Exploring Effective Methods for On-line Societal Risk Classification and Feature Mining

Nuo Xu[1,2] and Xijin Tang[1,2(✉)]

[1] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
xunuo1991@amss.ac.cn, xjtang@iss.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China

Abstract. China has to face lots of societal conflicts during periods of social and economic transformation. It is crucial to exactly detect societal risk for the mission to a harmonious society. On-line community concerns have been mapped into respective societal risks and support vector machine model has been used for risk multi-classification on Baidu hot news search words (HNSW). Different from traditional text classification, societal risk classification is a more complicated issue which relates to socio-psychology. Conditional random fields (CRFs) model is applied to access to societal risk perception more accurately. We regard the risks of all the terms throughout a hot search word as a sequential flow of risks. The experimental results show that CRFs model has superior performance with capturing the contextual constraints on HNSW. Besides, state features can be extracted based on CRFs model to study distributions of terms in each risk category. The distribution rules of geographical terms are found and summarized.

Keywords: Societal risk classification · HNSW · Paragraph Vector · Conditional random fields · Feature mining

1 Introduction

In the Web 2.0 era, Internet users are both content viewers and content producers. Search engines have been the most common tool to access to information. The contents of high searching volume of search engine reflect the netizens' attention. Baidu is now the biggest Chinese search engine. Baidu hot news search words (HNSW) are based on real-time search behaviors of hundreds of millions of Internet users and released at Baidu News Portal, reflecting the Chinese current concerns and ongoing societal topics. In such way, we utilize HNSW as a perspective to analyzing societal risk which refers to the risk problems raising the concern of the whole society. Traditional research on societal risk was studied from the angle of cognitive psychology based on the psychometric paradigm and questionnaires [1], which is generally expensive and time-consuming to be conducted. Zheng et al. constructed a framework of societal risk indicators including 7 categories which are national security, economy/finance, public morals, daily life,

X. Cheng et al. (Eds.): SMP 2017, CCIS 774, pp. 65–76, 2017.
https://doi.org/10.1007/978-981-10-6805-8_6

social stability, government management, and resources/environment [2]. Tang tried to map HNSW into either risk-free event or one event with risk label from the 7 risk categories and aggregate all risky events over the whole concerns as the on-line societal risk perception [3]. By labeling those HNSW with relevant societal risk categories, we may get a general perception of online societal risks. An automated way to carry out societal risk classification by machine learning is necessary. Moreover, the results directly affect the accuracy of evaluating the level of societal risk. It is of great significance to monitor societal risk timely and efficiently.

This paper focuses on two points of the societal risk classification problem. Firstly, societal risk classification is a more complicated issue which relates to socio-psychology compared with traditional text classification. Different individuals may have different subjective perception of risks. Meanwhile, more challenges are confronted including the emerging words with risks, the transfer of the word's risk and widely usage of argots and proverbs [3]. Besides, the data set of societal risks is seriously unbalanced. More than 50% of the hot words are labeled as "risk-free". Therefore, improve the performance of automatic risk identification by traditional machine learning methods is with a big challenge. Secondly, HNSW are short texts with no punctuations and spaces, which makes it more difficult to deal with. Relevant news texts are crawled and extracted simultaneously to provide corpus for machine learning. Experiments were conducted which carried out societal risk multiple classifications on news contents, while the accuracy was barely needed to be improved [4]. As a result of these two points, conditional random fields (CRFs) model is firstly applied to societal risk classification directly dealing with short texts without news texts compared with previous studies. We regard the risk classification as a sequence labeling problem and use CRFs model to capture the relations among terms in hot words. In this paper, support vector machine (SVM) based on Paragraph Vector is also introduced in order to get better results of risk classification. SVM based on bag-of-words (BOW) used in previous study is chosen as baseline [4].

This paper is organized as follows: Sect. 2 introduces different models for societal risk multi-classification of Baidu hot news search words. Section 3 presents the risk multi-classification experiments and carries out the results analysis. Section 4 illustrates feature terms analysis in each risk category according to state features of CRFs model. Conclusions and future work are given in Sect. 5.

