sciences **266**, 90–100 (2014)
5. Steunebrink, B.R., Dastani, M., Meyer, J.-J.C.: A formal model of emotion triggers: an approach for bdi agents. Synthese **185**(1), 83–129 (2012)
6. Che, W., Li, Z., Liu, T.: Ltp: A chinese language technology platform. In: International Conference on Computational Linguistics: Demonstrations, pp. 13–16 (2010)
7. Xu, L., Liu, H., Pan, Y., Ren, H., Chen, J.: Constructing the affective lexicon ontology. Journal of the China Society for Scientific and Technical Information **27**(2), 180–185 (2008)
8. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: 1st ICLR (2013)
9. Cui, A., Zhang, M., Liu, Y., Ma, S.: Emotion tokens: bridging the gap among multilingual twitter sentiment analysis. In: Salem, M.V.M., Shaalan, K., Oroumchian, F., Shakery, A., Khelalfa, H. (eds.) AIRS 2011. LNCS, vol. 7097, pp. 238–249. Springer, Heidelberg (2011)
10. Zhang, P., He, Z.: A weakly supervised approach to chinese sentiment classification using partitioned self-training. Journal of Information Science **39**(6), 815–831 (2013)
11. Gao, K., Zhou, E.-L., Grover, S.: Applied methods and techniques for modeling and control on micro-blog data crawler. In: Liu, L., Zhu, Q., Cheng, L., Wang, Y., Zhao, D. (eds.) Applied Methods and Techniques for Mechatronic Systems. LNCIS, vol. 452, pp. 171–188. Springer, Heidelberg (2014)