Design and Implementation of Weibo Sentiment Analysis Based on LDA and Dependency Parsing

Yonggan Li[1,2], Xueguang Zhou[3*], Yan Sun[4], Huanguo Zhang[1,2]

[1] Computer School of Wuhan University, Wuhan University, Wuhan 430079, China
[2] The Key Laboratory of Aerospace Information Security and Trust Computing, Ministry of Education, Wuhan 430072, China
[3] Department of Information Security, Navy University of Engineering, Wuhan 430033, China
[4] Unit Number of 92941, PLA, Huludao, Liaoning Province 125000, China

Abstract: Information content security is a branch of cyberspace security. How to effectively manage and use Weibo comment information has become a research focus in the field of information content security. Three main tasks involved are emotion sentence identification and classification, emotion tendency classification, and emotion expression extraction. Combining with the latent Dirichlet allocation (LDA) model, a Gibbs sampling implementation for inference of our algorithm is presented, and can be used to categorize emotion tendency automatically with the computer. In accordance with the lower ratio of recall for emotion expression extraction in Weibo, use dependency parsing, divided into two categories with subject and object, summarized six kinds of dependency models from evaluating objects and emotion words, and proposed that a merge algorithm for evaluating objects can be accurately evaluated by participating in a public bakeoff and in the shared tasks among the best methods in the sub-task of emotion expression extraction, indicating the value of our method as not only innovative but practical.

Keywords: information security; information content security; sentiment analysis; dependency parsing; emotion tendency classification; emotion expression extraction

I. INTRODUCTION

Google Research announced that they have open sourced the code of the world's most accurate natural language parser, SyntaxNet. The accuracy of the model trained with this parser by Google has achieved more than 90 percent, while the core of the parser, McParseface, is only effective in dealing with English text. If we wish to make use of the achievements of the open-source parser, secondary development must be made on this parser for Chinese Weibo sentiment analysis. At present, however, some difficulties arise in achieving automatic sentiment analysis for Chinese Weibo texts:

1) Chinese Weibo texts have some characteristics in aspects of opinion sentence use, expressing views of language, and the object of evaluation looming, which make it difficult to obtain satisfactory results when using traditional sentiment analysis methods, such as colloquial, negative, short sentences, strong emotion, non-standard language. Also, the evaluation objects in the sentence may not directly appear [1].
2) When extracting emotional factors from Chinese Weibo text, it is not enough that only emotional words be extracted. For example, in the Chinese Weibo expression

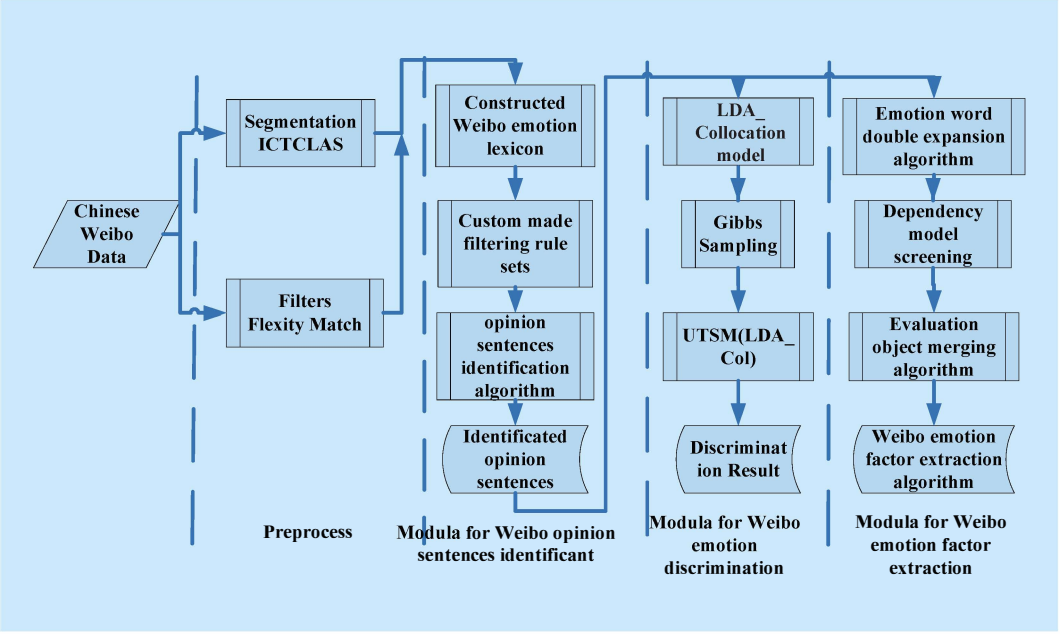

Fig.1 *Flow chart of Chinese Weibo sentiment classification*

"ta bi zhu hai zhu (He is more stupid than a pig), the first word "zhu" belongs to a subject word, holding the original meaning of pig, while the second word "zhu" escapes for emotional words and is transformed with stupid. Such parts of speech transformation cannot be extracted using traditional methods of emotional factors extraction [2].

3) A small amount of invalid reply-posts or active jamming reply-posts reside in Chinese Weibo posts. Artificial regulation can recognize and process them instantly. However, when automatically analyzed by computers, if the computers utilize the word frequency statistics as the basis of classification, then results of sentiment analysis may not be satisfactory [3] and [19].

In order to solve the above problems, combined with the public evaluation of Chinese Weibo emotional analysis, this paper explores the following:

1) Through the LDA collocation model, which represents the text with word order flow, we proposed an unsupervised topic sentiment mixture algorithm for automatic classification of Chinese Weibo texts [4].
2) Using dependency parsing, we divided the dependency relationships into two categories with subject and object, summarized six kinds of dependency models from evaluation objects and emotion words, and proposed a merge algorithm for evaluation objects, which can be used for automatically extracting emotional factors from Chinese Weibo texts.
3) To verify the effect of algorithm, we designed and implemented an evaluation system. The architecture of it is shown as Fig. 1. This evaluation system is used to carry out the experiments of Weibo sentiment classification, Weibo open evaluation, and the comparison of different public evaluation algorithms. Subject to space limitations, the Weibo pretreatment and module of point of view sentence recognition are not included in this paper.

II. Related Work

2.1 Weibo sentiment analysis and factor extraction

Weibo sentiment analysis can mainly be divided into two categories, Twitter analysis and Chinese Weibo analysis, respectively. The processing technology based on English texts of

Twitter is inconvenient for use with Chinese Weibo. This paper focuses on Chinese Weibo processing technology.

Reference [5] defined four networks: forward, reply, copy, and read, and proposed a random walk model called MultiRank in multiple networks, which is more efficient than TwitterRank. This model did not carry out the extraction of emotional words. With the help of the relationship between Weibo contacts, Reference [6] proposed a Weibo generation model named MB-LDA based on LDA, and with this model they can extract Weibo topics and the subjects that the contacts concern, and in this way help users find a sense of interest in social circles. Reference [7] proposed an online burst detection algorithm based on emotion symbol, and constructed an emotion symbol model, extracting the emotion symbol in the data stream. A heuristic algorithm was then carried out to detect unexpected events, and, finally, the events were combined. The above works mainly aimed at Weibo emotional symbol extraction and classification, while there is no specific processing of emotional words in Weibo text.

Affective factors extraction is the most important and difficult task, including the extraction of emotional words and evaluation objects. Considering that the coordinate relation or the turning point between the adjectives represent the syntactic relation, reference [8] extended this to the whole sentence, and used syntactic analysis for the acquisition of emotional words and evaluation objects. Kobayashi et al. obtained the evaluation object by means of syntactic analysis and finding (feature, evaluation), according to [9]. Qiu et al. obtained the evaluation object by using the relationship between the emotional words and evaluation objects in [10]. Reference [11] proposed an iterative algorithm based on a bidirectional graph to obtain implicit evaluation objects and evaluation words. Also, [12] described the relationship between the evaluation object and the evaluation of the word by automatic acquisition of the syntactic path, and then calculated the editing distance from the syntactic path to extract the evaluation unit.

