Expert Systems with Applications 41 (2014) 1742–1749

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Text-based emotion classification using emotion cause extraction

Weiyuan Li [a,b], Hua Xu [a,*]

[a] 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 100084, China
[b] Beijing University of Posts and Telecommunications, Beijing 100876, China

ARTICLE INFO

Keywords:
Emotion classification
Emotion cause extraction
Microblogging
Weibo

ABSTRACT

In recent years, increasing impact of social networks on people's opinions and decision making has attracted lots of attention. Microblogging, one of the most popular social network applications that allows people to share ideas and discuss over various topics, is taken as a rich resource of opinion and emotion data. In this paper, we propose and implement a novel method for identifying emotions in microblog posts. Unlike traditional approaches which are mostly based on statistical methods, we try to infer and extract the reasons of emotions by importing knowledge and theories from other fields such as Sociology. Based on the theory that a triggering cause event is an integral part of emotion, the technique of emotion cause extraction is used as a crucial step to improve the quality of selected features. First, after thorough analysis on sample data we constructed an automatic rule-based system to detect and extract the cause event of each emotional post. We build an emotion corpus with Chinese microblog posts labeled by human annotators. Then a classifier is trained to classify emotions in microblog posts based on extracted cause events. The overall performance of our system is very promising. The experiment results show that our approach is effective in selecting informative features. Our system outperformed the baseline noticeably in most cases, suggesting its great potential. This exploration should provide a new way to look at the emotion classification task and lay the ground for future research on textual emotion processing.

1. Introduction

Microblogging is a form of blogging that allows users to publish brief text posts (usually with a strict length limit of not more than 200 characters) (Kaplan & Haenlein, 2011). Microbloggers post about a wide range of topics including daily life, comments on movies or books and opinions on social events. Because of the simplicity and casualness, the number of microbloggers has been growing rapidly in recent years. Since users are able to update their content quickly, microblogging services also act as a hub of real-time news. Organizations such as companies, charity groups and departments of government use microblogging as a tool for marketing and public relations as well. Microblogging services are gradually becoming a platform where information, ideas and opinions converge. Nowadays, many people make their decisions under the influence of the microbloggers they follow. Microblog posts are considered as rich sources of emotion and opinion data (Pak & Paroubek, 2010). It is of great interest to mining user emotions in a microblogging community for the purpose of public opinion tracking, content filtering and customer relationship management (Zhang, Zeng, Li, Wang, & Zuo, 2009).

Emotion processing in text is currently a hot and active area in the field of Natural Language Processing (NLP). Textual emotion detection or classification is one task that many scholars and researchers concentrate on. Though the details may vary, the general goal is the same – to detect and recognize the type of emotion, for example, *happiness*, *anger* and *surprise*, conveyed by the target document (Mihalcea & Liu, 2006). Traditionally, due to the statistical classification nature, the most common practices adopted by researchers are mainly statistics-based models. Feature selection methods like InfoGain and χ^2 Test and classifier algorithms like Support Vector Machine (SVM) and k-nearest neighbor (k-NN) are some of the traditional ways for text classification tasks. However, those approaches are very limited in two ways. First, complicated sentences with negation or rhetorical questions cannot be handled well. Second, information of deeper levels, such as the why and how this specific emotion rises are neglected. However, they are very interesting information and sometimes can better reflect the emotion. If we think about how a typical person feels and understands the emotion within a piece of article, we will notice that it is often factors which are not statistically significant, such as the events, the reaction of people, that truly define

* Corresponding author at: 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 100084, China. Tel.: +86 18610022988.
E-mail address: xuhua@tsinghua.edu.cn (H. Xu).

0957-4174/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2013.08.073

our perception of emotion. It is intuitive to take such information into consideration when classifying emotions in posts people put on the web.

Among many elements regarding how emotions rise, expressed and perceived by others, the triggering event are often considered one of the most crucial ones. Many scholars in various disciplines have been studying the relation and interplay between emotions and cause events. From a Psychological point of view, there are theories (James, 1884) believe that the cause event itself should be an integral part of emotional experience. In the field of Sociology, Kleres (2011) proposed "narrative analysis" as a methodological approach to systematically analyze emotions by finding out what happens. Lee, Chen, and Huang (2010a) designed a rule-based system to detect emotion causes. Even though those researches do not concentrate on the task of automatic emotion classification, they lay the ground for us.

In this study, we take a fresh approach on classifying textual emotions. We propose a method to classify emotions using the emotion cause extraction technique, based on the combination of cross-disciplinary knowledge and careful investigation on microblog data. We consider it as a novel method because are unaware of any previous emotion classification works using this technique. We focus on identifying emotions based on posts extracted from the website Weibo, the most popular and influential Chinese microblogging community. We take emotion cause events as our entrance point to overcome some of the drawbacks of traditional approaches. Emotions and reactions triggered by the same event are assumed to be similar, so the errors caused by rhetoric can be reduced. Also, deep-level information is taken into account. The experiment results show that the our system can extract emotion cause events from microblog posts with a good accuracy. Based on an efficient cause event extraction, the emotion classification results of our system improves noticeably.

The rest of this paper is structured as follows. Section 2 discussed the related work on emotion analysis including traditional methods and new explorations. Section 3 gives a brief introduction of the Chinese microblogging platform Weibo, and describes the proposed method using emotion cause extraction technique. In Section 4, experiment results of the two stages of our approach are reported and discussed. Section 5 presents the conclusions and our future work.

