2015 IEEE International Conference on Systems, Man, and Cybernetics

The Challenges and Feasibility of Societal Risk Classification Based on Deep Learning of Representations

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 posts of Tianya Forum as the data source and adopting the socio psychology study results on societal risks perception, we analyze the challenges and feasibility of the document-level multiple societal risk classification of BBS posts. To effectively capture the semantics and word order of documents, a deep learning model as Post Vector is applied to realize the distributed vector representations of the posts in the vector space. Based on the distributed vector representations, cross-validated classification of the posts labeled by different annotators with KNN method and pairwise similarities comparisons of the posts between risk categories are implemented. The big variance of the results of cross validation shows the differences of individual risk perceptions, which reflects the challenges of societal risk classification. Furthermore, the higher similarities of posts in same societal risk category manifest the feasibility of the classification of societal risks, and indicate the possibility to improve the performance of the societal risk classification of BBS posts.**

Keywords-societal risk classification; Tianya forum; deep learning; individual risk perception; pairwise similarity

I. INTRODUCTION

Up to date, more and more Chinese people use social media (such as blog, micro-blog, BBS) as one way to express their opinions toward the daily phenomena and social events, so it is a better way to monitor the risk level based on these online opinions, as one supplement to the traditional investigations [1]. "Tianya Zatan board is one of the most popular and influential board of Tianya Forum, which is a famous Internet forum in China, and provides BBS, blogs, micro-blogs and photo album services etc." [2] The posts on Tianya Zatan board mainly cover the hot and sensitive topics of current society [3]. Therefore, Tianya Zatan board of Tianya Forum is selected as one of the data sources to explore effective strategies for online societal risk monitoring.

Based on comprehensive comparisons, the framework of societal risks with 7 categories and 30 sub categories that constructed by socio psychology researchers [4] before Beijing Olympic Games is chosen. For adaption, necessary modifications of sub categories are made [1]. To acquire the daily risk level timely, 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 classifying posts by human annotation. Since the effectiveness of machine learning method is proven in text classification, the machine learning method is a better approach for the classification task.

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]. However, based on traditional Bag-of-Words representation, the machine learning method (e.g. 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 was optimized [6]. 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 considerations of syntactic or semantic information and the loss of word order information [7].

To overcome the problems of Bag-of-Words representation, the distributed representation using deep learning method is proposed [8]. The distributed vector representation mitigates the curse of dimensionality problem. Moreover, the semantic and word order information are encoded in the distributed vector space. Recently, many prominent algorithms have been proposed for word vector construction, such as: SENNA [9], Word2Vec [10] and GloVe [11]. Stimulated by the effectiveness of the distributed vector representation of word, many models have been tried at word level up to phrase-level or sentence-level representations [12-13]. A more sensible deep learning method is proposed by Le et al. [14] to solve the issue of the distributed vector representation of paragraph or document, which is inspired by Work2Vec [10]. Combined with an additional paragraph vector, paragraph vector method build 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 paragraph [14]. To deal with the Chinese online documents, a deep learning method as Post Vector model is

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.

978-1-4799-8697-2/15 $31.00 © 2015 IEEE
DOI 10.1109/SMC.2015.110

IEEE computer society

proposed, the model shows improvements for Chinese document distributed representation [15].

With the effective Post Vector model, we focus on realizing the distributed vector representations of BBS posts, and verifying the improvements of the distributed vector representation in societal risk classification. Furthermore, based on the distributed vector representation, the differences of individual risk perceptions are studied through cross-validated classification of the posts labeled by different people. The pairwise similarities comparisons between societal risk categories are conducted to show the societal risk categories can be distinguishable.

The rest of this paper is organized as follows. Section 2 explains the deep learning method: Post Vector model. Section 3 addresses the data sets, experiment procedures and performance measures. Section 4 presents the results and discussions and Section 5 for the concluding remarks.

II. POST VECTOR MODEL

The deep learning method as Post Vector is mainly designed for the distributed representation of Chinese documents [15].