2.2 Language model and dependency analysis

The basic object of language processing is the document. The Vector space model (VSM) model uses document -> word mapping, in which way the model has the problem of identifying synonyms and ambiguous words. The latent semantic analysis (LSA) model adds a semantic dimension, which is using document -> semantic -> word mapping, which lays the foundation for the topic model. However, the LSA model does not have a good reconstruction of the original TF-IDF matrix, and the use of the singular value decomposition algorithm based on linear algebra is highly complex and difficult to be parallel. In this respect, Hofmann proposed a probabilistic latent semantic analysis (pLSA) model from the point of view of the generative model. The pLSA model looks for a generative model that is represented by the co-occurrence matrix of the document and the word, rather than looking for the generative model of the document itself. This presents the pLSA model with two shortcomings: first, the probability of document topic is related to the probability of the training documents, and there lacks the new documents outside the processing data set; second, since the estimated parameters grow with the number of documents, it is easy to cause the overfitting problem [13].

The core of dependency analysis method is to use links to express the relationship between words and words. In the 1970s, J J Robinson put forward four axioms about the dependency relation in dependency grammar. Because of the particularity of the structure of Chinese language, a fifth axiom is proposed [14]: The words on left and right sides of the center word are not dependent on each other. Pennsylvania Chinese tree is an early and popular Chinese structure grammar library. The Ctbparser open source toolkit can automatically obtain the dependency between words, which can be seen in [15] and [16]. The repre-

sentative application of dependency analysis is extensive, such as using the dependency syntactic structure and the dependency grammar to parse results, extracting the evaluation phrase from the syntactic tree, hierarchical dependency parsing and Chinese dependency parsing based on active learning, seen in [17].

2.3 Research foundation

The symbol definitions used in this paper refer to Table 1. The definitions of the dependency relation type used in this paper are shown in Table 2. We use the Pennsylvania Chinese tree tagging system as part of the speech tagging method.

III. System Mode Unsupervised Topic Sentiment Classification Algorithm Based on Topic Mode

3.1 A non-topic emotion model based on LDA-collocation

In 2003, through adding super parameter layer on the pLSA model to establish the probability distribution of potential variables z, Blei et al. proposed the LDA model. This model is a complete generation model, which uses a bag of words to represent text. In this model, the topic is the hidden variable in the document set, and the model calculates the generation probability of words from a given document. The bag of words method does not consider the order between words and words; what it can express is just a rough context of semantic information. For example, "gao" (high) and "tie" (iron) cannot express the meaning of "gaotie" (high railway). In order to make the topic model express more accurate semantic information, Griffiths et al. proposed the LDA-collocation model, based on the LDA model [18]. This uses a word order flow representation approach. In this method, words can express more accurate semantic information; for example, "bu xihuan"(dislike) can express more accurate semantic information than words "bu," "xihuan" (not like). Thus, compared with the LDA model, this model can express more accurate semantic information.

Table I *Symbol description*

Parameter	Definition
θ_d	Topic distribution of document d
φ_d	Emotional distribution of document d
$\phi_{z,m}$	Topic emotion~word distribution
m_s	The emotion of sentences
$(z,m)_n$	The topic and emotion of word
w_n	Words in sentence
α	The Dir parameter of Document-topic distribution
χ	The Dir parameter of document-emotion distribution
β	The Dir parameter of Topic emotion-word distribution
M	The number of documents
S	The number of sentences in the document
N	The number of words in the sentence
L	The number of emotions
K	The number of topics

Table II *Dependency type*

Type of dependency	Description	Examples
SUB	s-v relation	Ta de qiche hen piaoliang.(His car is very beautiful). (qiche,piaoliang).(car, beautiful)
OBJ	v-o relation	Wo you yige pingguo.(I have an apple). (pingguo,you).(apple,have)
PRD	L-P relation	You shi chehuo.(second traffic accident). (chehuo,shi). (traffic accident, is)
VC	c-v relation	Ta beiqi dongxi zou chuqu le. (He carried and went out). (beiqi,zou).(carried, went).
AMOD	Number relation	Wo you yige pingguo. (I have an apple). (yi, ge). (one,single).
NMOD	A-M relation	Keche yu da huoche xiangzhuang. (Bus collided with a big truck). (da,huoche).(big, truck)
PMOD	p-o relation	Ta zai meiguo. (He is stateside). (meiguo, zai).(America,stay)
VMOD	AD-M relation	Qiche kai de hen kuai. (The car is driven very fast). (hen,kuai).(very, fast)
DEP	Dominate "de(real), de(get),de(the earth)"	Wo de pingguo hen haochi.(My apple is delicious). (wo,de).(I,real)
SBAR	auxiliary word	Ni you namoduo qian a.(You wallow in money). (you,a).(have, Oh).
P	auxiliary symbols	Qiche kai de hen kuai.(The car is driven very fast). (kai,.).(drive,.)

Assuming that there are M documents in the corpus, denoted as $D = \{d_1,..., d_m\}$; a to-

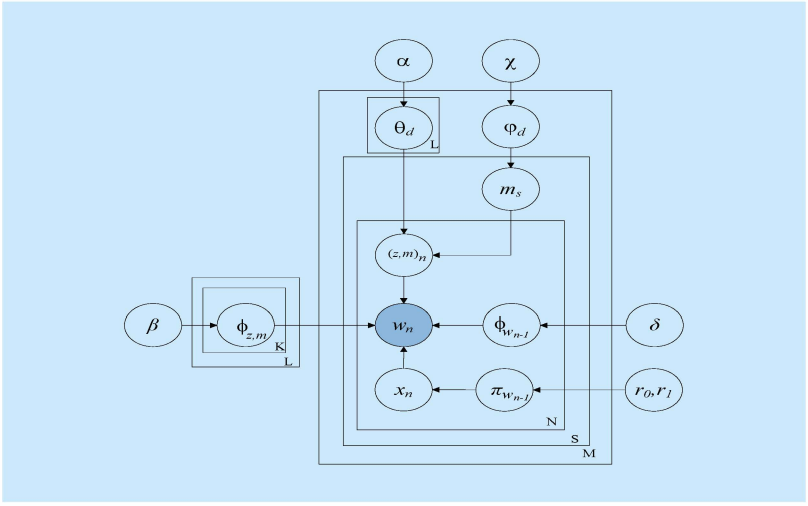

Fig.2 *UTSM(LDA-Col) model*

Algorithm 1: UTSM(LDA-Col) generation algorithm

Input: M documents in the corpus

output: Document topic distribution, document emotion distribution, subject emotion word distribution

process:

1) For every pair of topic z and emotion m:

2) From the Dirichlet distribution of the parameters for β, extract Multinomial distribution $\phi_{z,m}$, i.e. sampling $\phi_{z,m} \sim$ Dir(β)

3) For every document d:

3.1) From the Dirichlet distribution of the parameters for χ, extract Multinomial distribution φ_d, i.e. sampling $\varphi_d \sim$ Dir(χ)

3.2) For every emotion j:

From the Dirichlet distribution of the parameters for α, extract Multinomial distribution θ_{dj}, i.e. sampling $\theta_{dj} \sim$ Dir(α)

3.3) For every sentence s in the document:

3.3.1) extract a Multinomial variable emotion m_s from φ_d, i.e. sampling $m_s \sim$ Multi(φ_d)

3.3.2) For every word $w_{d,n}$ in s:

3.3.2.1) From the Beta distribution of the parameter for γ_0 and γ_1, extract binomial distribution $\pi_{w_{n-1}}$, i.e. sampling $\pi_{w_{n-1}} \sim Beta(r_0, r_1)$

3.3.2.2) Extract variable x_n from $\pi_{w_{n-1}}$, if $x_n = 0$

Extract a Multinomial variable topic $Z_{s,n}$ from θ_{dm}, i.e. sampling $Z_{s,n} \sim$Multi(θ_{dm})

Extract a Multinomial variable word w_n from $\phi_{z,m}$, i.e. sampling $w_n \sim$Multi($\phi_{z,m}$)

3.3.2.3) If $x_n = 1$

From the Dirichlet distribution of the parameters for δ, extract $\phi_{w_{n-1}}$, i.e. sampling $\varphi_{w_{n-1}} \sim Dir(\delta)$

Extract a Multinomial variable word w_{n-1} from $\phi_{w_{n-1}}$, i.e. sampling $w_{n-1} \sim$Multi($\phi_{w_{n-1}}$)

tal of K topics, denoted as $T = \{z_1, L, z_K\}$; a total of L kinds of emotions, denoted as $S_e = \{m_1, m_L\}$; the document D is composed of S sentences, each of which is composed of N words, and the number of the distinct words in the corpus is V. Based on the LDA-collocation model, this paper presents an unsupervised topic sentiment model (UTSM), denoted as UTSM (LDA-col). The model is shown in Figure 2. The symbols used in the model are shown in Table 1, partly referenced within [20].

