Sentiment Analysis of online reviews based on LDA and AP-Bert model

Menglin Yang, Yiqing Lu*

School of Information Management, Beijing Information Science &Technology University, Beijing, China

*Corresponding author e-mail: luyiqing@126.com

Abstract. The purpose of this paper is to construct a more accurate behavior matrix by using fine-grained aspect level emotion analysis method. Firstly, LDA topic extraction model is used to extract the topic of online review text, and the concerned attributes are extracted. According to the characteristics of online comments, a BERT emotion analysis model with enhanced pooling was proposed. Activation function layer and max-average pooling layer were designed to solve the over-fitting problem of BERT model in the process of emotion analysis. Finally, by combining LDA extraction results and AP-Bert sentiment analysis results, the proportion matrix is obtained. Experimental results show that the accuracy, recall rate and AUC value of AP-Bert model are better than those of the same type of model and original BERT model.

Keywords: Sentiment analysis, BERT, Online reviews, LDA.

1. Introduction

As one of the most popular tasks in the field of natural language processing, sentiment analysis was first proposed by Nasukawa in 2003. Text sentiment analysis is widely used in public opinion analysis, hot topic prediction, comment text classification and other fields. At present, there are three research methods commonly used in text sentiment analysis, which are based on emotion dictionary, machine learning and deep learning. The method based on the emotion dictionary is that the constructor constructs the emotion dictionary according to the prior knowledge. After the word segmentation of online comments, the score is obtained according to the comparison with the emotion dictionary to achieve the emotion analysis. Due to the great influence of the builder, the quality of the sentiment dictionary is difficult to be guaranteed, so it is not suitable for the need of mass online comment sentiment analysis. Method based on machine learning is first drawn from the comment text characteristic value, and then passed to the relevant machine learning model in the analysis of emotion, the performance of this method for the number of features and typical dependence is bigger, in the current online comments by higher orders of magnitude increases, this method on the characteristics of the engineering cost of manpower cost, time cost is increased, This method is not suitable for sentiment analysis of online comments. Deep learning solves this problem. Deep learning models realize automatic extraction of text features of comments, and are increasingly used in text sentiment classification. Online review texts are characterized by complex sentence patterns, rich semantics and numerous attributes. However, the deep learning model is not suitable for review texts with such features as multi-dependent word vector encoding or one-HOT encoding in the text embedding layer. Based on this, Google proposed the BERT (Bidirectional Encoder Representation from Transformers) pre-training model, which is different from the structures of convolutional neural networks and recurrent neural networks in previous deep learning. BERT is a deep bi-directional language model based on Transformer architecture, which can not only accurately identify the relationship between clauses in comments, but also fulfill the need of high-performance sentiment analysis for online comments.

He Yanxiang et al. (2017) added weibo emoji as an important feature of weibo comments and proposed the emotional-Semantics enhanced MCNN(EMCNN) Emotion classification model of weibo. By constructing feature representation matrix and combining it with deep learning model MCNN, A good effect is achieved in the emotion classification of microblog emotion comment data

set. Chen Pingping et al. (2020) extracted features from tF-IDF and Chi-square statistics from the perspective of word segmentation and double-word segmentation, and combined with a series of machine learning models, found that BP neural network and polynomial Bayesian algorithm were more suitable for the emotional orientation analysis of online movie reviews. Zhou Junqiang (2019) used BERT model for word vector representation. In order to solve the problem that words not included in the corpus have no corresponding emotional orientation information, text emotional labels were directly used as the emotion labeling of words, and word vectors were used as the input of classification algorithm to effectively improve the accuracy of the model. Shi Zhenjie and Dong Zhaowei (2020) combined the BERT model with CNN to mark the text of a mobile phone comment on JINGdong. The BERT model was used to express the semantic structure of sentences, and then the CNN model was used to extract the local features of sentences. The bert-CNN model achieved good performance.