2. Related work

We present and briefly introduce related work on the task of emotion processing in this section.

2.1. Emotion classes

There are mainly two types of classification, binary classification (coarse-grained classification of sentiment polarity) and multi-class classification (fine-grained classification multiple classes) (Plutchik, 1980; Turner, 2000).

Most prior research work concentrated on binary classification, i.e. positive and negative. However, a multi-class classification system that reveals more detailed information usually has more practical interest. For example, commercial advertisements will be pushed more accurately and less annoyingly if user's specific emotion status is known. Understanding more about users' current feelings will also help social network websites to create a atmosphere that is warmer and friendlier. Even though right now there is no common agreement among many theories on multi-class emotion classification, several basic emotion types are generally assumed. Some primary emotion classes such as joy and anger are intuitive and commonly-found in similar researches. In our research, to achieve a balance between performance and richness of classes, we adopted the Ekman and Friesen (1971)'s list of six basic emotions (happiness, anger, disgust, fear, sadness, surprise). It should be noted that there is no line separating right and wrong. The set-up of emotion classes should be designed case by case.

2.2. Textual emotion classification

Generally speaking, the research and application of emotion processing in text is still in a very preliminary stage. The inherent ambiguity and subtlety of natural languages are some of the many factors that make this task very challenging, especially in an social network environment where sentences are often incomplete or incoherent. There are many researches concerning recognizing and classifying emotions in different types of text, such as news reports, children's fairy tales, product reviews and customer feedbacks. Generally there are two common approaches to this task, namely a rule-based one and a machine-learning-based one. A rule-based system that tags emotions in news headlines was proposed and implemented by Chaumartin (2007). It computes word's sentiment polarity according to linguistic knowledge and predefined rules. Even though this system achieves a high accuracy, the recall is rather low. When it comes to the machine-learning-based approach, Tan and Zhang (2008) explored four feature selection methods (MI, IG, CHI and DF) and five learning methods (centroid classifier, K-nearest neighbor, winnow classifier, Naive Bayes and SVM) in an empirical study. The experiment results show that IG and SVM perform best. They also point out that classifiers severely depends on domains and topics. Tokuhisa, Inui, and Matsumoto (2008) adopts the k-nearest-neighbor (k-NN) method and a two-step classification model. Based on a very big amount of data extracted from the web, this system significantly outperformed the baseline. There are also variations in terms of the approach and dataset. Ghazi, Inkpen, and Szpakowicz (2010) compared hierarchical and flat classification. Tang and Chen (2011) modeled emotion mining from the writer perspective, reader perspective and the combined perspective using a Plurk dataset. Ye, Zhang, and Law (2009) incorporated sentiment classification techniques into review mining and reached accuracies of over 80% with a large training dataset. Kontopoulos, Berberidis, Dergiades, and Bassiliades (2013) proposed a more efficient sentiment analysis of Twitter posts using ontology-based techniques.

Our work is different from the others. We first import some knowledge of human emotions from other fields such as Sociology, and investigate the posts on Weibo to reveal some connections between the cause events and certain types of emotion. Based on the framework proposed by Lee et al. (2010a), we also developed a emotion cause extraction subsystem that applies well to microblog data. In addition, we do not stop at just detecting the emotion causes. The emotion cause extraction subsystem, which efficiently mines deep-level emotional information, is integrated as a part of our emotion classification system, resulting in an improved classification result. Our exploration will provide a brand new way to look at the problem of emotion processing.

3. Emotion classification using the emotion cause detection technique

The basic idea behind our approach is looking for features that are "meaningful" to emotions instead of simply choosing words with high co-occurrence degree. Fig. 1 depicts the general framework of our emotion classification method. In the following subsections, we expand on the important parts of our system to explain how we classify emotions in microblog posts with emotion causes.

