The Risk Level Estimation Based on Deep Learning Method for Tianya Forum

Jindong Chen **Xijin Tang**

Institute of Systems Science, Academy of Mathematics and Systems Science
Chinese Academy of Sciences, Beijing, 100190 P.R. China
j.chen@amss.ac.cn, xjtang@iss.ac.cn

Abstract

Using the societal risk indicators from socio psychology, a deep learning method is applied to estimate the risk level of Tianya Forum. Due to the effectiveness in semantic and word order information extraction for documents, a deep learning method Post Vector is used to generate the distributed representations of BBS posts. Through the experimental comparison on societal risk classification of BBS posts, the performance of kNN based on Post Vector is superior to kNN based on Bag-of-Words, edit distance or Lucene-based search method. Therefore, with kNN based on Post Vector method and the annotated data of Tianya Zatan broad, the risk level of Baixing Shengyin broad in different months is estimated, and the reasonability of the estimated results is analyzed.

Keywords: Tianya Forum, Societal Risk Classification, Deep Learning, kNN, Post Vector

1 General Instructions

Up to date, more and more Chinese people treat social media (such as blog, micro-blog, BBS, etc.) as one way to express their opinions toward the daily phenomena and social events, so it is a better way to monitor the relative societal risk level based on these online data [1]. Through measurement of the topics and their frequency expressed online, the current relative societal risk level can be estimated. "Tianya Forum is a famous Internet forum in China, and provides BBS, blogs, micro-blogs and photo album services etc." [1]. Tianya Forum includes multiple boards; the posts on Tianya Zatan board and Baixin Shengyin board etc. mainly cover the hot and sensitive topics of current society [2]. Therefore, the boards of Tianya Forum are selected as the data sources to explore effective strategies for online societal risk monitoring.

According to comprehensive analysis and comparison [1], the framework of societal risk indicators including 7 categories and 30 sub categories based on word association tests which is constructed by Zheng et al. [3] is chosen as risk categories. To evaluate the current risk level, the main challenge is to classify each post into one of multiple societal risk categories (7 main categories and 1 risk free category). However, the massive amount and negative effects of the posts lead to the impracticability of the classification of posts by human. Since the effectiveness of machine learning method in text classification, the machine learning method is a better approach for the classification task [4].

The basic principle of text classification is utilizing machine learning strategies to assign predefined labels to new documents based on the model learned from a trained set of labels and documents [5]. Generally, two main procedures affect the accuracy of text classification: document representation and classifier construction. The traditional document representation method is Bag-of-Words. For Bag-of-Words representation, the vector size equals to the vocabulary size, the vector elements at the indexes of the words occurred in the document are "word frequency" while the other elements are "0"s [6]. Bag-of-Words representation is mainly through extraction and selection of feature word to improve the quality of document vector [7]. There are many research works have proven the effectiveness of Bag-of-Words representation in text classification field, such as news classification [8] and personality classification [9]. The representative machine learning methods for text classification are K-Nearest Neighbor (KNN) [10], naïve Bayes [11] and support vector ma-

[1] http://en.wikipedia.org/wiki/Tianya_Club

chine (SVM) [5, 8], etc.

However, the document-level societal risk classification of BBS posts is different from the normal text classification. The distinctive features of this task includes: First, the human labeling of post is unstable. As the societal risk indicators are from socio psychology, but individual cognitions are inconsistent, which leads to the unstable labeling of post. Second, the risk classification of BBS posts is document-level classification. The document-level classification not only increases the input size, but also brings more challenges (such as the big variance of the text length, the complicated syntax, the involvement of multiple topics in one document). Third, the classification is multi-class. The multiple risk classes also increase the complexity of text classification. From the state-of-the-art performance in the similar fields (such as: multiple emotion classification and fine-grained sentiment classification): the accuracies of both tasks are less than 50%, we can also imagine the complexity of the document-level societal risk classification [12, 13]. In all, it can be inferred that the document-level societal risk classification of BBS posts is more challenging.

