Ensemble of SVM Classifiers with Different Representations for Societal Risk Classification

Jindong Chen and Xijin Tang(✉)

Institute of Systems Science, Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing 100190, People's Republic of China
j.chen@amss.ac.cn, xjtang@iss.ac.cn

Abstract. Using the posts of Tianya Forum as the data source and adopting the societal risk indicators from socio psychology, we conduct document-level multiple societal risk classification of BBS posts. Two kinds of models are applied to generate the representations of posts respectively: Bag-of-Words focuses on extracting the occurrence information of words in posts, and a deep learning model as Post Vector is designed to capture the semantics and word order of posts. Based on the different post representations, two types of support vector machine (SVM) classifiers are developed and compared in the societal risk classification of the posts. Furthermore, as the complementary information contained in the two different post representations, several SVM ensemble methods at the decision score level of the two SVM classifiers are proposed to improve the performance of societal risk classification. The experimental results reveal that the SVM ensemble method achieves better results in document-level societal risk classification than SVM based on single representation.

Keywords: Societal risk classification · Tianya forum · Deep learning · Bag-of-Words · Support vector machine

1 Introduction

To monitor the daily risk classes and level of Tianya Zatan Broad of Tianya Forum timely, societal risk classification of BBS posts is the main task. Since the framework of societal risks includes 7 main categories and 1 risk free category, societal risk classification of BBS posts is document-level multiple classification [1, 2]. The document-level multiple societal risk classification is a quite difficult task, since i) the document-level classification brings more challenges, such as the big variance of the text length, the complicated syntax, the involvement of multiple topics in one document; ii) the multiple risk classes also increase the complexity of text classification.

For text classification, the primary step is to represent text as vectors. The traditional method is Bag-of-Words (BOW), disregarding semantic and word order but keeping multiplicity. To overcome the issues of BOW representation, the distributed representation using deep learning method was proposed [3]. In this method, the semantic and word order features are encoded in the distributed vectors through sliding-window training mode. Recently, many prominent deep learning algorithms have been proposed for word vector construction, such as: SENNA [4], Word2Vec [5] and GloVe [6]. Le et al.

S. Zhang et al. (Eds.): KSEM 2015, LNAI 9403, pp. 669–675, 2015.
DOI: 10.1007/978-3-319-25159-2_61

[7] proposed a more flexible deep learning method to realize the distributed representation of paragraph or document [5]. Combined with an additional paragraph vector, the method includes two models: PV-DM and PV-DBOW for paragraph representation, where the paragraph vector contributes to predict the next word in many contexts sampled from the paragraph [7]. To realize the distributed representation of Chinese online documents, a deep learning method as Post Vector (PV) model was proposed, the model showed its effectiveness for Chinese document representation [8].

The representative classifiers for text classification are K-Nearest Neighbor, naïve Bayes and support vector machine (SVM), etc. Due to the good performance of SVM for societal risk classification of Baidu hot word [9], SVM method is chosen. However, based on BOW representation, SVM method hardly achieved the expected performance in societal risk classification, even though the training set was increased and the feature word selection method was optimized. Therefore, with the deep learning method as PV model, we focus on realizing the distributed representation of BBS posts, and developing SVM classifier based on the distributed representations. Furthermore, as the complementary information contained in BOW representation and the distributed representation, we construct an ensemble model at decision score level of SVM classifiers, for performance improvement in societal risk classification of BBS posts.

2 Post Vector Model

The deep learning method as PV is mainly designed for the distributed representation of Chinese documents [8]. 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 *D* and every word of post is also represented by a unique vector, which is a column in word matrix *W*, where *D* and *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 w_t are taken into input. The objective of the Post Vector model is to maximize the average log probability

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

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 updated across all 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 used 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.