3.2 Unsupervised topic sentiment algorithm

3.2.1 UTSM (LDA-Col) generation algorithm

The process of UTSM (LDA-Col) generation algorithm is shown in Algorithm 1.

3.2.2 UTSM (LDA-Col) solution

We use i to represent the index number of the lexical token, i = (D, s, n), and word symbol w_i = wd, s, N represents the word related to the location of the document, sentence position, and s_i, represents the sentence where w_i locates in. m_{s_i} represents the emotional distribution of the sentence in which the word w_i is located. $m_{-si\,S}$ indicates the emotional distribution of the sentence in which the other words are located except for the current word. Z_i represents the topic assignment for the w_i, and Z_{-i} represents the rest all topic assignment except for the current word w_i. Thus we have $w = \{w_i = t, w_{-i}\}$, $z = \{z_i = k, z_{-i}\}$, $m = \{m_{si} = j, m_{-si}\}$. By using the Gibbs sampling algorithm, the probability that current word w_i under topic z and emotion j can be calculated through formula (1).

$$p(z_i = k, m_{s_i} = j \,|\, z_{-i}, m_{-s_i}, w) \propto \begin{cases} \dfrac{n^{(t)}_{k,j,-i} + \beta}{\sum_{t=1}^{V}\left(n^{(t)}_{k,j,-i} + \beta\right)} \dfrac{n^{(j)}_{d,-s_i} + \chi}{\sum_{j=1}^{L}\left(n^{(j)}_{d,-s_i} + \chi_j\right)} \dfrac{n^{(k)}_{d,j,-i} + \alpha}{\sum_{k=1}^{K}\left(n^{(k)}_{d,j,-i} + \alpha\right)} & if\ x_i = 0 \\ \dfrac{n^{(j)}_{d,-s_i} + \chi}{\sum_{j=1}^{L}\left(n^{(j)}_{d,-s_i} + \chi\right)} \dfrac{n^{(k)}_{d,j,-i} + \alpha}{\sum_{k=1}^{K}\left(n^{(k)}_{d,j,-i} + \alpha\right)} & if\ x_i = 1 \end{cases} \quad (1)$$

In formula (1), $n^{(t)}_{k,j,-i}$ represents the number of words with the same content with w_i, and are assigned to the same topic k and emotion j in addition to the current word w_i. $n^{(j)}_{d,-s_i}$ indi-

cates the number of sentences assigned to the same emotion j except for the sentence that w_i locates in. $n^{(k)}_{d,j,-i}$ represents the number of words assigned to the same topic k and emotion j with w_i except for it. Note that all the statistics are carried out on the premise of the value of X_i. For example, when $X_i = 0$, $n^{(k)}_{d,j,-i}$ represents the number of the words which is equal to 0 and are assigned to the same topic k and emotion j with w_i except for it. When $X_i = 1$, $n^{(k)}_{d,j,-i}$ represents the number of the words which is equal to 1 and are assigned to the same topic k and emotion j with w_i except for it. X_i sampling from the distribution of the formula (2).

$$p(x_i \mid x_{-i}, w, j) \propto \begin{cases} \dfrac{n^{(t)}_{k,j,-i} + \beta}{\sum_{t=1}^{V}\left(n^{(t)}_{k,j,-i} + \beta\right)} \dfrac{n^{0}_{w_{i-1}} + \gamma_0}{\sum_{u=0}^{1} n^{u}_{w_{i-1}} + \gamma_0 + \gamma_1} & x_i = 0 \\ \dfrac{n^{(t)}_{w_{i-1}} + \delta}{\sum_{t=1}^{V}\left(n^{(t)}_{w_{i-1}} + \delta\right)} \dfrac{n^{1}_{w_{i-1}} + \gamma_1}{\sum_{u=0}^{1} n^{(u)}_{w_{i-1}} + \gamma_0 + \gamma_1} & x_i = 1 \end{cases} \tag{2}$$

In formula (2), $n^{(t)}_{k,j,-i}$ represents the number of the words which have the same content with w_i, and are assigned to the same topic k and emotion j in addition to the current word w_i. $n^{0}_{w_{i-1}}$ represents the number of words with the same content with w_{i-1}, but not a collocation of it in addition to w_{i-1} itself. $n^{1}_{w_{i-1}}$ represents the number of the words which have the same content with w_{i-1} and is a collocation of it in addition to w_{i-1} itself. $n^{(t)}_{w_{i-1}}$ represents the co-occurrence times of w_i and w_{i-1}.

The document topic distribution θ, and document emotion distribution φ, and topic emotion word distribution ϕ are shown from formula (3) to (5).

$$\hat{\theta}_{d,j,k} = \frac{n^{(k)}_{d,j} + \alpha}{\sum_{k=1}^{K}\left(n^{(k)}_{d,j} + \alpha\right)} \tag{3}$$

$$\hat{\varphi}_{d,j} = \frac{n^{(j)}_{d} + \chi}{\sum_{j=1}^{L}\left(n^{(j)}_{d} + \chi\right)} \tag{4}$$

$$\hat{\phi}_{k,j,w} = \frac{n^{(w)}_{k,j} + \beta}{\sum_{w=1}^{V}\left(n^{(w)}_{k,j} + \beta\right)} \tag{5}$$

$\hat{\theta}_{d,j,k}$ represents the probability estimation that the emotion of topic k in document d equals to j. $\hat{\varphi}_{d,j}$ represents the probability estimation of emotion j in document d. $\hat{\phi}_{k,j,w}$ represents the probability estimation that the topic of the word w is k and the emotion of it is k. $n^{(j)}_{d}$ represents the number of sentences assigned to emotion j in document d. $n^{(k)}_{d,j}$ represents the number of words assigned to topic k and emotion j in document d. $n^{(w)}_{k,j}$ represents the times that word w is assigned to topic k and emotion j in document d.

The distribution of conjunction ϕ_w and collocation distribution π_w is shown in formula (6) and (7). W_1 and W_2 are used to denote W_{I-1} and W_i in order to distinguish the W_i.

$$\hat{\phi}^{(w_2)}_{w_1} = p(w_2 \mid w_1, x_2 = 1) = \frac{n^{(w_2)}_{w_1} + \delta_{w_2}}{\sum_{w_i=1}^{V}\left(n^{(w_i)}_{w_1} + \delta_{w_i}\right)} \tag{6}$$

$$\hat{\pi}_{w_1} = p(x_2 = 1 \mid w_1) = \frac{n^{1}_{w_1} + \gamma_1}{\sum_{u=0}^{1} n^{(u)}_{w_1} + \gamma_0 + \gamma_1} \tag{7}$$

$\hat{\phi}^{(w_2)}_{w_1}$ represents the probability estimation of the co-currency of W_1 and W_2. $\hat{\pi}_{w_1}$ represents the probability estimation of $x_2 = 1$ given W_1. $n^{(w_2)}_{w_1}$ represents the times of the co-currency W_1 and W_2. $n^{1}_{w_1}$ represents the number of the words which is a collocation of W_1.