2 Societal Risk Classification Methods

Baidu hot news search words are provided in forms of 10 to 20 hot query news words updated every 5 min automatically which refer to bring the most search traffic. Each of HNSW corresponds to 1–20 news whose URLs are at the first page of hot words search results, as shown in Fig. 1. "HotWord Vision 2.0" was developed to hourly download HNSW and their corresponding news texts since November of 2011. HNSW serve as an instantaneous corpus to maintain a view of netizens' empathic feedback for social hotspots, etc. Therefore, we utilize HNSW as a perspective to analyze societal risk. The task for societal risk classification

is conducted from two perspectives. On one hand, we map these Baidu hot news into eight categories. One hot search word belonging to one risk category is determined by the votes of risk categories for hot news. On the other hand, we directly map one hot search word into one risk category. Two different approaches to the societal risk classification will be discussed in the following subsections.

Fig. 1. HNSW released at Baidu News Portal and the corresponding news texts

2.1 Societal Risk Classification Based on Baidu Hot News

We try to investigate multi-classification problem of societal risk through mapping Baidu hot news texts into eight categories. Generally the most common fixed-length vector representation for texts is the BOW. Hu and Tang carried out multiple classifications on hot news utilizing SVM algorithm based on BOW [4]. What kinds of risk categories the HNSW belong to are determined by the largest number of risk categories of Baidu hot news. However, with the volume of news accumulated, BOW method is prone to dimension disaster. Besides, BOW method does not take semantic of the sentence and word order into consideration. Neutral networks approaches have overcome these problems by implementing unsupervised word embedding for feature representations [5]. Paragraph Vector model was proposed as an unsupervised framework that learned continuous distributed vector

representations for pieces of texts [6]. The texts can be of variable-length, ranging from sentences to documents. Chen and Tang had applied Paragraph Vector model to societal risk classification on the corpus of posts crawled from Tianya Forum and the performance was better than basic machine learning methods [7,8]. Paragraph Vector model has demonstrated obvious superiority in the issue of text classification with its merits in capturing the semantics of paragraphs. Therefore, we adopt the learning algorithm Paragraph Vector for societal multi-classification on news contents.

As to the Paragraph Vector, the vector of a paragraph is concatenated with several word vectors from the paragraph and the following word is predicted in the given context [6]. The process of implementing societal risk classification based on Baidu hot news by Paragraph Vector is illustrated in Fig. 2. Take the hot search word "安徽多县遭遇虫灾" (Many counties had suffered pests in Anhui) for example. First, the Baidu hot news ID and the corresponding news text are fed to Paragraph Vector model. After the vector representations have been learned by Paragraph Vector model, n-dimensional vectors of Baidu hot news are acquired. Next, the risk categories are concatenated with the vectors of Baidu hot news which are extended to $n + 1$ dimensions. Finally, train SVM classifiers based on $(n + 1)$-dimensional vectors for prediction. The categories the hot search words belonging to are dependent on the votes of risk categories of Baidu hot news.

