J Inf Process Syst, Vol.16, No.2, pp.329~342, April 2020
https://doi.org/10.3745/JIPS.04.0165

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Topic Extraction and Classification Method Based on Comment Sets

Xiaodong Tan*

Abstract

In recent years, emotional text classification is one of the essential research contents in the field of natural language processing. It has been widely used in the sentiment analysis of commodities like hotels, and other commentary corpus. This paper proposes an improved W-LDA (weighted latent Dirichlet allocation) topic model to improve the shortcomings of traditional LDA topic models. In the process of the topic of word sampling and its word distribution expectation calculation of the Gibbs of the W-LDA topic model. An average weighted value is adopted to avoid topic-related words from being submerged by high-frequency words, to improve the distinction of the topic. It further integrates the highest classification of the algorithm of support vector machine based on the extracted high-quality document-topic distribution and topic-word vectors. Finally, an efficient integration method is constructed for the analysis and extraction of emotional words, topic distribution calculations, and sentiment classification. Through tests on real teaching evaluation data and test set of public comment set, the results show that the method proposed in the paper has distinct advantages compared with other two typical algorithms in terms of subject differentiation, classification precision, and F1-measure.

Keywords

Comment Text Set, Emotional Classification, LDA Topic Model, Support Vector Machine

1. Introduction

As an emerging field of data mining, text sentiment analysis, also called opinion mining, is to identify the implicit sentimental information through the analysis on comments from online users. It is a process of analyzing, processing, summarizing and reasoning on the sentimental texts [1,2]. At present, sentiment analysis is mainly applied in such fields as commodities and news comments. Hai et al. [3] applied sentiment analysis to the texts from Twitter commentary data and found that the data has unique linguistic characteristics, such as informal and creative. Several research [4,5] applied sentiment analysis to study the commodity comment data. In the field of sentiment analysis, the main task is to identify subjective views of texts and to judge positive and negative emotion tendencies. “Dictionary rule method” and “machine learning method” are two methods frequently used in sentiment analysis. The authors of [6-8] used emotional dictionary to classify by mainly considering negative words, uppercase and lowercase letters, etc., in order to calculate the emotional polarity value, so as to better classify them. Several research [9,10] applied supervised learning methods to emotion classification.

Manuscript received December 6, 2018; first revision February 27, 2019; accepted April 22, 2019.

Corresponding Author: Xiaodong Tan (370911641@qq.com)
* School of Mathematics and Computer Science, HeZhoul University, Guangxi, China (370911641@qq.com)

By comparing the characteristics of one-character features, binary features, adjective scoring, position and feature weight selection strategies, they obtained the effects of various machine learning classification algorithms. The latent Dirichlet allocation (LDA) model [11,12], also called a three-layer Bayesian probability model and proposed by Blei et al. [13] in 2010, is a document topic generation model that consists of three layers of words, topics, and documents. The LDA is an unsupervised machine learning technique that can be used to identify latent topic information in large document sets or corpora. Later, Jin et al. [14] applied the LDA model to social media user recommendations to mine user interests and then recommend new friends with the same interests and hobbies. Xia et al. [15] applied the LDA model to dynamic automatic annotation to elicit the latent topics of resources and related language chats so that the resources can be effectively marked according to latent topics. In recent years, the deep learning and reinforcement learning have been widely studied in some application such as bioinformatics mining, multimedia documents mining, etc. [16-18].

However, there still exist defects and application limitations in the methods mentioned above. Under this background and based on LDA and support vector machine (SVM), a high-precision sentiment classification method has been presented in this paper in order to effectively improve the topic distribution effect of LDA and the sentiment classification quality of the comment set.

With topic-related words submerged by high-frequency words, which leads to reduction of classification precision, my motivation is to study a weighted LDA (W-LDA) topic model to solve the above problem. In this topic model, we, for the first time, propose a measure of adding weighted value to the LDA. In general, the contributions in this paper include the following aspects:

- An average weighted value is introduced to avoid topic-related words from being submerged by high-frequency words, so as to improve the distinction of the topic. We use Gaussian function to weight topic-related feature words.
- On the basis of other affective dictionaries, a set of affective dictionaries suitable for teaching evaluation content is constructed by manual processing.
- We have conducted various experiments to prove effectiveness and efficiency of our method.

2. LDA Topic Model Building

2.1 Topic Model Idea

The main goal of the LDA topic model is to calculate the probability distribution of the subject-word from a document or multiple documents. The probability distribution here is a polynomial distribution, a typical bag of words model and an unsupervised learning algorithm [19]. Symbols and related descriptions of the LDA topic model mentioned in this paper are shown in Table 1.

Given a topic document set *M*, which contains m documents, that is $M = \{M_1, M_2, ..., M_m\}$. The set contains k independent topics, each topic presents random polynomial probability distribution characteristics, and there are several polynomial probability distribution words for each topic, and the polynomial probability distributions in here meet the Dirichlet distribution. Therefore, the manifestation of this document can be characterized as a LDA topic model, whose generation process is as follows:

(1) To generate the topic distribution $\vec{\theta}_m$ of the document M_m through sampling from the prior knowledge α of the Dirichlet distribution.