Based on traditional Bag-of-Words representation, the machine learning method (SVM) hardly achieved the expected performance in societal risk classification, even though the training set was up to ten thousands of posts and the feature word selection method is optimized [4]. According to the analysis of Qiu et al., the main issue that hinders the performance of text classification is the drawbacks of Bag-of-Words representation, such as: the curse of dimensionality, no syntactic or semantic information and the word order information loss [14].

To solve the issues of Bag-of-Words representation, Bengio et al. proposed the distributed vector representation of word. Instead of a Bag-of-Words vector, a word is represented by a real-valued vector with a much smaller size [6]. The distributed vector representation is without the curse of dimensionality problem. Moreover, the semantic and word order information are encoded in the distributed vector space. However, the computational complexity of this strategy is originally too high for real world tasks. To improve the efficiency of training, many prominent algorithms have been proposed recently, such as: SENNA [15], Word2Vec [16] and GloVe [17], which show great improvements on the quality and efficiency of word vector construction. According to the effectiveness of the distributed vector representation for word, it is natural to develop the distributed vector representation of document for text classification. Le et al. [18] proposes a more sensible deep learning method to realize the distributed representation of paragraph or document. Combined with an additional paragraph vector, the method includes two models: PV-DM and PV-DBOW for paragraph or document representation, where the paragraph vector contributes to predict the next word in many contexts sampled from the paragraph. For the Chinese online documents representation, a deep learning method as Post Vector (PV) model is proposed, the model shows improvements for Chinese document representation [19].

Therefore, with the deep learning method as PV model, we focus on realizing the distributed representation of BBS posts, and developing societal risk classifier based on the distributed representations. Furthermore, based on the classifier, the risk level of different board of Tianya Forum during different time is estimated. The rest of this paper is organized as follows. Section 2 presents the deep learning model Post Vector. The data sets, experiment procedures and performance measures are explained in Section 3. The effectiveness analysis of kNN based on PV is presented in Section 4.The risk level estimation of Baixing Shengyin board is presented in Section 5. Finally, concluding remarks are given in Section 6.

2 Post Vector Model

The deep learning method as PV is mainly designed for the distributed representation of Chinese documents [19]. In PV framework (Figure 1), the Chinese documents of posts are segmented into words using segmentation tools. The post ID which is treated as another word is concatenated with the segmented words of the post, and combined with other words sampled from the post to predict the next word of the post. To enhance the performance of PV model, the words after the predictive word are also taken into consideration. Each post is represented by a unique vector, which is a column in post matrix $\boldsymbol{D}$ and every word of post is also represented by a unique vector, which is a column in word matrix $\boldsymbol{W}$, where $\boldsymbol{D}$ and $\boldsymbol{W}$ are real matrix, the initial values

are [-0.5/l, 0.5/l], where l is the word vector size . For the random initialization of word matrix and post matrix, large corpus is preferred for training. After the training, the word matrix and post matrix can be obtained simultaneously.

Formally, PV model can be viewed as a three-layer network: input, hidden and output. Before the model training, set the word vector size as l and window size as k. For a given post, it can be viewed as a post ID and a sequence of words: w_{ID}, w_1, w_2, w_3,…w_T, T is the number of words in context. To predict the word w_t, t=1,2, …,T, k words before or after wt are taken into input. The objective of the Post Vector model is to maximize the average log probability

$$\frac{1}{T}\sum_{t=k}^{T-k}\log p(w_t|w_{ID},w_{t-k},\ldots,w_{t+k}) \qquad (1)$$

In the training process, input features are of fixed-length and sampled from a sliding window over the document of the post. The document vector (the vector of w_{ID}) is shared across contexts generated from the same post. Hence, the document vector acts as a memory that remembers what is missing from the current context or the topic of the post. The word vector matrix W is shared by all posts. *i.e.*, *vector*(拆迁, demolition) is the same for all posts. The training process of PV model can be regarded as the process of dimension reduction of document vector.

After the training convergence, the document vectors can be used as features for the post. These features can be directly fed to conventional machine learning methods such as kNN or SVM.