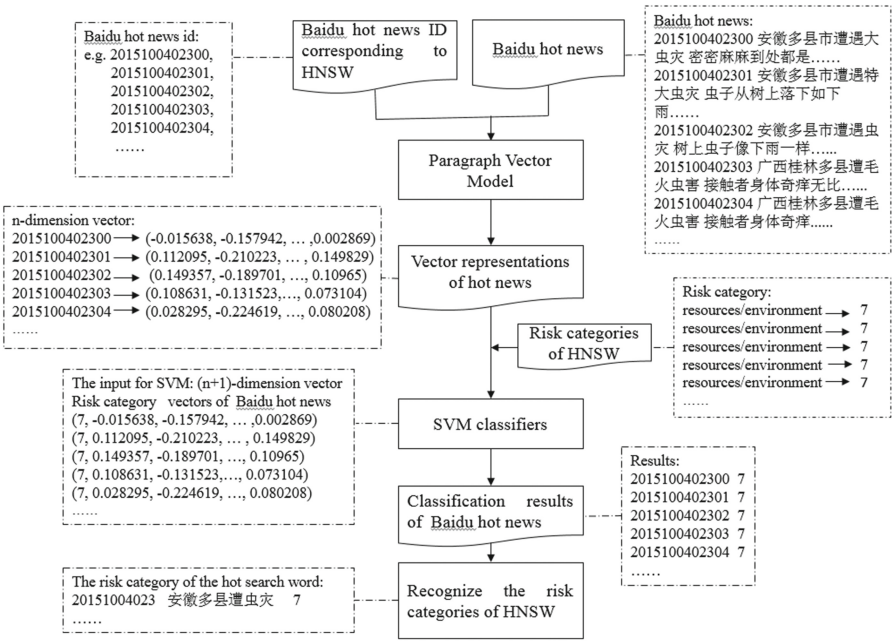

Fig. 2. Process for risk classification of HNSW by Paragraph Vector

2.2 Societal Risk Classification Based on Hot Words

Most of researchers focus on how to extract useful textual features for text classification using traditional machine learning algorithm as well as deep learning. Since HNSW consist of fewer words, traditional classification methods face the challenges of feature sparseness. Thus, CRFs model is adopted to deal with this problem.

CRFs model is an undirected graphical model used to calculate the conditional probability of a set of labels given a set of input variables [9], which has better performance in most natural language processing (NLP) applications, such as sequence labeling, part-of-speech tagging, syntactic parsing, and so on. Both maximum entropy and hidden Markov model, which are regarded as the theoretical foundations of CRFs model, have been successfully applied to text classification and achieved good performance [10,11]. CRFs model was previously used for short text classification and sentiment classification. The results proved that CRFs outperformed the traditional methods like SVM and MaxEnt [12–15]. In this paper, we utilize CRFs model for societal risk classification. For capturing the contextual influence, we treat original societal risk classification as a sequence labeling problem.

Linear chain conditional random field (LCCRF) is the most simply and commonly used form of CRFs model. We choose LCCRF to carry out societal risk classification. HNSW and their corresponding risk categories are respectively represented as the observed sequences and state sequences. In view of the risk classification process, $X = (x_1, x_2, \cdots, x_n)$ is a set of input random variables. $Y = (y_1, y_2, \cdots, y_n)$ is a set of random labels. We have a collection of hot search words sequences D where each hot search word sequence $d \in D$ is a sequence of tuples $[(x_1, y_1), (x_2, y_2), \cdots, (x_T, y_T)]$. Each tuple$(x_T, y_T)$ is respectively presented as segmented word x_T and risk label y_T. The sequence length T varies for each sequence. For example, the hot search word "河北连日强降雨" (There are heavy rainfalls in Hebei for days.) can be expressed as the observation sequence and state sequence. The observation sequence is X = (河北,连日,强,降雨). The state sequence is Y = (resources/environment, resources/environment, resources/environment, resources/environment). Hot search words and their corresponding risk categories can be turned into the risk tagging sequence. In a given observation sequence X, the probability distribution of generating the output sequence can be described as follows:

$$P_w(Y|X) = \frac{exp(w \cdot F(Y, X))}{\sum_Y exp(w \cdot F(Y, X))}. \tag{1}$$

Here, $F(Y, X) = (f_1(Y, X), f_2(Y, X), \cdots, f_K(Y, X))^T$ is the feature vector, where $f_i(Y, X)$ is a binary indicator feature function with $f_i(Y, X) = 1$ when both the feature and label are presented in a hot word and 0 otherwise; w is a learned weight for each feature function as well as the main parameter to be optimized. Figure 3 shows the framework for CRFs applied on risk multi-classification.

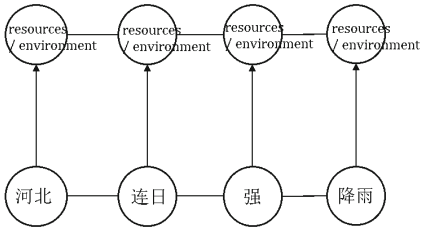

Fig. 3. Label sequences in CRFs model training

It is necessary to define the template for feature exaction to train LCCRFs model. We use the example above to illustrate the process of feature generation. Assume the current token is "强", the feature templates and corresponding features are defined as Table 1.

