Featuring, Detecting, and Visualizing Human Sentiment in Chinese Micro-blog

ZHIWEN YU, Northwestern Polytechnical University
ZHITAO WANG, Northwestern Polytechnical University; The Hong Kong Polytechnic University
LIMING CHEN, De Montfort University
BING GUO, Northwestern Polytechnical University
WENJIE LI, The Hong Kong Polytechnic University

Micro-blog has been increasingly used for the public to express their opinions, and for organisations to detect public sentiment about social events or public policies. In this paper we examine and identify the key problems of this field, focusing particularly on the characteristics of innovative words, multi-media elements and hierarchical structure of Chinese "Weibo". Based on the analysis we propose a novel approach and develop associated theoretical and technological methods to address these problems. These include a new sentiment word mining method based on three wording metrics and point-wise information, a rule set model for analyzing sentiment features of different linguistic components, and the corresponding methodology for calculating sentiment on multi-granularity considering emoticon elements as auxiliary affective factors. We evaluate our new word discovery and sentiment detection methods on a real-life Chinese microblog dataset. Initial results show that our new diction can improve sentiment detection, and demonstrate that our multi-level rule set method is more effective with average accuracy being 10.2% and 1.5% higher than two existing methods for Chinese micro-blog sentiment analysis. In addition, we exploit visualisation techniques to study the relationships between online sentiment and real life. The visualisation of detected sentiment can help depict temporal patterns and spatial discrepancy.

Categories and Subject Descriptors: H.2.8 [**Database Applications**]: Data Mining

General Terms: Algorithms; Experimentation

Additional Key Words and Phrases: Sentiment Detection, Sentiment Lexicon Expansion, Rule Set Based Model, Visualization

ACM Reference Format:
Zhiwen Yu, Zhitao Wang, Liming Chen, Bin Guo, Wenjie Li, 2015. Featuring, Detecting, and Visualizing Human Sentiment in Chinese Micro-blog. *ACM Trans. Knowl. Discov. Data.* 0, 0, Article 00 (April 2015), 23 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

This work is supported by the National Basic Research Program (973 Program) of China (No. 2012CB316400), National Nature Science Foundation of China (No. 61333005, No. 61373119, No. 61222209), the Specialized Research Fund for the Doctoral Program of Higher Education (No. 20126102110043), and Microsoft.
Author's addresses: Z. Yu and B. Guo, School of Computer Science, Northwestern Polytechnical University, Xi'an, China; Z. Wang and W. Li, Department of Computing, The Hong Kong Polytechnic University, Hong Kong; L. Chen, School of Computer Science and Informatics, De Montfort University, The Gateway, Leicester, UK.

DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

As a widely accepted information broadcast platform, micro-blog has become one of the mainstream social media. From the worldwide famous Twitter to Chinese Sina Weibo, micro-blog has attracted growing users and generates billions of opinion text data on a daily basis [Yu et al. 2015][Lu et al. 2014]. Using opinion texts is possible to discover users' sentiment, from which we can explore the temporal and spatial variation of users' moods, sense the effect of important events in different sphere of public views [Bollen et al. 2011a][O'Connor et al. 2010]. Sentiment analysis can also help identify potential trends in some applications such as predicting the stock market [Bollen et al. 2011b]. Compared to the traditional opinion text, micro-blog generated text has the following unique characteristics.

Unique Style: Micro-blog is limited in 140 words. Nevertheless, the colloquial and fashionable expression, usually user customised, is intractable.

New Expression Elements: Not only texts but also emoticons are transmitted with micro-blog net, which embody and convey affective factors.

Social Attribute: In addition to expressing one's own opinions, micro-blog allows users to retweet and comment others' opinions. With these interactions, sentiment of the original users would impact on their followers' opinion.

Spatiotemporal Attribute: Users' sentiment would fluctuate at different times and locations, which may result in a diversity of life styles. As such, spatiotemporal attribute is the key factor on the research of sentiment variation.

Lack of Labelled Data: Micro-blog data is unstructured without metadata. Labelling and managing vast users' data to make them useful is still insufficient and difficult.

The above characteristics present opportunities as well as challenges to sentiment research of social media. While research on English micro-blog analysis has made substantial progress, e.g., Twitter based sentiment analysis, sentiment detection based on Chinese character analysis in Chinese micro-blog such as Sina Weibo is still at early stage. Moreover, Chinese language is more obscure for textual analysis than English as a character-based language. For example: a Chinese tweet may have more information than an English tweet in the same length; word segmentation of Chinese can easily affect the meaning of tweet; the grammar of expression is quite different [Cui et al. 2013]. To address the unique characteristic of Chinese micro-blog analysis, we have identified three key research questions relating to Chinese micro-blog sentiment detection:

— *How to recognize the new words in Chinese micro-blog and interpret their sentiment implication?*

— *How to leverage other media modalities, e.g., emoticon to assist textual analysis?*

— *How to develop a hierarchical sentiment calculation method based on Weibo's linguistic features?*

The first question focuses on the identification of innovative uses of Chinese words or absolutely newborn vocabulary, which is a general phenomenon in Weibo. To recognize these new words in a large-scale text set we introduce a unsupervised learning method using three statistical parameters, namely: occurrence number, inside coupling and information entropy of neighboring character set, as wording standard. New words will be recognised when the wording metrics satisfy the threshold and the words are out of common dictionaries. Their sentiment orientations are determined by pointwise mutual information with known basic words. To address the second question, we transform emoticons into emotional words to construct a dictionary and take them into account in sentiment calculation. In addition, we estimate sentiment influence with interaction information in sentiment detection. Based on multi-level structure of Chi-

nese tweet, we propose a rule set based method to solve the third problem. According to the hierarchical structure we define specific rules for each level to analyze and extract the linguistic features. A tweet's sentiment is detected by a bottom-up calculation with the result of top-down analysis. This structural method takes into consideration the sentiment factors in variety of granularity. Furthermore, we develop and apply visualization techniques to visualise the sentiment analysis results of a large micro-blog dataset, which can discover and depict visually the relevance between sentiment of social network and people's life. Analysis results have shown two temporal patterns and some spatial discrepancy of Beijing users. Sentiment fluctuation is not completely regular, which may be impacted by some events. To further investigate this, we visualize hot words to connect sentiment with events. We employ a Bayesian average method to mine the hot words, and from the visualization of word cloud we find that some negative events surely pull the sentiment down at a same time point, which proves that there is a high correlation between social network sentiment and hot events.

The main contributions of our work are summarized as follows.

— We identify and extract the crucial elements related to sentiments in Chinese micro-blog, e.g., new sentiment words and emoticons. We develop a new sentiment mining method using three wording metrics and point-wise information, and we build a systematic sentiment lexicon for microblogging network.

— We propose a hierarchical rule-based model for analyzing sentiment features of different linguistic components, and the corresponding methodology for calculating sentiment on multi-granularity. Initial results of experiments on real-life micro-blog dataset show that our method is more effective with higher average accuracy than two existing methods for Chinese micro-blog sentiment analysis.

— We exploit visualisation techniques to results of sentiment detection, which helps to reveal and understand the temporal and spatial patterns of online sentiments as well as the influence of events on the online emotions.

The remainder of this paper is organized as follows. Section 2 presents the related work. In Section 3, we firstly introduce the sentiment lexicon we construct, then describe the new sentiment words mining method. Following this, we present the multi-level rule set method of sentiment detection. Our visualization work is introduced in Section 4. Experiments and evaluations are described in Section 5. We conclude the paper and discuss the future work in Section 6.

2. RELATED WORK

In terms of the limited length, micro-blog is most similar to sentence-level and phrase-level text. The general methods for this level text analysis such as [Kim and Hovy 2004][Wilson et al. 2005] provide us with some insights for tweet analysis. With the huge increase of online text, specialized research on opinion analysis has attracted growing attention and generated a plethora of literature in this domain. Existing work includes the mining and modelling of users' opinions based on online reviews [Dave et al. 2003][Titov and McDonald 2008], sentiment analysis on blogs [Melville et al. 2009] and sentiment extraction for cross-domain web texts [Su et al. 2008][Pan et al. 2010]. Different from online reviews which have given topics, and blogs which have adequate content, micro-blog tweet has its unique problems, i.e., topic dispersion, as well as content sparsity and incompleteness, which are considered in our work.

Research on automatically labelling and enriching blog data has been undertaken in English micro-blog, like Twitter. [Pak and Paroubek 2010] utilized text emoticons in users' inputs for automatic annotation, and [Barbosa and Feng 2010] collected labeled data from linked Twitter sentiment web sites. Linguistic features and resources

such as sentiment lexicons, which are also the foundation for our work, have been proven useful for Twitter sentiment detection in [Kouloumpis et al. 2011]. Some researchers focus their work on lexicon construction[Turney 2002][Neviarouskaya et al. 2011]. Some researchers try to extract more features with rich information. For example, [Saif et al. 2012] added tweets' semantic concept as an additional feature; [Tan et al. 2011] used the information about social relationships to improve sentiment analysis.