However, the expected scale of Wikipedia used by BERT is very different from the specific domain, and the downstream task of sentiment analysis of comments in the domain will cause serious over-fitting problems. Previous text sentiment analysis tasks are mainly aimed at coarse-grained sentiment analysis. Compared with coarse-grained text sentiment analysis, there are still few researches on fine-grained text sentiment analysis. For a review text library in a particular field, we often hope not only to know the emotional orientation of each comment, but also to dig out the aspect emotional orientation of different comment objects.

Based on the above reasons, this paper crawls jingdong Mall online comments, performs word segmentation and preprocessing, and then uses LDA topic extraction model to conduct topic modeling for comment prediction, extracts the main topics and attribute words in the comment prediction, defines the theme and initializes the attribute keywords, and uses Word2vec for word quantization processing. After the subject words are obtained, the aspect level sentiment tendency analysis is carried out based on BERT's improved model to solve the over-fitting problem of original BERT. Finally, the theme emotion proportion matrix is constructed.

2. The related theory

2.1 LDA（Latent Dirichlet Allocation）

LDA is a topic model proposed by Bel et al. The topic of each document in a document set can be given in the form of probability distribution, so that the topic distribution can be extracted by analyzing some documents, and topic clustering or text classification can be carried out according to the topic distribution. LDA is to construct the theme that words fail to express clearly, and it is also a means of mining the real expression intention of text content. We can summarize the results of LDA into a general theme, which is usually composed of several words. In the LDA model, it is assumed that the prior distribution of document and topic, topic and word is Dirichlet distribution, and the expression of Dirichlet distribution is

$$Dir(\vec{p}|\vec{\alpha}) \triangleq \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)}{\prod_{k=1}^{K}\Gamma(\alpha_k)}\prod_{k=1}^{K}p_k^{\alpha_k-1} \triangleq \frac{1}{\Delta(\vec{\alpha})}\prod_{k=1}^{K}p_k^{\alpha_k-1} \quad (1)$$

In the formula, Γ is the Gamma function. The LDA model is shown in Figure 1, k represents the number of extracted topics, k∈ [1, k]; N is the number of words in the comment, N ∈ [1, N]; M represents the number of documents, m∈ [1, M]; θm represents the topic distribution for document m; tm,n denotes the n topic of document n; φk represents the word probability distribution corresponding to each topic. Wm,n is the n word of the m document.

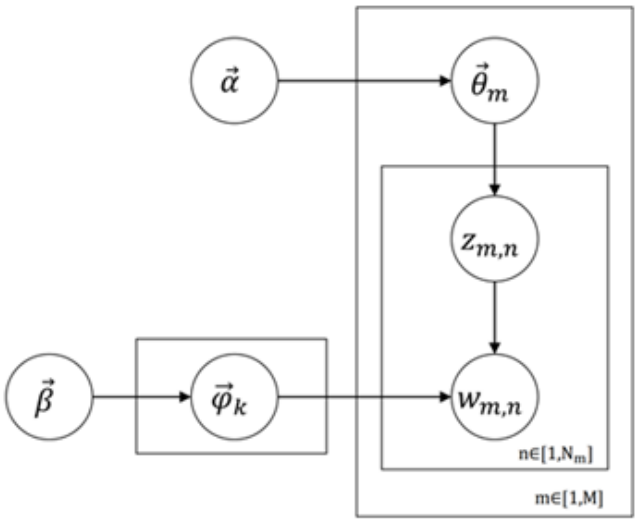

Figure 1. The LDA model diagram

By solving the parameters θm and φk, the topic distribution of each comment text and the distribution of words under each topic can be obtained, thus the online topic extraction of LDA can be realized.

2.2 BERT

Bert (Bidirectional Encoderrepresentation from Transformer) The model is a pre-training model in the field of Natural Language Processing (NLP) proposed by Google in 2018. As shown in Figure 2, BERT uses the encoder part of bidirectional Transformer to build the model, which is used to extract the features of input text information, remove the restrictions on the distance between each word and other words, and express the dependence on their semantic relations. At the same time, the structure of the model is different from that of the traditional text processing methods, and the circular structure such as RNN and LSTM is abandoned, which effectively solves the problem of parallel processing and text dependence.