A. Algorithm of Post Vector

In Post Vector 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 Post Vector 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 ***D*** and each word of post is also represented by a unique vector, which is a column in word matrix ***W***. Due to 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.

The Post Vector model is consisted of three layers: input, hidden and output. Before the model training, set the word vector length as l and window size as k. Given a post ID and a sequence of words of the post: 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, the vectors of the k words before or after w_t are taken into consideration.

During the training process, the input features are of fixed-length and sampled from a sliding window over the document of the post. The post document vector (the vector of w_{ID}) is shared across all contexts generated from the same post, but no occurrence in different posts. Hence, the post 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. After the training convergence, the post document vectors can be used as features for the post. These features can be directly fed to conventional machine learning methods such as logistic regression, support vector machines, or K-nearest neighbors (KNN) [16].

B. The Effectiveness Illustration of Post Vector

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

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

(1). "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...)"

(2). "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...)"

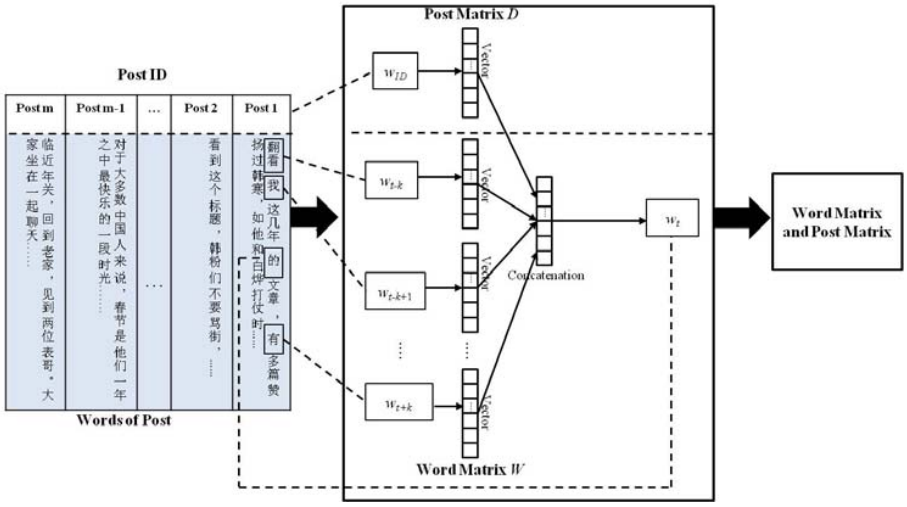

Figure 1. A framework for learning word vectors and post vectors

TABLE I. THE DISTRIBUTIONS OF POSTS FOR DIFFERENT MONTHS

Period / Risk Category	Dec.2011	Jan.2012	Feb.2012	Mar.2012	Jul.2012	Aug.2012	Sep.2012	Otc.2012	Nov.2012	Dec.2012
Total	12125	12032	20330	37946	31017	40655	37646	39614	42704	41016
Risk free	1278	2047	2645	14569	15348	16104	17513	24919	28133	19396
Government Management	3373	1809	3099	6879	4437	5490	4751	4807	5892	9122
Public Morality	3337	3730	8715	6065	2619	3871	2458	2656	2210	3738
Social Stability	954	1013	1746	2108	1787	4364	2326	1700	1412	2075
Daily Life	2641	3063	3142	6920	4566	5043	3494	3716	3805	5535
Resources & Environments	223	147	309	329	548	299	297	271	220	362
Economy & Finance	248	133	460	609	487	263	457	380	279	134
Nation's Security	71	90	214	467	1225	5221	6348	1165	753	645

The first document describes one traffic accident, thus its risk category is "daily life". The second document is about the subject's complaint of a bad luck, and labeled as risk free. However, after the segmentation of documents, the vectors of two Chinese documents that constructed by the Bag-of-Words are same, while their risk labels are different, which then confuses the classifier. In contrast to Bag-of-Words, Post Vector generates two different vectors for both documents, and makes it easier for classifier to learn the difference between the two documents. Hence, Post Vector can be more effective to represent the differences in documents.