Our work is completely conducted on Chinese micro-blog, such as Chinese Sina Weibo, and the features we consider distinguish from the Twitter research with different means of expression. There is a number of existing work on Chinese micro-blog analysis. For example, [Yuan et al. 2013] applied both lexicon-based and learning-based approaches with limited number of Chinese language features being analyzed. [Wang et al. 2013] focused on Weibo sentences' grammar and proposed a Chinese bag-of-opinions model. [Cui et al. 2013] paid attention into creative words of Weibo. [Zhao et al. 2012] took advantage of emoticons in Chinese Weibo, whereas combination with textual features may be more expected. Nevertheless, most of these existing works is conducted based on one or a couple of general features while neglecting others. Also, we mainly focus on rule based method, and there are numbers of criteria in order to generate rules, the training phase construct all the rules depending on these criteria. The most two common criteria are support and confidence [Ma 1998][Medhat et al. 2014]. Rather than using single level features or unstructured models, we propose a multi-level model containing features from words to sentence, new emoticon elements, as well as social attention factors in our work.

In the research area of micro-blog information visualization, Nokia Internet Pulse [Kaye et al. 2012] visualizes current discussions around a particular topic, and TwitInfo [Marcus et al. 2011] has the ability to summarize and visualize opinions of events. There is little work on Chinese Weibo data visualisation except the work [Jiayu et al. 2013] which proposed a visualization prototype on traffic theme. Our work develops techniques to provide an overall visualization of Chines Weibo sentiment considering temporal and spatial factors as well as social events.

3. SENTIMENT LEXICON

3.1. Components of Lexicon

Sentiment resources, especially sentiment lexicons play a critical role in sentiment analysis. There have been a number of work to construct English sentiment lexicons [Mohammad et al. 2013][Turney 2002]. Nevertheless, existing Chinese lexicons are insufficient and cluttered for micro-blog, whose vocabulary is diverse. Optimization and extension of sentiment lexicon are essential to adapt linguistic circumstance of Chinese micro-blog. We concentrate on the mining and sentiment recognition of the new words which make micro-blog apparently different from traditional text. Based on the current resources we build up a lexicon which consists of four components:

Basic Word Dictionary: We create a basic dictionary by integrating How-Net dictionary [Dong 2000], a canonical Chinese sentiment resource, and NTUSD[Ku and Chen 2007], an open sentiment diction from National Taiwan University. Both of them classify words into negative and positive categories. We further optimize the diction by eliminating obscure words, selecting those words with relatively obvious orientation and extending the scale with some common colloquial phrases.

Emoticon Dictionary: As a unique element of micro-blog, emoticons are frequently used to express emotions, and their sentiment properties are almost certain. In Chinese micro-blog emoticons can be converted into character codes, for example, in Sina an emoticon is transcoded to a pair of square brackets with a corresponding word

Table I: Emoticon Sentiment Lexicon

Level	Score	Word Count	Example
Strong Positive	2	12	
Positive	1	29	
Negative	-1	24	
Strong Negative	-2	15	

inside. With this transcode, we define the polarity of an emoticon and the level in that polarity according to its actual use. Based on their expressiveness and trans-word meaning, emoticons are classified into five levels. We assign each level with a sentiment score at the aggregate {-2, -1, 1, 2} according to the sentiment intensity criterion in our work, which will be introduced in Section 4.2.1. Whether being greater than zero or not represents a positive or negative sentiment orientation of the emoticon, and the value ± 2 expresses a stronger feeling than ± 1. This dictionary consists of 81 common-use emoticons.

Modified Word Dictionary: Modified words like privative and degree adverbs have a latent impact on the sentences' emotion. Privative might result in a polarity-reversal of the tweet's sentiment, while adverbs may strengthen or weaken emotional attitudes to some certain degrees. We refer to the How-Net's research about intensifiers of Chinese, which contains 219 degree adverbs and 19 denial words. According to its classification for adverbs, we integrate the levels with similar intensities and derive 3 intensity levels for the adverbs, i.e., "a bit", "very" and "extremely".

New Word Dictionary: The emergence of innovative words is a common phenomenon in Chinese micro-blog. Once widely accepted, these words would be adopted widely to express users' emotion on other things. Hence mining new words and identifying their sentiment orientations are key to emotional analysis of Chinese micro-blog. The next section will introduce our method for mining new sentiment words.

3.2. Expansion of Sentiment Words

3.2.1. New Word Mining. In Chinese micro-blog new words are used with the same flexibility as any normal words. In other words, new vocabulary has the same wording characteristic with normal words. Nevertheless, new words do not exist in any dictions in hand. If we can obtain all the words in micro-blogs, new words would be easily extracted by comparing with conventional dictionaries. The problem therefore could be transformed to how to mine words with wording characteristics. Unlike English words in Twitter, which are easily detected since they are separated by space, Chinese words need some wording features to be extracted from a long string without any separators. Based on this characteristic, we define three statistical metrics to describe the wording features and mine all possible words with a unsupervised method.

Occurrence Frequency: If a string recurs more than certain times, it's probably a word. And occurrence number could depict the repetitiveness of words. We let $F(w)$ denote the occurrence frequency of the string w in a preset text field of Chinese tweets.

Inside Coupling: Despite a high occurrence number, it's difficult to judge if the string is a word or not, it might be a phrase or an insignificant combination of some words. We define inside coupling to measure internal compactness of a string, in order to discern words from known phrases or meaningless combinations. Due to limited length of micro-blog, it is observed that most combinations are made up of two words. To calculate inside coupling of the string w, we divide w into two substrings in every possible cases, and get an aggregate: $\left\{\left(w_1{}', w_1{}''\right), ..., \left(w_i{}', w_i{}''\right), ..., \left(w_n{}', w_n{}''\right)\right\}$ where each $\left(w_i{}', w_i{}''\right)$ denotes a possible two-parts combination of w, and there are totally n

kinds of combinations. Let $IC(w)$ denote the inside coupling, and defined as follows:

$$IC(w) = \frac{1}{n}\sum_{i=1}^{n}\frac{P(w)}{P(w_i') \times P(w_i'')} \tag{1}$$

where $P(w)$, $P(w_i')$, $P(w_i'')$ represent occurrence probability of w, w_i', w_i'' in a text field, and they are calculated as $P(w) = N(w)/N$, in which N is the sum of all strings' occurrence number in the field. If $IC(w)$ has a relatively high value that means w is hard to be divided into shorter strings as a word.

Information Entropy of Neighboring Character Set: Not only internal wording of a string, but also external wording attribute need to be measured. Neighboring character set $\{c_1, c_2, ..., c_n\}$ of w, reflecting its external attribute, denotes all single characters appearing to the left or right adjacent to w in a tweet field. Information entropy can measure the uncertainty of one thing, the more uncertain, the more information and the higher entropy. If a candidate string with strong inside coupling is not a word, it may need to combine its neighbour character to become a word. And if its neighbouring character set has a great certainty, this string added with this character is more likely a word. Thereby w's left(right) information entropy of neighboring character set is defined to quantize its external flexibility:

$$IE_{left(right)}(w) = -\sum_{\mathrm{i}=1}^{\mathrm{n}}\frac{n_i}{n}\log_2\frac{n_i}{n} \tag{2}$$

Where n_i is the occurrence number of character c_i as a left(right) neighbor of w, while n is the sum occurrence number of all w's left(right) neighbors. We select the smaller set $\min\{IE_{left}(w), IE_{right}(w)\}$ as a wording standard.

Using the above three parameters, we can extract words from a huge micro-blog field and further recognise new words to build a general new word diction. We define a function $\delta(w, d)$ to decide whether w is a new word:

$$\delta(w, d) = \prod_{i=1}^{3}\delta_i(w, \sigma_i)\delta_d(w) \tag{3}$$

$$\delta_i(w, \sigma_i) = \begin{cases} 1 & f_i(w) \geq \sigma_i \\ 0 & f_i(w) < \sigma_i \end{cases} \quad \delta_d(w) = \begin{cases} 1 & w \notin d \\ 0 & w \in d \end{cases} \tag{4}$$

where $f_i\,(i = 1, 2, 3)$ denotes the three functions $N(w)$, $IC(w)$, $IE(w)$ respectively, $\sigma_i\,(i = 1, 2, 3)$ denotes the wording threshold of these three parameters, and d means a given diction. If $\delta(w, d)$ gets a value 1, w is a new word to d. We regard a text field of millions tweets as a long string, whose length is L, and set l as a length limitation of candidate words. Our new words mining algorithm can be described as Algorithm 1.

3.2.2. Sentiment Recognition for New Word. From the above method, new words are extracted without sentiment labelled. The following presents the method of new words' sentiment recognition. The idea of our method is to observe the similarity between new words and sentiment-known words. Point-wise Mutual Information(PMI)[Kaji and Kitsuregawa 2007] has been used to calculate semantic similarity. For two words w_1 and w_2, their PMI is calculated as follows:

$$PMI(w_1, w_2) = \log_2\frac{p(w_1, w_2)}{p(w_1) \times p(w_2)} \tag{5}$$

where $p(w_1, w_2)$ denotes co-occurrence probability of w_1 and w_2. $p(w_1)$ and $p(w_2)$ denote occurrence probability of w_1 and w_2 respectively.