The probability of the currency of W_1 and W_2 can be calculated from formula (6) and (7).

$$p(w_2 \mid w_1) = p(w_2 \mid w_1, x_2 = 1)p(x_2 = 1 \mid w_1) + p(w_2 \mid w_1, x_2 = 0)p(x_2 = 0 \mid w_1) \tag{8}$$

$p(w_2|w_1, x_2 = 0)$ represents the $p(w_2|w_1)$ in UTSM. It can be calculated through formula (9).

$$p(w_2 \mid w_1, x_2 = 0) = \frac{\sum_{(z,m)} p(w_2 \mid (z,m))\, p(w_1 \mid (z,m))}{\sum_{(z,m)} p(w_1 \mid (z,m))} = \frac{\sum_{(z,m)} \phi^{(z,m)}_{w_2} \phi^{(z,m)}_{w_1}}{\sum_{(z,m)} \phi^{(z,m)}_{w_1}} = \sum_{(z,m)} \phi^{(z,m)}_{w_2} \frac{\phi^{(z,m)}_{w_1}}{\sum_{(z,m)} \phi^{(z,m)}_{w_1}} \tag{9}$$

Thus $p(w_2|w_1)$ can be calculated through formula (10).

$$p(w_2|w_1) = \pi^{(w_1)}\phi_{w_2}^{(w_1)} + (1-\pi^{(w_1)})\sum_{(z,m)}\phi_{w_2}^{(z,m)}\frac{\phi_{w_1}^{(z,m)}}{\sum_{(z,m)}\phi_{w_1}^{(z,m)}} \quad (10)$$

3.3 Weibo sentiment classification and evaluation method

The probability estimation of the distribution of emotion j in document d can be calculated by using $\hat{\phi}_{d,j}$ in UTSM. The sentiment of document d can be calculated through the maximum value of the probability estimation of each sentiment tendency in document d. i.e.:

$$m_d = \arg\max_j \left\{\hat{\varphi}_{d,j} | j \in [1,\cdots,L]\right\} \quad (11)$$

Weibo sentiment classification is an application in text categorization, and precision, recall and F-measure is often used for evaluation.

$$\text{F\_measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (12)$$

IV. Weibo Emotion Factors Extraction Based on Dependency Model

The Weibo emotion factors extraction in this paper is implemented in three steps: emotional word double expansion, dependency model screening, and evaluation of object merging.

4.1 Emotional word double expansion algorithm

The basic idea of the emotional words double expansion algorithm is first to establish an emotion dictionary. The sentiment word is expanded through the seed emotion and the dependency of the emotion words. The following expansion is conducted through the dependency of the known emotion words and evaluation objects, as well as the dependency of the emotion words and the known evaluation objects, until no new emotional word and evaluation object generate.

{KW} is the known emotion lexicon, and {KT} is the known evaluation objects, and {JVV} represents adjectives or verbs, i.e., {JV} = {JJ,VA,VV}, in which JJ, VA, VV indicate adjectives, predicate adjectives, other verbs, respectively. {NP} represents noun, and {NP} = {NR,NT,NN,PN}, in which NR,NT,NN,PN indicates proper nouns, time nouns, other nouns, and pronouns. (w1, w2) is the dependency pairs that dependent order is not considered. This means that it can be w1 depends on w2, or w2 depends on w1. (w1,w, w2) indicates that w1 and w2 depend on w simultaneously, and the dependent order is same. The emotional word double expansion algorithm is shown in Algorithm 2.

Algorithm 2: Emotional word double expansion algorithm

Input: Seed emotion dictionary and the corpus to be tested
Output: New emotion dictionary
Process:

1) Extract all the dependency pairs (w_1,w_2) which satisfy $\text{pos}(w_1) \in \{\text{JV V}\} \cap \text{pos}(w_2) \in \{\text{N P}\} \cap w_1 \in \{\text{KW}\}$, put w2 into {KT}
2) Extract all the dependency pairs (w_1,w_2) which satisfy $\text{pos}(w_1) \in \{\text{JVV}\} \cap \text{pos}(w_2) \in \{\text{JVV}\} \cap w_1 \in \{\text{KW}\}$, put w2 into {KT}
3) Extract all the dependency pairs (w_1', w, w_2') which satisfy $\text{pos}(w_1') \in \{\text{JVV}\} \cap \text{pos}(w_2') \in \{\text{JVV}\} \cap w_1' \in \{\text{KW}\}$, put w_2' into {KW}
4) Extract all the dependency pairs (w_1,w_2) which satisfy $\text{pos}(w_1) \in \{\text{N P}\} \cap \text{pos}(w_2) \in \{\text{N P}\} \cap w_1 \in \{\text{KT}\}$ put w2 into {KT};
5) Extract all the dependency pairs (w_1', w, w_2') which satisfy $\text{pos}(w_1') \in \{\text{NP}\} \cap \text{pos}(w_2') \in \{\text{N P}\} \cap w_1' \in \{\text{KT}\}$, put w_2' into {KT}
6) Extract all the dependency pairs (w_1,w_2) which satisfy $\text{pos}(w_1) \in \{\text{N P}\} \cap \text{pos}(w_2) \in \{\text{JVV}\} \cap w_1 \in \{\text{KT}\}$, put w_2 into {KW};
7) Repeat steps above until no emotion word and evaluation object generate.

4.2 Dependency model screening

In order to avoid the noise that extracting emotion words and evaluation objects just based on dependency relations brings, when dependency analysis is conducting, we add restrictions on the part of speech of words, and establish the priority of evaluation unit dependency models, which is from high to low in turns from 1 to 6. In the evaluation unit dependence model, bold indicates emotional words and the evaluation objects.

(1) $\{\textbf{NP}\} \xleftarrow{SUB} \{\textbf{AV}\}$

In {NP} = {NR,NT,NN,PN}, NR,NT,NN,PN indicate proper nouns, time nouns, other

nouns and pronouns. In {AV} = {VA,VV}, VA and VV indicate predicate adjectives and other verbs. The model (1) indicates that the subject is the object of evaluation, and the subject depends on the verb. The dependency relationship between the evaluation object and the emotional words is direct, such as "WeiNan cheng guan zhen bian tai! (WeiNan city inspectors are so weird!)."The dependency analysis is shown in Fig.3.

The evaluation units are "bian tai, cheng guan," (weird, city inspector), and after merging "WeiNan" and "city inspector," we get evaluation object "bian tai, WeiNan cheng guan,"(weird,WeiNan city inspectors).

(2) $\{\mathbf{NP}\} \xleftarrow{SUB} \{\mathbf{VC}\} \xrightarrow{PRD} \{\mathbf{NP}\} \xrightarrow{NMOD} \{\mathbf{JV}\}$

{VC} represents verbs, and in {JV} = {JJ, VA}, JJ and VA indicate adjectives and the predicate adjective. It says predicate is the verb, and subject is the evaluation object, and the predicative attribute is emotional words. Between the subject and predicative attribute, there is the extended dependency relationship. For example, in "ta shi yige youxiu de xuesheng," (He is an excellent student), the dependency analysis is shown in Fig.4.

"$youxiu \xleftarrow{SBAR} de \xleftarrow{NMOD} xuesheng$" $(excellent \xleftarrow{SBAR} of \xleftarrow{NMOD} student)$ can get "$youxiu \xleftarrow{SBAR} xuesheng$" $(excellent \xleftarrow{SBAR} student)$ after pruning operation, and evaluation unit (excellent, he) can be got by using model (2).

(3) $\{\mathbf{AV}\} \xrightarrow{OBJ} \{\mathbf{NP}\}$

The model (3) indicates that the predicate is not a verb, and the object is the evaluation object. The predicate is the emotional word, and the object is directly dependent on the predicate. For example, "wo xihuan ipad3 de pingmu." (I love the screen of iPad3), the dependency analysis is shown in Fig.5.

The evaluation unit is (love(xi huan), screen(ping mu)), and after appling merging operation on "$\text{ipad3} \xleftarrow{DEP} de \xleftarrow{NMOD} pingmu$" $(\text{ipad3} \xleftarrow{DEP'} s \xleftarrow{NMOD} screen)$ we can get "ipad3 de pingmu" (the screen of ipad3), and finally we can get the evaluation unit "xihuan, ipad3 de pingmu,"(love,the screen of ipad3).

