Performance Comparison of TF*IDF, LDA and Paragraph Vector for Document Classification

Jindong Chen[1], Pengjia Yuan[2], Xiaoji Zhou[1], and Xijin Tang[2(✉)]

[1] China Academy of Aerospace Systems Science and Engineering, Beijing 100048, People's Republic of China
j.chen@amss.ac.cn, zh_xj@sina.com
[2] Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, People's Republic of China
ypjl992@126.com, xjtang@iss.ac.cn

Abstract. To meet the fast and effective requirements of document classification in Web 2.0, the most direct strategy is to reduce the dimension of document representation without much information loss. Topic model and neural network language model are two main strategies to represent document in a low-dimensional space. To compare the effectiveness of bag-of-words, topic model and neural network language model for document classification, TF*IDF, latent Dirichlet allocation (LDA) and Paragraph Vector model are selected. Based on the generated vectors of these three methods, support vector machine classifiers are developed respectively. The performances of these three methods on English and Chinese document collections are evaluated. The experimental results show that TF*IDF outperforms LDA and Paragraph Vector, but the high-dimensional vectors take up much time and memory. Furthermore, through cross validation, the results reveal that stop words elimination and the size of training samples significantly affect the performances of LDA and Paragraph Vector, and Paragraph Vector displays its potential to overwhelm two other methods. Finally, the suggestions related with stop words elimination and data size for LDA and Paragraph Vector training are provided.

Keywords: TF*IDF · LDA · Paragraph vector · Support vector machine · Document classification

1 Introduction

Text is an important source of information, which mainly includes unstructured and semi-structured information. In Web 2.0 era, Internet users are willing to express their opinions online, which accelerates the expansion of text information [1]. Owing to the increasing amount of text information, especially for the unstructured information, to extract useful information or knowledge efficiently, document classification plays an important role [2]. Normally, document classification is to assign the predefined labels to new documents based on the model learned from a trained set of labels and documents.

J. Chen et al. (Eds.): KSS 2016, CCIS 660, pp. 225–235, 2016.
DOI: 10.1007/978-981-10-2857-1_20

The process of document classification can be divided into two parts: document representation and classifier training. Compared to classifier training, document representation is the central problem for document classification. Document representation is tried to transfer text information into a machine understandable format without information loss, such as n-gram models. Unfortunately, if n is more than 5, the huge computation cost makes the transformation infeasible. Consequently, several frequently used types of n-gram models are unigram, bigram or trigram [3]. For academic document classification or news classification, owing to the difference of feature words in different categories, those kinds of methods are capable of meeting the requirements of practical application. Meanwhile, a comprehensive analysis of the performances of different classifiers on different data sets is conducted by Manuel et al., and reveals that support vector machine (SVM) and random forests are more effective for most classification tasks [4].

The rapid increase of text data brings new challenges to the available traditional methods [5]. Big corpus dramatically increases the dimension of the representations generated by the traditional methods. High-dimensional vectors take up more memory space, even cannot work on low-configuration computer. Furthermore, even if the transformation is available, the big time cost of classifier training on high-dimensional vectors is another issue for document classification. To meet the tendency of information expansion, it is an important task to reduce the dimension of the representation without much information loss for document classification.

Up to date, document classification is not limited for news classification or academic document classification, and expands to more areas, such as sentiment classification [6], emotion classification [7] and societal risk classification [8]. Different from traditional document classification, these types of document classification face two new challenges: one is that the category of document is related with syntax and word order, the other is different categories may use similar feature words. The traditional methods lack in semantic and word order information extraction, which affects their performances in these areas.

To improve the efficiency of document classification, from dimension reduction and semantic information extraction aspects, several strategies of document representation are proposed:

(1) Topic model. Topic model is not only increasing the efficiency by a more compact topic representation, but also capable of removing noise such as synonymy, polysemy or rare term use. The distinguished methods of topic model include: latent sematic analysis (LSA), probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA) [9]. LDA is a generative document model that is capable of dimension reduction as well as topic modeling, and shows better performance than LSA and PLSA. LDA models every topic as a distribution over the words of the vocabulary, and every document as a distribution over the topics, thereby one can use the latent topic mixture of a document as a reduced representation. Based on the representation of latent topic mixture, document clustering and document classification are conducted [10, 11].
(2) Neural network language model. Bengio et al. proposed a distributed vector representation generated by neural network language model [12]. Due to the fixed

and small size of document vector, the distributed representation of neural network language model eliminates the curse of dimensionality problem. Meanwhile, through sliding-window training mode, the semantic and word order information are encoded in the distributed vector space. Recently, based on the neural network language model proposed for word vector construction [13], Le and Mikolov [14] proposed a more sensible method Paragraph Vector (PV) 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.

