Message board information extraction based on LDA and RNN

Lili Meng, Zhaoshun Wang*
Department of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China

ABSTRACT

The message board text put forward by netizens on a certain issue is a suggestion or opinion, which is sparse and emotional. The traditional LDA model cannot solve the sparsity problem of the short message and ignores the emotional factor. In order to solve the above problems, a message board information extraction method based on LDA model and RNN model is proposed. First, eigenvalues are introduced to classify the text to solve the sparsity problem based on LDA model. Second, RNN model is used to realize the emotional features on the basis of message text vectorization. The experiment shows that the LDA model with eigenvalues has better topic extraction ability compared with the traditional model, and with the fusion of the RNN model, it can comprehensively display the potential information of the message text. As a result, the proposed method achieves information extraction maximize.

Keywords: LDA, topic model, RNN, message board, information extraction

1. INTRODUCTION

Nowadays, with the popularization of the Internet, the number of Chinese netizens continues to increase. The Internet not only plays an increasingly important role in people's economic and cultural life, but also slowly infiltrates into their political life.

Latent Dirichlet Allocation (LDA) model is an unsupervised method in practical application. Its problems mainly include two aspects:

- The LDA model may miss some existing topics, and it is usually solved by improving the algorithm structure of LDA model.
- Keywords representing topics are not clear, and it is usually solved by improving corpus and providing more prior knowledge.

In this paper, the LDA model is optimized by improving the corpus with the eigenvalues of message text.

In order to fully explore the latent information of message text in this paper, the LDA model combined Recurrent Neural Network (RNN) model is built. First, the topic information in message text is extracted based on LDA model which can mine the semantic topic features of message text in the field of web semantic annotation[1]. To conduct prior classification of the message text, the eigenvalues are introduced that can make the topic distribution more centralized and remove meaningless topics on the result. Next, the emotional tendency of message text is analysed by RNN model. To establish the connections of emotional tendency of the words sequence features, Word2Vec model is used to construct the word vector relationship of the short text and it can effectively solve the problem of semantic relationship extraction, which makes it possible to conduct fine semantic modeling of message text from word granularity level[2]. Experimental results and analysis show that the proposed method is easy to understand not only the topic content that netizens pay attention but also the attitude on it, which can broaden the breadth and depth of the problem.

2. RELATED WORK

LDA model is one of the commonly used topic models at present. It was first proposed by Blei et al.[3] and then improved by Giffiths et al.[4]. Some studies have improved the structure of LDA model. For example, Gao et al.[5] proposed a Co-LDA model, which improved the quality of LDA model in extracting topics from short texts. Other researchers

*menglili2019@126.com

International Conference on Computer Application and Information Security (ICCAIS 2021), edited by Yingfa Lu, Changbo Cheng, Proc. of SPIE Vol. 12260, 122600T · © The Authors. Published under a Creative Commons Attribution CC-BY 3.0 License · doi: 10.1117/12.2637403

conducted further studies on the application of LDA model. For example, Wang et al.[6] proposed the concept of Topic Over Time, which is used to analyse the evolution process of topics over time.

On the other hand, it is mainly aimed at the research of text emotion analysis. Irsoy et al.[7] used cyclic neural network to obtain text sequence features and compared the effects of RNN with different layers on performance. Kamp et al.[8] studied the most important synonym graph theory model in WordNet lexical semantic web and proposed a semantic localization method to determine the three elements of subjective meaning of adjectives.

In general, one of the biggest commonality of these methods is that they rely too much on dictionaries and resources created by manual way. With the development of the Internet, a variety of new vocabulary continues to emerge. Therefore, updating existing vocabulary resources and applying them to text analysis models have become an urgent problem to be solved.

3. MESSAGE BOARD INFORMATION EXTRACTION METHOD BASED ON LDA AND RNN

Based on the problems in the above research and the characteristics of message text, a method to extract message board information based on LDA model and RNN model is proposed. The research process is divided into four stages:

- Text pre-processing.
- Topic identification and extraction.
- Topic analysis.
- Emotion analysis.

The specific process is shown in Figure 1.