ALGORITHM 1: New Word Mining

Input: The text field of micro-blog data T; New words length limitation l; A general diction d;
Output: List of new words, $List_{nw}$;
initial T to a long string, calculate its length L;
for $i = 1$ *to* L **do**
  for $j = i + 1$ *to* $i + l + 1$ **do**
    Extract suffix-string w_{ij} from $Postion[i]$ to $Postion[j]$, add w_{ij} to candidate list $List_{cw}$;
  end
end
sort $List_{cw}$ by alphabet order;
for $k = 1$ *to* $List_{cw}.length$ **do**
  if *(*$\delta(w_k, d) == 1$*)* **then**
    Add w_k to $List_{nw}$;
  end
  if *(*$w_k \in BasicDiction \cup EmoticonDiction$*)* **then**
    Add w_k to $List_{bsw}$;
  end
end
sort $List_{bsw}$ by $N(w)$;
return $List_{nw}$;

If w_1 is a new word, w_2 is a basic word with a sentiment label(positive or negative), similarity calculated by PMI equation will suggest whether w_1 has same sentiment orientation with w_2. Note that in Algorithm.1 we collect sentiment-known words(including basic words and emoticons) from the current corpus and sort them by occurrence number. Selecting top 30 basic positive and negative words as well as top 5 emoticon words in $List_{bsw}$ we establish a positive word baseline set W_P and a negative one W_N. With these two sets, which are corpus adapted, we revise the PMI equation to estimate a new word's sentiment orientation

$$Sen\,(w) = \frac{1}{M}\sum_{i=1}^{M}\left(PMI\,(w, w_{pi}) - PMI\,(w, w_{ni})\right) \tag{6}$$

where w_{pi} and w_{ni} are elements of W_P and W_N, w is a new word. If $Sen\,(w) > 0$, w has a positive orientation, while if $Sen\,(w) < 0$, w is a negative word.

4. SENTIMENT DETECTION

4.1. Hierarchical Structure of Chinese Tweets

A Chinese tweet has a multi-level structure, and each level has its own linguistic features which have high correlation with sentiment. Therefore, we intend to formalise this structure to extract different language features that affect emotional expressions at each level. At top level, tweets can be segmented to a few composite sentences (Sentence Level in Fig.1) by Chinese terminative punctuation marks, e.g., "。", "！" and "？" and further divided to several sub-sentences (Sub-sentence Level in Fig.1) by other punctuation marks, e.g., ",", "；" or "、" . We number them in sequence: s_i denotes the ith sub-sentence of the tweet. A composite sentence may contain a few sub-sentences, so we let $S^{(ij)} =< \{s_i, s_{i+1}, ..., s_j\}, \mathbf{P} >$ denote a composite sentence, where $\{s_i, s_{i+1}, ..., s_j\}$ is set of sub-sentences in $S^{(ij)}$ and $\mathbf{P}$ is a set of the terminative punctuation. If there are m composite sentences, the tweet text can be denoted as $\{S_1, S_2, ..., S_m\}$. At the phrase level we focus on the internal structure of sub-sentences. To extract a fine-granularity structure, we conduct syntactic analysis to find out frequent patterns of syntactic path [Zhao et al. 2011] and dependency rela-

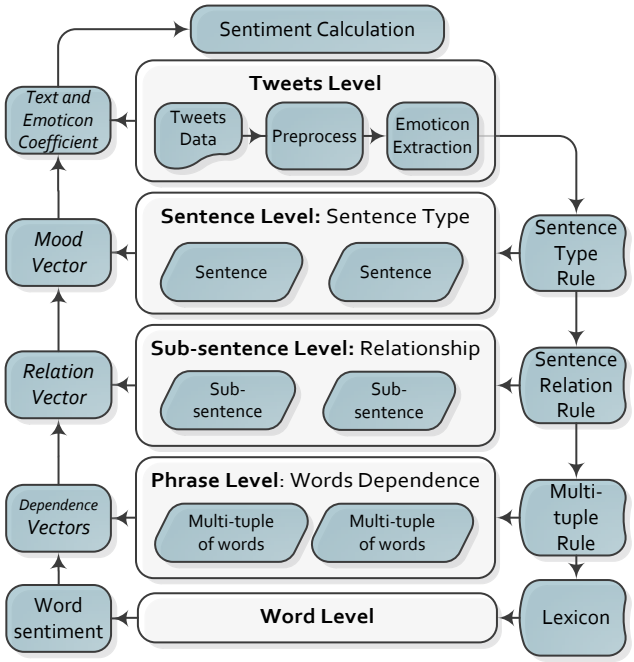


Fig. 1: Sentiment Calculation Procedure

tionships of words in Chinese tweets. Based on frequent patterns, we construct a word multi-tuple to indicate the sentiment expression group of a sub-sentence. We define $MT_i =< \{obj\}, (\{v\}, \{a\}), (\{adv\}, \{neg\}) >$ (Phrase Level in Fig.1) as the words multi-tuple of sub-sentence s_i. $\{obj\}$ denotes a collection of sentiment expression objects. $(\{v\}, \{a\})$ denotes the basic sets of sentiment words, i.e. emotional verbs v and adjectives a, because most Chinese sentiment words belong to these two part-of-speech patterns. $(\{adv\}, \{neg\})$ denotes the modified unit of sentiment words, where $\{adv\}$ represents a collection of degree modifier and $\{neg\}$ represents a collection of denial relations.

4.2. Sentiment Calculation based on Multi-level Rule Set

Previous studies paid little attention to this structure. They only considered features in some specific levels, especially word level, or sometimes integrate features from different levels in parallel, which ignores relationships among these levels. To bridge this knowledge gap, we propose a sentiment calculation method based on multi-level rule set as shown in Figure 1, in which the structural characteristics are kept and considered. Features of each level are mapped to a parameter space. This method has two main procedures: a top-down feature analysis and a bottom-up sentiment calculation.

Top-down feature analysis: This procedure is executed from top level, in which tweet data is firstly preprocessed with noises' filtration and segmentation. Noise means useless strings to tweet's sentiment, such as URL code or automatically added words by system. Segmentation includes partitioning sentences to sub-sentences and sub-sentences to words. Following this, emoticons are extracted by matching their trans-code's regular expressions. Structures in different granularities are analyzed in the corresponding level based on their own feature rules, namely: the type of a complete sentence at sentence level, the relationship among sub-sentences at sub-sentence level and dependence of words within sub-sentences at phrase level. This rule-based

process outputs parameter vectors, which would be used in the bottom-up procedure. And at the bottom, lexicon acts as a large scale rule set for sentiment related words.

Bottom-up sentiment calculation: This procedure starts from the word level, where lexicon gives a basic value of sentiment words and modified words. Every subsequent step upwards receives a vector reflecting feature of the current level. Level by level, the calculation will finally climbs back to the tweet level considering all elements in different granularities. The detail of sentiment calculation will be presented later.

4.2.1. The Criterion of Sentiment Intensity. The rules defined in the above method would map the sentiment features of each level to a parameter space. The values of the parameters in this method reflect influences of features on tweet's sentiment polarity or intensity. The sentiment polarity can be classified into positive and negative. In terms of sentiment intensity, we establish 5 levels criterion, i.e., "no emotion", "a bit", "normal", "very" and "extremely" according to the adverb classification in How-Net knowledge database [Dong 2000]. Generally, positive words are labelled with 1 and negative words with -1 in sentiment lexicon, which represents the value of "normal". And if there is no emotional expressions in tweet, we set this message as "no emotion" with the value 0. Between "normal" and "no emotion", there is the level "a bit", so we choose mid-value 0.5 between the scores of these two levels. Accordingly, we set a 0.5 dissimilarity value among each level of intensity and form the corresponding sentiment score set$\{0, \pm 0.5, \pm 1, \pm 1.5, \pm 2\}$. The rules generate parameters to change the sentiment scores in continuous value interval [-2,2]. We set the initial value of parameter as 1 to keep the original sentiment of the corresponding expression unit. The negative value is used to denote the opposite sentiments. The sentiment calculation method is as follows.

4.2.2. Multi-level Rules Set Method. As shown in Figure 1, we define the three sets of rules, i.e., Sentence Type Rules, Sentence Relation Rules and Multi-tuple Rules, at the three levels of the language structure. The following elaborates the rules at the three levels with the corresponding feature analysis procedure.

Composite Sentence Level

At the top level, we regard Chinese terminative punctuations, e.g. "。", "！" and "？", as the symbol of complete sentence or composite sentence containing a few subsentences. The punctuation marks of a composite sentence denoting different sentence type bring different changes in sentiment intensity. We let α_i denote sentence type factor of sentence S_i, and derive a vector $\vec{\alpha} = (\alpha_1, \alpha_2, ..., \alpha_m)$ for tweet T, where α_i has an initial value 1. If the sentiment score without considering sentence type of the composite sentence S_i is $Sen(S_i)$, sentence type rules mainly consider changes on $Sen(S_i)$ brought by different types of sentences. Generally, we can judge sentence type by using terminative punctuations. But sometimes users do not type in punctuation when posting so we need to consider the symbol words which can indicate the sentence type. For example, an exclamatory sentence usually has an interjection at the end, e.g., "Aaa" or "Wa", and a rhetorical question sentence has symbol words, e.g., "Nandao" or "Qibushi". So we collect symbol words often used by Weibo users to effectively identify these two kinds of sentences. Considering above conditions, the sentence type rules are defined as follows:

For single punctuation

— ***Rule 1.1***: In the case that p of S_i is "！", if $|Sen(S_i)| < 1.5$, then $\alpha_i = \frac{|Sen(S_i)|+0.5}{|Sen(S_i)|}$; and if p of $|Sen(S_i)| \geq 1.5$, then $\alpha_i = \frac{2}{|Sen(S_i)|}$, denoting a sentiment strengthening by exclamatory sentence;

— ***Rule 1.2***: If p of S_i is "？ " and there are no symbol words of rhetorical question, then let $\alpha_i = 0$, denoting sentiment hiding by interrogative sentence type;

— ***Rule 1.3***: If p of S_i is "？ " and there are symbol words of rhetorical question, then let $\alpha_i = -1$, denoting sentiment reversal;