(2) To generate the topic $Z_{m,n}$ of the nth vocabulary of the document M_m through sampling from the topic distribution $\vec{\theta}_m$ obtained in step (1).

(3) To generate a vocabulary distribution $\vec{\varphi}_{Z_{m,n}}$ corresponding to topic $Z_{m,n}$ through repeatedly sampling from prior knowledge $\vec{\beta}$ of Dirichlet distribution.

(4) Combining the topic $Z_{m,n}$, to generate a vocabulary $W_{m,n}$ through sampling from the word distribution $\vec{\varphi}_{Z_{m,n}}$.

(5) From above steps, repeat (2), (3), and repeat (2), (3), (4) until all vocabularies are selected.

Table 1. Symbols descriptions of LDA topic model

Symbol	Explanation
K	The number of topics
m	The number of documents
M_m	The mth document
N_m	The number of words contained in the mth document
V	The total number of all words
$\vec{\alpha}$	The Dirichlet prior hyper parameter to the document set, which is a k-dimensional vector
$\vec{\beta}$	The Dirichlet prior hyper parameter to any topic, which is a v-dimensional vector
$Z_{m,n}$	The topic of the nth word of the mth document
$\vec{\varphi}_{Z_{m,n}}$	The word distribution to the topic $Z_{m,n}$
$\vec{\theta}_m$	The topic distribution of the mth document
$W_{m,n}$	The nth word of the mth document

The intuitive representation of the above LDA generation process is illustrated as Fig. 1. The rectangle in Fig. 1 is represented as a loop, the data in the lower right corner of the rectangle as the number of loops, the hollow circles as latent variables, and the shaded circles as observable variables. Their dependencies are indicated by the arrow directions.

The joint distribution probability of all parameters of the LDA model shown in Fig. 1 is calculated as shown in Expression (1). Assume that there is an existing document set M, $\vec{\alpha}$ is the document topic, and $\vec{\beta}$ is prior distribution parameters to the vocabulary of the topic and, i.e., they are $\phi = \{\vec{\varphi}Z_{m,n}\}$ and $K \times V$ *matrix* . The joint distribution probability of all parameters of the model corresponding to Fig. 1 is calculated as shown in Expression (1).

$$P(\vec{W}_m,\vec{Z}_m,\vec{\theta}_m,\phi)=\prod_{n=1}^{N_m}P(W_{m,n}\mid\vec{\varphi}_{Z_{m,n}})P(Z_{m,n}\mid\vec{\theta}_m)P(\vec{\theta}_m\mid\vec{\alpha})P(\phi\mid\vec{\beta}) \tag{1}$$

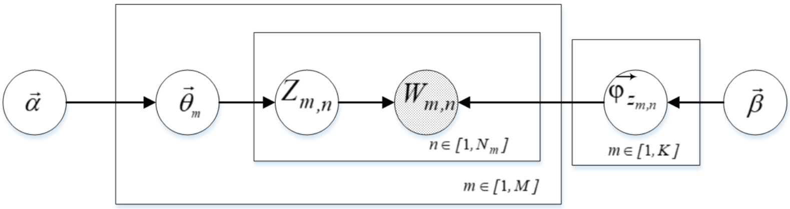


Fig. 1. Generation process of LDA topic model.

2.2 W-LDA Improved Topic Model Derivation

The data samples studied obtains from the university teaching evaluation and public comment data sets. In the research, it's found that, if using only the traditional LDA model, most of the words that represent the topic would be submerged by high frequency words. Thus expression capabilities of the issue would be significantly affected. Although the elimination of high frequency words can currently be achieved in some algorithms by stopping words or setting thresholds. There are no stop words in some fields, and unsatisfactory results are produced even if it is processed manually by domain experts. Give this situation, a Gaussian function has been used in this paper to perform weighting procedures to feature words to construct an improved W-LDA topic model. The specific processing method is as follows: First, using the Expression (2) to weight the word m in the document:

$$a_m = exp\left[-\frac{(f_m - f_i)^2}{2\sigma^2}\right] \tag{2}$$

Among them: the variance $\sigma^2 = \frac{\sum_{m=1}^{V}(f_m - f_i)^2}{V-1}$, f_m is the frequency of word m, f_i is the frequency of word i that lies in the middle of the word frequency. In order to make the total number of words approximately consistent before-and-after the data set being weighted, the average value is treated with Expression (2). The processing formula is shown in Eq. (3), where the average weight named $weight_m$ will be applied to Expression (8) to calculate the expectations of topics and their word distribution.

$$weight_m = \frac{n}{\left[\sum_{m=1}^{V} f_m \cdot a_m\right]} \cdot a_m \tag{3}$$

In the derivation of the LDA topic model, a sampling algorithm or variational method is usually used. As a fast and effective sampling algorithm, Gibbs sampling algorithm is used here to generate Markov chain, and then obtain a complex multivariate distribution. As is shown in Expression (1) and Fig. 1, that the topic polynomial distribution θ is generated by a priori parameter α, and the word polynomial distribution ϕ under topic $Z_{m,n}$ is generated by β, then Expression (1) can be transformed into a union probability distributions as shown in Expression (4).