The purpose of this research is to study the efficiency of different methods for document classification. TF*IDF, LDA and PV have been proposed for a while, and Andrew et al. [15] has compared these three methods on two big datasets: Wiki documents and arXiv articles, each contains nearly 1 million documents, but there is no comprehensive comparative study on these methods for Chinese documents and different sizes of datasets, and no result is reported concerning their classification performances on semantic classification etc. Therefore, to further analyze the performances of these three methods, three datasets: Reuters-21578[1], Sogou news dataset[2] and the posts of Tianya Zatan Board[3] are selected, which includes English and Chinese documents, and aims for news classification and societal risk classification tasks. Based on the document representations generated by these three methods, SVM is adopted for document classification respectively [8], and the performances of each method are compared.

Afterward, LDA relies on the occurrence of words to extract topics, and PV model generates document vector based on word semantic and word order, so stop words present different impacts to LDA and PV. Hence, to clarify the impacts of stop words to LDA and PV, on Sogou news dataset, the influences of stop words elimination operation to LDA and PV model training are analyzed. Next, due to the iterative learning process of PV model, the size of training samples affects the performance of PV. Therefore, on Reuters-21578, Sogou new dataset with repeated data, the performances of PV-SVM are analyzed.

Therefore, the rest of this paper is organized as follows. The data sets and experimental procedures are explained in Sect. 2. The results and discussions are presented in Sect. 3. Finally, concluding remarks are given in Sect. 4.

2 Data Sets and Experimental Procedure

This section introduces data sets and experimental procedures for the different classification algorithms.

[1] http://ronaldo.cs.tcd.ie/esslli07/data/reuters21578-xml/.

[2] www.sogou.com/labs/dl/c.html.

[3] http://bbs.tianya.cn/list-free-1.shtml.

2.1 Data Sets

Reuters-21578. Reuters document collection is applied as our experimental data. It appeared as Reuters-22173 in 1991 and was indexed with 135 categories by personnel from Reuters Ltd. in 1996. For convenience, the documents from 4 categories, "agriculture", "crude", "trade" and "interest" are selected. In this study, 626 documents from agriculture, 627 documents from crude, 511 documents from interest and 549 documents from trade are assigned as our target data set.

Sogou. Sogou news dataset used in experiments of this paper are from Sogou Laboratory Corpus. Sogou Laboratory Corpus contains roughly 80,000 news documents, which are equally divided into 10 categories. The categories are Cars, Finance, Education, IT, Healthy, Sport, Recruitment, Culture, Military and Tour.

Tianya Zatan. With the spider system of our group [16], the daily new posts and updated posts are downloaded and parsed. According to the framework of societal risks constructed by socio psychology researchers [17] before Beijing Olympic Games, the new posts of Tianya Zatan in 2012 are almost labeled. To reveal the effectiveness of different methods for societal risk classification of BBS posts, 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. Different from previous two datasets, 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 total number of these categories is more than 85 % of all posts.

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
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
Total	12125	12032	20330	37946

2.2 Experimental Procedures

On the three datasets, three kinds of experiments are tested here: (1) SVM based on TF*IDF method (TF*IDF-SVM), (2) SVM based on LDA method (LDA-SVM), (3) SVM based on Paragraph Vector model (PV-SVM). The desktop computer for all experiments are 64-bit, 3.6 GHz, 8 cores and 16 GB RAM.