For some punctuation combinations

— ***Rule 1.4***: In the case that p of S_i is "！ ？ " or a string of repeated "？ " and there are symbol words of rhetorical question, if $|Sen(S_i)| < 1.5$ then let $\alpha_i = -\frac{|Sen(S_i)|+0.5}{|Sen(S_i)|}$; and if p of $|Sen(S_i)| \geq 1.5$, then $\alpha_i = -\frac{2}{|Sen(S_i)|}$, denoting a strong reversal;

— ***Rule 1.5***: If the case that p of S_i is a string of repeated "！ " then $\alpha_i = \frac{2}{|Sen(S_i)|}$, denoting a strong strengthening;

For no punctuation

— ***Rule 1.6***: In the case that there are interjections in S_i, if $|Sen(S_i)| < 1.5$, then $\alpha_i = \frac{|Sen(S_i)|+0.5}{|Sen(S_i)|}$; and if p of $|Sen(S_i)| \geq 1.5$, then $\alpha_i = \frac{2}{|Sen(S_i)|}$, denoting a sentiment strengthening by exclamatory sentence;

— ***Rule 1.7***: If there are symbol words of rhetorical question, then let $\alpha_i = -1$, denoting sentiment reversal;

Sub-Sentence Level

In sub-sentence level, relationships among short sentences become an important feature, and we summarize common relationships in Chinese micro-blogs. These relations result in reciprocal sentiment impact among sub-sentences. We let β_k denote relation factor of sub-sentence s_k in $S^{(ij)}$, and its initial value is 1. We can derive a relation factor vector $\vec{\beta} = (\beta_i, \beta_{i+1}, ..., \beta_j)$ for each composite sentence $S^{(ij)}$ in a tweet. The rules would give a sentence relations' vector $\vec{\beta}_T = \left(\vec{\beta}_1, \vec{\beta}_2, ..., \vec{\beta}_m\right)$ of a tweet with m composite sentences. From calculation on bottom level, we can derive the sentiment score $Sen(MT_k)$ of the multi-tuple in sub-sentence s_k, which means the internal sentiment of s_k. Sentence Relation rules mainly analyze external sentiment changes on $Sen(MT_k)$ brought by different relations among sub-sentences when considering the overall sentiment of a composite sentence.

Transition relation always brings a reversal to sentiment, which are always considered. The sub-sentences of transition have a predominance in sentiment expression. These two relations have a kind of symbol words called adversative conjunction. In modern Chinese, adversative conjunctions can be divided into two categories: "Sui Ran" (means "although") category introducing non-transition part of the sentence and "Dan Shi" (means "but") category introducing transition part. And in Chinese sentence, a complete transition relation generally has a pair of them. It is a commonplace in Chinese micro-blogs users always prefer omitting either of them. Based on this special style, we give the following definitions:

— ***Rule 2.1.1***: If there is a single conjunction belonging to "Dan Shi" category in s_k of $S^{(ij)}$, set $\beta_i = 0, \beta_{i+1} = 0, ..., \beta_{k-1} = 0$ and $\beta_k = 1, \beta_{k+1} = 1, ..., \beta_j = 1$;

— ***Rule 2.1.2***: If there is a complete pair of adversative conjunctions in $S^{(ij)}$ and the conjunction belonging to "Dan Shi" category appears in s_k, set $\beta_i = 0, \beta_{i+1} = 0, ..., \beta_{k-1} = 0$ and $\beta_k = 1, \beta_{k+1} = 1, ..., \beta_j = 1$;

— ***Rule 2.1.3***: If there is a single conjunction belonging to "Sui Ran" category in s_k $(k \neq i)$ of $S^{(ij)}$, set $\beta_i = 1, \beta_{i+1} = 1, ..., \beta_{k-1} = 1$ and $\beta_k = 0, \beta_{k+1} = 0, ..., \beta_j = 0$;

— ***Rule 2.1.4***: If there is a single conjunction belonging to "Sui Ran" category in s_i of $S^{(ij)}$, set $\beta_i = 0$ and $\beta_{i+1} = 1, ..., \beta_j = 1$;

Hypothesis relation in Chinese is similar to the adverbial clause of condition in English. For instance, the symbol words "Ru Guo" and "Chu Fei", who belongs to normal hypothesis relation category, are equivalent to English words "if" and "unless". Besides, there is always a corresponding set of auxiliary words for symbol words of this category, e.g., "Na" or "Jiu", which is a sign introducing the main sentence after hypothesis part. Like adverbial clause of condition, the assumed condition in sentence is foundation of the whole sentence's emotion. Therefore, we highlight the contribution of hypothesis for part sub-sentences' sentiment determination. Besides, there is another category – negative assumption, of symbol words(e.g., "Ruo Guo Bu") in hypothesis relation. Also this category also has a set of auxiliary words, e.g., "FouZe". We cope with it as an antonymy to real emotion. The rules about hypothesis relation are defined as:

— ***Rule 2.2.1***: If $S^{(ij)}$ contains a symbol word belonging to normal hypothesis relation category and auxiliary words appear in s_k, set $\beta_i = 1$, $\beta_{i+1} = 1$,..., $\beta_{k-1} = 1$, and $\beta_n = \frac{|Sen(MT_n)|-0.5}{|Sen(MT_n)|}$ (n=$k, k+1, ..., j$) when $|Sen(MT_n)| > 0.5$ or $\beta_n = 0$ (n=$k, k+1, ..., j$) when $|Sen(MT_n)| \leq 0.5$;

— ***Rule 2.2.2***: If $S^{(ij)}$ contains a symbol word belonging to negative hypothesis category and auxiliary word appears in s_k, set $\beta_i = -1$, $\beta_{i+1} = -1$,..., $\beta_{k-1} = -1$, and $\beta_n = -\frac{|Sen(MT_n)|-0.5}{|Sen(MT_n)|}$ (n=$k, k+1, ..., j$) when $|Sen(MT_n)| > 0.5$ or $\beta_n = 0$ (n=$k, k+1, ..., j$) when $|Sen(MT_n)| \leq 0.5$;

Cause-effect relation may have little influence in sentiment, so we just keep the initial value when it comes across this relation(***Rule 2.3***).

Progressive relation is used to express gradually strengthening emotions. There is a few symbol words indicating this relation in Chinese micro-blogs. In most cases, progressive relation is only related to the prior sub-sentence. So the rule is described like this:

— ***Rule 2.4***: If $S^{(ij)}$ contains progressive relation and symbol word is in s_k, set $\beta_k = \frac{|Sen(MT_{k-1})|+0.5}{|Sen(MT_k)|}$ when $|Sen(MT_{k-1})| < 1.5$ or set $\beta_k = \frac{2}{|Sen(MT_k)|}$ when $|Sen(MT_{k-1})| \geq 1.5$

Phrase Level

While the top level rules analyze the external feature of sub-sentences, we focus on its internal structure at the phrase level. As mentioned above, we construct a word multi-tuple $MT_i =< \{obj\}, (\{v\}, \{a\}), (\{adv\}, \{neg\}) >$ to indicate the sentiment expression group of a sub-sentence, and all elements in the tuple have one of frequent dependency relationships as shown in Table II. Figure 2 shows the multi-tuple model, which embodies its structure and reciprocal dependencies among elements. The multi-tuple rules are defined to focus on the contributions of these dependencies to the internal sentiment of a sub-sentence, which have expanded the rules in the work [Wu et al. 2009].

Degree modifier rules are defined for the component $\{a\}$, since $\{adv\}$ only has dependency "*advmod*" with $\{a\}$. We set a degree vector $\vec{\gamma}$ for all adjectives in $\{a\}$, where element γ_i is the modify degree of a_i. The degree d_j of adv_j is searched from Modified Word Diction and the rules are:

— ***Rule 3.1.1***: If a_i in $\{a\}$ has a "*advmod*" with adv_j, γ_i=d_j;

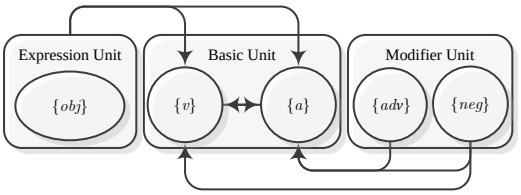


Fig. 2: Multi-tuple Model

Table II: Words Dependencies

Tag	Dependency	Involved Tuple	Example
nsubj	nominal subject	$\{a\}\{v\}\{obj\}$	我打扫 I swap
dobj	direct object	$\{v\}\{obj\}$	打扫房间 swap room
amod	adjectival modifier	$\{a\}\{obj\}$	房间脏乱 dirty room
advmod	adverbial modifier	$\{a\}\{adj\}$	非常辛苦 very hard
comod	coordination	$\{a\}\{adj\}$	辛苦而费尽 hard and tired
neg	negation modifier	$\{a\}\{v\}\{neg\}$	不高兴 not happy

— ***Rule 3.1.2***: If a_i in $\{a\}$ has a "*advmod*" with adv_j or adv_k and adv_j has a "*comod*" with adv_k, $\gamma_i = \max\{d_j, d_k\}$;

— ***Rule 3.1.3***: If a_i in $\{a\}$ has no "*advmod*" with elements in $\{adv\}$, $\gamma_i = 1$

Negative relation rules are defined for $\{a\}$ and $\{v\}$. Normally, a single negative dependency "*neg*" changes sentiment words' polarity, whereas a number of "*neg*" and dependencies between $\{a\}$ and $\{v\}$ would make it complex. The rules are to talk about this intricate situation. We define $\vec{\varphi}_v = (\varphi_{v_1}, \varphi_{v_2}, ..., \varphi_{v_m})$ and $\vec{\varphi}_a = (\varphi_{a_1}, \varphi_{a_2}, ..., \varphi_{a_n})$, where φ_{v_i} and φ_{a_j} are the negative factors of v_i and a_j. We consider up to two negative relations over verbs or adjectives, ***Rules 3.2*** are summarized in Table III.