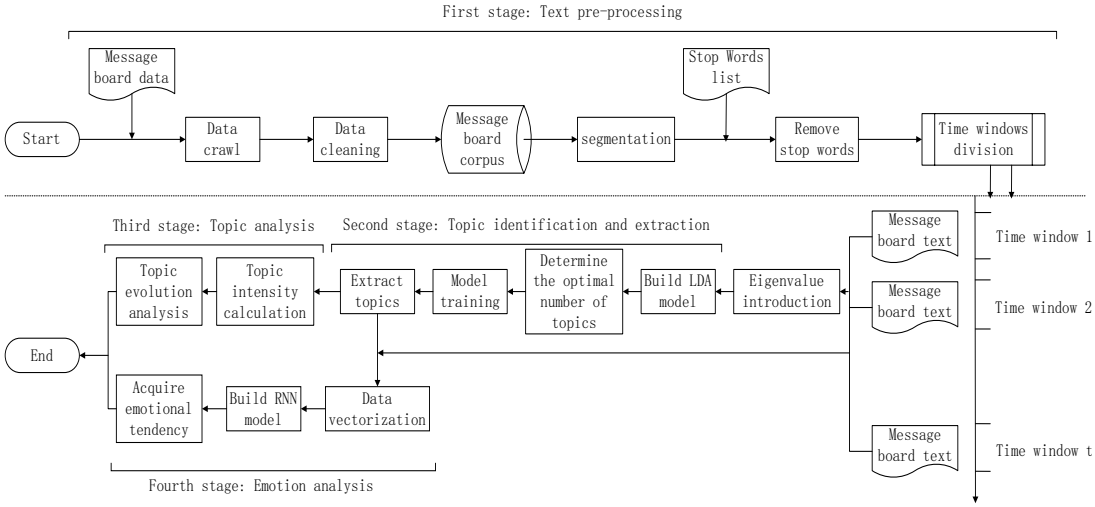

Figure 1. The flow chart of message board information extraction.

2.1 Text pre-processing

The used data comes from the module of "Leaders Message Board" built by People's Daily. The data quantity distribution is shown in Figure 2.

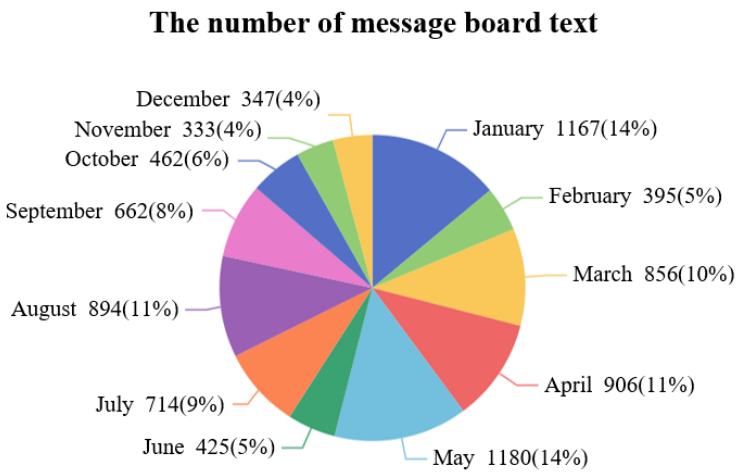

Figure 2. Distribution of the number of message board text.

According to the data size, the collected message text is divided to five-time windows: January and February are the first-time window; March and April are the second time window; May and June are the third time window; July and August are the fourth time window; September to December is the fifth time window.

With the known of time slice, the collected message text can be divided into several words by Jieba. At the same time, in the process of LDA model construction, the words that cause interference to the clustering results are removed.

To suppress the interference of similar topics in other time windows, eigenvalues are introduced to classify the texts. As a result, 46 eigenvalues are separated from the collected messages, which corresponds to 46 ministries.

2.2 Topic identification and extraction

After introducing eigenvalues into the corpus, topics need to be identified and extracted by LDA model. Gibbs sampling method is used to calculate parameter estimation of LDA model, and the perplexity is used to determine the number of topics in each time window.