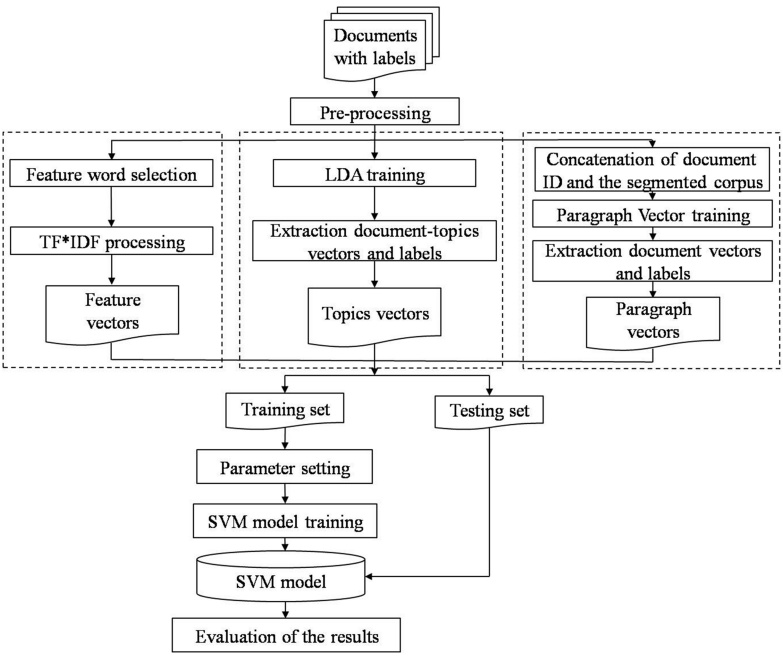

Fig. 1. The process of TF*IDF-SVM, LDA-SVM and PV-SVM for document classification

The pre-processing of English document includes: tokenizing, elimination of stop words and stemming. Meanwhile, the main pre-processing step of Chinese document is word segmentation, and the elimination of stop words is depending on different requirements. Word segmentation tool is Ansj-Seg[4], the stop words dictionary are from Harbin Institute of Technology.

The processes of TF*IDF-SVM, LDA-SVM and PV-SVM for document classification are illustrated in Fig. 1.

The main steps of TF*IDF-SVM include: preprocessing, feature word selection, TF*IDF processing, SVM training and testing and results evaluation. The CHI-square test is adopted for feature word selection. Considering the multi-class classification issue in this field, the One-Against-One approach is adopted.

The main difference of LDA-SVM is the LDA training and topic vectors extraction. The parameters a and β of LDA are set as 1.0/(number of topics). Based on the mixture topic vectors, SVM is also used for document classification. SVM training adopts the same strategy used by TF*IDF-SVM.

For PV-SVM, after the pre-processing of document, an extra document ID is concatenated with the segmented corpus. The processed corpus is fed into PV model to generate the paragraph vector of document. SVM classifier training is based on the generated paragraph vector.

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

3 Experiment Results and Discussions

According to the experimental procedures of Sect. 2.2, through the unsupervised training of TF*IDF, LDA and PV model, the document vectors of each document collection are generated. Based on the generated vectors, the classification experiments are conducted on the three datasets.

3.1 Performances Comparison of Different Methods

The kernel function for SVM is chosen as RBF. The parameters of SVM of TF*IDF-SVM are $C = 2$ and $g = 0.5$, and the parameters of SVM of LDA-SVM and PV-SVM are $C = 2$ and $g = 0.1$. 5-fold cross-validations are implemented on the three datasets. The performances are measured by the macro average and micro average on precision, recall and F-measure [7].

3.1.1 TF*IDF-SVM

For χ^2-test, the ratio is set as 0.4. Through feature extraction and selection, feature vectors of the documents in the three datasets are generated by TF*IDF method. According to the procedures of Sect. 2.2, the classification results of TF*IDF-SVM on the three data sets are shown in Table 2.

From the results in Table 2, it can be found that, owing to the significant difference of feature words in different news categories, TF*IDF-SVM shows better performances on news classification. The low-quality corpus of Tianya Zatan and the semantic understanding of societal risk classification decrease the performance of TF*IDF-SVM significantly.

Table 2. The *Macro_F* and *Micro_F* of TF*IDF-SVM

Reuter	1st fold	2nd fold	3rd fold	4th fold	5th fold	Mean
Macro_F	96.23 %	96.36 %	97.87 %	97.62 %	97.30 %	**97.07 %**
Micro_F	95.90 %	96.98 %	96.54 %	95.90 %	96.31 %	**96.32 %**
Sogou	1st fold	2nd fold	3rd fold	4th fold	5th fold	Mean
Macro_F	92.83 %	92.47 %	92.53 %	91.95 %	88.91 %	**91.74 %**
Micro_F	89.49 %	89.46 %	88.73 %	88.71 %	85.99 %	**88.47 %**
Tianya Zatan	1st fold	2nd fold	3rd fold	4th fold	5th fold	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 %**

3.1.2 LDA-SVM

Through the unsupervised training of LDA model, the mixture topic representations of the documents in the datasets are yielded. To reveal the influences of the number of topics, the performances of the numbers of topics: 50, 100, 150, 200, 250, 300 are tested and compared. According to the procedures of Sect. 2.2, the classification results of LDA-SVM on the three data sets are shown in Table 3.