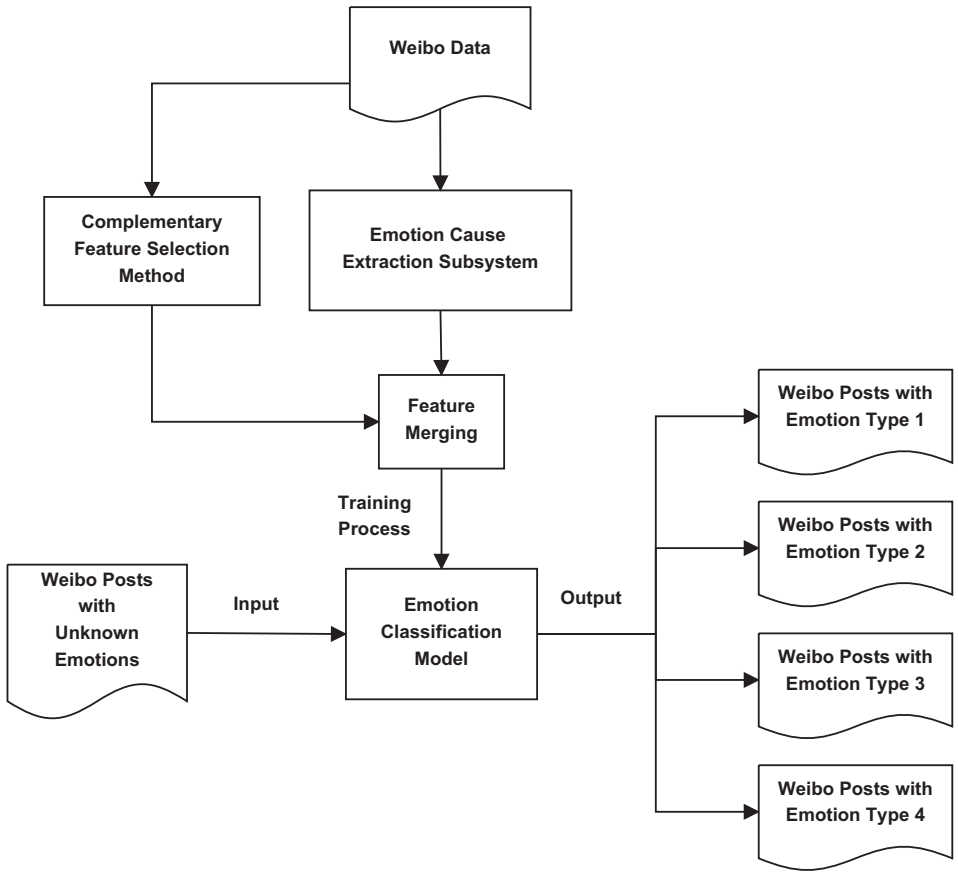

Fig. 1. General workflow of emotion classification with emotion cause detection.

3.1. The Weibo microblogging community

Weibo, the Chinese name for "microblog" is a web-based social network service in China. The content on Weibo is in a format similar to Twitter. There is a length limit of 140 characters for each post. Users may mention or talk to others using the "@UserName" format, mark the topics or key points using the "#TagName#" hashtag format. Besides, Weibo also provides some advanced features. Users can insert various graphical emoticons and web links in their posts or attach video or pictures. Posts can be commented and 're-tweeted' using the "//@UserName" format. It can also serve as a instant messaging service for users to exchange real-time messages. Launched in 2009, Weibo has quickly attracted a huge amount of users and is now one of the most popular websites in China, in use by more than 30% of Internet users. Fig. 2 shows the search interest of Weibo.

Typical Weibo posts are short and informal pieces of text submitted by users. Weibo was chosen as the target source of test data of our study for the following reasons. First, it is the most popular platform where China's Internet users make voices online. So it is safe to assume the amount of posts containing emotions and emotion-provoking events is relatively big. Second, exploring emotion classification in an environment of modern social network will have much practical interest. Third, there are few prior researches on Chinese posts from the web.

After automatically obtaining random posts from Weibo, some necessary preprocessing steps are performed. Distractions like web links and geotags are removed. Then we use the ICTCLAS tool-kit to parse, segment and tag the Chinese posts with proper part-of-speech (POS) tags.

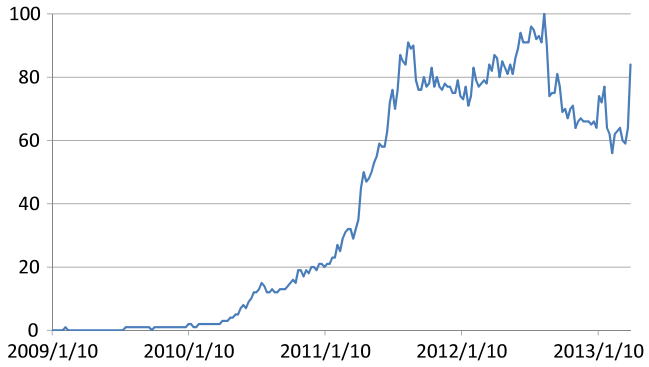

Fig. 2. Google trends for Weibo.

3.2. Feature selection subsystem using emotion causes

From a sociological point of view, narratives are inextricably emotionally structured and narrative analysis can be used as an method to study and examine emotions (Kleres, 2011). The emotion experience of human has a crucial narrative dimension. The key idea behind this technique is that emotion can be inferred and analysed if "who acts how to whom and what happens" (Sarbin, 1989). In addition to this, external events are often regarded as the triggers of certain emotions (Wierzbicka, 1999). Kleres (2011) also listed a number of linguistic manifestations, on lexical and syntactical levels, that could provide great access to how emotions are expressed and perceived. For example, direct references to emotions and double propositions, where a neutral sentence is

embedded in an emotive one are some common expressions at the level of entire sentences.

Beyond its meaning for psychologists and sociologists, this methodological approach also describes an analytical framework for us to look at the emotion processing problem in a different way. Since there is a strong connection between emotions and events, it is natural to come up with the idea that finding out the events would potentially help mining the emotions. So in our work we attempt detect emotions by finding out why emotions are generated and how they are felt by human readers. Firstly, we adopt the notion of emotion cause event. As described by Talmy (2000), 'cause events' should be the reason of emotions. It should be noted that the cause event is not always the actual trigger event we usually understand. It could be the event that cause the person to have the emotion, or the perception of that event. Here are some examples where the cause event parts are marked with underlines.

1. *Shi3 Zai4 Tai4 Gao1 Xing4 Le, Wo3 De Fen3 Si1 Chao1 Guo4 Bai3 Wan3 Le!*
   (I am so thrilled that I have more than one million followers now!)
2. *Zu3 Qiu3 Zhi3 Shi4 Yi1 Xiang4 You3 Xi4. Yan3 Bian4 Cheng2 Ru2 Ci3 Xue4 Xing1 De4 Si1 Sha1, Bei1 Ai1 A4!*
   (Soccer is just a game. How could it turn into such a violent fight. What a tragedy!)
3. *Zui4 Ai4 Zhou1 Mo1 De4 Hao3 Shi2 Guang1 Le. He2 Qin1 You3 Yi4 Qi3 Chi1 Fan4, He1 Cha2. Zhen1 Shi4 Tai4 Qie4 Yi4 La.*
   (Love the good time of weekend most. It is such a pleasure to have dinner and drink some tea with family and friend.)