2.2.1 LDA model. LDA model is an unsupervised machine learning technique whose essence is fuzzy clustering of words which represents a potential topic[9], so it can identify the underlying topic information in a corpus. Its characteristic is that the topic is a probability distribution of words, and the text is randomly mixed by the topic[10]. In LDA model, a text generation process is as follows:

- A sample from the Dirichlet distribution $\vec{\alpha}$ generates a topic distribution $\vec{\theta}_i$ for text *i*.
- A sample from the polynomial distribution $\vec{\theta}_i$ of the topic generates the topic $\vec{z}_{i,j}$ of the *j*th word in text *i*.
- A word distribution $\vec{\varphi}_{\vec{z}_{i,j}}$ for topic $\vec{z}_{i,j}$ is generated by sampling the Dirichlet distribution $\vec{\beta}$.
- The word $w_{i,j}$ is sampled from the polynomial distribution $\vec{\varphi}_{\vec{z}_{i,j}}$ of the word.

Among them, $\vec{\alpha}$ represents the weight distribution of each topic before sampling, and $\vec{\beta}$ represents the prior distribution of each topic. They are the prior parameters of LDA model. Table 1 shows the symbol description of LDA model.

Table 1. LDA model symbol description.

Symbol	Description	Symbol	Description
$\vec{\alpha}$	The Dirichlet distribution of text and word	w	Word
$\vec{\beta}$	The Dirichlet distribution of topic and word	K	The number of words
$\vec{\theta}$	The distribution of text and topic	M	The number of texts
$\vec{\varphi}$	The distribution of topic and word	N	The number of words per text
z	Topic		

Gibbs Sampling. The purpose of Gibbs Sampling is to extract samples, which closes to a probability distribution from Markov chains by solving the probability of sampling the current word, and the solution formula is:

$$p(z_i^d \mid z_{-i}^d, w, .) \propto \frac{C_{dj}^{DK} + \alpha}{\sum_{k=1}^{K} C_{dk}^{DK} + K\alpha} \frac{C_{ij}^{VK} + \beta}{\sum_{k=1}^{V} C_{kj}^{VK} + V\beta} \tag{1}$$

C_{ij}^{VK} is the *ij*th item in the C^{VK}, that is, the number of times the *i*th word is assigned to the *j*th topic. C_{dj}^{DK} is the *dj*th item in the C^{DK}, that is, the number of words assigned to the *j*th topic in the text *d*.

Gibbs sampling process in LDA model can be described as follows:

- Initialization: Initialize each topic z_i to a random number between 1-T (T is the total number of topics), generating the initial Markov chain.
- Iteration: *i* cycles from 1 to N (N is the total number of words) according to equation (1), assigning words to topics, and generating the next state of the Markov chain.
- Calculate the values of θ and φ: repeat previous step until Markov chain closes the target distribution, and the posterior estimates of θ and φ are calculated according to equation (2).

$$\hat{\theta}_j^d = \frac{C_{dj}^{DK} + \alpha}{\sum_{k=1}^{K} C_{dk}^{DK} + K\alpha}, \hat{\varphi}_i^j = \frac{C_{ij}^{VK} + \beta}{\sum_{k=1}^{V} C_{kj}^{VK} + V\beta} \tag{2}$$

Perplexity. In the application of LDA model, the number of topics needs to be set in advance. Perplexity is used to determine the optimal number of topics. Perplexity is the reciprocal of the geometric mean of the similar sentences contained in the text set, which gradually decreases with the increase of sentence similarity[11].

The calculation formula of perplexity is:

$$perplexity(D) = \exp(-\frac{\sum \log p(w)}{\sum_{d=1}^{M} N_d}) \tag{3}$$

The denominator is the sum of all words in the training text. p(w) refers to the probability of each word in the text of the training.

LDA Model Training. LDA model training refers to the process of generating parameters of LDA model by using training text sampling. The training process of LDA model can be inferred as follows:

- After pre-processing the training text, each word is assigned a topic number randomly.