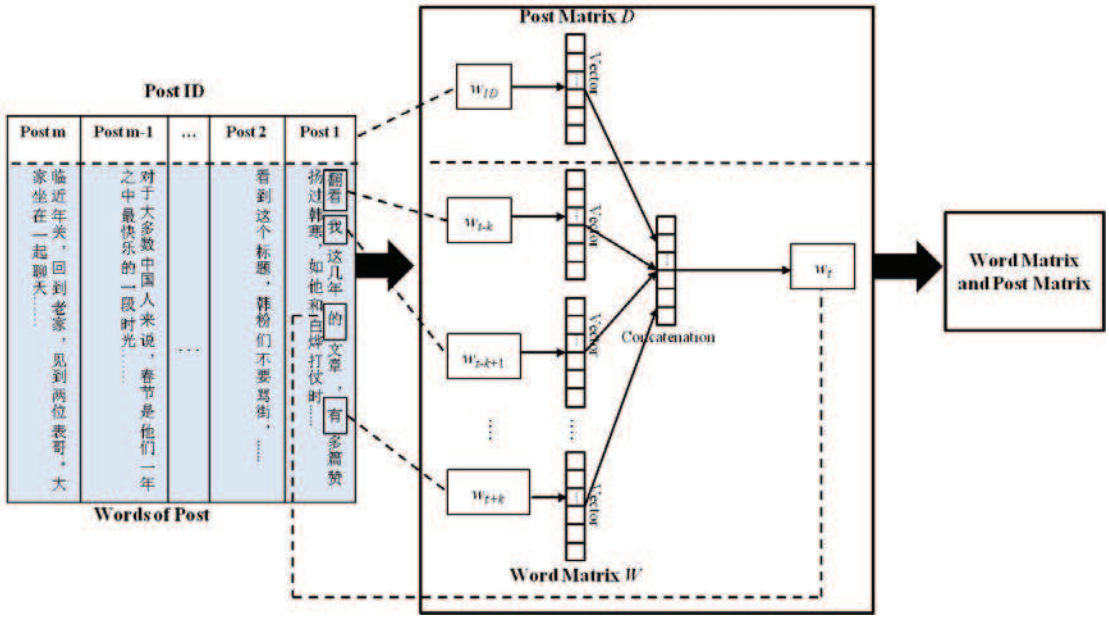

Figure 1. A framework for learning word vectors and post vectors

How does this kind of model help in societal risk classification? Two examples of Chinese post documents are presented to show the advantages of PV model in document representation for societal risk classification.

Example 1: Same words with different order, but different societal risk

i) “我^走在^路上^,^汽车^不小心^撞了^我^一下^.^真^倒霉…(I was walking on the street, I was hit by a car accidentally, what a bad luck...)”

ii) “我^走在^路上^,^我^不小心^撞了^汽车^一下^.^真^倒霉…(I was walking on the street, I hit a car accidentally, what a bad luck...)”

The first document describes a transportation accident, so its risk category is daily life. The second document is the author complains the bad luck, and labeled as risk free. However, after the segmentation of documents, the vectors of two Chinese documents that constructed by the BOW are same, while their risk labels are different, which then confuses the classifier. In contrast to BOW, PV generates two different vectors for both documents, and makes it easy for classifier to learn the difference between the two documents. Hence, PV is effective to represent the differences in documents.

Example 2: Similar words, same societal risk

i) “企业^偷排^污水^,^造成^大量^鱼^死亡…(An enterprise surreptitiously drains polluted water, which causes massive death of fishes…)”

ii) “公司^非法^倾倒^废料^,^污染^严重… (A company illegally dumps wasted materials, which causes heavily pollution…)”

Both documents mention about the environment pollution, and are labeled to the same societal risk: Resources & Environments. After the segmentation of documents, it can be found that no words are shared between the two documents. Based on BOW, the vectors of two documents are totally different, which makes the classifier lose the similarity between the two documents. However, by PV model, the similar words shared more information in word vectors. After the iterative training, the similarities of words are embodied in the post document vectors. The vectors of these two documents share more information, which is easier to classify these two documents into one risk category. Hence, PV is more useful to identify the similarity of documents.