Lee et al. (2010a) demonstrated the feasibility to mine deep information such as emotion cause events, which coincided with the Sociology theory of narrative nature of emotion. They presented a well-designed and versatile framework for emotion cause event detection. This framework consists of following key components:

- Marker List: a list of keywords to mark the occurrence of emotion cause events.
- Emotion Keyword List: a list of words and phrases commonly used to express emotions or feelings.
- Linguistic Pattern Set: a table of linguistic patterns that describe how emotions are expressed and how elements are arranged.

Because of the similar goal to detect and extract the emotion cause events in text, we decide to import the general structure of this rule-based method. However, Lee et al. (2010a)'s work was based on the "Academia Sinica Balanced Corpus of Modern Chinese", simplified as Sinica Corpus (Chen, Huang, Chang, & Hsu, 1996). Texts collected in this corpus range from news reports to novels and poets, which are usually long and formal. Our work, on the contrary, focuses on short (sometimes even incomplete) informal posts in a microblogging community. The obvious difference between document style means that the original setting of the algorithm no longer applies to our task.

To overcome this problem, we conducted a thorough investigation to re-design the system. First, we examined a separate development dataset containing more than 1000 random emotion entries based on Weibo. The list of linguistic cue words (Lee, Ying, & Huang, 2010b) were run through to investigate each word's effectiveness in identifying the occurrence of emotion cause events. Some formal words and phrases are extremely rare in a microblog context and some casual expressions are more likely to be used by Internet users. Based on the manual examination, we removed several words that are less highly collocated with cause events from the Marker List and added some new ones. Table 1 is the marker list after adaptation to microblog posts.

For the same reason, a traditional list of emotion words such as the one provided by HowNet does not work with microblog data well. Instead, we use a emotion word list developed through our long time of work. It contains a total of 1845 words and short phrases commonly used to express feelings, opinions and emotions. The list covers many words and phrases popular among Chinese Internet users like "Niu2" (superb) and "Gei3Li4" (awesome). It also includes synonyms and common variations of most keywords.

Then we independently generalize a set of linguistic patterns to pinpoint emotion causes after taking typical expression habits of Internet users and structural patterns described by Kleres (2011) into consideration. Because of the length limit, each post normally concentrates on one topic. This concise and expressive characteristic indicates that if an event is extracted from a post, it is very likely to be the emotion cause event. In addition, we remove the clause constraints to make the pattern matching process more flexible.

Our final goal is to classify the emotion types. Emotion cause extraction is an attempt to improve the traditional feature selection. It is obvious that if there is too much noise in the selected features, it will be difficult for the classifier to achieve good performance. So the precision is a higher priority than recall when we were designing the rule set. Table 2 lists the linguist patterns

Table 1
List of Linguistic markers.

Group No.	Cue Words
I	'let/make': Rang4, Ling4, Shi3, Gao3De2, Nong3De2, Shi3De2
II	'to think about': Xiang3, Xiang3Dao4, Xiang3Qi3, Yi1Xiang3 'to talk about': Shuo1Dao4, Jiang3Dao4
III	'to feel': Gan3Dao4, Jue2De2, Gan3Jue2
IV	'to see': Kan4Dao4, Kan4Jian4, Jian4Dao4 'to hear': Ting1Dao4, Ting1Shuo1 'to know': Zhi1Dao4, De2Zhi1, Fa1Xian4
V	'for': Wei4Le, Dui4Yu2
VI	'because': Yin1, Yin1Wei4, You2Yu2
VII	'is': Shi4, De4Shi4 'can': Neng2, Neng2Gou4, Ke3Yi3

Table 2
Linguistic patterns for emotion cause detection.

No.	Linguistic pattern
1	C + I + E + K E: the nearest Noun after I C: the nearest (Noun, Verb, Noun) before I
2	E + K + II/III/IV + C E: the nearest Noun before II/III/IV C: the nearest (Noun, Verb, Noun) after II/III/IV
3	V/VI + C + E + K E: the nearest Noun before K C: the nearest (Noun, Verb, Noun) after V/VI
4	E + "yue4Cyue4K" E: the nearest Noun before "yue4Cyue4K" C: the nearest (Noun, Verb, Noun) in "yue4Cyue4K"
5	E + K + V/VI + C E: the nearest Noun before K C: the nearest (Noun, Verb, Noun) after V/VI
6	IV + C + [E] + K C: the nearest (Noun, Verb, Noun) after IV
7	E + V/VI + C + K E: the nearest Noun before V/VI C: the (Noun, Verb, Noun) between V/VI and K
8	C + E + IV + K E: the nearest Noun before IV C: the nearest (Noun, Verb, Noun) before E
9	[E + K] + VII + C + [E + K] one of the two 'E + K' must present C: the nearest (Noun, Verb, Noun) after VII

generalized to describe the association of elements such as emotional expressions, keywords and emotion causes within a Chinese microblog post.

The abbreviation and symbols in Table 2 are explained below.

- C: Cause event
- I/II/III/IV/V/VI/VII: Linguistic markers in the corresponding group
- E: Experiencer
- K: Emotion Keyword
- "[]": Optional linguistic components.