$$P(\vec{W},\vec{Z} \mid \vec{\alpha},\vec{\beta}) = P(\vec{Z} \mid \vec{\alpha})P(\vec{W} \mid \vec{Z},\vec{\beta}) \tag{4}$$

It can be seen that the solution of Expression (4) is divided into two processes, one of sampling topics $\vec{Z}$ by prior parameters and the other of sampling words based on topic $\vec{Z}$ and prior parameters $\vec{\beta}$. To the first factor $P(\vec{Z}|\vec{\alpha})$ in Expression (4), according to the topic θ generated by α, and Z generated by θ, we can get Expression (5) as follows:

$$P(\vec{Z} \mid \vec{\alpha}) = \int P(\vec{Z} \mid \theta)P(\theta \mid \vec{\alpha})d\theta = \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m + \vec{\alpha})}{\Delta(\vec{\alpha})}, \vec{n}_m = \{ n_m^{(k_i)} \}_{K=1}^{K} \tag{5}$$

Among them: $n_m^{(k_i)}$ represents the number of occurrences of topic *k* in the document *m*.

Similarly, the expansion formula of the second factor in Expression (5) can be obtained as shown in Expression (6).

$$P(\vec{W} \mid \vec{Z},\vec{\beta}) = \int P(\vec{W} \mid \vec{Z},\phi)P(\phi \mid \vec{\beta})d\phi = \prod_{Z=1}^{K} \frac{\Delta(\vec{n}_Z+\vec{\beta})}{\Delta(\vec{\beta})}, \vec{n}_Z = \{n_Z^{(v)}\}_{v=1}^{V} \tag{6}$$

Among them: $n_Z^{(v)}$ represents the number of occurrences of the word v in the topic *Z*, and *K* is the product of the model.

When comprehensively using Expressions (5) and (6), the joint distribution formula for words and topics can be obtained as:

$$P(W,Z \mid \alpha,\beta) = \prod_{Z=1}^{K} \frac{\Delta(\vec{n}_Z+\vec{\beta})}{\Delta(\vec{\beta})} \cdot \prod_{m=1}^{M} \frac{\Delta(\vec{n}_m+\vec{\alpha})}{\Delta(\vec{\alpha})} \tag{7}$$

Based on (7), through the conditional distribution of the latent topic variable *Z* of the observable variable $\vec{W}$ and the characteristics of the Dirichlet distribution, the topic of the document *m* and the expectation formula of the word distribution of the topic can finally be obtained as :

$$\theta_{m,k} = \frac{n_m^{(k)}+\alpha_k}{\sum_{k=1}^{K} n_m^{(k)}+\alpha_k}, \varphi_{k,v} = \frac{n_k^{(v)}+\beta_v}{\sum_{v=1}^{V} n_k^{(v)}+\beta_v} \tag{8}$$

According to the calculation method of the average weighted value of Expression (3), during Gibbs samples words processes, when the topics are distributed to document *m*, then, to $n_m^{(k)}$, not 1 is simply added but the average weighted values are accumulated.

3. Sentiment Analysis Based on Comment Corpus

Abundant sentiments implied in the comment corpus on which the method of analysis is used in this section. The technique is to overall extract sentiment words from the comment sets. They are then using the W-LDA topic model to obtain the topic distribution of the comment corpus and the word probability distribution table of the topic. Finally, adopting SVM to make two classifications of positive and negative polarity for comments, and to make a scientific evaluation of the classification results. The sentiment classification framework based on comment sets is illustrated as Fig. 2.

3.1 Sentiment Word Analysis

Text sentiment analysis, including the detection, analysis and mining of the text subjectivity as the user's opinions, hobbies, and emotions, is a multidisciplinary research area involving natural language processing, computational linguistics, information retrieval, machine learning and artificial intelligence.

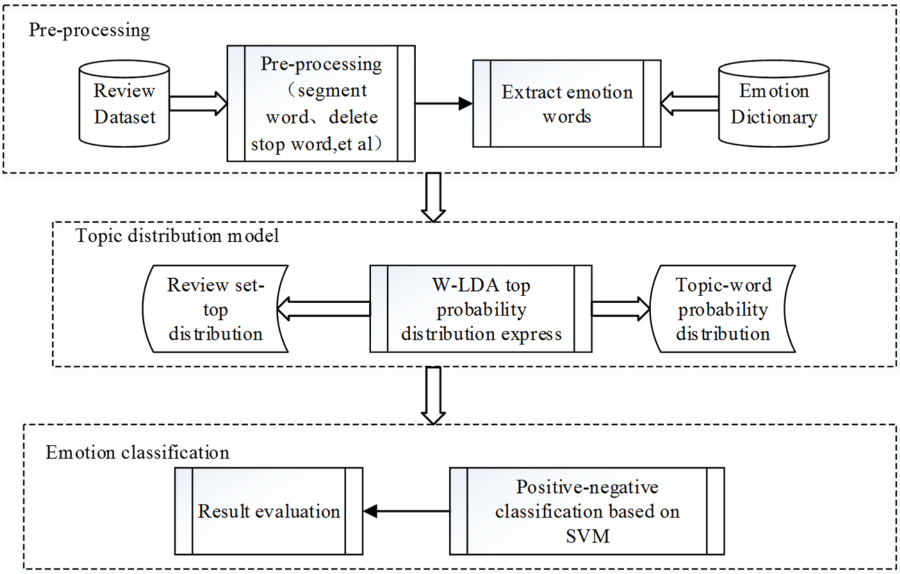

Pre-processing
Review Dataset
Pre-processing（segment word、delete stop word,et al）
Extract emotion words
Emotion Dictionary
Topic distribution model
Review set-top distribution
W-LDA top probability distribution express
Topic-word probability distribution
Emotion classification
Result evaluation
Positive-negative classification based on SVM

Fig. 2. Emotional classification processing framework based on comment sets.