Table III: Negative relation rule (*Rules 3.2*)

#*neg* of v_i	#*neg* of a_j	Dependency of v_i with a_j?	φ_{v_i}	φ_{a_j}
1	0	No	-1	1
0	1	No	1	-1
1	0	Yes	-1	-1
0	1	Yes	1	-1
1	1	No	-1	-1
1	1	Yes	-1	1
0	2	Yes or No	1	1
2	0	Yes or No	1	1

Object relevance rules are built upon dependencies between sentiment words $\{a\}$ or $\{v\}$ and expression objects $\{obj\}$. If a sentiment word has no dependency with objects, this word is deemed to be irrelevant to emotion that this sub-sentence expresses. We give vectors $\vec{\eta}_v = (\eta_{v_1}, \eta_{v_2}, ..., \eta_{v_m})$ and $\vec{\eta}_a = (\eta_{a_1}, \eta_{a_2}, ..., \eta_{a_n})$ to indicate object relevance of $\{v\}$ and $\{a\}$. The rules are as follows:

— ***Rule 3.3.1***: If $v_i(a_j)$ has dependency with elements in $\{obj\}$, then $\eta_{v_i} = 1(\eta_{a_j} = 1)$;

— ***Rule 3.3.2***: If $v_i(a_j)$ has no dependency with elements in $\{obj\}$, then $\eta_{v_i} = 0(\eta_{a_j} = 0)$

4.2.3. Sentiment Calculation. At the bottom level, sentiment words $\{v\}, \{a\}$, which may be basic or new sentiment words, are fundamental unit of multi-tuple level, and we can obtain their sentiment polarity from Basic Word Dictionary and New Word Dictionary to form vectors $\vec{p}_v$ and $\vec{p}_a$. In terms of the above rules, we have multi-tuple vectors $\vec{\varphi}_v$, $\vec{\eta}_v$ for $\vec{p}_v$ and $\vec{\gamma}$, $\vec{\varphi}_a$, $\vec{\eta}_a$ for $\vec{p}_a$. Subsequently we derive sentiment scores of one multi-tuple through the following equation:

$$Sen\left(MT\right) = \frac{\vec{p}_v \cdot \left(\vec{\varphi}_v \circ \vec{\eta}_v\right)}{n_v} + \frac{\vec{p}_a \cdot \left(\vec{\gamma} \circ \vec{\varphi}_a \circ \vec{\eta}_a\right)}{n_a} \tag{7}$$

where "$\circ$" denotes the Hadamard Product operation:

$$\vec{\varphi}_v \circ \vec{\eta}_v = \left(\varphi_{v_1} \cdot \eta_{v_1}, \varphi_{v_2} \cdot \eta_{v_2}, ..., \varphi_{v_m} \cdot \eta_{v_m}\right) \tag{8}$$

While in sub-sentence level, factor β is a consideration about external relationship among sub-sentences in a same composite sentence. And for composite sentences, there is a sentence type factor α. we can calculate the sentiment of a composite sentence S_i consisting of N_i sub-sentences as:

$$Sen\left(S_i\right) = \frac{1}{N_i} \sum_{j=1}^{N_i} \beta_{ij} \cdot Sen(MT_{ij}) \tag{9}$$

where $\beta_{ij} \cdot Sen\left(MT_{ij}\right)$ represents sentiment score of the jth sub-sentence in the ith composite sentence. By combining the rules from the three levels, we can give a synthetic calculation of a tweet text containing m composite sentences:

$$Sen\left(text\right) = \frac{1}{m} \sum_{i=1}^{m} \frac{\alpha_i}{N_i} \sum_{j=1}^{N_i} \left(\beta_{ij} \cdot Sen\left(MT_{ij}\right)\right) \tag{10}$$

At tweet level, we can obtain the sentiment scores of each emoticons from Emoticon Dictionary and we can calculate the sentiment of all emoticons in one tweet as:

$$Sen\left(emo\right) = \frac{1}{k} \sum_{i=1}^{k} Sen\left(e_i\right) \tag{11}$$

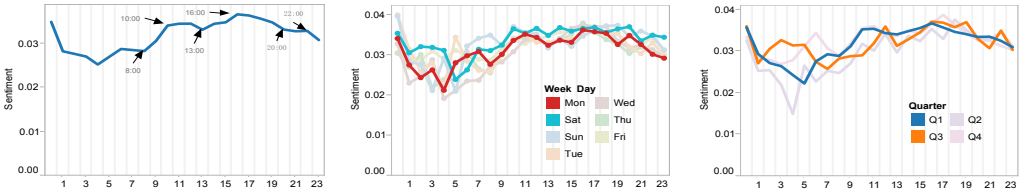
Then we combine emoticons' sentiments and text's sentiments with the following equation:

$$Sen\left(tweet\right) = \frac{N_S \cdot Sen(text)}{N_S + N_E} + \frac{N_E \cdot Sen(emo)}{N_S + N_E} \tag{12}$$

where N_S and N_E are respectively the number of composite sentence and the number of emoticons in this tweet.

5. VISUALIZATION

To efficiently review and understand the sentiment detected from large scale Chinese micro-blogs, we present sentiment visualization based on a Sina Weibo dataset. The data set was collected from Jan. 2011 to Feb. 2012 including 15 million original tweets of around 30,000 active users in Beijing. We used the data of Beijing users for its denseness and completeness. We sample randomly during data collection to guarantee density distribution on time and space of the dataset conforming to real usage fact. Our visualization is conducted to show the sentiment distribution in three aspects: Temporal Patterns, Spatial Characteristics and Sentiment & Hot Events.



(a) Global diurnal pattern (b) Diurnal pattern in a week (c) Diurnal pattern of seasons

Fig. 3: Diurnal Pattern

5.1. Sentiment of Tweets' Set

When visualizing the sentiments, we need to evaluate the sentiments of tweets' sets. Because viewpoints can be diffused in micro-blogs network through users interactions which lead to the transmission of sentiment, we should consider the influence of each tweet when evaluating the sentiment of tweets' collections. The intensity of interactions, such as the numbers of retweeting or commenting, can reflect a tweet's sentiment influence. The more a tweet is retweeted or commented, the wider its sentiment spreads, and the more significant its influence is. Therefore, we define attention degree of $Tweet_i$: rc_i, the sum of retweet number and comment number. Given a large-scale tweet field T we can derive the average sentiment value of T by the equation:

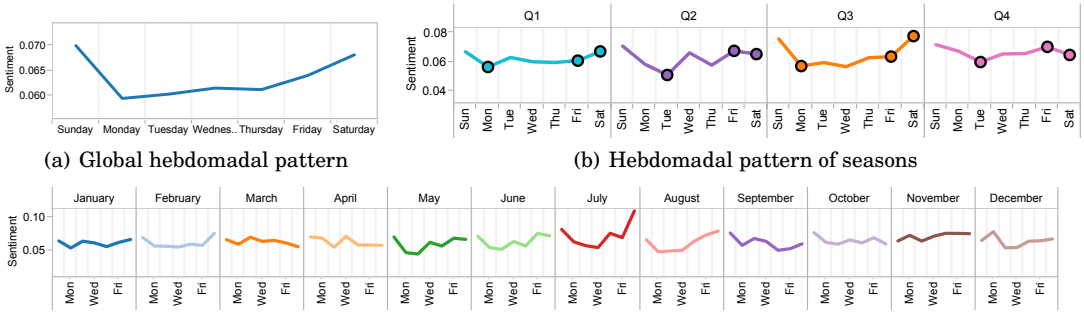
$$Sen\,(T) = \frac{1}{M}\sum_{i}^{M}\frac{(rc_i - \min\{rc_t|t \in T\}) \cdot Sen\,(Tweet_i)}{\max\{rc_t|t \in T\} - \min\{rc_t|t \in T\}} \tag{13}$$

5.2. Temporal Patterns

We present visualization of two types of temporal patterns: diurnal pattern and hebdomadal pattern. We visualize the two patterns on multiple time density and explore temporal variations of people's sentiment.

Diurnal patterns are shown in Figure 3. Figure 3(a) suggests that people would have first positive peak at 0:00 before sleep. After a low value from 1:00 to 5:00, which may be caused by unexpected insomnia or overtime work, people's sentiment rises at 6:00. From 7:00 to 8;00, people usually have a poor mental state after wake-up, as such sentiment scores decline. Then, it keeps rising until 10:00 and stays steady at 11:00-12:00. At 13:00 it fluctuates a little bit down. It ascends again at 14:00, and reaches the second peak of the day at 16:00. After that point it descends slowly, and stabilizes from 20:00 to 22:00. To probe into this pattern, we explore weekly and seasonal variations of this pattern as shown in Figure 3(b) and Figure 3(c). Both the holistic trends on these two time divisions are in line with the global diurnal patterns with little difference of details. We highlight the Monday pattern and the Saturday pattern in Figure 3(b). It shows clearly that the overall sentiment of Saturday is higher than Monday. Besides, Saturday's sentiment pattern is much higher than Mondays from 1:00 to 3:00 and becomes lower from 5:00 to 7:00, which could be explained that people may prefer to enjoy a night life on weekends and sleep much late. This change also occurs in seasons that quarter 3 has a sentiment peak from 1:00 to 3:00 while quarter 1 has an intrinsic drop in this period, as shown in Figure 3(c). The seasonal change may be caused by temperature and day length, that people fall asleep earlier at quarter 1.