Both verb and noun can be used to narrate an event in natural language. Lee et al. (2010a) represent events with the structure of (Noun, Verb, Noun), but used the clause instead due to some practical difficulties. However, because microblogs are short and sometimes composed of incomplete sentences, a clause covers more than half a post in many cases. In order to pinpoint the emotion cause in a post more precisely, we design our representation scheme as follows. First, the system looks for the structure of (Noun, Verb, Noun) as the basic frame. Because the emotion cause can be expressed as a proposition or a nominal (Lee et al., 2010a) and because of the informality of microblogs, it is not necessary for the two nouns and one verb to present at the same time. In practice, the cause (C) part of a pattern is considered as a match if at least one of the three is found. After locating the frame, text outside this frame is discarded. In other words, only the substring within this frame is saved and returned as the emotion cause, instead of keeping the whole clause containing the structure. Two examples are given as follows.

1. **Original Post:** *"Zhong1 Guo2 Hao3 Sheng1 Yin1" Rang4 Wo3 Jing1 Tan4, Zhe4 Shou3 Ge1 Ba3 Wo3 Ye3 Chang4 De2 high Qi3 Lai Le, Shui4 Bu4 Zhao2 Jiao4.*
   ("The Voice" surprised me. This song really made me excited and unable to sleep.)
   Linguistic Pattern: [C (Noun):*("Zhong1 Guo2 Hao3 Sheng1 Yin1")*]+[I:*Rang4*]+[E: *Wo3*] + [K: *Jing1 Tan4]*
   ([C (Noun): ("The Voice")] [K:surprised] [E:me])
   Emotion Cause Event: *"Zhong1 Guo2 Hao3 Sheng1 Yin1"*
   ("The Voice")
2. **Original Post:** *Zai4 Jia1 Li3 Dou1 You3 Ming2 Xian3 Zhen4 Gan3, Ke3 Shi1 Di4 Zhen4 Ju2 Que4 Mei2 Neng1 Fa1 Bu4 Yu4 Gao4, Nong4 De2 Ren2 Men2 Hen3 Shi4 Fen4 Nu4.*
   (We can even felt the shake in our apartments, but the Seismological Bureau failed to send any warning, which made people very angry.)
   Linguistic Pattern: [C (Noun,Verb,Noun):*(Di4 Zhen4 Ju2, Mei2 Neng1 Fa1 Bu4, Yu4 Gao4)*] + [I: *Nong4 De2*] + [E: *Ren Men2*] + [K: *Fen4 Nu4]*
   ([C (Noun, Verb, Noun): (Seismological Bureau, failed to send, warning)] [I: made] [E:people] [K: angry])
   Emotion Cause Event: *Di4 Zhen4 Ju2 Que4 Mei2 Neng1 Fa1 Bu4 Yu4 Gao4*
   (Seismological Bureau failed to send any warning)

For both examples, linguistic pattern 1 is applied to extract the cause. In Example 1, only one noun is found, so it is saved as the emotion cause. In Example 2, the entire triples of (Noun, Verb, Noun) is found so the substring within this structure is saved as the emotion cause.

After the emotion causes have been extracted from the original posts in the training data, all the strings that represent cause events are saved. Then we remove all the stopping words from the collection of emotion causes. The reason that we move the process of "removing stop words"[1] from the pre-processing step to here is that our emotion cause extraction subsystem is partly dependent on the stop words (linguistic markers). Then we combine all the remaining words and emotion keywords (K) used in the extraction stage to form the emotion cause event set. That is, the set of emotion causes is an aggregation of (C) s and (K) s.

There are three kinds of situations regarding if the posts contain explicit emotions and corresponding causes:

1. Neutral posts (including posts with emotion types too vague to decide);
2. Posts with emotions and emotion causes.
3. Posts with emotion but without explicitly expressed emotion causes.

Clearly, if we only use the output of the emotion cause extraction subsystem as selected features, situation 1 and 3 would be indistinguishable because they both contain no extractable emotion causes. To separate those posts with emotions but without emotion causes from neutral ones, a feature selection methods that could cover situation 3 must be adopted as well. We choose χ^2 (Chi-squared) test. χ^2 test is a classic test for dependence. It compares observed data with data expected to obtain according to a specific hypothesis. Chi-square is the sum of the squared difference between observed and expected value divided by the expected value in all classes. The definition of χ^2 test measure is:

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where χ^2 is Chi-square, O is the observed frequency in each class, E is the expected frequency.

With χ^2 test as complement, our feature selection strategy is designed in a straightforward way. The two sets of features produced by emotion cause extraction and χ^2 test are mixed to generate the final feature set. Algorithm 1 describes the merging process, where *emocause_set* is the set of emotion causes, *Chi_list* is the list of words in descending order by χ^2 test score, and *final_set* the final feature set.

Algorithm 1. Feature set merging

```
d ← predefined dimension of final feature set
d_cause ← sizeof (emocause_set)
r ← d − d_cause
final_set ← emocause_set
for each w in Chi_list do
  if r == 0 then
    break
  else
    if w is not in final_set then
      add w to final_set
      r ← r − 1
    else
      continue
    end if
  end if
end for
```

3.3. SVM and SVR for emotion classification

Support Vector Machine (SVM) (Cortes & Vapnik, 1995) has been proven to be an effective model for text-driven sentiment

[1] Stop words:words that have the same likelihood of occurring in those documents not relevant to a query as in those documents relevant to the query (Wilbur & Sirotkin, 1992).

classification (Tan & Zhang, 2008). Based on the principle of risk minimization, SVM makes decisions according to the selected elements from the training data. It seeks a hyperplane to separate the training data into two classes, positive and negative. Fig. 3 depicts the basic mechanism of SVM.