There are two methods for sentiment analysis, one based on sentiment dictionary and the other on machine learning [19]. The technique of sentiment dictionary is firstly used in this paper to perform the simple topic distribution calculation of the teaching comment corpus, and then the machine learning algorithm SVM to do the sentiment analysis. As is known, that there are many English sentiment dictionaries and few Chinese ones. The frequently-used Chinese sentiment dictionaries are Li Jun's Chinese derogatory, meaning dictionary of Tsinghua University in China, sentiment dictionary of HowNet, and the simplified Chinese sentiment dictionary of Taiwan University. Among them, the HowNet sentiment dictionary, with each sentiment word as a benchmark to find degree adverbs and negative words in a positive direction, is used to find sentiment words in text corpus. Based on these three kinds of sentiment dictionaries, a set of emotional dictionaries for teaching, electronic products, books, and other commentary contents have been constructed in this paper through manual processing. Specific descriptions are illustrated as Table 2.

Table 2. Chinese emotional dictionary

Category	Number of pieces	Classification standard	Part of the words
Claim word	97	self-definition, feeling	feel, find, etc.
Positive emotion word	1,381	love to things	like, so nice, etc.
Positive evaluation	10,841	sure, support to things	not bad, earnest, etc.
Negative emotional word	1,453	disgust, dislike to things	disgust, uncomfortable, etc.
Negative evaluation word	12,271	oppose, negate to things	poor, difficult to comprehend, etc.
Degree word	524	decorate, restrict for adjective or adverb	very, extremely, etc.

It's found that the polarity of sentiment words extracted from the comment set is more evident through experiments on teaching evaluation data based on the new sentiments established, which is very favorable to the feature representation of the topic model. Table 3 shows the results of the emotional word extraction of a piece of teaching evaluation content.

Table 3. Emotional word extraction results of partial evaluation

Original evaluation corpus	Emotional word extraction result
New teaching methods, fascinating, easy to understand, body language is very rich, Explain vivid, I really like it, It is better that the teacher will reduce the speed of lectures.	Novel, attractive, into the victory, very, abundant, vivid, really, like, old, more, good, a little

3.2 SVM Classification Model

The SVM model [9] proposed by Corinna Cortes and others has many unique advantages in small sample, nonlinear and high dimensional pattern recognition. Its primary purpose is to find the hyperplane with the most significant separation as the boundary of the classification. The kernel function is introduced in the SVM to subtly solve problems in the inner product operation and the nonlinear classification. Frequently-used kernel functions include linear kernel functions, polynomial kernels, and radial basis kernels. The radial basis function (RBF) kernel function is relatively stable, and the polynomial kernel function is less stable. In this paper, the RBF kernel function is chosen to implement the SVM model. The kernel function manifests as:

$$K(X_i,X_j)=exp(-\gamma\,||\,X-X_e\,||^2) \tag{9}$$

Eq. (9) is a monotonic function of the Euclidean distance from any point X to center X_e in space. The kernel function of the SVM has two parameters: the penalty factor C and the kernel parameter γ. The penalty factor represents the margin of error with C higher and no tolerance of errors. The kernel parameter γ implicitly determines the distribution of data mapping to the new feature space. The higher γ is, the smaller the support vector is.

Next, the feasibility of choosing SVM and RFB kernels will be expounded. The purpose of the kernels is to map the feature data from low-to-high dimensions so that the data can be linearly separable. In practice, if any data can be linearly separable, it is only necessary to change the parameter σ of the RBF function to be sufficiently small. The separability of features becomes possible when the comment data set is mapped into a high-dimensional space. Therefore, it is feasible to use SVM and select RFB to classify the comment data set in this model.

Based on the SCKIT-learning machine learning toolkit, the grid search method is used in this paper to determine the value of C and γ. The process of building the SVM classification model is shown in the following steps and illustrated as Fig. 3.

(1) SVM training process: Select the RBF kernel function to map the W-LDA topic feature distribution to the high dimensional feature space. Then, finding the optimal classifying hyperplanes in the high dimensional feature sample space. Training them as support vectors and VC credibility until a classification discriminant function of the commentary material rate is finally obtained.

(2) SVM classification judgment process: SVM uses the RBF kernel function to map the feature vector of the comment set to be classified to the high-dimensional feature space. Using it as the input parameter of the classification discriminant function obtained in the training stage. Finally, outputs the emotional two-category results of the comment set by the discriminant function.