Hebdomadal pattern coincides with our common knowledge that people would have happier emotion on weekends. As shown in Figure 4(a), hebdomadal pattern takes on a "W" or "U" shape, where we set x axis begins from Sunday. It is obtrusive that Monday has the lowest value, and we could call this phenomenon "Black Monday", which is a state switching point of human beings from rest to work. We draw this pat-



(a) Global hebdomadal pattern

(b) Hebdomadal pattern of seasons

(c) Hebdomadal pattern of months

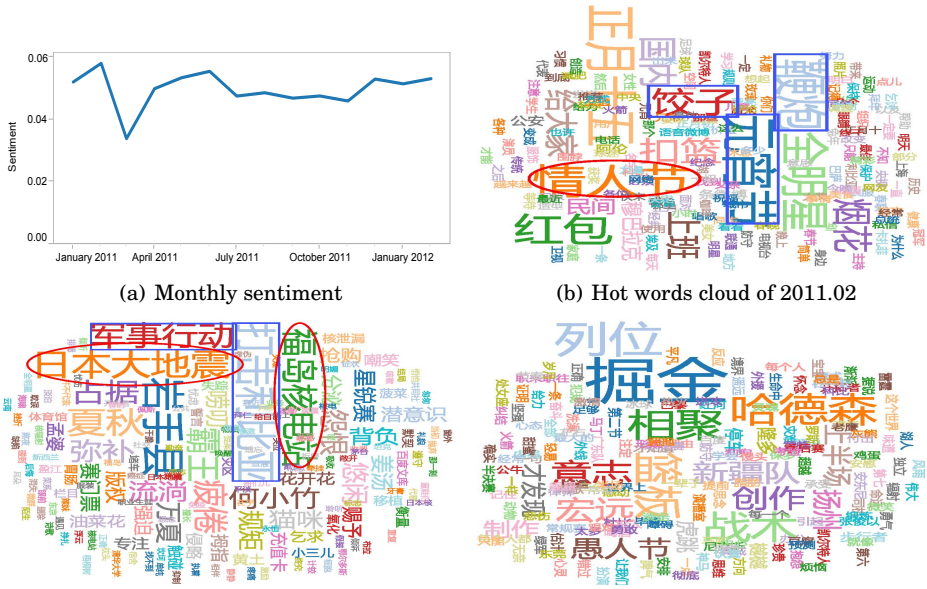
Fig. 4: Hebdomadal Pattern

Fig. 5: Spatial Visualization of Different Districts

tern in seasonal and monthly as well, and in most cases it has a similar shape to the global hebdomadal pattern. Whereas, the pattern has seasonal changes in Figure 4(b) that quarter 2 and quarter 4 have the lowest value on Tuesday, which is different from the pattern of quarter 1 and quarter 3. We make a further step to illustrate monthly changes of this pattern as shown in Figure 4(c).

5.3. Spatial Characteristics

To visualize sentiment in the spatial dimension, we select the top 4 districts in terms of the number of tweets in Beijing. We transform the hebdomadal patterns of these districts to pie charts, which are divided equally into 7 pieces as shown in Figure 5. The seven days of week arrange in clockwise order from Sunday to Saturday. The piece's color means the overall sentiment score of this district on a certain day, the redder the higher the score, and the greener the lower the score. The size of the pie reflects the active degree of user engagement in each district. The pie charts show that Sunday, Saturday and Friday have a redder color, while Monday or Tuesday is the greenest one. The four districts have an obvious discrepancy in sentiment value. Dongcheng district is similar as Haidian in global view while Chaoyang is much higher and Xicheng is the lowest, which is shown by the color of pies. In addition, Haidian has a small fluctuation during a week which is different from the substantial changes in Dongcheng and Xicheng. These discrepancies are possibly decided by main properties of these districts. Xicheng and Dongcheng are mainly residence zones. Chaoyang is a commercial area, and Haidian has many schools. The special function of different districts may lead some special patterns, like high work enthusiasm or calm study atmosphere.



(a) Monthly sentiment

(b) Hot words cloud of 2011.02

(c) Hot words in 2011.03

(d) Hot words in 2011.04

Fig. 6: Sentiment and Hot Words

5.4. Sentiment & Hot Event

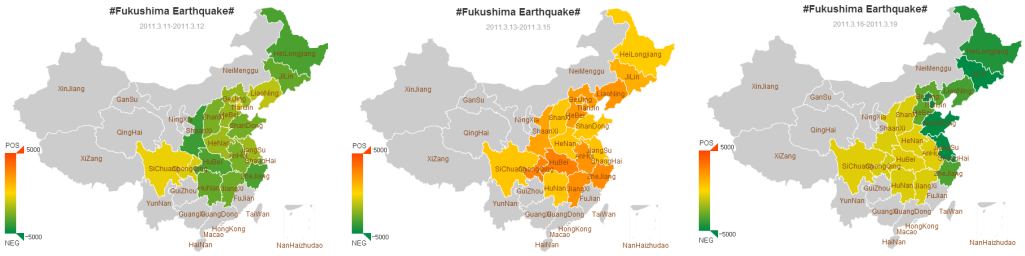
In social network, a big event would bring about an information explosion that users express enormous opinions on this event. The nature of the event may cause different sentiment tendency. A disaster would result in overwhelming negative emotion, while a festival would cause a widespread positive atmosphere. Momentous events can be reflected by hot words. We present visualization on hot words for linking sentiment and events.

Hot words are mined from a set of tweets. We adopt Bayesian average, which is also used in IMDB scoring (www.imdb.com/chart/top), to evaluate the "hot" degree of words in a certain period of time. Given a micro-blog field F divided into m subfields, we evaluate a word w_i's hot rating $R\left(w_i^{(j)}\right)$ in subfields F_j as follows:

$$R\left(w_i^{(j)}\right) = \left(N_i^{(j)} + \overline{N} \cdot \overline{P}^{(j)}\right) / \left(\sum_{k=1}^{m} N_i^{(k)} + \overline{N}\right) \tag{14}$$

where $N_i^{(j)}$ is the occurrence number of w_i in F_j, $\overline{N}$ denotes the mean occurrence number of all words in F, while $\overline{P}^{(j)}$ is the average occurrence probability of all words in F_j. By this way the word with a higher frequency in one subfield and lower in the other ones would get high score.

To investigate the impact of big events on monthly sentiment, we limit time span of hot words to a month. As shown in Figure 6(a), sentiment has the highest score in Feb. 2012 and the lowest in Mar. 2011. Meanwhile, hot words of every month are visualized on word cloud, where the size and the brightness are positively correlated with their hot scores. Figure 6(b) shows the hot words of Feb. 2012, the words with a blue mark are related to the Chinese Spring Festival and Lanterns Festival, and the word in the red circle means the Valentine's Day. As shown in Figure 6(c), the words with a red mark describing the earthquake, nuclear leak of Fukushima in Japan and the ones



(a) 2011.03.11-2011.03.12 (b) 2011.03.13-2011.03.15 (c) 2011.03.16-2011.03.19

Fig. 7: Sentiment Distribution in Fukushima Earthquake Topic

with a blue mark reflecting the war in Libya are conspicuous on the Mar. 2011 word cloud. The other months such as Apr. 2011(Figure 6(d)) contain normal words and their sentiment is near to average line. Hence the visualization of sentiment and hot words shows that the lowest value is possibly derived from people's negative emotions like panic, worry about the disaster and anger, complaint on the war, while the higher month is mainly effected by the gala atmosphere, which proves that there is a higher correlation between social network sentiment and hot events.

We take one step further to explore the temporal and spatial characteristics of sentiment in hot issues. We collect a 9-day data discussing Fukushima earthquake through hashtag, visualize the sentiment distribution in different provinc of China based on the daily data and classify the sentiment distribution into 3 stages as shown in Figure 7. The first two days(Fig.7(a)) people show a overwhelming negative sentiment in which shock and worry are dominant. In the second stage(Fig.7(b)) people express their positive energy like wish and prayer. But in the third stage, sentiment in coastal areas becomes negative again seriously(Fig.7(c)). That is possibly because the message that nuclear leak resulted by this earthquake would contaminate the water and air in some coastal provinces in China began to spread quickly in public and the residents in these areas reveal their panic via Weibo.

From above visualizations, we can detect the crucial event, also track sentiment temporal fluctuation and spatial discrepancy in the event, which is significant to grasp the emotional trend in public and control extreme cases.