There are several variations of SVM in terms of the selected kernel function $K(x,y)$. We limit the discussion on the linear kernel function in our work, for its relatively high performance. In addition, because our work concentrates on short and informal posts randomly crawled from a microblogging website, we choose support vector regression (SVR) for the classification task. SVR is a version of SVM for regression proposed by Drucker, Burges, Kaufman, Smola, and Vapnik (1997). It uses the same principles as SVM, but the output of SVR is a real number. It is a better choice for handling an imbalanced data set.

4. Experiments and discussion

Our experiment is conducted in a two-stage manner:

1. Extracting emotion cause events;
2. Training and testing the classifier.

In this section, we first introduce our dataset based on Weibo. Then we describe the two steps and corresponding evaluation measures separately.

4.1. The Weibo emotion dataset

Firstly, more than 20000 posts were randomly crawled down from the Weibo website. After removing duplicates, 16485 posts remained in our database. Two human annotators labeled all items in the corpus with the type of emotion and, the cause of emotion if exists. To avoid ambiguity and to simplify the problem, only items that express a certain type of emotion explicitly are labeled as "has emotion" and the tag of corresponding emotion. When multiple emotions exist in the same entry, only the dominate one is picked and labeled. If the emotion types are unclear or too hard to decide, those entries will be taken as neutral. Posts annotated inconsistently are forwarded to a third annotator for final decision. In addition, all posts are confirmed by the author after being annotated.

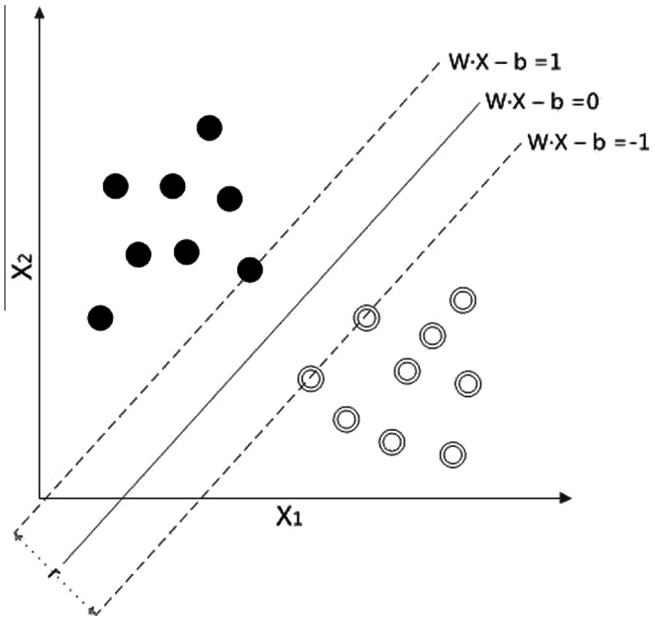

Fig. 3. A SVM max hyperplane with margin.

The emotion cause events are represented by a substring of the target document. For example, in the sentence of *"Zhen1 Gao1 Xing4, Jin1 Tian1 Dao4 Gong1 Si1 Hou4 Ling3 Dao3 Gei3 Wo3 Men Mei3 Ge4 Ren2 Dou1 Fa1 Le4 Hong2 Bao1."* ("So happy that today after we got to company our boss gave everyone some bonus"), the string of *"Ling3 Dao3 Gei3 Wo3 Men Mei3 Ge4 Ren2 Dou1 Fa1 Le4 Hong2 Bao1."* ("boss gave everyone some bonus") is copied and tagged as the emotion cause event. The annotators were instructed to follow the rule of minimal length, which means only to mark shortest length that is enough to cover the cause. We assume that in a microblogging website where no posts are longer than 140 characters, complicated situations in which emotion causes are multiple long events are extremely rare and thus safe to neglect. Table 3 presents the actual number of instances of each class in our corpus. We keep the data set in its imbalanced state, because we believe it reflects the actual situation. One important cause is that a large proportion of posts are re-posts without comments and automatically generated by 3rd-party applications. Also, news headlines which usually do not contain emotions contribute a lot to the neutral posts.

4.2. Emotion cause extraction

4.2.1. Evaluation standard

Like other NLP tasks, we evaluate the experiment results with three key values: precision, recall and f-score. In our evaluation for the emotion cause extraction subsystem, the precision and recall are defined as follows:

$$Precision(G,S) = \frac{\sum_{P_i \in G}\sum_{k_j \in P_i} ListP(SList_j, GList_j)}{\sum_{P_i \in S}\sum_{k_j \in P_i}} \quad (1)$$

$$Recall(G,S) = \frac{\sum_{P_i \in G}\sum_{k_j \in P_i} ListR(SList_j, GList_j)}{\sum_{P_i \in G}\sum_{k_j \in P_i}} \quad (2)$$

$$ListP(SList, GList) = \frac{\sum_{SStr_j \in SList} \max_{GStr_i \in GList} StrP(GStr_i, SStr_j)}{|SList|} \quad (3)$$

$$ListR(SList, GList) = \frac{\sum_{GStr_j \in GList} \max_{SStr_i \in SList} StrR(GStr_i, SStr_j)}{|GList|} \quad (4)$$

where G is the set of human-annotated gold-standard causes, S is the set of causes produced by our system, P_i is a microblog post in the dataset, k_j is a emotion keyword in this post, $SList_j$ and $GList_j$ are system output and gold-standard lists of emotion causes of the corresponding keyword k_j in P_i, $SStr_i$ and $GStr_i$ are strings that represent the emotion cause.