- According to Gibbs sampling method, the current topic is excluded for each word. From all the remaining words and topics, the probability distribution of the current word to each topic is calculated according to equation (1), and a topic is assigned to the current word.
- Repeat previous step until the distribution converges between each text and topic, and between each topic and word.
- The average value of training results sampled by Gibbs for several iterations is taken as the final parameter estimation value.

The distribution of new text topics can be inferred by the trained LDA model.

From above analyses, it can be seen that the parameters of LDA model are estimated by Gibbs sampling, and the number of topics is determined by perplexity.

2.3 Topic analysis

To transform the topic information into valuable social information, the extracted topics need stability analysis.

The changing trend of topic heat over time is the most direct standard to measure the stability of a topic. The topic heat can be reflected by the number of comments on the topic. The time information of the text is used to retrieve the distribution probability of a topic in different time slices, which can measure the intensity of the topic. For topic z_i, its intensity value in each time slice is calculated as:

$$\delta_i^t = \frac{1}{D_t} \sum_{t_d \in t} \theta_{dt} \tag{4}$$

D_t represents the number of texts belonging to time window *t*.

By calculating the intensity of the topic in time window, it is easy to observe the intensity change of the topic in the whole time series.

2.4 Emotion analysis

In order to analyse the emotional tendency of the message text after extracting topics, it is necessary to mine the emotional characteristics. Word2Vec model is used to vectorize the message text that belongs to different topics in each time window. The vectorized data is the input for RNN model to obtain the value of emotional tendency.

Word2Vec. In order to better predict the contextual content of message text, it can vectorize the input of RNN model with Word2Vec model. Word2Vec is the by-product of RNN model training: model parameters (the weight of RNN neural network). These parameters are regarded as vectorized representation of the input values of RNN model. The Word2Vec model is usually divided into two models: continuous bag-of-words (CBOW) model and Skip-Gram model, as shown in Figure 3. During the training of CBOW model, the input is the word vector related to the context of a certain keyword, and the output is the word vector of this keyword. The Skip-Gram model is opposite to the CBOW model.

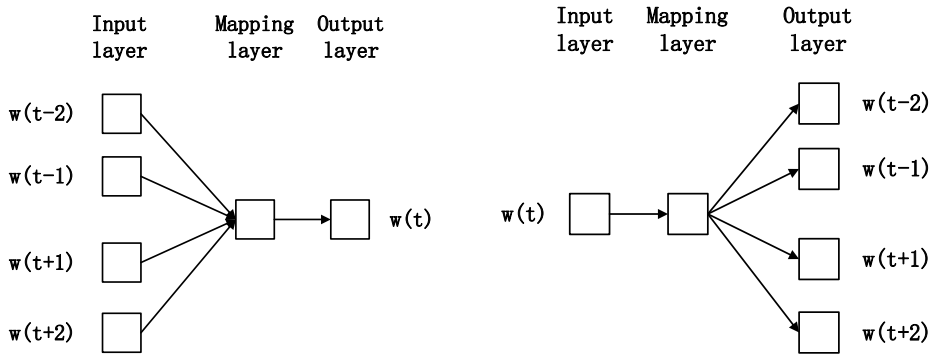

Figure 3. CBOW model and Skip-gram model.

RNN Model. In order to deal with the sentences which are related to each other in the message text, RNN model is used to process serialization information. The RNN cyclic neural network is expanded by time shown in Figure 4. Compared

with the traditional feedforward neural network, it introduces the ring structure into the network to establish the connection between neurons and themselves[12].

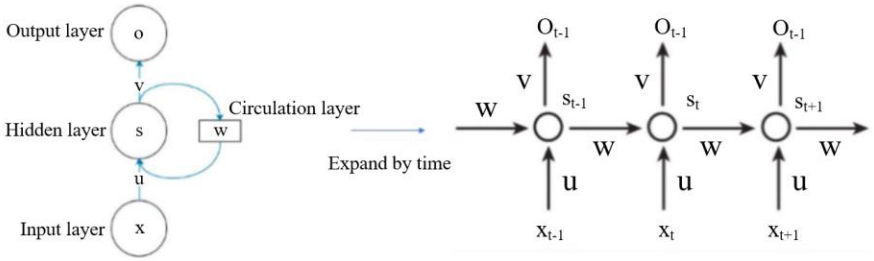

Figure 4. Cyclic neural network is expanded by time.