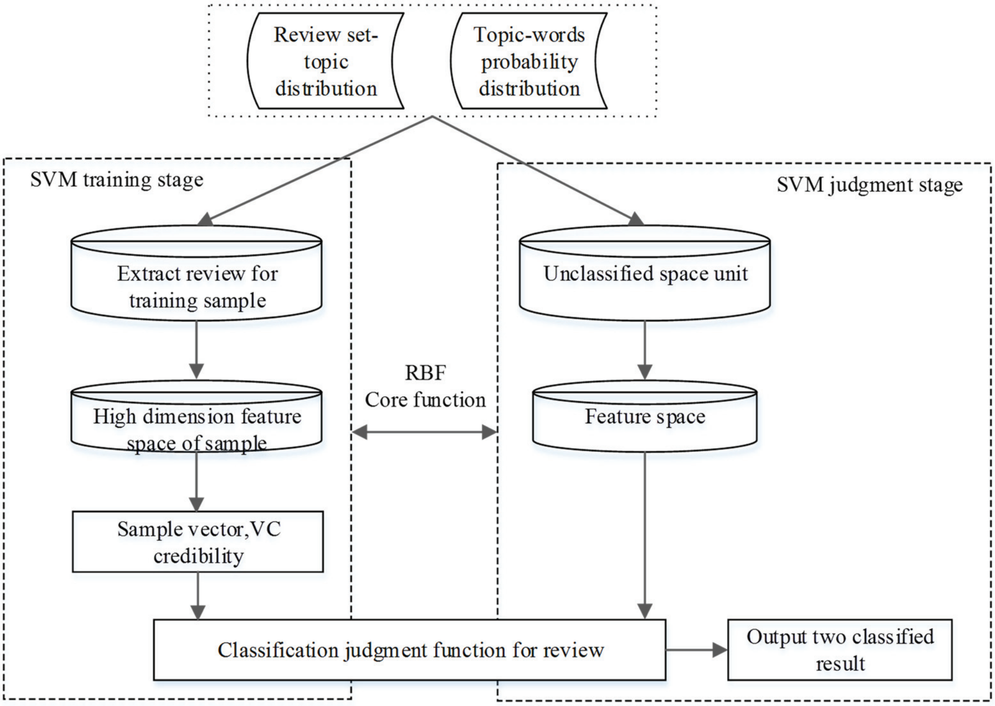

Review set-topic distribution
Topic-words probability distribution
SVM training stage
SVM judgment stage
Extract review for training sample
Unclassified space unit
RBF Core function
High dimension feature space of sample
Feature space
Sample vector,VC credibility
Classification judgment function for review
Output two classified result

Fig. 3. SVM-based comment sentiment classification process.

3.3 Example of Sentiment Analysis Combining W-LDA Feature Representation with SVM Classification

As shown in the W-LDA improved topic model described in Section 2.1, Gibbs sampling has been used in this paper to estimate the parameters of the model. And it's found that better output results can be obtained with less Gibbs sampling calculation. At the same time, considering that the content of the commentary documentation is small, a particular topic number variable T can be set for the W-LDA model, and the topic distribution range of the commentary document is $[2,T]$. After W-LDA obtains the topic feature representation vector, SVM is used for classification, so that all records of the comment set are better classified. Here, the results of the word segmentation of six teaching evaluation data in a teaching evaluation system in an individual university are given first. The specific situation is illustrated as Table 4.

Table 4. Sentiment feature representation of teaching evaluation data

Data	Positive emotion word	Degree word	Negative evaluation word
Teaching evaluation text 1	positivity, creativity, passion, infection	especially	no
Teaching evaluation text 2	good, distinct, responsible	very, already	null
Teaching evaluation text 3	High, level	sufficient, then	no, less
Teaching evaluation text 4	passion, strict, responsible	usual	little, no
Teaching evaluation text 5	encourage, help, greatest love	repeatedly	null
Teaching evaluation text 6	blame, love, friendship	often, fully	nor

Through emotional words, degree adverbs, and negative words, weights are calculated to obtain positive and negative sentiment tendencies. As illustrated in Fig. 4, there are five evaluation criteria: teaching attitude, interaction, professional skills, teaching quality, and language expression. In this paper, the review materials of teachers' courses summarize five aspects. Teachers have excellent communicative competence and teaching technique from the results but show moderate levels in teaching attitude, professional skills, and language expression.

Next, the effectiveness of the SVM classification model will be further validated. This research utilizes score data of 387 real teaching evaluation indicators. The W-LDA topic model is applied first after the data processing. The relevant parameters of the model are calculated as follows: a priori parameters α=2, β=0.01, k=25. Experiments show that the precision of the model is the highest when k=25, then four sets of tests are performed with different kernel functions. The results show that the SVM classifier based on the RBF kernel function works well. The specific situation is illustrated in Fig. 5.

From the above analysis and test results, it can be included that the W-LDA topic model proposed in this paper takes into account the importance of topic-related words and that it has been paid attention to in sentiment analysis. At the same time, the parameters of the W-LDA topic model proposed in this paper still need to be continuously trained and optimized, combined with the selection of SVM for the kernel function of the comment data set, to achieve higher precision of sentiment classification.