Above two examples illustrate that PV model can extract useful features for societal risk classification, due to the improvements in the word order and semantic understanding of document.

3 Data Sets, Experimental Procedure and Performance Measures

This section introduces the data sets, experiment procedures, as well as the performance measures for the classification results of cross-validations.

3.1 Data sets

With Tianya Forum spider system of our group [20], the daily new posts and updated posts are downloaded and parsed. To train post vector model, the new posts (title + text) during November of 2011 to March of 2013, 16-month new posts every day, more than 470 thousands posts, are used.

To compare the effectiveness of different methods in societal risk classification, the labeled posts of Dec. 2011-Mar. 2012 are used. The amount of posts of these four months and the amount of posts in different societal risk categories of each month are presented in Table 1.

Table 1. The risk distribution of posts on Tianya Zatan board of different months

Period / Risk Category	Dec.2011	Jan.2012	Feb.2012	Mar.2012
Total	12125	12032	20330	37946
Risk free	1278	2047	2645	14569
Government Management	3373	1809	3099	6879
Public Morality	3337	3730	8715	6065
Social Stability	954	1013	1746	2108
Daily Life	2641	3063	3142	6920
Resources & Environments	223	147	309	329
Economy & Finance	248	133	460	609
Nation’s Security	71	90	214	467

The figures in Table 1 show the risk distributions of the posts are unbalanced. The posts on Tianya Zatan mainly concentrate on risk free, government management, public morality and daily life, the posts of these categories are more than 85% of all posts.

3.2 Experimental procedure

Generally, considering the most opinions of people is a good approach to minimize the influence of the inconsistency of individual attitudes. To directly synthesize the most opinions of people, KNN method is adopted for societal risk classification. Here is the procedure of KNN classification based on Post Vector:

1) All the post corpuses are segmented with Ansj tool2.
2) The post ID is concatenated with the segmented words of the post.
3) All the corpuses are fed into Post Vector model to generate the post matrix.
4) The post document vectors are extracted from the post matrix.
5) The risk label of each post is combined with the post document vector to consist of the testing data set for kNN.

[2] Ansj tool is a JAVA package based on inner kernel of ICTCLAS. https://github.com/ansjsun/ansj_seg

6) kNN classification method is applied to document-level multi-class societal risk classification on the data set.

3.3 Performance measures

The precision, recall and F-measure are used for performance measurement. For multiple classes, the macro average and micro average are used for evaluation. The macro average and micro average on precision, recall and F-measure are computed as follow [13].

$$\begin{cases} Macro\_Precision = \frac{1}{M}\sum_{i=1}^{M}\frac{T_i}{C_i} \\ Macro\_Recall = \frac{1}{M}\sum_{i=1}^{M}\frac{T_i}{N_i} \\ Macro\_F = \frac{2\times Macro\_Precision\times Macro\_Recall}{Macro\_Precision + Macro\_Recall} \end{cases} \quad (2)$$

$$\begin{cases} Micro\_Precision = \sum_{i=1}^{M}T_i \Big/ \sum_{i=1}^{M}C_i \\ Micro\_Recall = \sum_{i=1}^{M}T_i \Big/ \sum_{i=1}^{M}N_i \\ Micro\_F = \frac{2\times Micro\_Precision\times Micro\_Recall}{Micro\_Precision + Micro\_Recall} \end{cases} \quad (3)$$

where T_i is the number of posts correctly classified in each category, C_i is the number of posts classified in each category, N_i is the number of posts manually annotated in each category, M=8 means 8 classes of societal risks are taken into performance measure.

4 The Effectiveness Analysis of Post Vector

Several experiments are carried out to verify the effectiveness of post vector. The computer is Dell Optiplex 9020, CPU is 8*3.4GHz, the memory is 8G, and all the experiments are run on Ubuntu 12.02.

To improve the performance of kNN based on Post Vector, three parameters are optimized: vector size, k, window size. Through 5-fold cross validation on the data set of Table 1, the parameters of kNN based on Post Vector are set as: *vector size*=250, k=40 and *window size*=3.