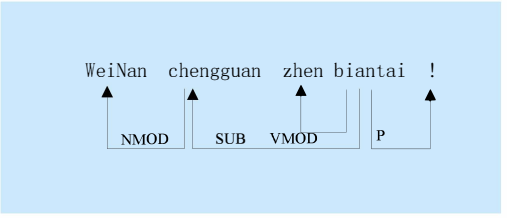

Fig.3 *Example of appraisal expression dependency mode 1*

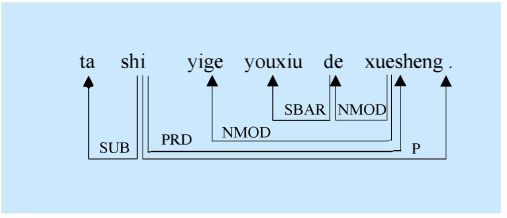

Fig.4 *Example of appraisal expression dependency mode 2*

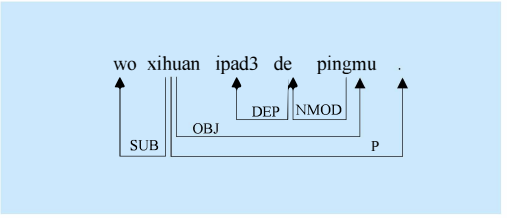

Fig.5 *Example of appraisal expression dependency mode 3*

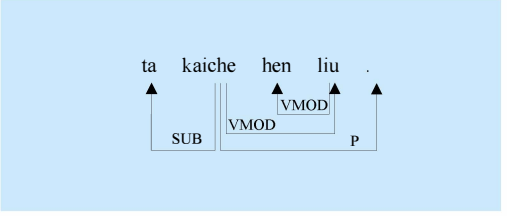

Fig.6 *Example of appraisal expression dependency mode 4*

(4) $\{\mathbf{NP}\} \xleftarrow{SUB} \{\mathbf{VV}\} \xrightarrow{VMOD} \{\mathbf{AV}\}$

The model (4) indicates that the subject is the object of evaluation, and the complement of the verb is an emotional word. The subject and verb complement is the dependent relation of the dependency relation or the extension. For example, in "ta kai che hen liu" (He is good at driving), the dependency analysis is shown in Fig.6.

The evaluation unit is "liu, ta" (is good at,

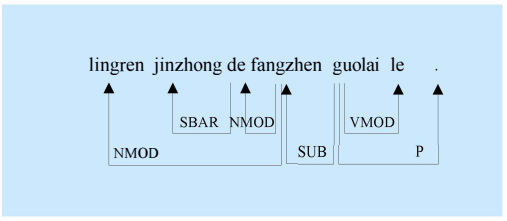

Fig.7 *Example of the appraisal expression dependency mode 5*

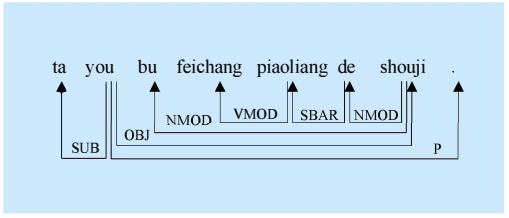

Fig.8 *Example of appraisal expression dependency mode 6*

he).

(5) {**JV**} $\overset{NMOD}{\longleftarrow}$ {**NP**} $\overset{SUB}{\longleftarrow}$ {VV}

The model (5) indicates that the subject is the object of evaluation, and the modifier of the subject is an emotional word, and the subject and subject attribute are directly or indirectly dependent. For example, in "lingren jingzhong de fangzhen guolai le" (The admirable square is coming), the dependency analysis is shown in Fig.7.

The evaluation unit is "jingzhong, fang zhen" (admirable, square).

(6) {VV} $\overset{OBJ}{\longrightarrow}$ {**NP**} $\overset{NMOD}{\longrightarrow}$ {**JV**}

The model (6) indicates that the object is the evaluation object, and the attribute of the modified object is the emotional word. The object and object attribute is in the direct or indirect dependency relation. For example, in "ta you bu fei chang piaoliang de shouji" (he has a very nice cellphone), the dependency analysis is shown in Fig.8.

"*piaoliang* $\overset{SBAR}{\longleftarrow}$ *de* $\overset{NMOD}{\longleftarrow}$ *shouji*" (*nice* $\overset{SBAR}{\longleftarrow}$ *of* $\overset{NMOD}{\longleftarrow}$ *cellphone*) can get "*piaoliang* $\overset{SBAR}{\longleftarrow}$ *shouji*" (*nice* $\overset{SBAR}{\longleftarrow}$ *cellphone*) after pruning operation, and evaluation unit "piaoliang, shouji" (nice, cellphone) can be got by using model (6).

In addition to the above six modes, for the sentence containing only the partial positive structure phrases and punctuation, we use model (5) to process, such as "fengkuang de dacong" (crazy allium fistulosum) and "nao xin de linshigong" (upset temporary worker), and so on. Note that the priority of the evaluation unit dependency model is gradually reduced from 1 to 6. This means that we firstly match the evaluation unit dependency with model 1, and if the match is successful the candidate evaluation unit is selected; or model 2 will be used, and so on.

4.3 Evaluation object merging algorithm

Among the six dependency models of emotion words and evaluation objects, the evaluation object of model 1, model 2, model 4, and model 5 is subject, while the evaluation object of model 3 and model 6 is object. In order to extract the evaluation objects as completely and clearly as possible, a merging operation is needed for application to obtain a complete subject and object. For example, "wo xihuan ipad3 de pingmu" (I love the screen of iPad3), "ipad3 de pingmu" (the screen of iPad3) but not "pingmu" (screen) should be extracted. In order to distinguish the original evaluation object and the evaluation object, the original evaluation object is called the evaluation object reference in this paper. The evaluation object merging operation starts from the leftmost term of the evaluation object reference words, judging from right to left, and if the father node is a benchmark or is a right word of the reference, then we merge it to the evaluation object; if the sentence or its parent node are not standard words or is not the right word, then we stop and return.

The evaluation object merging algorithm is shown as Algorithm 3 on the next page.

4.4 Weibo emotion factor extraction algorithm

The algorithm is shown as Algorithm 4 on the next page.

V. Experiments and Analysis

5.1 Experiment environment and architecture

Experiment environment. We use the open source toolkit "ctbparser" for dependency analysis. Penn Chinese Treebank tag set is used to do word segmentation and POS tagging. The experiment is implemented in C++ on Visual Studio 2010.

Experiment architecture. Chinese Weibo sentiment analysis includes four modules (Figure 1).

5.2 Chinese Weibo sentiment module function experiment

Two datasets are involved in this experiment. We use "Star of the East sinking boat event" as topic 1 and extract 500 comments from it; we use "Philippine warship collision accident" in 2012 as topic 2 and extract 500 comments from it. Comments from each topic contain key words showing subjective feelings. Topic 1 and topic 2 represent "sadness" and "anger," respectively.

5.3 Sentimental words scatter distribution experiment

Topic 1 contains 25 kinds of key sentimental words and topic 2 contains 30 kinds. We selected seven kinds of words (happiness, preference, anger, disgust, fear, sadness, astonishment). Corresponding words appear in ascending order in the sentence set. We presented Figure 9 and Figure 10 based on the word distribution. Word type has sentiment intensity and POS has a corresponding sentiment weight coefficient. Word frequency and sentiment intensity with attached weight decided the sentimental tendency of the word together. We can see in the figure that the higher the word frequency and the word's sentimental intensity, the closer is the distribution to the bottom-right of the scatter plot. In Fig. 9, the word "too horrible to look" embodies the Weibo sentiment tendency of topic "Star of the East sinking boat event" the best. In Fig. 10, the word "anger" is the best word to show the sentiment tendency of "Philippine warship collision accident."