Example 2: Similar words, same societal risk

(1). "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…)"

(2). "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 category: Resources & Environments. After the segmentation of documents, it can be found that no words are shared between the two documents. Based on Bag-of-Words, the vectors of two documents are totally different, which makes the classifier lose the similar information between the two documents. However, by Post Vector 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. Consequently, the vectors of these two documents share more information, which is easier to classify these two documents into one risk category.

Above two examples illustrate that Post Vector model can extract more useful features from documents for societal risk classification, which benefits from the improvements in the word order and semantic understanding of document. Next, based on the effective representation of posts produced by Post Vector, the challenges and feasibility of societal risk classification are discussed.

III. DATA SETS, EXPERIMENTAL PROCEDURES AND PERFORMANCE MEASURES

This section introduces the data sets, experimental procedures for the training of Post Vector, cross validation and similarity comparison, as well as the performance measures for the classification results of cross-validations.

A. Data Sets

To train Post Vector model, we select the new posts (title + text) of Dec. 2011-Mar. 2013, more than 470 thousands posts, which were crawled by the daily working Tianya Forum spider system of our group [17]. To exhibit the challenges and feasibility of societal risk classification of BBS posts, the labeled posts published in Dec. 2011-Mar. 2012 and Jul.2012-Dec.2012 are used. The amount of those 10-month posts and the amount of the monthly posts in different societal risk categories are presented in Table I.

The figures in Table I show the distributions of the posts in different risk categories are unbalanced. The societal risks of posts from Tianya Zatan mainly distribute in risk free, government management, public morality and daily life, the total number of these categories is more than 85% of all posts.

B. Experimental Procedures

The experiments mainly contain three parts: the Post Vector model training, cross-validated classification of the posts annotated by different people and pairwise similarities comparisons of the posts between risk categories. The desktop computers for all the experiments are 64-bit, 3.6GHz, 8 cores and 8GB RAM.

Due to the Chinese corpus, the segmentation of corpus is required. The training steps of Post Vector are: i) all the post corpuses are segmented with Ansj_Seg tool[1]; ii) post ID is concatenated with the segmented words of the post; iii) all the corpuses are fed into Post Vector to generate the post matrix.

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

For cross-validated classification of the posts annotated by different people, KNN method is adopted for classification. Here is the procedure of KNN classification based on the post document vectors: i) the posts labeled by different annotators are chosen; ii) the post document vectors are extracted from the post matrix; iii) the risk label of each post is combined with the post document vector; iv) using KNN, cross-validation of all data sets; v) comparing with human annotation, the performances are evaluated.

For similarities comparisons, the pairwise similarities of the posts of same risk category or between different risk categories are calculated. The main steps are as follows:

1) Extract the post document vectors from the post matrix;
2) Combine the risk label of each post with the post document vector;
3) **if** the pairwise similarities of posts in the same category
   for ***post$_i$*** in the category
   Calculate the cosine similarities of ***post$_i$*** to all other posts in the same category;
   end
   else if the pairwise similarities of posts between two categories
   for ***post$_i$*** in one category
   Calculate the cosine similarities of ***post$_i$*** to all posts in another category;
   end
4) Calculate the mean and variance of the similarities.

C. Performance Measures

The precision, recall and F-measure are used for performance measurement of cross-validated classification. 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 [18].

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

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

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.

IV. Eeperimental Results And Discussions

According to the experimental procedures, through the unsupervised training of Post Vector model, the post document vector of each post in the training set is generated.

A. The Challengs of Societal Risk Classification

The complexity of post corpus is just one issue of societal risk classification. In this part, based on post document vectors, we address the issue of individual risk perception differences, which increases the difficulty of societal risk classification. Two kinds of experiments are carried out: i) cross-validated classification of the posts annotated by different people in different group; ii) cross-validated classification of the posts annotated by different people in same group.