Table 3. The *Macro_F* and *Micro_F* of LDA-SVM

Reuters	**The number of Topics**	**50**	**100**	**150**	**200**	**250**	**300**
	Macro_F	93.64 %	**94.48 %**	91.52 %	93.61 %	94.41 %	93.52 %
	Micro_F	91.53 %	**92.91 %**	91.92 %	92.78 %	92.35 %	93.00 %
Sogou news datatset	**The number of Topics**	**50**	**100**	**150**	**200**	**250**	**300**
	Macro_F	75.95 %	80.09 %	85.79 %	86.15 %	85.35 %	**87.23 %**
	Micro_F	75.06 %	79.26 %	80.78 %	81.62 %	81.42 %	**82.11 %**
Tianya Zatan	**The number of Topics**	**50**	**100**	**150**	**200**	**250**	**300**
	Macro_F	36.56 %	42.26 %	42.19 %	41.17 %	40.15 %	**43.50 %**
	Micro_F	52.51 %	54.33 %	54.30 %	54.65 %	54.44 %	**54.73 %**

From Table 3, it can be found that, with the increase of the number of topics, the improved performances of LDA-SVM are shown on the three datasets. A significant improvement is appeared from 50 to 100, and the differences of other cases become smaller. A similar result is obtained by Andrew [15].

3.1.3 PV-SVM

Through the unsupervised training of PV model, the distributed representations of the documents in the data set are generated. Except for Tianya Zatan dataset, only the labeled documents are used for PV model training. To train PV model on Tianya Zatan dataset, the new posts (title+text) of Dec. 2011–Mar. 2013, more than 470 thousands posts, are used.

To reveal the influences of vector sizes, the performances of the vector sizes: 50, 100, 150, 200, 250 and 300 are tested and compared. According to the procedures of Sect. 2.2, the classification results of PV-SVM on the three data sets are shown in Table 4.

Table 4. The *Macro_F* and *Micro_F* of PV-SVM

Reuters	**Vector size**	**50**	**100**	**150**	**200**	**250**	**300**
	Macro_F	85.06 %	88.14 %	88.30 %	**88.75 %**	88.66 %	88.42 %
	Micro_F	85.52 %	88.02 %	88.46 %	**88.59 %**	88.54 %	88.41 %
Sogou news datatset	**Vector size**	**50**	**100**	**150**	**200**	**250**	**300**
	Macro_F	63.10 %	70.16 %	75.29 %	79.25 %	83.34 %	**86.16 %**
	Micro_F	61.27 %	68.09 %	72.13 %	75.35 %	78.17 %	**80.40 %**
Tianya Zatan	**Vector size**	**50**	**100**	**150**	**200**	**250**	**300**
	Macro_F	35.77 %	44.79 %	46.25 %	47.26 %	47.86 %	**48.20 %**
	Micro_F	53.48 %	55.16 %	55.85 %	56.36 %	56.74 %	**57.03 %**

From Table 4, it can be found that, on the three datasets, with the increase of vector size, the performances of PV-SVM are improved. However, the improvements of *Macro_F* and *Micro_F* are declined, but the improvement tendencies are different for different data sets.

From the results of Tables 2, 3, 4, TF*IDF-SVM obtains overall best performance. Toward Reuters-21578 and Sogou news dataset, the performances of LDA-SVM are better than PV-SVM. Although LDA and PV extract semantic information from documents, the reduced dimension of the two representations loses much information, which leads to the decrease of general performance of LDA-SVM and PV-SVM. However, the dimension of BOW is at least 10 thousands, and the computation and time cost of TF*IDF -SVM are much bigger than the two other methods. Meanwhile, the parameters of SVM are also important to document classification, while this study does not consider the parameter optimization, and the parameters are set by experiences.

3.2 The Influence of Stop Words to LDA and PV

To test the influence of stop words elimination to LDA and PV, Sogou news dataset is selected. Two kinds of experiments are required: (I) the training corpus with stop words; (II) the training corpus without stop words. As the results presented in Sect. 3.1, the performances of LDA and PV model training without stop words have been compared.

In this section, only the experiments of model training with stop words are conducted. To fully compare the performance of LDA-SVM and PV-SVM, two more cases: the number of topics or vector size of 400 and 500 are implemented. The results are shown in Table 5.