There are two ways to calculate the StrP and StrR: relaxed match 1 and relaxed match 2 (Lee et al., 2010a). Relaxed match 1 only considers whether there is any overlap between the system output string and our gold-standard string. Relaxed Match 2 takes the overlapped length into consideration. Since we intend to give a brief evaluation over the improvement of the cause extraction system, we simplified this part and adopted the scheme of Relaxed Match 1.

Table 3
Summary of corpus.

Emotion	Number of posts	Number of posts with causes
Happiness	513	354 (69.0%)
Anger	478	452 (94.6%)
Disgust	153	137 (89.5%)
Fear	125	61 (48.8%)
Sadness	313	255 (81.5%)
Surprise	102	46 (45.1%)
Neutral	14801	N/A
Total	16485	1305

$$StrP(GStr, SStr) = \begin{cases} 1 & if\ GStr\ and\ SStr\ overlaps \\ 0 & else \end{cases} \quad (5)$$

$$StrR(GStr, SStr) = \begin{cases} 1 & if\ GStr\ and\ SStr\ overlaps \\ 0 & else \end{cases} \quad (6)$$

F-score is calculated normally as the harmonic average of precision and recall.

4.2.2. Result of cause extraction

To evaluate our work of re-designing the rule-based system for a microblogging environment, we conduct the cause extraction experiment on the same microblog dataset using the same rules and linguistic cues proposed by Lee et al. (2010a) as well.

Because the difference between types of text, we decide to adopt a same baseline for a better comparison. This naive baseline can be described as one rule: C (V) + K, meaning that the clause that contains the first verb to the left to the emotion keyword is the clause of emotion cause.

Table 4 shows the performance of the baseline and the original rule-set. The original rule-set significantly outperformed the baseline in terms of precision. However, its recall is much lower than the baseline's. One main reason for this result is that the original rule-set is not designed based on microblog posts, which also proves the necessity of our adaptation work.

Table 5 shows the overall performance of our system. With a list of linguistic markers and a rule-set designed for microblog posts, both precision and recall improve greatly.

4.3. Emotion classification

Since our work is a multi-class problem, we conduct our experiments in a "one-vs-rest" manner. For each round of the experiment, only one class of emotion is classified.

The goal of our work is to examine whether using the emotion cause extraction technique will help finding a more informative feature set. To serve that goal, it is necessary for us to compare our system with a baseline system that is purely based on traditional feature selection method. As mentioned earlier, to cover emotional posts that do not contain emotion cause events, χ^2 test is adopted as a complementary measure. So, naturally, the baseline is designed as follows: only χ^2 test technique is used and words with highest χ^2 score are selected as our feature words. Based on the same training set, all the feature words are selected according to their χ^2 score. In order to fully reflect whether the emotion cause extraction technique can select more informative features from posts, the baseline's feature vectors are all of the same dimensions.

We use 75% posts in our corpus as the training dataset, and 25% as the test dataset. The linear kernel is adopted for SVR and parameters are set by default.

Table 4
The performance of the original cause extraction rule-set.

	Precision	Recall	F-score
Baseline	45.06	51.24	47.95
The original rule-set	71.63	41.51	52.56

Table 5
The performance of the redesigned algorithm.

	Precision	Recall	F-score
Redesigned algorithm	75.70	51.59	61.30

Table 6
The performance in emotion classification.

Class	Feature construction strategy	Precision	Recall	F-score
Happiness	Baseline	0.8541	0.4795	0.6142
	~with EC	0.8736	0.4853	0.6240
Anger	Baseline	0.7262	0.7428	0.7344
	~with EC	0.7388	0.7600	0.7493
Disgust	Baseline	0.6455	0.4146	0.5049
	~with EC	0.5033	0.6097	0.5514
Fear	Baseline	0.6032	0.6712	0.6354
	~with EC	0.6145	0.6011	0.6077
Sadness	Baseline	0.7157	0.7362	0.7258
	~with EC	0.6945	0.7204	0.7072
Surprise	Baseline	0.7171	0.6291	0.6702
	~with EC	0.7252	0.6322	0.6755

Table 6 shows the performance of the baseline and our emotion-cause-integrated system in emotion classification. The notation of" with EC" means the result of emotion is obtained with emotion cause extraction technique. For most emotion classes (Happiness, Anger, Fear and Surprise), our system has higher precisions while maintaining recalls on the same level. For the class of disgust, though the precision of our system is lower, the recall is higher. In terms of F-score, which evaluates both precision and recall, our system outperforms the baseline noticeably in most cases. Specifically, the f-score improves by 1.6% for happiness, 2.0% for anger, 9.2% for disgust and 0.8% for surprise. It indicates that the technique of emotion cause extraction is effective in selecting better, more informative features. For the class of sadness, our system fails to outperform the baseline. Our hypothesis is that people are less willing to state the reason when they are feeling sad. People are more keen to describe the situation under positive or strong emotions like happiness or surprise. So the emotion causes extracted from sad posts are less effective in separating them from other emotions.

Even though generally the precision and recall are yet not very high, the improved results do demonstrate the great potential of our proposed method. Also, as the size of training dataset becomes larger, the computational complexity of classic feature selection methods like χ^2 test increases exponentially while the rule-based system only has a linear growth.