Table 1. Feature template and corresponding features

Template	Implication	Feature
U00：%x[-2,0]	the second term before current token	河北
U01：%x[-1,0]	the previous term	连日
U02：%x[0,0]	current token	强
U03：%x[1,0]	the previous term	降雨
U04：%x[2,0]	the previous term	/

We define two variables, namely L and N. L represents the number of categories including 7 risk categories and risk-free category, N represents the number of features generated by the template. There are $L * N$ feature functions, that is to say, there are 80 feature functions in the above example. The training of CRFs is based on maximum likelihood principle. The log likelihood function is

$$L(w) = \sum_{Y,X}[\widetilde{P}(Y,X)w \cdot F(Y,X) - \widetilde{P}(Y,X)log\sum_{Y} exp(w \cdot F(Y,X))]. \quad (2)$$

Limited-memory BFGS (L-BFGS) algorithm is used to estimate this nonlinear optimization parameters.

3 Societal Risk Classification of HNSW

3.1 Data Description and Data Processing

In this paper, we perform risk multi-classification respectively on Baidu HNSW collected from November 1, 2011 to December 31, 2016 and corresponding news corpus collected from April 1, 2013 to December 31, 2016 based on "HotWord

Table 2. Descriptive statistics of hot words and hot news

Risk category	Train dataset		Test dataset	
	#hot words	#hot news	#hot words	#hot news
National security	2258	18472	178	1568
Economy/finance	1222	8403	119	1205
Public morals	3368	25004	399	3440
Daily life	4920	32037	656	5870
Social stability	5342	58890	364	3087
Government management	5552	52748	339	3428
Resources/environment	1716	14653	358	3156
Risk-free	24587	274669	4855	42978

Vision 2.0". Table 2 shows the quantity distribution of each risk category respectively on hot search words and hot news.

We choose Ansj[1] as the segmentation tool to deal with hot words and corresponding news texts. We then remove stopwords and only reserve verbs, nouns, adjectives and adverbs. In the experiment, CRFs model, SVM based on Paragraph Vector and SVM based on BOW are compared as follows:

(1) **SVM-BOW:** We use SVM model based on vector representation BOW for text. The feature extraction method Chi-square is chosen, and the top 20% features are selected. We use LinearSVC in sklearn package for SVM model, whose parameters are set as default values. We then choose the news texts from April 1, 2013 to December 31, 2015 as the training set while all the news in 2016 as the testing set. The votes of risk categories of hot news identify which categories the hot search words belong to.

(2) **SVM-PV:** We perform news texts from April 1, 2013 to December 31, 2016 to learn vector representations. We also choose LinearSVC in sklearn package for SVM, whose parameters are set as default values. Once the vector representations have been learned, we feed them to the SVM to predict the risk label. The process is as shown in Fig. 2. The parameters are set as follows: the learned vector representations are set 100 dimensions, the optimal window size is 8, CBOW is chosen for vector representations. The votes of risk categories of hot news identify which categories the hot search words belong to.

(3) **CRFs:** Each hot word is represented as a label sequence. The template defined in Sect. 2.2 is chosen for feature extraction. We then choose the hot words from November 1, 2011 to December 31, 2015 as training set, while all the hot words in 2016 as testing set. L-BFGS algorithm is introduced to optimize the objective function. We use *sklearn_crfsuite*[2] package for CRFs model. We set the iteration number to 100 in the training process of the method based on CRFs.

[1] http://www.demo.ansj.com/.
[2] https://pypi.python.org/pypi/sklearn-crfsuite/0.3.3/.

3.2 Results

We utilize accuracy, macro-average and micro-average as the evaluation metrics to evaluate the overall performance of each model. Precision, recall and F-measure are used for performance measurement of each societal risk category. The accuracy of CRFs model, SVM based on Paragraph Vector and SVM based on BOW are 0.78, 0.68 and 0.74 respectively. CRFs have achieved the best performance. Table 3 shows the results of three models.