To show the effectiveness of Post Vector, the results of kNN classification based on Bag-of-Words, Edit Distance and Lucene-based search method are compared. The k value of KNN method is set as 40, and 5-fold cross validation is applied to all the experiments.

The procedures of kNN classification based on Bag-of-Words: first, all post texts are segmented by Ansj; second, to avoid the big document-word matrix, the cosine similarity based on Bag-of-Words representation is obtained by one-against-one fashion with R "tm" package; third, sorting the similarities of the testing post to all voting posts, the top 40 similar posts are selected; fourth, the majority vote of the 40 posts is used to label the testing post. Each fold of kNN classification using Bag-of-Words distributes on 15 R Consoles, but the running time is still more than 10 days.

To accelerate the speed of finding top k similar posts, two other approaches are tried: one approach is to try Edit Distance; the other is to adopt the available search tool, such as Lucene, directly.

Edit distance: the minimum operations required to transfer the source post to the target post only by edit including insert, delete and substitute. Normally, through dynamic programming, the minimum edit step can be achieved. Without multiplication and square root calculation in the process, the speed of Edit Distance is much faster than Bag-of-Words, each fold of kNN classification takes almost 3 days on 5 R Consoles.

Lucene-based search method: Lucene is a free open source information retrieval software library. We apply Lucene to search the top 40 similar posts of the testing post from the index. The main steps of Lucene-based search method: segmentation of post text, creation of index, and retrieving post. The main time cost of Lucene-based search method is indexing. Considering the time of indexing and retrieving, each fold of kNN classification takes almost 6 hours.

The *Macro_F* and *Micro_F* comparisons of KNN classification based on these four methods are presented in Table 2.

Table 2. The performance comparisons of different methods

Method	Post Vector	Bag-of-Words	Edit Distance	Lucene-based Search
Macro_F	47.72%	37.76%	39.17%	45.32%
Micro_F	55.43%	44.35%	47.80%	48.45%

Table 2 shows the performance of the KNN based on Post Vector is much better than the KNN based on Bag-of-Words, the increment of Macro_F and Micro_F are 9.96% and 11.08% respectively. Both Edit Distance and Lucene-based Search perform better than Bag-of-Words, while not up to Post Vector. The KNN computational speed based on Post Vector is much faster than Bag-of-Words, Edit Distance and Lucene-based Search. Again, we say that post vector is more effective in document-level multiple societal risk classification.

Bag-of-Words only contains the appearance and frequency of words, ignoring the semantics and the word order information of documents, so the performance of KNN based on Bag-of-Words is worst. Through edit operations, Edit Distance measures the difference of the character order of documents, but without semantic information and segmentation of Chinese document, the improvement of Edit Distance is limited. Lucene-based Search is actually based on vector space model and Boolean model, and Tf-idf processing is also adopted for vector space model. The performance of KNN based on Lucene-based Search is much improved, but the loss of word order information of paragraph, the performance is still worse than Post Vector. Post Vector is trained through fixed-length and sampled from a sliding window over the paragraph. By this training fashion, the semantics and word order information are extracted and mapped into the fix-length post document vector. Therefore, KNN classification based on Post Vector produced the best results.

5 The Risk Level Estimation of Baixing Shengyin Board

To estimate the daily risk level of Baixing Shengyin board during February 2015 to March 2015, the risk categories of new posts on Baixing Shengyin board should be classified first. To classify the risk categories of new posts on Baixing Shengyin board, the labeled new posts of Tianya Zatan board during December 2011 to March 2012 are selected as training set, and kNN based on Post Vector is chosen as the classification method. The process of the post document vector generation for new posts of Baixing Shengyin board using Post Vector model includes: i) the new posts of Baixing Shengyin board during February 2015 to March 2015 and Tianya Zatan board during December 2011 to March 2013 are combined as the training set; ii) all the post corpuses are segmented with Ansj tool and assigned unique ID; iii) the corpuses are fed into Post Vector model for training; iv) extraction of the post document vectors.