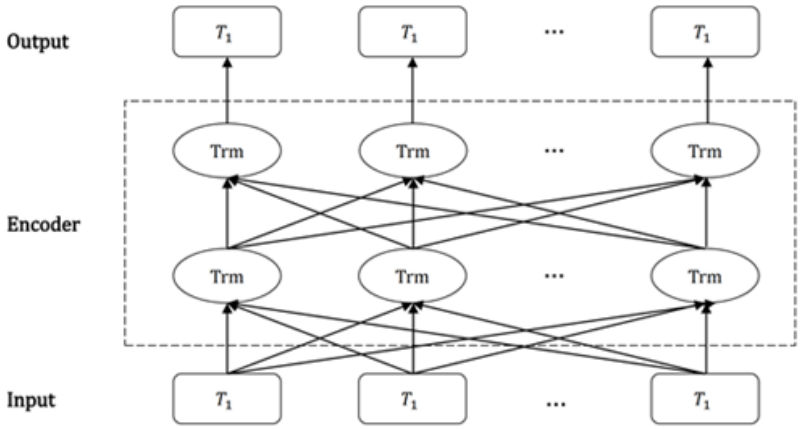

Figure 2. The BERT model diagram

The model is mainly innovated in pre-train, namely, Masked LM and Next sentence prediction, which are used to capture expression and sentence-level representation. Firstly, the Masked method is used for model training, that is, 15% words are randomly selected to process the input text. For this part of words, 10% words are replaced with other words, 10% remain unchanged, and the remaining 80% are replaced with [mask] symbol. The final loss function only calculates the token that is Masked out. Because the process is random, the model pays attention to each word.

3. Online Comment Sentiment Analysis Model Based on LDA and AP-Bert

3.1 Data Preprocessing

Data preprocessing is the primary task of sentiment analysis, and the validity of data directly affects the accuracy of sentiment analysis. Data cleaning is a checksum review of crawl data. In crawl comments, there are a large number of repeated statements and irrelevant texts, as well as a large number of acronyms and emoticons. Therefore, data cleaning includes deduplication, missing value processing, unified format, etc., to ensure the standardization and consistency of data.

3.2 Identify user concern topics

The preprocessed comment text is input into LDA model for topic extraction, and the topic distribution of LDA output is used to determine the attributes of consumer concern. Input the corpus comment text into LDA and assign attributes to the topic; Then, the theme is traversed, and the attribute distribution of each word is re-sampled and updated according to Gibbs formula until convergence. Finally, the LDA output topics are categorized and attribute names are given to each class.

3.3 Emotion analysis model based on pooled enhanced BERT model

The "MASK" word vector training model of BERT, Transformer is used as the semantic feature extractor, so that the longer context information is taken into account and the learned token word vector is used by downstream tasks, instead of just intercepting the token of window length and using shallow fully connected neural network for training. It can better adapt to the needs of online sentiment analysis of massive comments. However, the scale of the online comment corpus is too far from that of the Wikipedia corpus used by BERT, and serious over-fitting will occur in the downstream task of emotion analysis. Therefore, an emotion analysis model for BERT (AP-BERT) online reviews with enhanced pooling is proposed.

The model is shown in Figure 3, where: E is the position feature layer of input representation layer, sentence level feature layer of word and corresponding embedded representation of word vector layer; [CLS] is the start token in BERT model; Tok input tokens for words in BERT model; Trm is a Transformer encoder, which contains multiple semantic extraction layers in BERT model. C and T are the [CLS] and word token output by the semantic extraction layer respectively.