Algorithm 3: Evaluation object merging algorithm

Input: evaluation object reference TRRinRR
Output: New evaluation object TRRoutRR
process:
1) W(i) = TRRinRR;
2) j = i-1;
3) WHILE(j> = 0) {//all the left words W(j) of W(i)
4) IF(W(j).parent = i|| W(j).parent = j+1) THEN j--;
5) ELSE {
6) FOR(k = j;k< = i; k++) TRRoutRR = TRRoutRR+W(k);
7) BREAK;}
8) }

Algorithm 4: Weibo emotion factor extraction algorithm

Input: Determined view sentence set ValueSentenceID[ ]
Output: New complete evaluation object
Process:
1) Expand emotion words using emotional word double expansion algorithm;get emotion words and evaluation objects;
2) By using the evaluation model, the candidate evaluation unit is selected according to priority;
3) Get complete evaluation object by using evaluation object merging algorithm

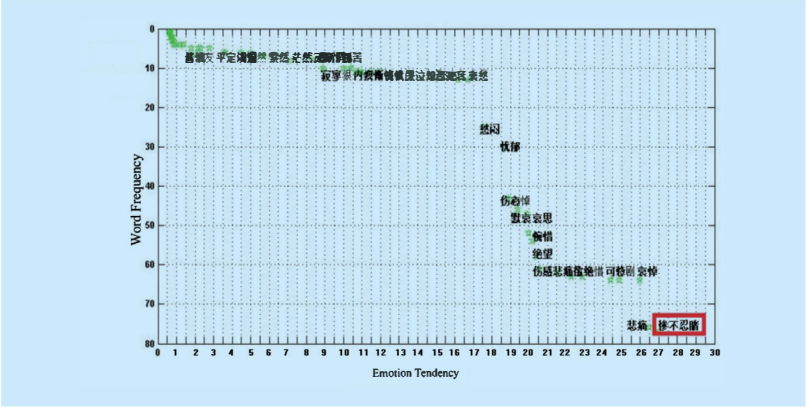

Fig.9 *Scatter diagram for microblog sentiment classification of shipwreck of Star of the East*

5.4 Fine-grained sentimental analysis experiment on Weibo

When the extracted word is low in amount and in its matching ratio to sentimental dictionary,

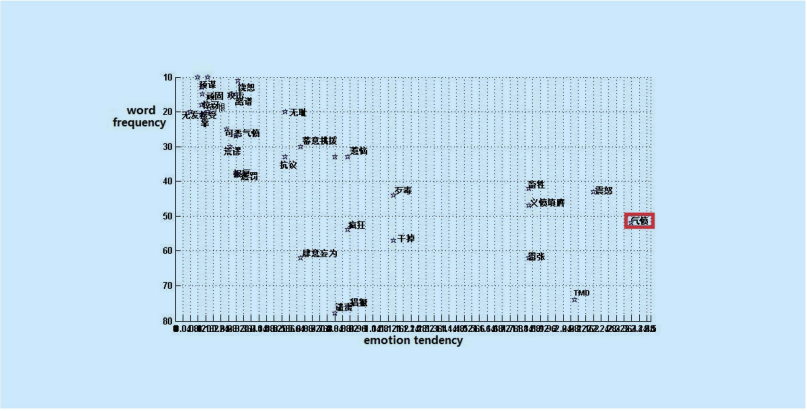

Fig.10 *Scatter diagram for microblog sentiment classification of impact trouble of Philippine warship*

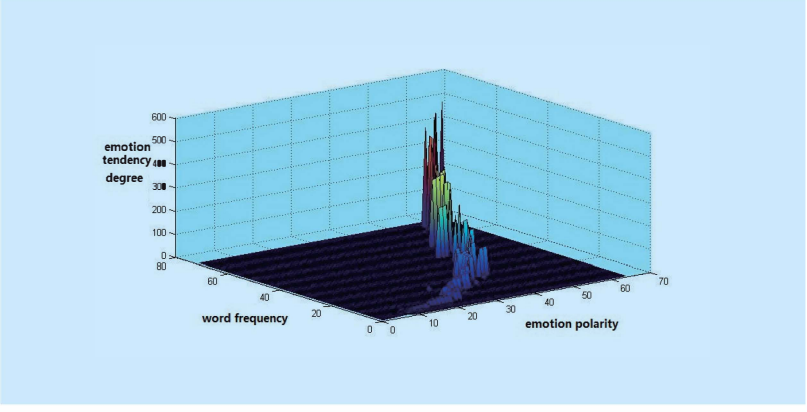

Fig.11 *Diagram for microblog sentiment fine grain classification of shipwreck of Star of the East*

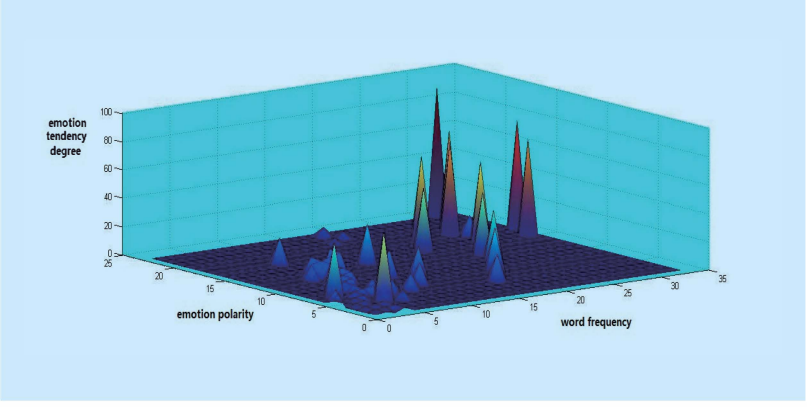

Fig.12 *Diagram for microblog sentiment fine grain classification of impact trouble of Philippine warship*

it is difficult to judge the sentimental tendency, with a scattered distributed and nonlinear three-dimensional spike. Fig. 11 and Fig. 12 show fine-grained analysis results of the two topics. The three-dimensional distribution is strong in its regularity and the spikes are more linear. Judgment on sentimental tendency is more intuitive. The spikes of topic 2 are more dispersed, resulting from its complexity in its sentimental words class, and thus it is more divergent in its sentimental tendency. However, we can still fix the sentiment of the topic on (31, 23, 90), (33, 17, 75) and (32, 15, 70), which represents "anger", "TMD" and "arrogance," respectively.

5.5 Random judgment on Weibo sentiment

Based on the method above, we performed an analysis on "happiness, preference, disgust, astonishment, fear," with 100 Weibo comments selected randomly and each independent of each other. By finding key words in the sentence and doing quantized analysis, the program judged them one by one and obtained their distribution on seven sentiment dimensions (Fig. 13). In Fig. 13, lines with different colors represent different sentiment types. In the sentiment analysis about a sentence, there is always a sentiment type with an outstanding spike value. Then the sentiment tendency can be estimated as the sentiment of the sentence.

To present sentiment distribution characteristics of different sentences better and more intuitively, Fig. 14 shows the distribution in three dimensions; the x axis corresponds to the serial number of the sentence, the y axis corresponds to seven sentiment types, and the z-axis corresponds to sentiment tendency. The distribution of the spikes is sparse and each of the sentences has a most outstanding spike, which indicates that each comment belongs to a sentiment type and the magnitude of this sentiment type can override other ones.

5.6 Public evaluation on Chinese Weibo sentiment analysis

China Computer Federation (CCF) Conference on Natural Language Processing &

Chinese Computing (NLP&CC) is the annual conference of the CCF Technical Committee of Chinese Information (TCCI). And we took part in the first conference, NLP&CC2012. Our experimental results below root in the public evaluation of the conference.

5.6.1 Sentiment tendency analysis algorithm

Fig. 15 shows algorithms used in Chinese Weibo sentiment tendency analysis and their results; simplified analysis is as follows:

(1) After comparison, we can see that conditional random field models (CRF) perform well on accuracy, recall and F-value. The accuracy of SVM, CRF, Tree, NN, and OURs is close. The accuracy of SVM is the highest, which is 0.034 higher than CRF, but it is less in recall for 0.113, which leads to a worse F-value than CRF.

(2) The accuracy of maximum entropy model (ME) performed on sentiment tendency is close to the accuracy of CRF, but its recall is far lower than CRF, which leads to a lower F-value. The results of other models are far behind CRF. We can conclude from the results that CRF has an outstanding performance in Chinese Weibo sentiment analysis.

(3) Reasons for the good performance of CRF include the fact that CRF is an undirected graph model. It has a strong reasoning ability and it can use complex, overlapping and non-independent features for training and reasoning. By using context information and other extra features, CRF can acquire rich information. Meanwhile, CRF solves the problem "label bias" which exists in ME. Theoretically, CRF is quite suitable for Chinese POS tagging.

(4) CRF has shortcomings. Feature selection and optimization has a major influence on the results. Besides, it requires more time than maximum entropy (ME) to train the model and the model is so large that it cannot run on a regular PC.