where, *t*-1, *t*, *t*+1 represent time series. *x* is the sample of input. S_t is the memory of the sample at time t. *W* represents the weight of the input, *U* represents the weight of the sample input at the moment, and *V* represents the sample weight of the output. At time *t*, the calculation formula of RNN is as follows:

$$\begin{aligned} h_t &= U * x_t + W * s_{t-1} \\ s_t &= f(h_t) \\ o_t &= g(V * s_t) \end{aligned} \tag{5}$$

where, $f(h_t)$ and $g(V*s_t)$ are sigmoid and softmax function, respectively.

RNN Model Training. RNN model training refers to the process of generating parameters of RNN model by using training text sampling. The training process of RNN model is as follows:

- One-hot encoder is used as the original input form of x, and a vector contains only one 1. All other 0 is used to uniquely represent the word.
- The input vector is mapped to a weighted vector VX through a hidden layer.
- After the output layer, the weight of the words from the hidden layer to the output layer is activated, and the vector VY is obtained, which is another vectorized representation of the input words.
- The result is taken as the input value of the next iteration of RNN model until the accuracy of the result tends to be stable. As a result, the parameter estimates of *W*, *U* and *V* are obtained.

The emotion value of the new text can be referred by the trained RNN model.

According to these processes, Word2Vec model is used to express message text in vectorization, and RNN model is used to realize emotion analysis on the basis of message text vectorization.

4. EXPERIMENTAL RESULTS AND ANALYSIS

4.1 Topic analysis

A total of 385 topics were extracted in the experiment. Table 2 shows part of the extraction results of message text topics with the eigenvalue of "National Health Commission" in the second time window. The three topics with the largest distribution probability are listed in Table 2, and the top five keywords for each topic are listed.

Table 2. The topics of "National Health Commission" under the second time window.

Topic	Topic content	Topic words
$T_{2,1}$	"Government regulation"	"Government" "Policy" "Information" "Standard" "Management"
$T_{2,2}$	"Medical development"	"TCM" "Doctor" "Cure" "Public health" "Student"
$T_{2,3}$	"Viral transmission"	"COVID-19" "Detection" "Pneumonia" "Nucleic acid" "Virus"

Figure 5 shows the result of evolutionary analysis of the topics in Table 2 over five-time windows. The intensity value of the topic is calculated by equation (4). The evolution of topic intensity can predict the future trend of the topic.

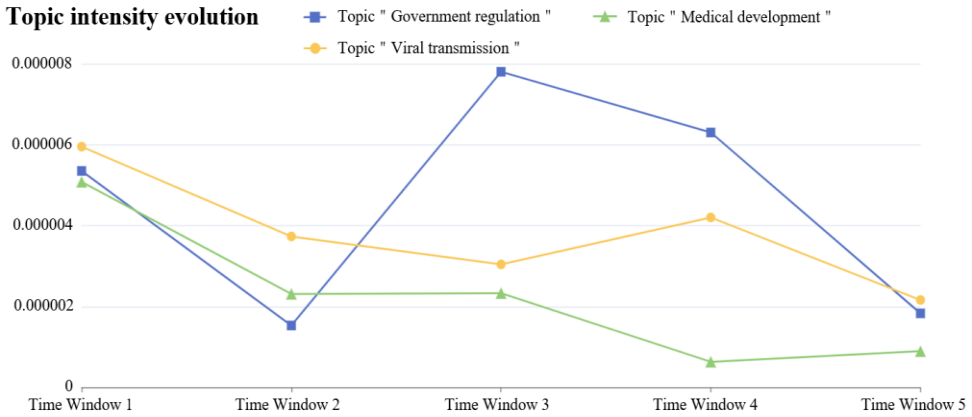

Figure 5. The evolutionary analysis result of part of topics.

After topic extraction in different time windows, message texts are grouped by the topics. Table 3 shows the top three results of topic classification by LDA model.

Table 3. Topic classification results.