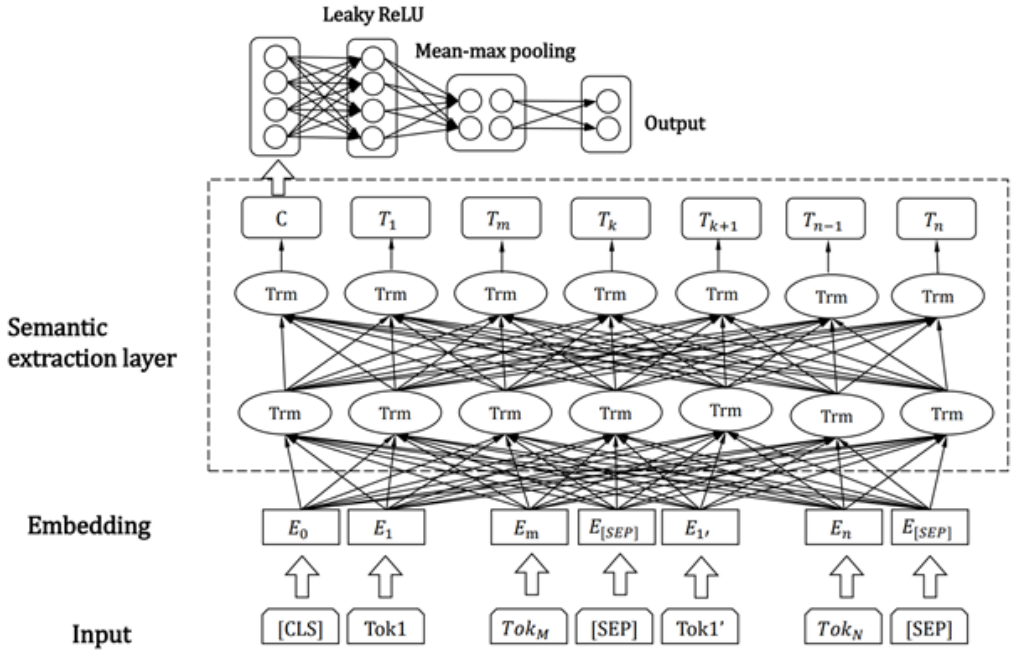

Figure 3. The AP-BERT module frame diagram

3.3.1 Input layer.

In order to make it more suitable for large number of online comment text with complex features, WorldPiece embedding model is used to encode according to double-byte encoding, which reduces the complexity of encoding to a large extent. As shown in Figure 4, the Masked LM model is used to construct short comment text. According to the principle, 15% words are randomly selected for processing. For these words, 10% are replaced with other words, 10% remain unchanged, and the remaining 80% are replaced with [mask] symbols. For comments on long sentences, add the token [SEP] at the position where the semantic logic distinguishes the next sentence. Context-relevant and context-independent comments were given as inputs in a 1:1 ratio so that the model understood the relevance between sentences.

3.3.2 Semantic extraction layer.

The semantic extraction layer of BERT model uses Transformer encoder with powerful feature extraction ability, which has the ability of extracting long distance dependency relationship with RNN and the ability of CNN parallel computation. These two abilities mainly benefit from the self-attention structure of the Transformer-encoder, which uses the words in its context when calculating the current words so that it can extract the long-distance dependence relationship between words. Since the calculation of each word is independent and not dependent on each other, the features of all words can be calculated simultaneously and in parallel.

Encoder consists of multiple identical layers (sub-coding layers), each structurally identical, but they do not share parameters. Each layer consists of two sub-layers (self-attention and feedforward network), which are multi-head self-attention mechanism and fully connected feed-forward network respectively. Each sub-layer has a residual connection around it and is followed by a "layer-normalization", so the output of the sub-layer can be expressed as

$$sub_{layer_{output}} = LayerNorm\left(x + \left(SubLayer(x)\right)\right) \tag{2}$$

The multi-head self-attention layer helps the encoder to pay attention to other words in the input sentence while encoding each word. According to the principle of attention, attention can be expressed in the following form

$$attention\_output = Attention(Q, K, V)$$

Multi-head attention is to project Q, K and V through H different linear transformations (H: the number of "attention heads"), and finally combine different attention results (self-attention means that Q, K and V are the same, but W of linear transformation is different).