5. Conclusion and future work

Emotion classification has a broad range of applications. An accurate and efficient classification system is of great interest. In this study, we imported the knowledge regarding the relation between emotions and narratives from Sociology and explored the task of textual emotion classification with the technique of emotion cause extraction. We conducted the experiment on a Chinese microblogging platform. First, as many as 16485 potentially emotion-provoking posts were automatically collected. Then our work was done in two sub-steps: feature selection using the technique of emotion cause extraction and emotion classification using SVR. We designed a rule-based subsystem to detect and extract cause events from the original posts based on the common social network characteristics and other carefully-generalized linguistic patterns. The output of this subsystem is then used to generate the feature set for training the classifier. A baseline scenario is also presented. The experiment results showed that with the emotion cause extraction technique, our system evidently outperformed the baseline system in most cases. The promising result demonstrated the potential to improve traditional emotion classification models with knowledge and methods from other disciplines.

Even though we managed to obtain fairly good results using the cause extraction technique, our work is far from being thorough.

There are still many factors that have big impact on human emotions waiting to be explored. Posts with complicated linguistic patterns are challenging to deal with. In addition, the process of generalizing and designing linguistic patterns still requires trained human experts which is both labor-consuming and time–costing.

For future work, we first need to examine other elements and factors to see their advantages and disadvantages in our task. We are also planning to design a more sophisticated cause extraction system to better tackle challenging situations and reduce errors. More linguistic patterns need to be investigated as well. Finally, we need to find a way to generate and modify the pattern set automatically so that human labor would be kept to a minimum.

Acknowledgments

This work is supported by National Natural Science Foundation of China (Grant No.: 61175110) and National Basic Research Program of China (973 Program, Grant No.: 2012CB316305).

References

Chaumartin, F. -R. (2007). Upar7: A knowledge-based system for headline sentiment tagging. In *Proceedings of the fourth international workshop on semantic evaluations* (pp. 422–425). Association for Computational Linguistics.

Chen, K.-J., Huang, C.-R., Chang, L.-P., & Hsu, H.-L. (1996). Sinica corpus: Design methodology for balanced corpora. *Language, 167*, 176.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning, 20*, 273–297.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, 155–161.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology, 17*, 124.

Ghazi, D., Inkpen, D., & Szpakowicz, S. (2010). Hierarchical versus flat classification of emotions in text. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 140–146). Association for Computational Linguistics.

James, W. (1884). Ii. What is an emotion? *Mind, 3*, 188.

Kaplan, A. M., & Haenlein, M. (2011). The early bird catches the news: Nine things you should know about micro-blogging. *Business Horizons, 54*, 105–113.

Kleres, J. (2011). Emotions and narrative analysis: A methodological approach. *Journal for the Theory of Social Behaviour, 41*, 182–202.

Kontopoulos, E., Berberidis, C., Dergiades, T., & Bassiliades, N. (2013). Ontology-based sentiment analysis of twitter posts. *Expert systems with applications, 40*, 4065–4074.

Lee, S. Y. M., Chen, Y., & Huang, C. -R. (2010). A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text* (pp. 45–53). Association for Computational Linguistics.

Lee, S.Y.M., Ying, C., & Huang, C. (2010b). Emotion cause events: Corpus construction and analysis. In *Proceedings of the seventh international conference on language resources and evaluation (LREC 2010)* (pp. 19–21).

Mihalcea, R., & Liu, H. (2006). A corpus-based approach to finding happiness. In *Proceedings of the AAAI spring symposium on computational approaches to Weblogs* (p. 19).

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of LREC* Vol. 2010.

Plutchik, R. (1980). *Emotion, a psychoevolutionary synthesis*. New York: Harper & Row.

Sarbin, T. R. (1989). Emotions as narrative emplotments. *Entering the Circle: Hermeneutic Investigation in Psychology*, 185–201.

Talmy, L. (2000). *Toward a cognitive semantics. Concept structuring systems* (Vol. 1). The MIT Press.

Tang, Y.-j., & Chen, H.-H. (2011). Emotion modeling from writer/reader perspectives using a microblog dataset. *Sentiment Analysis Where AI Meets Psychology (SAAIP)*, 11.

Tan, S., & Zhang, J. (2008). An empirical study of sentiment analysis for chinese documents. *Expert Systems with Applications, 34*, 2622–2629.

Tokuhisa, R., Inui, K., & Matsumoto, Y. (2008). Emotion classification using massive examples extracted from the web. In *Proceedings of the 22nd international conference on computational linguistics* Vol. 1 (pp. 881–888). Association for Computational Linguistics.

Turner, B. M. (2000). Histone acetylation and an epigenetic code. *Bioessays, 22*, 836–845.

Wierzbicka, A. (1999). *Emotions across languages and cultures: Diversity and universals*. Cambridge University Press.

Wilbur, W. J., & Sirotkin, K. (1992). The automatic identification of stop words. *Journal of Information Science, 18*, 45–55.

Ye, Q., Zhang, Z., & Law, R. (2009). Sentiment classification of online reviews to travel destinations by supervised machine learning approaches. *Expert Systems with Applications, 36*, 6527–6535.

Zhang, C., Zeng, D., Li, J., Wang, F.-Y., & Zuo, W. (2009). Sentiment analysis of chinese documents: From sentence to document level. *Journal of the American Society for Information Science and Technology, 60*, 2474–2487.