Table 5. The *Macro_F* and *Micro_F* of LDA-SVM and PV-SVM for Sogou with Stop Words

LDA-SVM	**The number of topics**	**50**	**100**	**150**	**200**	**250**	**300**	**400**	**500**
	Macro_F	73.82 %	77.26 %	79.50 %	80.46 %	84.42 %	85.02 %	83.50 %	85.71 %
	Micro_F	73.32 %	76.27 %	78.20 %	78.57 %	79.15 %	79.69 %	79.56 %	80.33 %
PV-SVM	**Vector size**	**50**	**100**	**150**	**200**	**250**	**300**	**400**	**500**
	Macro_F	71.72 %	77.28 %	80.73 %	83.80 %	86.36 %	88.18 %	91.28 %	**92.71 %**
	Micro_F	66.69 %	73.06 %	76.95 %	79.96 %	82.33 %	84.23 %	87.42 %	**89.79 %**

As it can be found in Table 5, without stop words elimination, the performance of PV-SVM is more effective than LDA-SVM on Sogou news dataset. However, the results presented in Sect. 3.1, the performance of LDA-SVM is more effective than PV-SVM on Sogou news dataset with stop words elimination. Considering the performances of LDA-SVM and PV-SVM on Sogou with/without stop words, PV-SVM on Sogou news with stop words shows dominant superiority. Meanwhile, the performance of PV-SVM on 500-dimension is also better than TF*IDF-SVM, so PV-SVM may generate better performance than LDA-SVM or TF*IDF-SVM with the increase of dimension.

For LDA, if keeping all stop words, these stop words show similar possibility to all topics, which will decline the clarity of each topic, and affect the performance of LDA-SVM. For PV model, the paragraph token acts as a memory that remembers what is missing from the current context – or the topic of the paragraph. The contexts are fixed-length and sampled from a sliding window over the paragraph for PV model training. In this mode, stop words bring useful information to different documents, and improve the performance of PV-SVM. Therefore, for LDA model training, stop words elimination of the training is necessary, but for PV model training, keeping all words will be more effective.

3.3 The Influence of Data Size to PV

From the previous results, it can be found that LDA model performs better on small datasets: Reuter and Sogou, and PV-SVM obtains better performance on the big dataset: Tianya Zatan dataset, due to almost 50 thousands posts for training. For this reason, to reveal the influence of data size to PV training, the documents of Reuter and Sogou are repeated one and two times for PV training, the results are shown in Table 6 and Table 7.

From Tables 6 and 7, on repeated Reuters-21578 dataset, compared with the non-repeated dataset, the *Macro_F* and *Micro_F* of PV-SVM are significantly increased. A tiny growth of performance is shown from the dataset repeated once to the dataset repeated twice. Conversely, a decrease of *Macro_F* and *Micro_F* on Sogou news dataset is shown, and the more the data repeated, the bigger decrease of performance is generated. As can be found, the data sizes of Reuters-21578 dataset and Sogou news dataset are different, and the size of Reuters-21578 is much smaller than Sogou news dataset. It can be concluded that the training process of PV on Reuters-21578 dataset is under-fitting, so the repeated dataset improves the performance of classification. While the training samples of Sogou news dataset is enough for PV model training, so the repeated data will lead over-fitting to PV model, which only makes worse results. Hence, a proper size of training samples is important to the performance of PV model.

Table 6. The *Macro_F* and *Micro_F* of PV-SVM for Reuters

Reuters repeated once	**Vector size**	**50**	**100**	**150**	**200**	**250**	**300**
	Macro_F	92.06 %	92.12 %	92.45 %	92.87 %	**93.04 %**	92.29 %
	Micro_F	91.57 %	91.96 %	92.69 %	92.69 %	**93.00 %**	92.65 %
Reuters repeated twice	**Vector size**	**50**	**100**	**150**	**200**	**250**	**300**
	Macro_F	92.92 %	92.65 %	93.04 %	93.60 %	**93.20 %**	93.19 %
	Micro_F	92.78 %	92.65 %	93.17 %	93.47 %	**93.56 %**	93.52 %