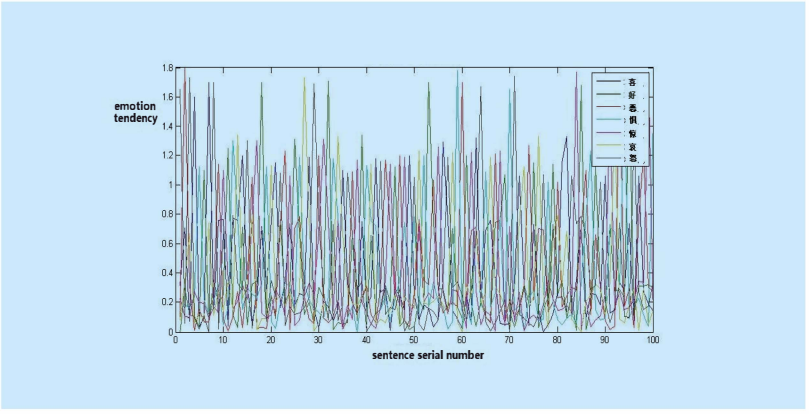

Fig.13 *Diagram for microblog sentiment classification of 100 stochastic sentences*

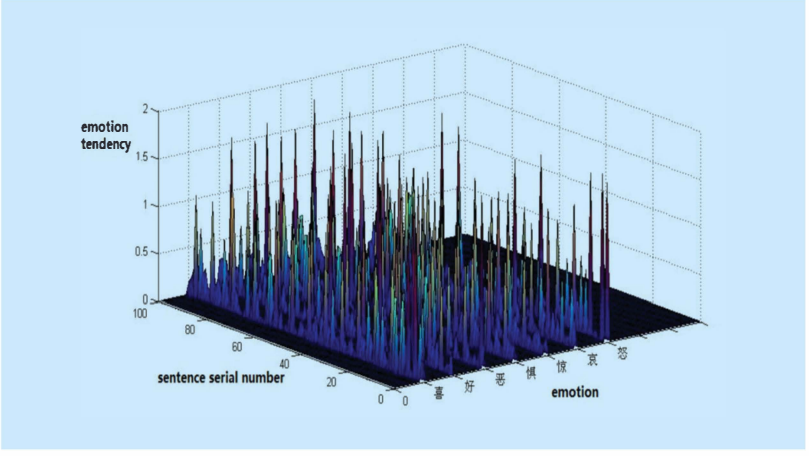

Fig.14 *Three-dimensional diagram for microblog sentiment classification of 100 stochastic sentences*

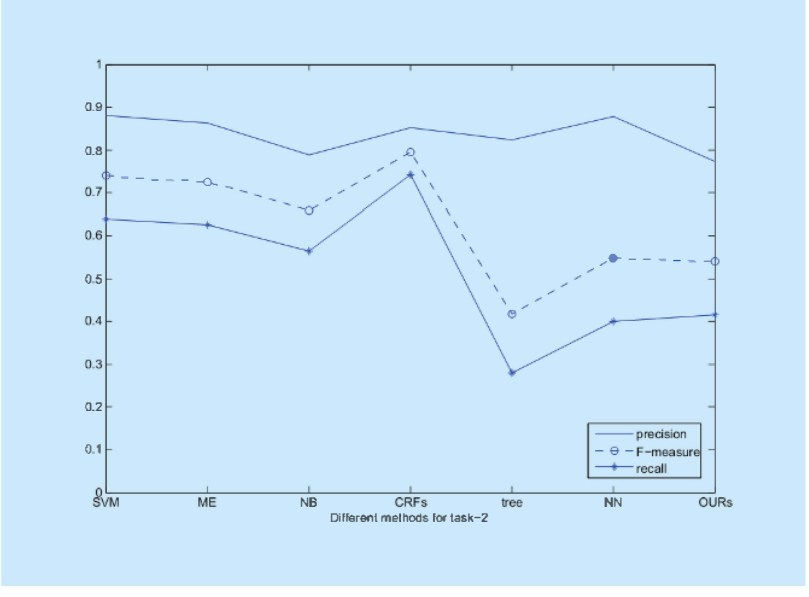

Fig.15 *Open evaluating results and algorithms of emotion tendency classification*

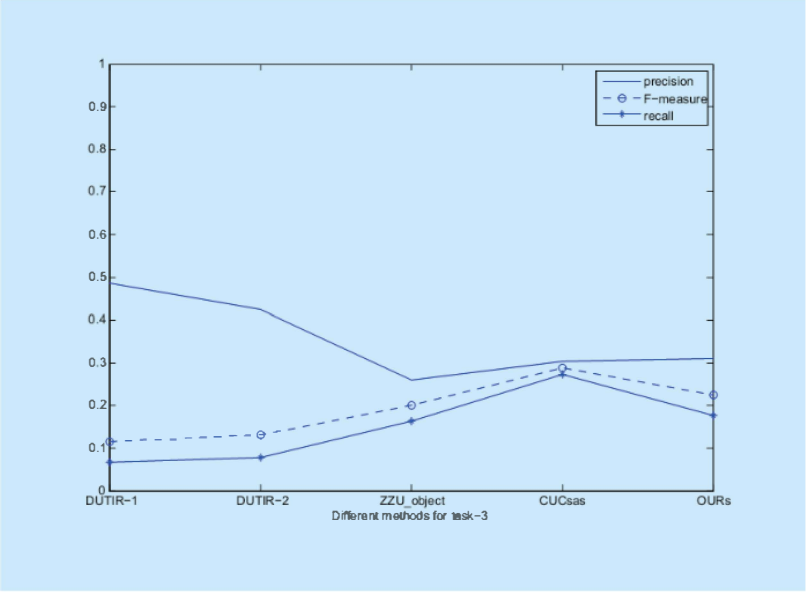

Fig.16 *Comparison results of different methods for emotion expression extraction*

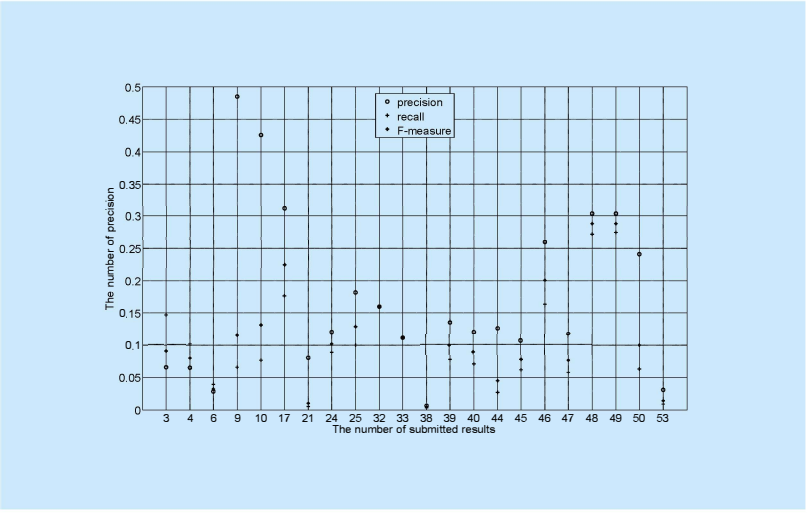

Fig.17 *Strict open evaluating results of emotion expression extraction*

5.6.2 Contrast experiments on sentiment main factor extraction algorithm

Four firms that carried out Weibo sentiment main factor extraction attended the meeting. The analysis result is shown in Fig. 16. Corresponding analysis is as follows:

(1) DUTIR-1 and DUTIR-2 are two results submitted by the same company. They all apply techniques based on rules and CRF. Features include high accuracy, low recall, and a low F-value.

(2) ZZU-object adopts dependency analysis based on the Chinese syntactic analyzer of Harbin Institute of Technology. The performance of the algorithm is middle-ranking in the five algorithms; based on the result of key sentence recognition and sentiment tendency, we can conclude that they focus on methods based on rules.

(3) CUCsas establishes a sentiment dictionary based on short terms by concluding Weibo language features and decides semantic polarity with a short term rule focused on negative format. To do Weibo sentiment analysis, it applies strategies with OBJ form. This method adopts manual empirical value and predicts 80% of the Weibo negative. However, it fails to do research on batch processing in Weibo sentiment analysis.

(4) The accuracy of OURs ranks third in accuracy and its recall and F-value are all in the second place. The results show that it is effective to define six evaluation unit dependency models, merge sentiment objects, and extract sentimental key factors. This algorithm is also applicable to perform Chinese Weibo automatic processing.