Based on the post document vectors, the societal risk classification of new posts on Baixing Shengyin board is realized through kNN method. The k value is set as 40. After the classification, the daily ration of risk posts is computed, and treated as the risk level. The amount of daily posts and daily risk level of Baixing Shengyin board are presented in Figure 2. The amount of daily posts in each risk categories of Baixing Shengyin board are described in Figure 3.

From the results of Figure 2, the amount of daily posts on Baxing Shengyin board during February 2015 to March 2015 is between 100 and 400, the risk level is relatively high, almost 0.7. From the results of Figure 3, the risk categories of new posts of Baixing Shengyin board concentrate on risk free, government management, public morality and daily life, but less in social stability, resources & environment, economy & finance and nation's security. Normally, the topics of Baixing Shengyin board are related to the issues or troubles of common people life, so the posts mentioned about the affairs of nation or society is very few. Based on this point, it can be said that the classification results of Baixing Shengyin using kNN based on Post Vector is reasonable, and the risk level of Baixing Shengyin has reference value.

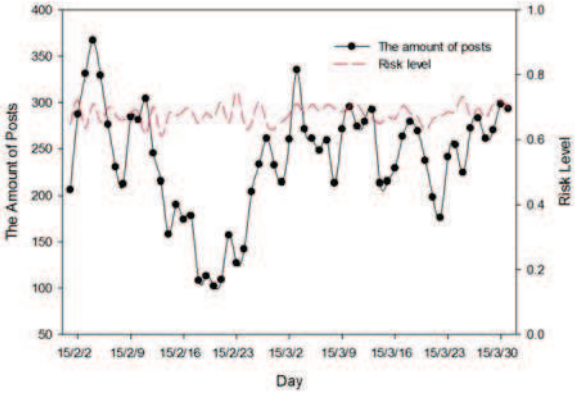

Figure 2. The daily amount of posts and risk level estimation of Baixing Shengyin board

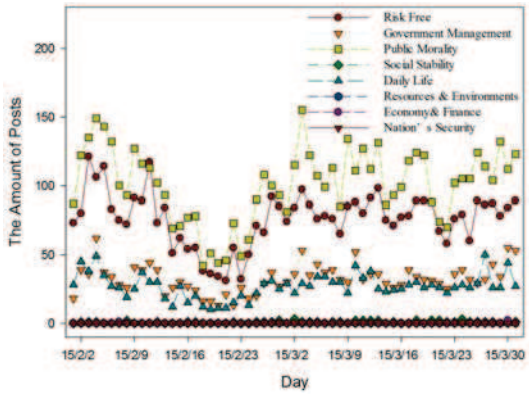

Figure 3. The daily amount of posts in each risk categories of Baixing Shengyin board

6 Conclusions

In this study, the new posts of Tianya forum are selected as the data source for societal risk monitor, but owing to the big amount and negative effect of posts, it is impractical to label the posts only relied on human. Due to the effectiveness of machine learning, it is adopted for societal risk classification of new posts, and then obtain the risk level of BBS. Several results are obtained in this study:

1) An effective deep learning method Post Vector for the distributed representation of Chinese BBS posts is applied in this study;
2) Through performance comparison, the effectiveness of kNN based Post Vector model is validated;
3) Based on the labeled pots and deep learning method, the risk level of the board of Tianya Forum can be effective estimation.

Although the improvement of Post Vector is remarkable, the performance of societal risk classification towards BBS posts is still unsatisfied. In the future, we intend to improve our classification accuracy through two ways: i) Extract more effective information from documents to represent post; ii) Develop more powerful machine learning methods in societal risk classification area.

Acknowledgment

This research is supported by National Basic Research Program of China under Grant No. 2010CB731405 and National Natural Science Foundation of China under Grant Nos.71171187, 71371107 and 61473284. The authors would like to thank other members of our team for their effort in data collection and post labeling.