Topic	Number	Ratio (%)
“Epidemic prevention”	985	14
“Transportation”	763	12
“Rural development”	692	10

4.2 Emotion analysis

The experimental data comes from the topic classification data output by LDA model. The specific data is shown in Table 3 in Section 4.1. RNN model is used to analyse the emotional tendency of message texts under these different topics.

The result of emotion analysis in the experiment is the value of -1 to 1, and the final result is the average value of emotional tendency of messages with different eigenvalues under each topic. With the value is closer to -1, the emotion is the more negative. While with the value is closer to 1, the emotion is more positive. Taking the topic "Rural Development" as an example, the result of emotion analysis is shown in Figure 6.

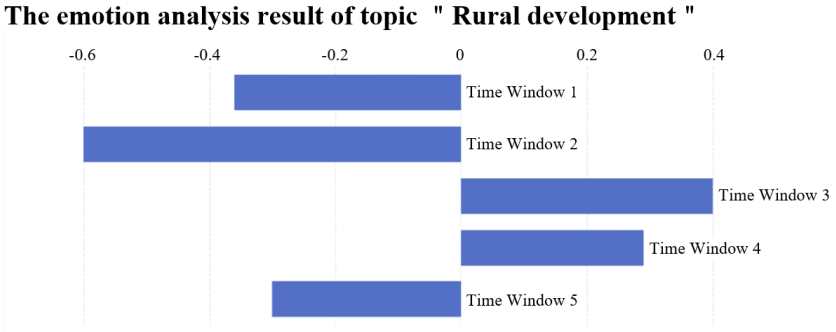

Figure 6. The emotion analysis result of topic “Rural development”.

4.3 Comparative experimental analysis

Compared with traditional method, the topic extraction method based on LDA model used in this paper introduces the eigenvalues. Therefore, the difference between the two methods is that the topic intensity of the proposed method is

under a certain eigenvalue, while the topic intensity of the traditional method is under the whole-time window. So the topic intensity of the proposed method is lower and the variation is small. Figure 7 shows the topic extraction comparison between the proposed method and the traditional method.

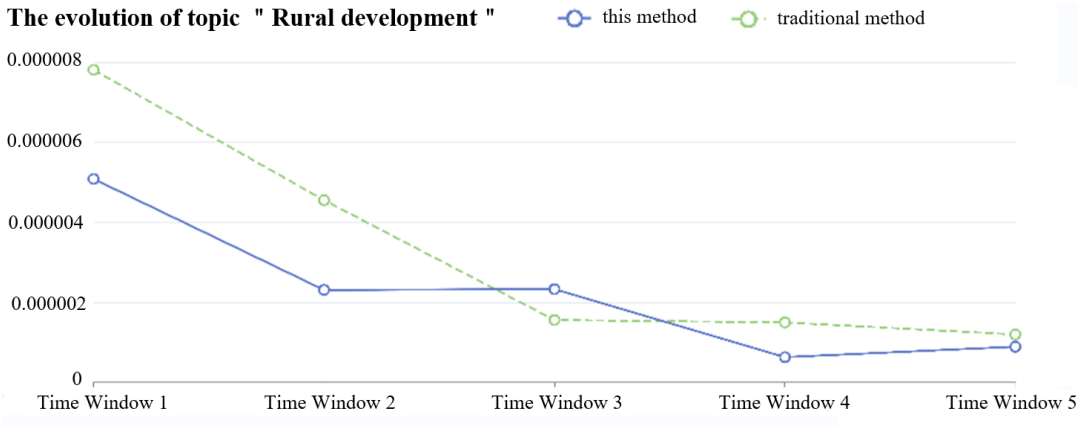

Figure 7. The comparison of intensity change of topic "Rural development".

Figure 8 shows the information extraction of the topic "Rural Development" after the fusion of RNN model. The result enriches the content of information extraction. As can be seen from Figure 8, the emotional attitude is weak when netizens pay less attention to a certain topic, and it is accord with the theoretical analysis.