Table 3. Comparison results with different models

Risk category	Precision			Recall			F-measure		
	BOW	PV	CRFs	BOW	PV	CRFs	BOW	PV	CRFs
National security	0.56	0.00	0.66	0.29	0.00	0.43	0.38	0.00	0.52
Economy/ finance	0.58	0.00	0.68	0.16	0.00	0.34	0.25	0.00	0.46
Public morals	0.43	0.00	0.54	0.14	0.00	0.23	0.21	0.00	0.32
Daily life	0.63	0.89	0.70	0.25	0.01	0.51	0.36	0.03	0.59
Social stability	0.47	0.40	0.56	0.49	0.37	0.51	0.48	0.39	0.54
Government management	0.49	0.52	0.58	0.35	0.20	0.56	0.41	0.29	0.57
Resources/ environment	0.82	0.87	0.93	0.54	0.12	0.56	0.65	0.22	0.70
Risk-free	0.78	0.70	0.81	0.94	0.98	0.93	0.85	0.82	0.87
Macro-average	0.60	0.42	0.68	0.40	0.21	0.51	0.47	0.28	0.58
Micro-average	0.74	0.68	0.78	0.74	0.68	0.78	0.74	0.68	0.78

As is shown, SVM-PV has got rather poor performance on both precision and recall especially for the risk category "national security", "economy/finance" and "public morals". The phenomena are found in Table 2 that the corpus generated by the netizen's online search behavior is severely unbalanced, the "risk-free" category takes the absolute majority in the corpus. Besides, there is little difference between the semantic information of two corpora from different kinds of societal risk categories, such as "public morals" and "risk-free", which leads to a high probability classifying a hot search word to the majority category. For the risk category "national security", "economy/finance" and "public morals", although the precision and recall of SVM-PV on hot news are not zero, the votes of risk categories of hot news causes no sample to be correctly labeled on hot search words. As far as the task of societal risk classification is concerned, it is essential to find out risky words as many as possible. In other words, we pay more attention to recall for evaluation. The recall of SVM based on Paragraph Vector on "risk-free" category is 0.98, tending to find hot words whose categories are risk-free. In contrast, the recall of

CRFs model on risk category "national security", "economy/finance" and "public morals" are respectively 0.43, 0.34 and 0.23. Meanwhile, the three values are in turn increased by 0.48, 1.10 and 0.64 compared with the SVM-BOW. In other words, CRFs model tends to capture risky hot words. As can be seen from the overall scores of the whole data for three methods, CRFs method achieves better performances in each risk category than the other two methods apparently. Overall, CRFs model shows the discriminatory power of predictive models in societal risk multi-classification. Moreover, it has obvious superiority in data processing which is relatively easy and captures comprehensive text semantics.

4 Analysis of Feature Terms on Societal Risk

CRFs model has demonstrated its superiority for risk multi-classification. Besides, we obtain state features after CRFs model completing parameters learning on the training set. The state features can be expressed as the distribution of terms' weight values in each risk category. The magnitudes of the weight values represent their contribution to predicting which risk categories the hot words belonging to. As a result, we could select terms with greater weight values in each risk category as the factors or characteristics on behalf of each risk.

4.1 Analysis of Feature Weight

We now perform feature terms analysis in each risk category according to state features and their corresponding weight values. The corpus is chosen from November 1, 2011 to December 31, 2016 including 56,233 hot news search words for training. When the training process is completed, the state features and weight values will be expressed as the distribution of terms' weight values in each risk category. The occurrence frequency of term "雾霾" (haze) under "daily life", "government management", "resources/environment" and "risk-free" are respectively 1, 2, 105, 7. And the corresponding weight values are -0.35, 0.71, 6.65, -0.05. The significance of these weight values can be explained from an aspect of CRFs formula. Since $\sum_Y exp(w \cdot F(Y, X))$ is the normalization factor, values of $P_w(Y|X)$ depend on values of $exp(w \cdot F(Y, X))$. Take "haze" for example, we assume that the sequence only has one word "haze" for simplicity.

$$P_{w_3}(resources/environment|haze) > P_{w_2}(governmentmanagement|haze)$$

$$> P_{w_4}(risk - free|haze) > P_{w_1}(dailylife|haze).$$

As is seen, weight values represent the contribution to the prediction of risk category. The higher the weight values of terms in one risk category, the greater the contributions of terms to conditional probability. For instance, the weight values of "房价" (house prices) in "finance/economy" and "daily life" are respectively -0.82 and 4.88. The larger weight value of "房价" (house prices) contributes greater to conditional probability on "daily life". Then we try to investigate the distribution pattern of place names in feature terms in each risk category.