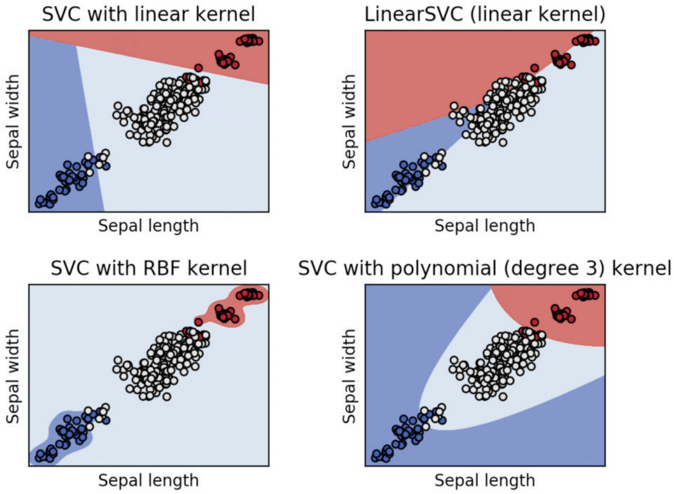


Fig. 5. Comparison of SVM classification effect to different kernel functions.

4. Experiment and Result Analysis

The sentiment classification algorithm of this paper and the hardware environment for series of experiments are: Intel Core i7-6700 3.40 GHz CPU, 8 G RAM and 1 TB hard disk. The operating system is Windows 10. The program code for the algorithm was developed by Python 3.

The word segmentation tool used in this section of the experiment is Jieba. The topic model adopts the improved W-LDA topic model algorithm proposed in the paper, and the classification method is the SVM algorithm of RBF kernel function. The proposed algorithm is contrasted with the single SVM classification and the classification method in [20] to verify its effectiveness, advantages, and disadvantages. The performance evaluation indicators of the test algorithm are accurate, recall ratio, and F1 test. The equations for the three symbols are:

$$Precision = mc / mt, Recall = mc / ml \ ,$$

$$F1\text{-}Measure = 2 * Precision * Recall /(Precision + Recall) \ ,$$

where *mc* is denoted by the number of correctly classified records, *mt* is denoted by the total number of records, and *ml* is denoted by the number of labeled records in samples. The corpus, collected from the database of Datatang, includes two types of comment sets of the hotel and of the electronic product. Their polarity distributions are illustrated in Table 5.

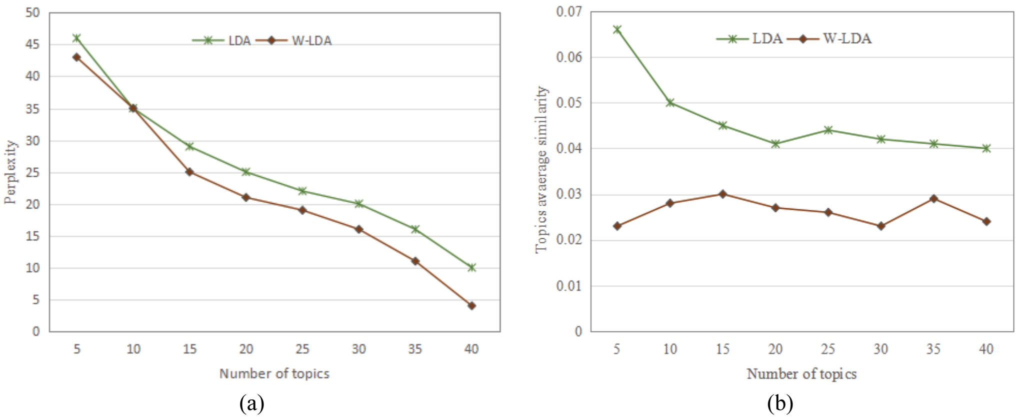
Table 5. Polarity distribution of experimental corpus

Corpus category	Positive comment number	Negative comment number
Comment set of electronic product	2,730	2,650
Comment set of hotel	1,080	1,100

4.1 Performance Experiment of Sentiment Theme Distribution

This experiment uses the general electronic product review data set shown in Table 4. The circumstance of the emotional dictionary, as shown in Section 2.1. A priori parameters and α=50/K and β=0.01, using half-off cross-validation test results, with each training iteration 1000 times, the experimental results are illustrated as Fig. 6(a) and (b).

It can be found that-compared to the traditional LDA topic model from Fig. 6. The improved W-LDA topic model proposed in this paper has more considerable advantages in terms of the complexity value of the topic distribution computation and the average similarity between topics. It's shown that the W-LDA topic model has a more reliable predictive power of distribution of the topics of the new comment set, a higher quality distribution of the topics obtained, and a high degree of differentiation of topic types.