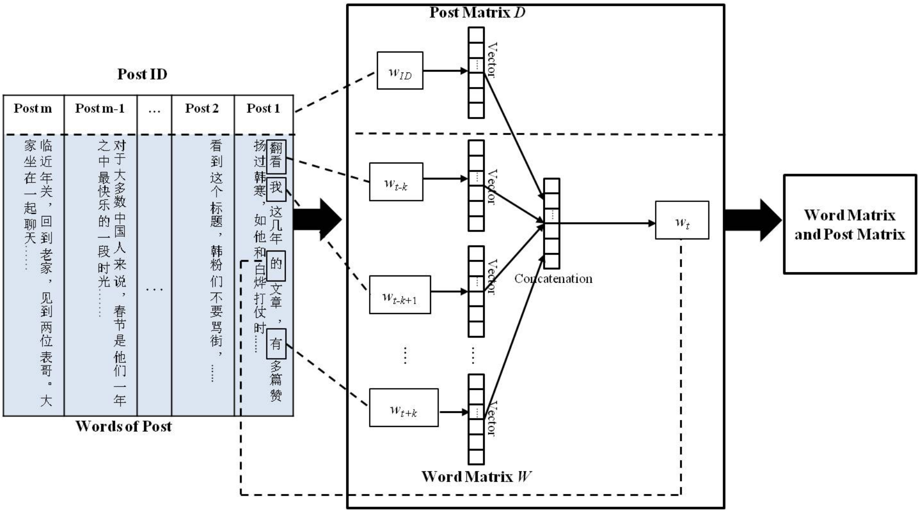
Post ID
Post m
Post m-1
…
Post 2
Post 1
临近年关，回到老家，见到两位表哥。大家坐在一起聊天……
对于大多数中国人来说，春节是他们一年之中最快乐的一段时光……
…
看到这个标题，韩粉们不要骂街，……
翻看我这几年的文章，有多篇赞扬过韩寒，如他和白烨打仗时……
Words of Post
Post Matrix D
w_{ID}
Vector
w_{t-k}
w_{t-k+1}
w_{t+k}
Concatenation
w_t
Word Matrix W
Word Matrix and Post Matrix

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

3 Data Sets, Experimental Procedure

3.1 Data Sets

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

3.2 Experimental Procedures

The process of SVM based on BOW representations for societal risk classification toward BBS posts includes: word segmentation, feature selection, feature weight and SVM training and test. The word segmentation tool is Ansj, the stop words are from HIT (Information Retrieval Laboratory, Harbin Institute of Technology), the χ^2-test is adopted for feature selection and *tf-idf* is used for feature weight.

A category membership score is applied to calculate the decision value of the classifier for each category. The category membership score is computed by Eq.2.

$$score = \frac{\sum S_i}{2*k} + \frac{k}{2*n} \tag{2}$$

where k is the number of voters supporting a certain category; n is the number of categories; S_i is the decision score of each supporting voter.

The process of SVM based on the distributed representations includes: word segmentation, Post Vector model training and SVM training and test. The word segmentation tool is same as before, and all the words are kept and fed into PV model to generate the post document vectors (the vectors of post ID). SVM training adopts the same strategy as SVM based on BOW representations, and the category membership score is also applied in this method.

The ensemble method is implemented at the decision score level. For a new post p_i, due to the One-Against-One training strategy, SVM classifier based on BOW representations or the distributed representations outputs 28 decision scores respectively. Based on the decision scores, the category membership scores of each SVM classifier are calculated. Using weighted or softmax regression method, the decision scores and category membership scores are combined to improve the performance of societal risk classification.

4 Experiment Results and Discussions

4.1 SVM Based on BOW Representations

For χ^2-test, the ratio is set as 0.4. The kernel function for SVM is chosen as RBF. After parameter optimization, the parameters of SVM are C=1.4 and g=0.5. 5-fold cross-validations are implemented on the data set. The classification results are presented in Table 2. The performance measures of classification results for each fold are computed as Ref. [10]. 8 classes of societal risks are taken into consideration.

Table 2. The *Macro_F* and *Micro_F* of SVM based on BOW representations

ith fold	1	2	3	4	5	Mean
Macro_F	53.89%	54.85%	53.66%	53.45%	54.84%	54.14%
Micro_F	60.52%	60.91%	60.30%	60.60%	61.15%	60.69%

4.2 SVM Based on the Distributed Representations

For Post Vector model training, the training set is the posts during November of 2011 to March of 2013, 16-month new posts every day, more than 470 thousands posts. Through the unsupervised training of Post Vector model, the distributed representations of the posts in the data set are yielded. Based on the distributed representations, SVM method is applied for societal risk classification of posts.

The kernel function for SVM is chosen as RBF. Through parameter optimization, the parameters of PV are *window size*=3 and *vector size*=250, the parameters of SVM are *C*=2 and *g*=0.5. 5-fold cross-validations are implemented on the data set. The classification results are presented in Table 3.