4.2 Distribution of Feature Terms

We first use Ansj to do the Chinese hot news search words segmentation and part-of-speech tagging. Terms that are tagged "ns" (geographical name) and "nt"(institutional name) are selected. Then we build the dictionary of Chinese regional areas which has a total of 34 provinces, including provinces, municipalities and autonomous regions. At last, the geographical terms with their weight values in each category are picked out according to the dictionary. The distribution pattern of geographical terms in each category is as shown in Fig. 4. The horizontal axis is the geographical terms, while the vertical axis is the eight risk categories. Each small colored cell in the figure represents weight values of the geographical terms in each category. The deeper the color is, the higher the weight value is. Here we list and analyze geographical terms results for illustration. By the visualized results, we summary the distribution patterns of geographical terms in each category as follows:

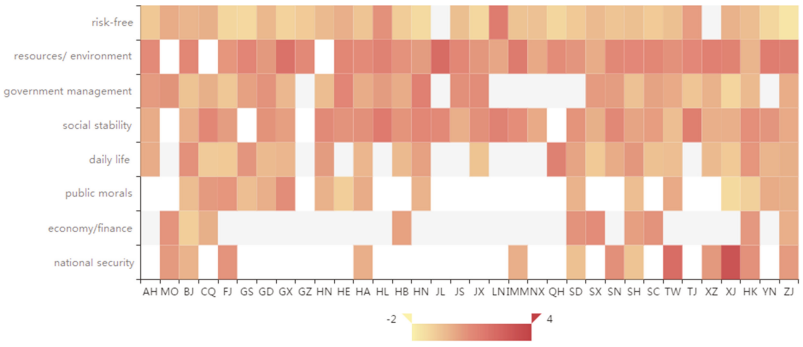

Fig. 4. Distribution pattern of regional terms in each category

(1) As to the risk of national security, the highest weight value of feature terms is Xinjiang, with Taiwan, Hong Kong, Fujian and Tibet decreasing in turn. This is because there have been hundreds of terrorist attacks happened in Xinjiang in recent years, including hijacking plane and attacking the police station so on. Terrorist attacks pose a great threat to social stability of Xinjiang and national security. In addition, there are patriotic movements like protecting the Diaoyu Islands and the South China Sea happened in Taiwan. Major political events such as sanctions against the Philippines and impeaching Chun-ying Leung have occurred in Hong Kong;
(2) As to the risk of finance/economy, weight values of Shanxi, Sichuan, Hong Kong and Shanghai decrease accordingly. Since anti-corruption movement in Shanxi, the economy of Lvliang city had crashed. Sichuan Province launched four trillion investment plans. Hong Kong and Shanghai are often mentioned in the risk of finance/economy owing to the Shanghai Composite Index and the Hong Kong Hang Seng Index, both of which may reflect the situation of stock market volatility to some extent;

(3) As to the risk of public morals, the issues such as integrity and social mode in Guangxi, Fujian, Chongqing and Henan are more salient and could not be neglected;
(4) As to the risk of daily life, Beijing and Shanghai mainly focus on housing issues such as property restriction and the rising price;
(5) As to the risk of social stability, weight values of Heilongjiang, Tianjin and Liaoning are higher relative to other areas in China. That is because the events such as coal mine explosion, the explosion of Tianjin harbor and the school bus rollover accident occurred respectively in Heilongjiang, Tianjin and Liaoning;
(6) As to the risk of government management, a number of top officials from provinces including Hunan, Hebei, Jiangxi, Guangdong and Shanxi were investigated by the commission for discipline inspection of the central committee due to the tough anti-corruption policy;
(7) As to the risk of resources/environment, there are earthquakes frequently occurred in Jilin, Yunnan and Tibet. And the snowstorm occurred in Inner Mongolia in November, 2012. As is known, haze pollution in Beijing is also prominent.