After parameter tuning, the parameters of KNN based on Post Vector are fixed: *vector_size*=250, *window_size*=3 and *k*=40, where *vector_size* is the size of post document vector, *window_size* is the size of words as the input of Post Vector, *k* is the parameter of KNN.

1) Cross Validation of the Posts Labeled by Different Groups: 5-month posts, Dec.2011, Jan.2012, Feb.2012, Jul.2012 and Aug.2012, are selected. Those posts were labeled by three different groups: Jan.2012 and Feb.2012 by one group, Jul.2012 and Aug.2012 also by one group, and Dec.2011 by the other group. The cross-validated classification results of KNN based on Post Vector model for those posts are presented in Table II.

From the results in Table II, it can be found, using the labeled posts of Dec.2011 as the training data, similar classification results are obtained for different data sets labeled by same group (the classification results of Jan.2012 and Feb.2012 are similar, the same situations occurred in Jul.2012 and Aug.2012), but a big difference between the data sets labeled by different group (the classification results of Jan.2012 and Jul.2012, etc.). Normally, the annotators in same group are from similar background, and they can communicate with each other. Conversely, the backgrounds of the annotators in different groups are different, and they hardly communicate with the annotators in different groups. Consequently, the annotators in same group show more consensus than the annotators in different groups. Therefore, if we use same data set as the training data set to classify the data sets labeled by same group, the similar results will be obtained, but a gap will emerge between different groups.

Moreover, as the annotators from the same group exchanges opinions easier, that more consensus in the same group in risk labeling may be displayed than that from different groups may be displayed by cross validation results of the data sets annotated by same group and by different group. As listed in Table II, using the labeled posts of Jan.2012, Feb.2012, Jul.2012 or Aug.2012 as the training data set, the best performance of cross validation for each case is obtained by the data set labeled by the same group. The results reveal that people from different groups display larger differences in societal risk labeling than the people from the same group.

TABLE II. CROSS VALIDATION OF THE POSTS LABELED BY DIFFERENT GROUPS

(*Macro_F*, *Micro_F*)	Dec.2011	Jan.2012	Feb.2012	Jul.2012	Aug.2012
Dec.2011		(38.07%, 53.00%)	(39.53%, 54.10%)	(39.72%, 33.12%)	(42.23%, 33.15%)
Jan.2012	(36.06%, 46.90%)		**(37.84%, 54.31%)**	(37.24%, 37.05%)	(39.31%, 36.57%)
Feb.2012	(39.76%, 49.53%)	**(41.34%, 52.99%)**		(40.45%, 33.38%)	(42.81%, 34.03%)
Jul.2012	(40.25%, 32.30%)	(37.74%, 37.43%)	(38.95%, 33.19%)		**(44.87%, 56.77%)**
Aug.2012	(39.43%, 35.09%)	(37.40%, 39.61%)	(40.44%, 34.69%)	**(45.25%, 60.49%)**	

TABLE III. CROSS VALIDATION OF THE POSTS LABELED BY DIFFERENT PEOPLE IN SAME GROUP

(*Marco_F, Micor_F*)	A-data1	A-data2	B-data1	B-data2
A-data1		**(34.76%, 60.94%)**	(30.95%, 46.22%)	(29.68%, 49.50%)
A-data2	**(30.28%, 64.45%)**		(30.67%, 49.79%)	(29.45%, 51.03%)
B-data1	(34.28%, 57.50%)	(32.67%, 58.73%)		**(49.03%, 61.35%)**
B-data2	(34.55%, 62.73%)	(33.36%, 61.14%)	**(47.24%, 62.76%)**	

2) Cross Validation of the Posts Labeled by Different People in Same Group: Although the people in the same group show more consensus in posts risk labeling, it is still unclear about whether the individual risk perceptions of the people in same group are different.