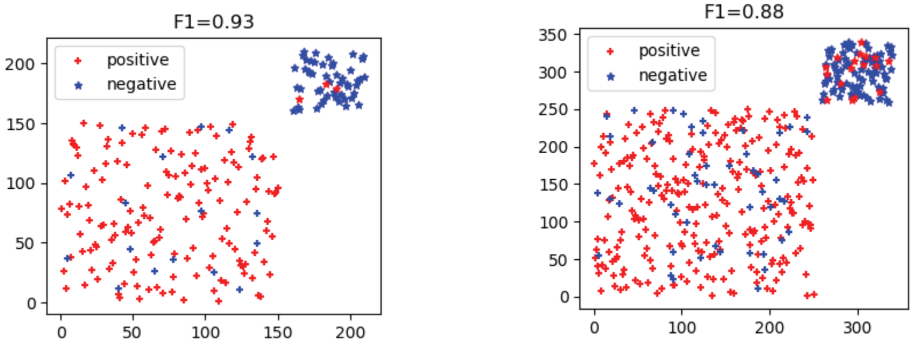


(a) (b)

Fig. 6. Comparison of experimental results of two topic models: (a) complexity contrast and (b) topic independence contrast.

4.2 Emotional Classification Experiment

The tagged data have been manually processed based on the experimental data of the hotel comment set shown in Table 5 to verify the flexibility and advantages of the algorithm. From the tagged data, what have been extracted are 200 records for 10 topics, 400 records for 20 topics, 800 records for 30 topics, 1,600 records for 40 topics. Fig. 7 is shown the classification of 200 comment records for 10 topics and 200 comment records for 20 topics. Each comment record is set with a unique serial number during classifying experiments to intuitively display whether the classification of each records is correct or not. The positive and negative polarity record numbers are distributed in different areas. The blue labels in the red dot area and the red label in the blue area are all the wrong classification records, which is illustrated in Fig. 7(a) and (b).



(a)

(b)

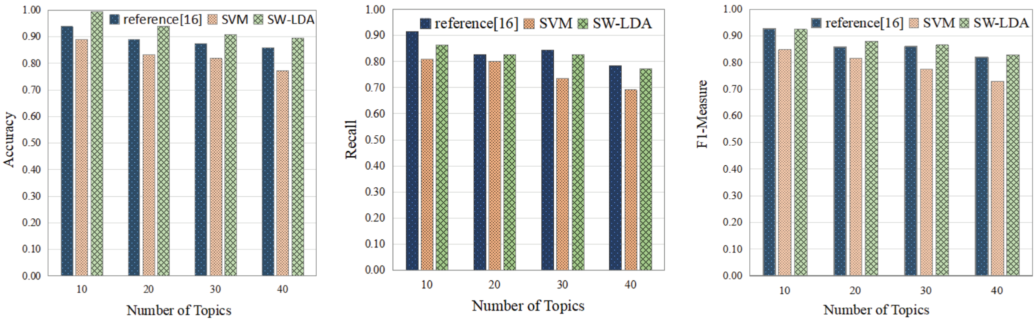
Fig. 7. Sentiment two-category results based on comment set: (a) classification of 10 topics, 200 comment records, and (b) classification of 10 topics, 400 comment records.

Table 6. Experimental results comparison of sentiment classification algorithm

# of topics	Method	# of marked records	Classification result	Correct number	Wrong number	Precision	Recall	F1 value
10	FESCSC	200	195	183	12	0.94	0.92	0.93
	LDA with SVM	200	182	162	20	0.89	0.81	0.85
	SW-LDA	200	174	173	1	0.99	0.87	0.93
20	FESCSC	400	372	331	41	0.89	0.83	0.86
	LDA with SVM	400	384	320	64	0.83	0.80	0.82
	SW-LDA	400	352	331	21	0.94	0.83	0.88
30	FESCSC	800	773	676	97	0.87	0.85	0.86
	LDA with SVM	800	720	589	131	0.82	0.74	0.78
	SW-LDA	800	727	661	66	0.91	0.83	0.87
40	FESCSC	1,600	1461	1255	206	0.86	0.78	0.82
	LDA with SVM	1,600	1433	1106	327	0.77	0.69	0.73
	SW-LDA	1,600	1381	1235	146	0.89	0.77	0.83

Four sets of experiments were performed on these experimental data to compare the sentiment classification method proposed in this paper with the algorithm in [20] (name after FESCSC) and the traditional SVM algorithm (name after LDA with SVM). The four sets of experiments were based on the distribution of different topics number—our method (SW-LDA), named after W-LDA with SVM. After the processing operation shown in Fig. 2, the final experimental results illustrated in Table 6 and the visual representation is shown in Fig. 8.

The experimental results given above show that the algorithm proposed in this paper is superior to the other two methods in precision and total measured values of F1 and is advantageous on accuracy. The precision of our approach is 0.04 and 0.1 higher than that of FESCSCH and LDA with SVM, respectively. However, in terms of the recall ratio of the algorithm, our method is lower than FESCSC. The reason is that the weighted processing operation causes to change a minimal number of sentiment feature parts of speech, which leads to a weakening of the sentiment degree of feature words. As a result, W-LDA with SVM makes some missed judgments, which will lead to a slightly lower recall rate over the FESCSC algorithm. In general, the sentiment classification method about comment set proposed in this paper has achieved good experimental results, and it has also been shown that the process has excellent flexibility and robustness.



(a) (b) (c)

Fig. 8. Experiment results comparison of various algorithms with the change of top number: (a) precision, (b) recall, and (c) F1 value.