6. EXPERIMENTAL RESULTS

6.1. Results of New Words Mining

6.1.1. Threshold Set. We select randomly 400,000 original tweets from our data set collected through Sina API to conduct our new words mining experiment. We first set thresholds of three wording standards introduced in Section 3.1. We calculate the values of N of all sentiment words in this field and take the mean values and minimum values of them as reference thresholds. We examine candidate values between these two values and find that the mean values give a little higher new words rate, however, with more rigorous σ_N the number of new words is much less than that of minimum values. So we set minimum value as the threshold of N: $\sigma_N = 10$. Under this σ_N, we test different threshold combinations of σ_{IE} and σ_{IC} and evaluate their performance by new word rate. As shown in Figure 8 , we can find that when σ_{IE} is less than 0.75, the new word rate always increases when σ_{IC} increases. But when σ_{IE} is set as 0.75, the new word rate starts to stay stable when σ_{IC} is larger than 2.5. On the other hand, like σ_N, the number of new words is monotone decreasing with both σ_{IC} and σ_{IE}. Therefore, in order to extract more new words we need to keep the thresholds as small

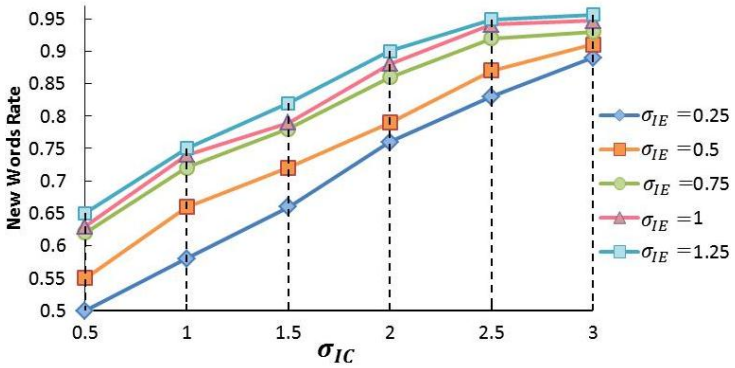


Fig. 8: New word rate with different threshold

Table IV: Mining Result of Top 5 New Words

Rank	New String	*N*	*IC*	*IE*
1	微博(Weibo)	9885	147.94	2.68
2	给力(Awesome)	8729	5.18	3.45
3	卖萌(Cute)	8266	66.65	4.03
4	有木有(Isn't it?)	7890	4.89	3.39
5	请关注(Please follow)	7507	45.27	1.74

Table V: Sentiment Lexicon

Components	#Pos Word	#Neg Word	#Neu Word	# Total
Basic	7376	9636	0	17012
New	97	241	749	1087
Emoticon	41	39	7	87
Whole Lexicon	7514	9916	756	18186

as possible. Both considering new word count and new word rate, we finally set σ_{IE} as 0.75 and σ_{IE} as 2.5.

6.1.2. New Words Results. After mining, a noise reduction is done to filter out some fixed collocations in our knowledge base, and 1865 candidate strings are detected and ranked by their occurrence number. Table IV shows the top 5 candidate strings and their values of three wording parameters. A number of word combinations have a strong "micro-blog" style, for example, No.5 string means "please follow". These fixed strings reflect the language features of tweet and enrich the diversity of new words dictionary.

With this method, we achieve a performance of 90% new words rate and build up a 1087 words' dictionary. Table V shows the details of sentiment lexicon in our work. Here in the new diction, neutral words take a large proportion due to a large amount of emerging new nouns, which are derived from new things or events, and most of their emotions tend to be neutral. Negative emotional words are more than positive ones perhaps because users prefer to use new negative words or satirize negative events.

To examine the performance of this new word diction, we compare it with another new words diction OVD[Cui et al. 2013] in sentiment detection. Figure 11(a) shows that the results of sentiment detection with our new diction (**ND**) are more accurate than the results without using the new diction, and the results by using OVD, which confirms that our new diction performs better.

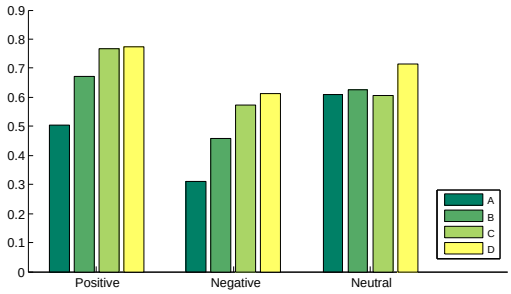


Fig. 9: Rules Evaluation by F-1

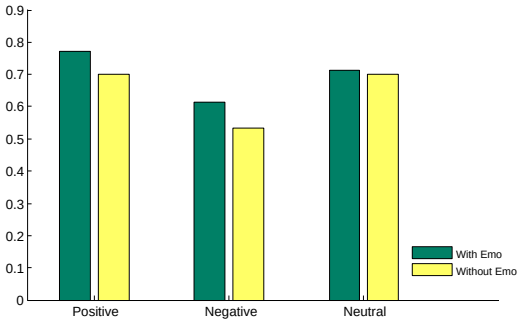


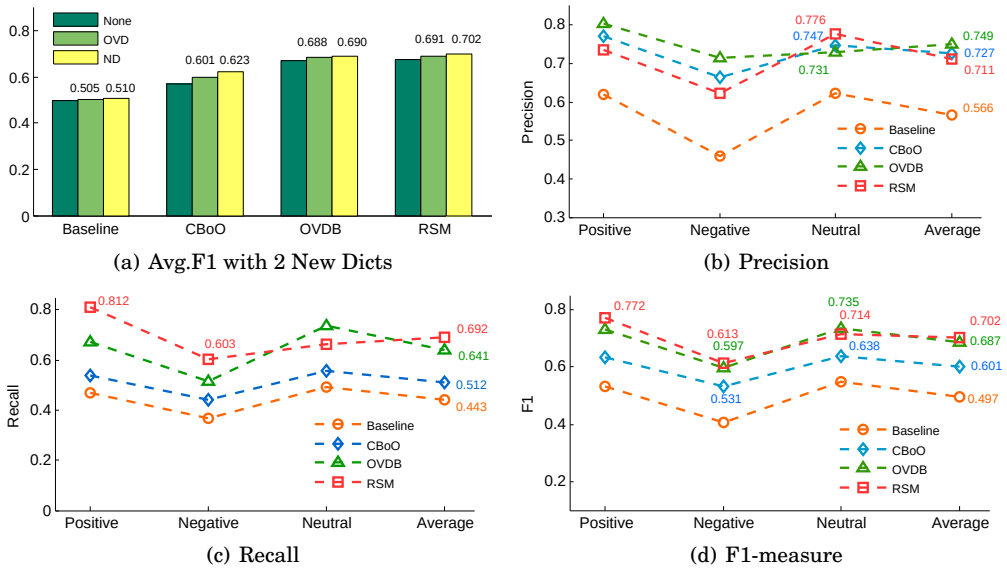
Fig. 10: Emoticons Evaluation by F-1

6.2. Results of Sentiment Detection

6.2.1. Experimental Data. In this experiment, we extract tweets of Xi'an users (registration site is Xi'an). The collected data contains a total of 5,366 users and 87,064 tweets. We find that most tweets (over 70%) are retweeting text. We believe that the original tweets are likely to reflect the sentiment of Xi'an users. We screened 3,163 original tweets describing emotions or recording experiences as experimental data. The tweets that some Weibo machines post automatically, such as advertisements, weather forecast and news reports, etc are filtered out from our data. The reason why we choose Xi'an users is that our three volunteers for manual sentiment annotation are all students of Xi'an colleges, who are familiar with the life of Xi'an and can label the sentiment of tweets accurately. The volunteers are told to make sentiment labels separately, and then discuss together to eliminate the disunity. Finally the selected 3,163 tweets were manually grouped into three sentiment categories: 1300 positive tweets, 1035 negative, and 728 neutral ones.

6.2.2. Evaluation of Rules. To examine the contributions of different rules, we setup an experiment considering different combinations of rules. We exclude rules of a specific level in each combination. We derive 3 combinations, which are A excluding Multi-Tuple Rules, B excluding Sentence Relation Rules and C excluding Sentence Type Rules. D represents our multi-level rule set method considering all rules defined above. We measure the performance of these combination with F-1 scores and the results are shown in Figure 9. The difference between the combination and D indicates the contribution of rules on a specific level. In all three categories, Multi-Tuple Rules prove the greatest contribution because sentiment words are the most crucial part in our method as well as in other methods. However, for neutral tweets the contributions of Sentence Relation Rules and Sentence Type Rules are similar to Multi-Tuple Rules due to the lack of emotional words in these tweets. Additionally, Sentence Type Rules provide less contribution to sentiment detection than Sentence Relation Rules both in positive and negative tweets, but perform better in the neutral tweets. This is mainly because a part of neutral tweets are interrogative sentences.

6.2.3. Contributions of emoticons. To examine the contributions of emoticons on sentiment detection, we setup another experiment considering the influence of emoticons. We consider and ignore emoticons respectively in the comparative experiments. We also utilize F-1 to measure the results. As shown in Figure 10, the importance of emoticons on sentiment detection is apparent since the F-1 scores of our method without using emoticons decline about 7% and 8% on detecting positive and negative tweets respectively. Additionally, emoticons have a smaller influence on neutral tweets for users used no emoticons or those having no emotional meaning.



(a) Avg.F1 with 2 New Dicts (b) Precision

(c) Recall (d) F1-measure

Fig. 11: Results of Sentiment Detection

6.2.4. Comparison Methodology. We compare the performance of our rule set model (**RSM**) with the following methods.

Baseline: We adopt the traditional count of sentimental word based method as the baseline of experiment. The idea of this method is to compare the numbers of positive and negative words to determine the polarity of Chinese tweets. We also revise this method to improve the performance of baseline: setting a sliding window of each sentiment word, privatives and considering degree modifiers as coefficients in sentiment decisions.

CBoO: The Chinese Bag-of-Opinion model in [Wang et al. 2013] based on dependency grammar representing Weibo sentences defines opinion as the minimum independent sentimental unit in a sentence. The bag of opinion is composed of a sentimental indicator, a set of modifiers, and dependency relationships between them. This method calculates sentiment polarity score for every opinion and gets a weighted summation sentiment evaluation for each sentence. To enhance the performance of this method, we combine it with a new words diction OVD introduced next.