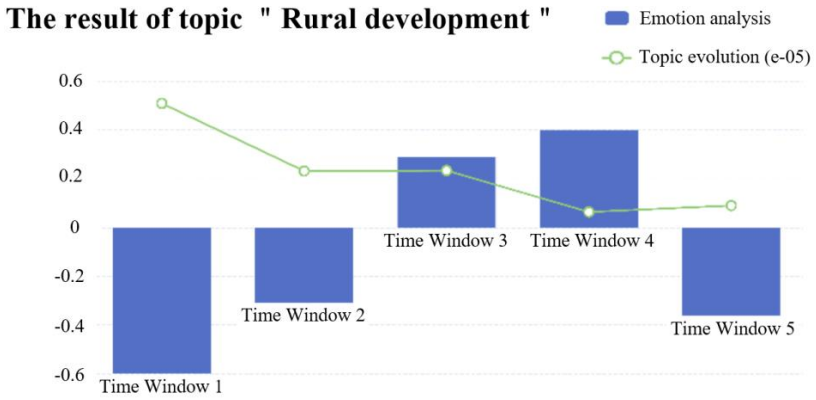

Figure 8. Message board information extraction result with the fusion of RNN model.

5. CONCLUSION

In view of topic extraction and emotion analysis of message short text, the following conclusions are drawn:

- In the research of message board text, it mainly establishes a text topic extraction and evolution model based on LDA model. In this model, the eigenvalues are introduced as the classification basis.
- According to the emotional characteristics of message text, RNN model is integrated to analyse the emotion tendency of extracted message text under different topics.
- Combined with the experimental results of this paper, the further research direction is the prediction of topic and emotion.
- The government and relevant departments should give priority to solve the problems that netizens attracted high attention, especially those with negative emotional tendency, and it can help the government have better service for people.

REFERENCES

[1] Tian, Y. and Gong, T. T., "Research on topic mining of online teaching demand data based on LDA model," Information Science, 39(9), 111 (2021). (in Chinese)

[2] Xi, X. W., Guo, Y., Song, X. N. and Wang, J., "Research on the technical similarity visualization based on word2vec and LDA topic model," Journal of the China Society for Scientific and Technical Information, 40(9), 976 (2021). (in Chinese)

[3] Blei, D. M., Ng, A. Y. and Jordan, M. I., "Latent Dirichlet allocation," Journal of Machine Learning Research, 3(7), 993-1022 (2003).

[4] Griffiths, T. L., Stevvers, M. and Tenenbaum, J. B., "Topics in semantic representation," Psychological Review, 114(2), 211-244 (2007).

[5] Gao, H. Y., Liu, J. W. and Yang, S. X., "Online medical comment topic excavation based on improved LDA," Journal of Beijing Institute of Technology, 39(4), 427-434 (2019). (in Chinese)

[6] Wang, X. and McCallum, A., "Topic over time: A non-markov continuous-time model of topical trends," Proc. of the 12th ACM SIGKDD Inter. Conf. on Knowledge Discovery and Data Mining, 426-433 (2006).

[7] Irsoy, O. and Cardie, C., "Opinion mining with deep recurrent neural networks," Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing, 720-728 (2014).

[8] Kamps, J., Max, M., Mokken, R. J., et al., "Using wordnet to measure semantic orientation of adjectives," Proc. of the 4th Inter. Conf. on Language Resources and Evaluation (LREC 2004), 1115-1118 (2004).

[9] Chen, Y. and Sheng, J. G., "Weibo tag generation algorithm based on LDA and Word2vec," Computers and Modernization, 12(6), 38 (2021). (in Chinese)

[10] Yang, X., Li, B. L. and Jin, M. J., "Hotspots and trend analysis based on LDA model," Computer Technology and Development, 22(10), 66-69 (2012).

[11] Li, X. D., Zhang, J. and Yuan, M., "On topic evolution of a scientific journal based on LDA model," Journal of Intelligence, 33(7), 115-121 (2014). (in Chinese)

[12] Lin, D. P. and Wang, H. J., "Comparison of news text classification based on BERT and RNN," Journal of Beijing Institute of Graphic Communication, 29(11), 157 (2021). (in Chinese)