5. Conclusions

This paper has proposed an improved average weighted topic model for the comment set with more emotional components. Some conclusion can be drawn as follows:

- The W-LDA model can highlight the topic expression vocabulary to avoid being overwhelmed by high frequency words, which improves the distinction of the topic extracted from the topic model.
- These highly differentiated topic types can be used as input parameters of the SVM algorithm, which can effectively improve the SVM algorithm's sentiment classification precision.
- The results have shown that the method proposed in the paper has certain advantages in terms of topic differentiation, classification precision, and classification F1 value through experimental testing of real teaching evaluation data and public comment test set.

However, there is still room for improvement in this method. For example, in constructing a topic model, the topic-and-topic and topic-and-word relationships are not considered yet, and the sentiment lexicon is still not perfect. Further researches need to be done in the above-proposed fields.

Acknowledgement

This work was partially supported by these Projects: Hezhou University 2018 School-level scientific research projects-Research on similarity join algorithm on large-scale uncertain graph data (No. 2018ZZK05), Hezhou University 2015 School-level Teaching Quality and Teaching Reform Project-Hezhou University Software Engineering Practice Teaching Center (No. hzxysfzx201506), and External Cooperation Project of Fujian Province in China (No. 2017I0016).

References

[1] L. Zhuang and D. Ye, "Text sentiment classification based on CSLSTM neural network," *Computer Systems & Applications*, vol. 27, no. 2, pp. 230-235, 2018.

[2] H. Zhou and Y. Wu, "Extracting and clustering features of evaluation object in Chinese user reviews," *Microcomputer & its Applications*, vol. 2014, no. 7, pp. 72-75, 2014.

[3] Z. Hai, K. Chang, and J. J. Kim, "Implicit feature identification via co-occurrence association rule mining," in *Computational Linguistics and Intelligent Text Processing*. Heidelberg: Springer, 2011, pp. 393-404.

[4] W. Wang, H. Xu, and W. Wan, "Implicit feature identification via hybrid association rule mining," *Expert Systems with Applications*, vol. 40, no. 9, pp. 3518-3531, 2013.

[5] T. C. Chinsha and S. Joseph, "A syntactic approach for aspect based opinion mining," in *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, Anaheim, CA, 2015, pp. 24-31.

[6] Q. Zhang and X. Liu, "Research on text emotion classification based on deep belief networks," *Journal of Northwestern Polytechnical University (Social Science Edition)*, vol. 36, no. 1, pp. 62-66, 2016.

[7] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, Lisbon, Portugal, 2015, pp. 1422-1432.

[8] X. Tang, J. Zhu, and F. Yang, "Research on sentiment classification of online reviews based on emotional ontology and kNN algorithm," *Information Studies: Theory & Application*, vol. 39, no. 6, pp. 110-114, 2016.

[9] H. Liu, Z. Zhao, B, Qin, T. Liu, "Comment target extraction and sentiment classification," *Journal of Chinese Information Processing*, vol. 2010, no. 1, pp. 84-88, 2010.

[10] P. Yin and H. Wang, "Sentiment classification for Chinese online reviews at product feature level through domain ontology method," *Journal of Systems & Management*, vol. 25, no. 1, pp. 103-114, 2016.

[11] Y. X. He, S. T. Sung, F. F. Niu, and F. Li, "A deep learning model enhanced with emotion semantics for microblog sentiment analysis," *Chinese Journal of Computers*, vol. 40, no. 4, pp. 773-790, 2017.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.

[13] D. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," *IEEE Signal Processing Magazine*, vol. 27, no. 6, pp. 55-65, 2010.

[14] J. Jin, Y. Liu, P. Ji, and H. Liu, "Understanding big consumer opinion data for market-driven product design," *International Journal of Production Research*, vol. 54, no. 10, pp. 3019-3041, 2016.

[15] H. Xia, J. Liu, and H. Zhu, "A comparative study on key technologies of the Chinese sentiment classification preprocessing," *Journal of Intelligence*, vol. 30, no. 9, pp. 160-163, 2011.

[16] K. Lan, D. Wang, S. Fong, L. Liu, K. Wong, and N. Dey, "A survey of data mining and deep learning in bioinformatics," *Journal of Medical Systems*, vol. 42, no. 8, article no. 139, 2018.

[17] W. B. A. Karaa and N. Dey, *Mining Multimedia Documents*. Boca Raton, FL: CRC Press, 2017.

[18] S. Xiao, S. Liu, F. Jiang, M. Song, and S. Cheng, “Nonlinear dynamic response of reciprocating compressor system with rub-impact fault caused by subsidence,” *Journal of Vibration and Control*, vol. 25, no. 11, pp. 1737-1751, 2019.

[19] S. Li, Q. Ye, Y. Li, and R. Law, “Mining features of products from Chinese customer online reviews,” *Journal of Management Sciences in China*, vol. 12, no. 2, pp. 142-152, 2009.

Xiaodong Tan **https://orcid.org/0000-0001-5775-1815**

He received master’s degree in School of Information Engineering from East China University of Technology in 2008. He currently is working for School of Mathematics and Computer Science, Hezhou University, Guangxi, China. His current research interests include Computer Algorithm and its Application.