References

[1] Tang X J. Exploring On-line Societal Risk Perception for Harmonious Society Measurement. Journal of Systems Science and Systems Engineering, 2013, 22(4): 469-486.
[2] Cao LN and Tang X J. Topics and Threads of the Online Public Concerns Based on Tianya Forum. Journal of Systems Science and Systems Engineering, 2014, 23(2): 212-230.
[3] Zheng R, Shi K and Li S. The influence factors and mechanism of societal risk perception. *Proceedings of the First International Conference on Complex Sciences: Theory and Application (Shanghai, China, J. Zhou eds.)*. Springer Berlin Heidelberg, 2009, 2266-2275.
[4] Chen J D and Tang X J. Exploring Societal Risk Classification of the Posts of Tianya Club. International Journal of Knowledge and Systems Science, 2014, 5(1): 36-48.
[5] Zhang W, Yoshida T and Tang X J. Text classification based on multi-word with support vector machine. Knowledge-Based Systems, 2008, 21(8), 879-886.
[6] Bengio Y, Ducharme R, Vincent P and Jauvin C. A neural probabilistic language model. Journal of Machine Learning Research, 2003, 3:1137–1155.
[7] Zhang W, Yoshida T and Tang X J. A comparative study of TF* IDF, LSI and multi-words for text classification. Expert Systems with Applications, 2011, 38(3), 2758-2765.
[8] Hu Y and Tang X J. Using support vector machine for classification of Baidu hot word. *In Knowledge Science, Engineering and Management* (KSEM2013, Dalian, China. M. Wang, et al eds.). LNCS, 8041, Springer, 2013, 580-590.
[9] Nie D, Guan Z D, Hao B B, Bai S T and Zhu T S. Predicting Personality on Social Media with Semi-supervised Learning. *The 2014 IEEE/WIC/ACM International Conference on Web Intelligence* (Warsaw, Poland). Washington: IEEE Computer Society, 2014, 158-165.
[10] Cover T M and Hart P E. Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory, 1967, 13(1):21-27.
[11] Lewis D D. Naive (Bayes) at Forty: the Independence Assumption Information Retrieval [C]. *Proceedings of the 10th European Conference on Machine Learning* (ECML'98, Chemnitz, Germany). London: Springer, 1998: 4-15.

[12] Wen S Y and Wan X J. Emotion Classification in Microblog Texts Using Class Sequential Rules. *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (Québec, Canada). 2014,187-193.
[13] Richard S, Alex P, Jean W, Jason C, Chris M, Andrew Y N and Chris P. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *EMNLP 2013: Conference on Empirical Methods in Natural Language Processing* (Seattle, Washington). Stroudsburg: Association for Computational Linguistics, 2013, 1631-1642.
[14] Qiu L, Cao Y, Nie Z Q and Rui Y. Learning Word Representation Considering Proximity and Ambiguity. *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence* (Québec, Canada). 2014,1572-1578.
[15] Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K and Kuksa P. Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research, 2011, 12:2461-2505.
[16] Mikolov T, Chen K, Corrado G and Dean J. Efficient Estimation of Word Representations in Vector Space. *ICLR 2013: International Conference on Learning Representations* (Scottsdale, Arizona, US). 2013, 1-12.
[17] Jeffrey P, Richard S and Christopher M. Glove: Global vectors for word representation. *EMNLP 2014: Proceedings of the Empirical Methods in Natural Language Processing* (Doha, Qatar). Stroudsburg: Association for Computational Linguistics, 2014, 1532-1543.
[18] Le Q, and Mikolov T. Distributed Representations of Sentences and Documents. *Proceedings of the 31st International Conference on Machine Learning* (ICML-14, Beijing). JMLR Workshop and Conference Proceedings, 2014, 1188-1196.
[19] Chen J D, Tang X J. Societal Risk Classification of Post Based on Paragraph Vector and KNN Method. *Proceedings of the 15th International Symposium on Knowledge and Systems Sciences*(Sapporo, November 1-2, 2014 ISBN: 978-4-903092-39-3, Wang S Y, Nakamori Y & Huynh V N, eds.). JAIST Press, 2014, 117-123.
[20] Zhao Y L and Tang X J. A Preliminary Research of Pattern of Users' Behavior Based on Tianya Forum. *The 14th International Symposium on Knowledge and Systems Sciences.* (Ningbo, P.R.China., Oct. 25-27, 2013). JAIST Press, 2013, 139-145.