Table 7. The *Macro_F* and *Micro_F* of PV-SVM for Sogou

Sogou repeated once	**Vector size**	**50**	**100**	**150**	**200**	**250**	**300**
	Macro_F	65.08 %	70.71 %	75.49 %	79.39 %	82.96 %	**85.22 %**
	Micro_F	62.83 %	67.97 %	71.44 %	74.25 %	76.47 %	**78.41 %**
Sogou repeated twice	**Vector size**	**50**	**100**	**150**	**200**	**250**	**300**
	Macro_F	65.01 %	70.71 %	75.30 %	78.72 %	82.39 %	**84.67 %**
	Micro_F	62.65 %	67.65 %	70.84 %	73.50 %	75.62 %	**77.61 %**

4 Conclusions

In this paper, experiments are conducted to examine the performances of three document representation methods: TF*IDF, LDA and PV for document classification. Basically, two kinds of metrics should be considered: speed and accuracy. Hence, the contributions of this paper can be summarized as follows.

(1) According to the performance comparison of these three strategies on Reuters21578, Sogou news and Tianya Zatan datasets, TF*IDF-SVM shows overall best performance, and LDA-SVM generates better results on small datasets than PV-SVM;
(2) The stop words elimination shows different effects to the performances of LDA-SVM and PV-SVM, and PV-SVM generates much better results when keeping all words, even better than TF*IDF-SVM;
(3) Through the experiments on the repeated training data, it is seen that a proper size of training samples is also important to PV model.

Although we have obtained some preliminary conclusions of TF*IDF, LDA and PV methods, more experiments are required for a comprehensive study. Furthermore, based on the conclusions of this research, how to improve the performance of document classification based on these methods is the future task of this research.

Acknowledgements. This research is supported by National Natural Science Foundation of China under Grant Nos. 61473284, 61379046 and 71371107. The authors would like to thank other members who contribute their effort to the experiments.

References

1. Cao, L.N., Tang, X.J.: Topics and threads of the online public concerns based on Tianya forum. J. Syst. Sci. Syst. Eng. **23**(2), 212–230 (2014). doi:10.1007/s11518-014-5243-z
2. Korde, V., Mahender, C.N.: Text classification and classifiers: a survey. Int. J. Artif. Intel. Appl. **3**(2), 85–99 (2012). doi:10.5121/ijaia.2012.3208
3. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. **34** (1), 1–47 (2002). doi:10.1145/505282.505283

4. Manuel, F.D., Eva, C., Senén, B., Dinani, A.: Do we need hundreds of classifiers to solve real world classification problems? J. Mach. Learn. Res. **15**(1), 3133–3181 (2014)
5. Zhang, W., Yoshida, T., Tang, X.J.: A comparative study of TF*IDF, LSI and Multi-words for text classification. Expert Syst. Appl. **38**(3), 2758–2765 (2011). doi:10.1016/j.eswa.2010.08.066
6. Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning, C.D., Ng, A.Y., Potts, C.: Recursive deep models for semantic compositionality over a sentiment treebank. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing *(EMNLP),* pp. 1631–1642. ACL (2013)
7. 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, Canada), pp. 187–193. AAAI (2014)
8. Tang, X.J.: Exploring on-line societal risk perception for harmonious society measurement. J. Syst. Sci. Syst. Eng. **22**(4), 469–486 (2013). doi:10.1007/s11518-013-5238-1
9. Blei, D., Ng, A., Jordan, M.: Latent Dirichlet allocation. J. Mach. Learn. Res. **3**(5), 993–1022 (2003)
10. Tang, X.B.: Fang XK (2013) Research on Micro-blog topic retrieval model based on the integration of text clustering with LDA. Info. Stud. Theory Appl. **8**, 85–90 (2013). (in Chinese)
11. Li, K.L., Xie, J., Sun, X., Ma, Y.H., Bai, H.: Multi-class text categorization based on LDA and SVM. Procedia Eng. **15**, 1963–1967 (2011). doi:10.1016/j.proeng.2011.08.366
12. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**, 1137–1155 (2003)
13. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceeding of International Conference on Learning Representations (ICLR2013, Scottsdale), pp. 1–12 (2013)
14. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning (Beijing). JMLR Workshop and Conference Proceedings, pp. 1188–1196 (2014)
15. Andrew, M.D., Christopher, O., Quoc, V.L.: Document embedding with paragraph vectors. arXiv:1507.07998 (2015)
16. Zhao, Y.L., Tang, X.J.: A preliminary research of pattern of users' behavior based on Tianya forum. In: Wang, S.Y. (eds.) The 14th International Symposium on Knowledge and Systems Sciences, Ningbo, pp. 139–145. JAIST Press (2013)
17. 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)