Table 3. The *Macro_F* and *Micro_F* of SVM based on the distributed representations

ith fold	1	2	3	4	5	Mean
Macro_F	50.20%	51.01%	50.28%	50.73%	52.36%	50.91%
Micro_F	58.20%	58.99%	58.14%	58.51%	58.98%	58.56%

4.3 The SVM Classifiers Ensemble

To further improve the performance of risk classification, the SVM classifiers ensemble methods are proposed. According to the description of Section 3.2, at the decision scores level, the two SVM classifiers are combined. Three kinds of methods are developed and tested:

I) Softmax regression. The 8 category membership scores of the two SVMs are concatenated as the input of softmax regression. 2000 labeled posts from the training set are used to identify the parameters of softmax regression. After the training, the softmax regression model is used to predict the risk category of the testing samples.

II) Max_Voter. Based on the 56 decision scores, the supporting votes of each risk category can be counted. The risk category with the maximum votes will be the risk label of the testing post. If more than one risk category gets the maximum supporting votes, the risk category with a bigger sum of the category membership scores of the two SVM classifiers will be applied to label the testing post.

III) Max_Score. If two SVM classifiers classify the testing post into the same risk category, the risk category of the testing post is confirmed. Otherwise, the label of testing post is as same as the risk category with the highest category membership score of the two SVM classifiers.

All the results of the three ensemble methods are present in Table 4.

Table 4. The *Macro_F* and *Micro_F* of the SVM classifiers ensemble methods

Methods	ith fold	1	2	3	4	5	Mean
Softmax Regression	*Macro_F*	50.77%	53.51%	51.25%	52.00%	52.86%	52.08%
	Micro_F	59.43%	60.25%	59.43%	59.13%	60.28%	59.70%
Max_Voter	*Macro_F*	52.42%	53.47%	51.98%	53.05%	54.56%	53.09%
	Micro_F	61.05%	61.23%	60.57%	61.24%	61.39%	61.10%
Max_Score	*Macro_F*	53.53%	54.47%	52.64%	53.31%	54.37%	53.66%
	Micro_F	61.05%	61.60%	60.92%	61.28%	61.41%	61.25%

From the results of Table 4, although the logistic regression is the most popular stacking method, softmax regression method gets better performance than SVM based on the distributed representations, but worse than SVM based on BOW representations. Max_Voter method improves *Micro_F*, but with the larger decrease in *Macro_F*, then the whole performance is still worse than SVM based on BOW representations. Although the decrease of *Macro_F* is still found, the improvement of *Micro_F* is more obvious, the entire performance of the Max_Score method is better than SVM based on single representation: BOW or the distributed representation. Therefore, the ensemble method Max_Score achieves state-of-the-art performance in societal risk classification.

5 Conclusions

The contributions of the paper can be summarized as follows.
An effective deep learning method Post Vector for the distributed representation of Chinese BBS posts is applied in this study;

1) SVM based on the distributed representation method are tested in societal risk classification, through cross validation, SVM based on the distributed representation method do not show its improvement in societal risk classification;
2) Three ensemble methods of the two SVM classifiers are tested, and Max_Score method gets the state of the art performance in societal risk classification.

Acknowledgements. This research is supported by 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. Zheng, R., Shi, K., Li, S.: The influence factors and mechanism of societal risk perception. In: Zhou, J. (ed.) Complex 2009. LNICST, vol. 5, pp. 2266–2275. Springer, Heidelberg (2009)
2. Tang, X.J.: Exploring On-line Societal Risk Perception for Harmonious Society Measurement. Journal of Systems Science and Systems Engineering **22**(4), 469–486 (2013)
3. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. Journal of Machine Learning Research **3**, 1137–1155 (2003)
4. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research **12**, 2461–2505 (2011)
5. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: International Conference on Learning Representations (ICLR 2013), Scottsdale, pp. 1−12 (2013)
6. Jeffrey, P., Richard, S., Christopher, M.: Glove: Global vectors for word representation. In: Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014), pp. 1532−1543. Association for Computational Linguistics, Stroudsburg (2014)

7. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning (ICML 2014). JMLR Workshop and Conference Proceedings, Beijing, pp. 1188−1196 (2014)
8. Chen, J.D., Tang, X.J.: Societal risk classification of post based on paragraph vector and KNN method. In: Wang, S.Y., Nakamori, Y., Huynh, V.N. (Eds.) Proceedings of the 15th International Symposium on Knowledge and Systems Sciences, Sapporo, November 1−2, pp. 117−123. JAIST Press (2014). ISBN: 978-4-903092-39-3
9. 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)
10. Wen, S.Y., Wan, X.J.: Emotion classification in microblog texts using class sequential rules. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, Québec, pp. 187−193 (2014)