To test the discrepancy of individual risk perceptions of people in same group, two annotators in same group are selected, named as A and B. Two-week posts labeled by these two people are selected: Jul. 23-31 2012 (A-data1) and Sep. 23-30 2012 (A-data2), Aug. 1-7 2012 (B-data1) and Aug. 8-14 2012 (B-data2). The cross validation results of these data sets with same classification method are presented in Table III.

As depicted in Table III, for each case, cross-validated classification of the data sets labeled by same annotator is better than the data sets labeled by different annotators. From this point, it can be found that the individual risk perceptions of people in same group are also different.

Even the people in same group show more consensus in posts risk labeling, the different majors, knowledge structures and personal experiences may be reasons of the difference in societal risk labeling. Therefore, the annotation of the posts is definitely affected by the individual cognition, which is not always consistent, may bring more noises and thus makes the societal risk classification more challenging.

B. The Feasibility of Societal Risk Classification

The societal risk classification of BBS posts faces two aspects of challenges: i) the complexity of online corpus; ii) the difference of individual cognition. Both the challenges are difficult, and lead to consider whether the framework of societal risks is feasible. To check the feasibility of the classification of societal risks, the pairwise similarities in same societal risk category and the pairwise similarities between two societal risk categories are computed.

As the memory limitation of our computer, all the labeled posts are divided into three data sets: Dec.2011-Mar.2012, Jul.2012-Sep.2012 and Oct.2012-Dec.2012. According to the experimental procedure, pairwise similarities of different data sets are calculated. Due to the space limitation, only the results of Dec.2011-Mar.2012 are presented in Table IV. The parameters of Post Vector are fixed: *vector_size*=250, *window_size*=3.

From the results of Table IV, it can be found that, for any societal risk category in all three data sets, the mean of the similarities of the posts in one societal risk category is bigger than the mean of the similarities of the posts between the societal risk category and one of other societal risk categories. We also compare the variance (Table V) of similarities of each case, no obvious difference is found for the variances. From this point, it can be concluded that, even with the influence of complex text and the difference of individual risk perceptions, the differences among the societal risk categories are still clear, which means the societal risk indicators are feasible.

V. CONCLUDING REMARKS

Based on the document vectors, the challenges and feasibility of document-level multi-class societal risk classification are discussed in this paper. The main contributions are summarized as follows.

- An effective deep learning method Post Vector for the distributed representation of Chinese BBS posts is applied in this study;
- Through cross-validated classifications of different data sets labeled by different people in different group, the difference of individual risk perception is revealed;
- According to the pairwise similarities comparisons between societal risk categories, the feasibility of the societal risks is verified.

Based on the results of this study, later we may 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. For the real time on-line societal risk monitoring, hybrid strategies may need to be considered.

TABLE IV. THE MEAN OF PAIRWISE SIMILARITIES COMPARISONS OF RISK CATEGORIES

Similarities (Mean, Dec.2011-Mar.2012)	Risk Free	Government Management	Public Morals	Social Stability	Daily Life	Recourses & Environment	Economy & Finance	Nation's Security
Risk Free	**0.043**	0.015	0.036	0.021	0.030	0.026	0.023	0.026
Government Management		**0.049**	0.017	0.039	0.023	0.033	0.032	0.03
Public Morals			**0.048**	0.002	0.027	0.017	0.019	0.032
Social Stability				**0.054**	0.026	0.029	0.023	0.022
Daily Life					**0.039**	0.029	0.029	0.015
Recourses & Environment						**0.075**	0.041	0.044
Economy & Finance							**0.080**	0.041
Nation's Security								**0.106**

TABLE V. THE VARIANCE OF PAIRWISE SIMILARITIES COMPARISONS OF RISK CATEGORIES

Similarities (Mean, Dec.2011-Mar.2012)	Risk Free	Government Management	Public Morals	Social Stability	Daily Life	Recourses & Environment	Economy & Finance	Nation's Security
Risk Free	0.007	0.007	0.007	0.007	0.007	0.007	0.007	0.007
Government Management		0.008	0.007	0.008	0.007	0.007	0.008	0.007
Public Morals			0.009	0.007	0.007	0.007	0.007	0.007
Social Stability				0.009	0.008	0.008	0.008	0.007
Daily Life					0.009	0.008	0.008	0.007
Recourses & Environment						0.01	0.008	0.007
Economy & Finance							0.01	0.008
Nation's Security								0.009

REFERENCES

[1] X. J. Tang, "Exploring On-line Societal Risk Perception for Harmonious Society Measurement," Journal of Systems Science and Systems Engineering, vol.22, no.4, 2013, pp469-486.