OVDB: Out Of Vocabulary Dictionary (OVD) is a new word construction method[Cui et al. 2013]. Out-of-vocabulary words are discovered with context entropy gain and mutual information. A co-occurrence graph is constructed to propagate polarity scores to the words in the lexicon. As the method uses OVD for sentiment classification, we name the method as Out Of Vocabulary Dictionary Based (OVDB). For sentiment analysis, OVDB applied a classic algorithm SVM as a classifier whose features are OVD words in addition to traditional n-gram features.

6.2.5. Results. We first assess the contribution of our new diction by comparing its performance with OVD in sentiment detection. We apply the four methods to three cases, namely: without new diction, adding OVD, and adding our new diction ND. The average F1-measure of 4 methods with different new dictions in Figure 11(a) indicates that new dictions improve the result in all methods. Although our new diction ND gets a similar improvement as OVD in Baseline(0.5% higher) and OVDB(0.2% higher), it gives a 1.2% higher accuracy to CBoO and 1.1% higher to RSM. In general, the new

diction we constructed makes a greater contribution to sentiment detection than other dictions.

Subsequently, we evaluate these methods in different measurements: *Precision*, *Recall* and *F1-measure*. As shown in Figure 11(b), RSM method obtains a little lower precision for positive and negative sentiment, while gets a higher precision 77.6% for neutral category than the other methods. From Figure 11(c), we can observe that RSM has the highest 81.2% recall when detecting positive sentiment that is 34.3% higher than CBoO and 27.6% higher than OVDB, and it gets also 16% and 9% higher recall on negative sentiment. Despite a 7.8% lower recall than OVDB on neutral category, the average recall of RSM attains 18% higher than CBoO and 5.1% higher than OVDB. Although our method obtained a little lower precision than other two methods, it achieved a much higher recall than other methods when detecting positive and negative tweets. Higher recall can help us to detect more positive and negative tweets, in which some important emotional tweets may be detected out. Additionally, considering precision and recall together, we get F1-measure of the 4 methods in Figure 11(d). The F1 value produced by RSM in recognizing positive tweets is 13.9% higher than CBoO also 4.1% higher than OVDB. RSM method also performs better in detecting negative sentiment. When identifying the neutral sentiment, OVDB gets a little bit better performance than RSM. In terms of average F1, RSM method reaches a 70.2% accuracy, which is 10.2% higher than CBoO and 1.5% higher than OVDB.

The overall results demonstrate that our method can detect sentiment successfully and performs better than existing two approaches. However, our method as well as CBoO and OVDB are not good at detecting negative sentiment as Chinese users prefer to using the ironic style to express their negative emotion in Weibo.

7. CONCLUSIONS

In this paper, we have analysed the unique characteristics of Chinese micro-blog text data based on linguistic and social features. Built upon this, we introduced a unsupervised new words mining method to expand the existing lexicon resources. Based on the optimized lexicon, we have developed a rule-set method for sentiment detection which takes into account the additional elements of emoticons and the hierarchical structure of Chinese micro-blog. In addition, we have made use of visualization techniques to review and understand the relationship and patterns of the detected sentiment between online emotion and real life. The approach has been evaluated in a dataset of original micro-blog tweets by comparison with several existing methods based on a number of performance measures. The experiment results have proved that our approach is effective in sentiment detection. Our research has shown that users' context information and other media elements (images, videos) are important attributes which could provide additional determinants for sentiment detection, as such, a possible future work would concentrate on combining basic model with social and media features. Sentiment transmission and a in-depth study on social media emotional influence will be another future work.

REFERENCES

Luciano Barbosa and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, 36–44.

Johan Bollen, Huina Mao, and Alberto Pepe. 2011a. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena.. In *ICWSM*.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011b. Twitter mood predicts the stock market. *Journal of Computational Science* 2, 1 (2011), 1–8.

Anqi Cui, Haochen Zhang, Yiqun Liu, Min Zhang, and Shaoping Ma. 2013. Lexicon-Based Sentiment Analysis on Topical Chinese Microblog Messages. In *Semantic Web and Web Science*. Springer, 333–344.

Kushal Dave, Steve Lawrence, and David M Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*. ACM, 519–528.

Zhendong Dong. 2000. Hownet Knowledge Database. (2000). http://www.keenage.com

Wu Jiayu, Fu Zhiyong, Liu Zhiyuan, Lin Xu, Tang Jiayu, Pan Jiajia, and Zhao Chen. 2013. Creating reflections in public emotion visualization: prototype exploration on traffic theme. In *Proceedings of the 9th ACM Conference on Creativity & Cognition*. ACM, 357–361.

Nobuhiro Kaji and Masaru Kitsuregawa. 2007. Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents.. In *EMNLP-CoNLL*. 1075–1083.

Joseph'Jofish' Kaye, Anita Lillie, Deepak Jagdish, James Walkup, Rita Parada, and Koichi Mori. 2012. Nokia internet pulse: a long term deployment and iteration of a twitter visualization. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*. ACM, 829–844.

Soo-Min Kim and Eduard Hovy. 2004. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 1367.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The Good the Bad and the OMG!. In *ICWSM*.

Lun-Wei Ku and Hsin-Hsi Chen. 2007. Mining opinions from the Web: Beyond relevance retrieval. *Journal of the American Society for Information Science and Technology* 58, 12 (2007), 1838–1850.

Xinjiang Lu, Zhiwen Yu, Bin Guo, and Xingshe Zhou. 2014. Predicting the content dissemination trends by repost behavior modeling in mobile social networks. *Journal of Network and Computer Applications* 42 (2014), 197 – 207. DOI:http://dx.doi.org/10.1016/j.jnca.2014.01.015

Bing Liu Wynne Hsu Yiming Ma. 1998. Integrating classification and association rule mining. In *Proceedings of the fourth international conference on knowledge discovery and data mining*.

Adam Marcus, Michael S Bernstein, Osama Badar, David R Karger, Samuel Madden, and Robert C Miller. 2011. Twitinfo: aggregating and visualizing microblogs for event exploration. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 227–236.

Walaa Medhat, Ahmed Hassan Yousef, and Hoda Korashy Mohamed. 2014. Combined algorithm for data mining using association rules. *arXiv preprint arXiv:1410.1343* (2014).

Prem Melville, Wojciech Gryc, and Richard D Lawrence. 2009. Sentiment analysis of blogs by combining lexical knowledge with text classification. In *Proc. of KDD '2009*. ACM, 1275–1284.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*. Atlanta, Georgia, USA.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2011. SentiFul: A lexicon for sentiment analysis. *Affective Computing, IEEE Transactions on* 2, 1 (2011), 22–36.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. *ICWSM* 11 (2010), 122–129.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining.. In *LREC*.

Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. 2010. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th international conference on World Wide Web*. ACM, 751–760.

Hassan Saif, Yulan He, and Harith Alani. 2012. Semantic sentiment analysis of twitter. In *The Semantic Web–ISWC 2012*. Springer, 508–524.

Qi Su, Xinying Xu, Honglei Guo, Zhili Guo, Xian Wu, Xiaoxun Zhang, Bin Swen, and Zhong Su. 2008. Hidden sentiment association in chinese web opinion mining. In *Proceedings of the 17th international conference on World Wide Web*. ACM, 959–968.

Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li. 2011. User-level sentiment analysis incorporating social networks. In *Proc. of KDD '2011*. ACM, 1397–1405.

Ivan Titov and Ryan McDonald. 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th international conference on World Wide Web*. ACM, 111–120.

Peter D. Turney. 2002. Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 417–424. DOI:http://dx.doi.org/10.3115/1073083.1073153

Jingang Wang, Dandan Song, Lejian Liao, Wei Zou, Xiaoqing Yan, and Yi Su. 2013. The Chinese Bag-of-Opinions Method for Hot-Topic-Oriented Sentiment Analysis on Weibo. In *Semantic Web and Web Science*. Springer, 357–367.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 347–354.

Yuanbin Wu, Qi Zhang, Xuanjing Huang, and Lide Wu. 2009. Phrase dependency parsing for opinion mining. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*. Association for Computational Linguistics, 1533–1541.

Zhiwen Yu, Zhu Wang, Huilei He, Jilei Tian, Xinjiang Lu, and Bin Guo. 2015. Discovering Information Propagation Patterns in Microblogging Services. *ACM Trans. Knowl. Discov. Data* 10, 1, Article 7 (July 2015), 22 pages. DOI:http://dx.doi.org/10.1145/2742801

Bo Yuan, Ying Liu, Hui Li, Thao Thi Thanh PHAN, Ghazala Kausar, Cindy NGAI Sing-Bik, Rita Gill SINGH, Juan Carlos Olmos Alcoy, Udita Sawhney, Preet Hiradhar, and others. 2013. Sentiment Classification in Chinese Microblogs: Lexicon-based and Learning-based Approaches. *IPDER* 68, 1 (2013).

Jichang Zhao, Li Dong, Junjie Wu, and Ke Xu. 2012. Moodlens: an emoticon-based sentiment analysis system for chinese tweets. In *Proc. of KDD '2012*. ACM, 1528–1531.

Yanyan Zhao, Bing Qin, Wanxiang Che, and Ting Liu. 2011. Appraisal expression recognition with syntactic path for sentence sentiment classification. *International Journal of Computer Processing Of Languages* 23, 01 (2011), 21–37.

Received April 2015; revised 2015; accepted 2015