5.6.3 Results and analysis of the open evaluation of emotion factors

Weibo emotion factors extraction evaluation uses two methods: strict evaluation and loose evaluation. Figure 17 shows all the results of NLP&CC2012 public evaluation of Weibo emotion factors extraction.

From Figure 17, we can see two methods, No. 9 and 10, sacrifice recall rate exchanging for high accuracy, but the F value is not ideal. While No. 48 and 49 (same unit) methods have a small difference between precision rate and recall rate, the F value is the best. This paper presents the results of the evaluation of the emotion factor extraction method (No. 17) with an average of strict evaluation results, and the recall rate and the F value in all participating units ranked 2.

The results of this method are not satisfac-

tory, and reasons are as follows:

(1) In this paper, the method is not combined with context to carry out demonstrative pronoun resolution. For example, in "Xiaoming studies in Peking University, he is an excellent student. (xiaoming jiu du yu bei jing da xue, ta shi yi ming hao xue sheng)," (he (ta), excellent (you xiu)) is extracted as the evaluation unit, while the true unit should be (Xiaoming (xiao ming), excellent (you xiu)).

(2) In the context of the non-dependence of the emotion words, if they are not in the emotion dictionary, then they cannot be obtained through dependency association expansion, which affects the recall rate.

In view of the low precision of Chinese Weibo word segmentation and the dependence analysis results, the accuracy rate is affected. As determining the emotion factors needs to extract the evaluation phrase, Weibo emotion factor extraction accuracy (< 70%) is much lower than the existing Chinese word segmentation accuracy (> 90%). In addition, due to instances where the Weibo author may write Chinese Weibo in a way that does not comply with Chinese characteristics and rules, it is difficult to improve Chinese Weibo automatic analysis efficiency with human operation.

VI. Conclusions

In order to solve the problem of Weibo content security regulatory issues, this paper proposes two algorithms: a Weibo emotion classification algorithm and Weibo open evaluation algorithm. We carried out evaluation experiments, and participated in the NLP&CC2012 public evaluation. Our model ranks second in two tasks: the identification of the point of view as well as emotional elements extraction. It is proved that our algorithm has advantages of high accuracy, high automation, and high system efficiency.

Acknowledgement

The authors would like to thank the anonymous reviewers for their detailed reviews and constructive comments, which have helped improve the quality of this paper. This work has been performed in the Project supported by National Key Basic Research Program of China (No. 2014CB340600), and partially supported by National Natural Science Foundation of China (Grant Nos. 61332019, 61672531), and partially supported by National Social Science Foundation of China (Grant No. 14GJ003-152).

References

[1] Zhang Huanguo, Han Wenbao, Lai Xuejia, et al, "Survey on cyberspace security," *Sci. China Inf. Sci.*,vol.58, no.11, pp.1-43,2015.

[2] Wang Xiang, Zhang Zhilin, Yu Xiang, et al. "Finding the Hidden Hands: A Case Study of Detecting Organized Posters and Promoters in SINA Weibo," *China Communications*. vol.12, no.11,pp. 143-155, 2015.

[3] Hou Min, Teng Yongling, Li Xueyan, et al. "The Characteristic of topic Microblogging and its analysis,"[DB/OL]. In: Proc. of the NLPCC2012, http://tcci.ccf.org.cn/conference/2012/dldoc/NLPCC2012papers/workshop papers/sen/003.pdf, 2016-4-10.

[4] Li Yang, Chen Yiheng, Liu Ting. "Survey on predicting information propagation in microblogs," *Ruan Jian Xue Bao/Journal of Software*, vol.27,-no.2,pp.247-263, 2016. (in Chinese with English abstract). http://www.jos.org.cn /1000-9825/4944.htm.

[5] Ding Zhaoyun, Zhou Bin, Jia Yan, et al. "Topical Influence Analysis Based on the Multi-Relational Network in Microblogs," *Journal of Computer Research and Development*, vol.50, no.10, pp. 2155-2175, 2013. (in Chinese with English abstract).

[6] Zhang Chenyi, Sun Jianlin, Ding Yiqun. "Topic Mining for Microblog based on MB-LDA Model," *Journal of Computer Research and Development*. vol.48, no.10,pp.1795-1802, 2011. (in Chinese with English abstract).

[7] Zhang Lumin, Jia Yan, Zhou Bin, et al." Online Bursty Events Detection based on Emoticons," *Chinese Journal of Computers*, vol.36, no.8, pp.1659-1667,2013. (in Chinese with English abstract).

[8] Hatzivassiloglou V, Mckeown K. "Predicting the Semantic Orientation of Adjectives," In Proceedings of the eighth conference on European chapter of the Association for Computational

Linguistics. Stroudsburg: ACL, pp. 174-181, 1997.
[9] Kobayashi N, Inui K, Matsumoto Y. "Extracting Aspect-evaluation and Aspect-of Relations in Opinion Mining,". In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg, ACL pp.1065-1074, 2007.
[10] Qiu G, Liu B, BU JJ, et al. "Opinion Word Expansion and Target Extraction through Double Propagation," *Computational Linguistics*, vol. 37, no.1, pp. 9-27, 2009.
[11] Zhang L, Liu B. "Extracting Resource Terms for Sentiment Analysis," In Proceedings of the 5th International Joint Conference on Natural Language Processing, Chiang Mai: Thailand, pp. 1171-1179, 2011.
[12] Zhao Yanyan, Qin Bing, Che Wanxiang, et al. "Appraisal Expression Recognition Based on Syntactic Path," *Journal of Software*,vol.22, no.5, pp.887-898, 2011. (in Chinese with English abstract).
[13] Zhao Xin, Li Xiaoming. "Research on the application for topic model in text mining," In: Technical Report PKU-CS-NCIS-TR2011XX, 2011(in Chinese with English abstract).
[14] Huang Changning, Yuan Chunfa, Pan Shimei. "Corpus, Knowledge Acquisition and Parsing," *Journal of Chinese Information Processing*,vol. 6, no.3, pp.3-8, 1992. (in Chinese with English abstract).
[15] Ni Maoshu, Lin Hongfei. "Mining Product Reviews based on Association Rule and Polar Analysis," In: Zhu QM, et al. eds., Proceedings of the NCIRCS 2007. pp. 628-634(in Chinese with English abstract).
[16] Jian Pin, Zong Chenqin, "Layer Based Dependency Parsing by Sequence Labeling Models," *Journal of Chinese information processing*, vol. 24, no.6, pp. 14-22, 2010 (in Chinese with English abstract).
[17] Che Wanxiang, Zhang Meishan, Liu Ting. "Active Learning for Chinese Dependency Parsing," *Journal of Chinese information processing*, vol.26, no.2, pp. 18-22, 2012 (in Chinese with English abstract).
[18] Griffiths T L, Steyvers M, Tenenbaum J B. "Topic in Semantic Representation," *Psychological review*,vol.114, no.2, pp. 211-244, 2007.
[19] Wang Hao, Li Yiping, Feng Zhuonan, et al. "Retweeting Analysis and Prediction in Microblogs: An Epidemic Inspired Approach," *China Communications*, vol.10, no.3, pp. 13-24, 2013.
[20] Sun Yan. "Research on Problems for Text Sentiment Analysis Oriented to Content Security," [Dissertation for Ph. D]. Wuhan: Naval University of Engineering ,2012 (in Chinese with English abstract).

Biographies

Yonggan Li, is currently a Ph.D. candidate at the School of Computer, Wuhan University, Wuhan, China. His current research interests include Weibo analysis, social network analysis and informatics. Email: lyglcn@163.com.

Xueguang Zhou, the corresponding author, e-mail: zxg196610@hotmail.com. He received his PhD in information security from Wuhan University. He is a professor and PhD supervisor of the Department of Information Security, Naval University of Engineering. Senior member of China Computer Federation. His current research interests include information content security and social network analysis.

Yan Sun, received her PhD in Communication Engineering from Naval University of Engineering. Member of China Computer Federation. Her current research interests include Weibo analysis and communication security.

Huanguo Zhang, Professor and PhD supervisor of the school of Computer, Wuhan University. Senior member of China Computer Federation. His current research interests include information security and trusted computing.