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

[3] L. N. Cao and X. J. Tang, "Topics and Threads of the Online Public Concerns Based on Tianya Forum," Journal of Systems Science and Systems Engineering, vol. 23, no.2, 2014, pp212-230.

[4] R. Zheng, K. Shi and S. Li, "The InfluenceFactors and Mechanism of Societal Risk Perception", Proceedings of the 1st International Conference on Complex Sciences: Theory and Application (Shanghai, J. Zhou eds.). Springer Berlin Heidelberg, 2009, pp2266-2275.

[5] W. Zhang, T. Yoshida and X. J. Tang, "Text Classification Based on Multi-word with Support Vector Machine," Knowledge-Based Systems, vol.21, no.8, 2008, pp879-886.

[6] J. D. Chen and X. J. Tang, "Exploring Societal Risk Classification of the Posts of Tianya Club," International Journal of Knowledge and Systems Science, vol.5, no.1, 2014, pp36-48.

[7] L. Qiu, Y. Cao, Z. Q. Nie and Y. Rui, "Learning Word Representation Considering Proximity and Ambiguity," Proceedings of the 28th AAAI Conference on Artificial Intelligence (Québec). 2014, pp1572-1578.

[8] Y. Bengio, R. Ducharme, P. Vincent and C. Jauvin, "A Neural Probabilistic Language Model," Journal of Machine Learning Research, vol.3, 2003, pp1137–1155.

[9] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu and P. Kuksa, "Natural Language Processing (Almost) from Scratch," Journal of Machine Learning Research, vol.12, 2011, pp2461-2505.

[10] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space," ICLR 2013: International Conference on Learning Representations (Scottsdale). 2013, pp1-12.

[11] P. Jeffrey, S. Richard and M. Christopher, "Glove: Global Vectors for Word Representation," EMNLP 2014: Proceedings of the Empirical Methods in Natural Language Processing (Doha). Stroudsburg: Association for Computational Linguistics, 2014, pp1532-1543.

[12] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, "Distributed Representations of Words and Phrases and Their Compositionality," Advances in Neural Information Processing Systems 26 (NIPS 2013, Lake Tahoe). 2013, pp3111-3119.

[13] S. Richard, C. L. Cliff, Y. N. Andrew and M.Chris. "Parsing Natural Scenes and Natural Language with Recursive Neural Networks," Proceedings of the 28th International Conference on Machine Learning (ICML-11, Bellevue). JMLR Workshop and Conference Proceedings, 2011, pp129–136.

[14] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," Proceedings of the 31st International Conference on Machine Learning (ICML-14, Beijing, China). JMLR Workshop and Conference Proceedings, 2014, pp1188-1196.

[15] J. D. Chen and X. J. Tang, "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, pp117-123.

[16] T. M. Cover and P. E. Hart, "Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory," vol.13, no.1,1967, pp21-27.

[17] Y. L. Zhao and X J. Tang, "A Preliminary Research of Pattern of Users' Behavior Based on Tianya Forum," The 14th International Symposium on Knowledge and Systems Sciences. (Ningbo, Oct. 25-27, 2013). JAIST Press, 2013, pp139-145.

[18] S. Y. Wen and X. J. Wan, "Emotion Classification in Microblog Texts Using Class Sequential Rules," Proceedings of the 28th AAAI Conference on Artificial Intelligence (Québec). 2014, pp187-193.