5 Conclusions

Societal risk refers to the risk problems raising the concerns of the whole society. The subjective perception of societal risk reflects the public attitudes to social issues as well as government decision-making. It is of great significance for government management and decision-making to monitor either the potential or the ongoing societal risk events. In this paper, CRFs and SVM-PV model are applied to obtain the subjective societal risk perception automatically and timely.

We conduct the research on societal risk perception based on HNSW. According to the current research, CRFs model is more effective in response to "subjective perception of societal risks" and "short texts". The main contributions are summarized as follows.

(1) CRFs model is first applied to societal risk classification directly dealing with short text, which tackles the challenge of feature sparseness and improves the performance.
(2) CRFs model is used to capture the contextual constraints on HNSW with obvious superiority in text processing.
(3) The geographical distribution rules of societal risks are found and summarized by studying distributions of place terms in each risk category by state features.

Lots of works need to be done. In the future, the combination of feature representation and CRFs will be developed to improve the performance. Besides, terms with greater weight values in each risk category may also be picked out either as the factors on behalf of risk events or as feature words to construct the risk lexicon.

Acknowledgments. This research is supported by National Key Research and Development Program of China (2016YFB1000902) and National Natural Science Foundation of China (61473284 & 71371107).

References

1. Xie, X.F., Xu, L.C.: The study of public risk perception. Psychol. Sci. **6**, 723–724 (2002). (in Chinese)
2. Zheng, R., Shi, K., Li, S.: The influence factors and mechanism of societal risk perception. In: Zhou, J. (ed.) Complex 2009. LNICSSITE, vol. 5, pp. 2266–2275. Springer, Heidelberg (2009). doi:10.1007/978-3-642-02469-6_104
3. Tang, X.J.: Exploring on-line societal risk perception for harmonious society measurement. J. Syst. Sci. Syst. Eng. **22**, 469–486 (2013)
4. Hu, Y., Tang, X.: Using support vector machine for classification of baidu hot word. In: Wang, M. (ed.) KSEM 2013. LNCS, vol. 8041, pp. 580–590. Springer, Heidelberg (2013). doi:10.1007/978-3-642-39787-5_49
5. Mikolov, T., Sutskever, l., Chen, K., Corrado, G., Dean, J.: Distributed representations of words and phrases and their compostitionality. In: Advances in Neural Information Processing Systems, pp. 3111–3119 (2013)
6. Le, Q.V., Mikolov, T.: Distributed representations of sentences and documents. Comput. Sci. **4**, 1188–1196 (2014)
7. Chen, J.D., Tang, X.J.: The challenges and feasibility of societal risk classification based on deep learning of representations. In: 2015 IEEE International Conference on Systems, Man, and Cybernetics, pp. 569–574 (2015)
8. Chen, J.D., Tang, X.J.: Societal risk classification of post based on paragraph vector and KNN method. In: Proceedings of the 15th International Symposium on Knowledge and Systems Sciences, pp. 117–123. JAIST Press (2014)
9. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: probabilistic models for segmenting and labeling sequence data. In: Proceedings of The Eighteenth International Conference on Machine Learning, ICML, pp. 282–289 (2001)
10. Nigam, K., Lafferty, J., Mccallum, A.: Using maximum entropy for text classification. In: Proceedings of the IJCAI-99 Workshop on Information Filtering, pp. 61–67. San Fransisco (1999)
11. Yi, K., Beheshti, J.: A hidden Markov model based text classification of medical documents. J. Inform. Sci. **35**, 67–81 (2009)
12. Zhao, J., Liu, K., Wang, G.: Adding redundant features for CRFs-based sentence sentiment classification. In: Conference on Empirical Methods in Natural Language Processing, pp. 117–126 (2008)
13. Sudhof, M., Goméz Emilsson, A., Maas, A. L., Potts, C.: Sentiment expression conditioned by affective transitions and social forces. In: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1136–1145 (2014)
14. Li, T.T., Ji, D.H.: Sentiment analysis of micro-blog based on SVM and CRF using various combinations of features. Appl. Res. Comput. **32**, 978–981 (2015). (in Chinese)
15. Zhang, C.Y.: Text categorization model based on conditional random fields. Comput. Technol. Dev. **21**, 77–80 (2011). (in Chinese)