$$head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right) \tag{3}$$

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4}$$

$$MuiltiHead(Q, K, V) = Concat(head_1, head_2, \dots, head_h)W^O \tag{5}$$

3.3.3 Pooling layer.

In order to solve the problem of over-fitting in the process of sentiment analysis of online comments, two steps have been added to [CLS] of BERT model through semantic extraction layer. Step 1, the Leaky ReLU function layer has been added for nonlinear mapping to improve the convergence speed of the model. In the second step, the scale of data processed based on the original BERT model is very different from that currently solved. Pooling layer is added again to reduce the model scale, and output features are integrated to further solve the over-fitting problem of the original model.

In the pooling layer, the maximum-average pooling method is adopted, as shown in Figure 4. Firstly, the maximum and mean values are calculated along the length and dimension of the comment text, and the two are combined into a vector to convert the hidden sequence into a vector. The process formula is as follows

$$\max\_pooled = max(X_{hidden}, d_{imension} = seq_{len}) \quad (6)$$

$$\text{mean}\_pooled = mean(X_{hidden}, d_{imension} = seq\_len) \quad (7)$$

$$\max\_\text{mean}\_pooled = concatenate(\max\_pooled, \text{mean}\_pooled, d_{imension} = embedding\_dim) \quad (8)$$

$$\max\_pooled, mean\_pooled, \max\_mean\_pooled \in R^{batch\_size \times embedding\_dim \times 2}$$

In the formula, Max_pooled is maximum pooled, mean_pooled is average pooled, and max_mean_pooled is max-mean pooled.

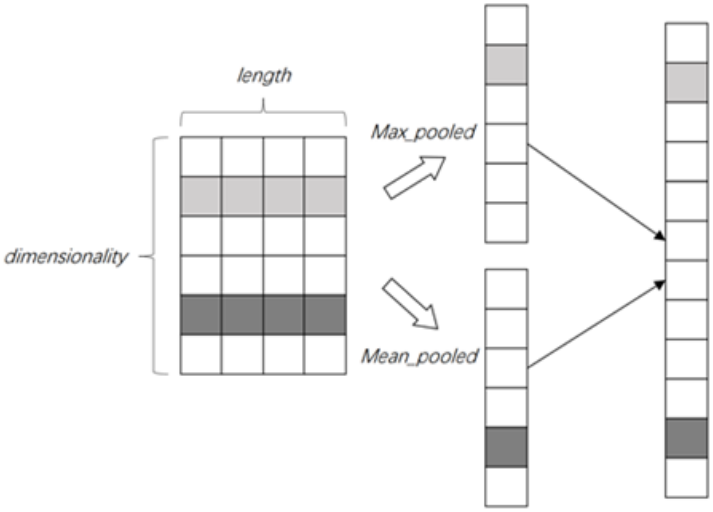

Figure 4. The max_mean_pooled layer.

3.3.4 Output layer of emotion classification.

The max_mean_pooled output from the pooling layer is input to the emotion classification layer, and the Sigmoid function is used to calculate, so as to carry out the emotion analysis of the online comment text. The formula is as follows

$$\hat{y} = sigmoid\big(Linear(\max\_mean\_pooled)\big), \ \hat{y} \in (0,1) \quad (9)$$

AP-BERT model to encode the input representation layer three parts, the construction of characteristic vector, after the input layer semantic extraction of feature vector to study, use the transformer encoder for semantic information learning and capturing, extracting information input to pooling layer information further options, in the end, the input output layer emotion classification to classify emotions.

4. Experimental results and analysis

4.1 Data Preprocessing

In order to verify the effectiveness of the proposed method, this paper selected real reviews from JD Store for verification, and used Bazhuayu tool to climb online reviews of five categories of commodities. After data cleaning, there were 57,162 valid reviews. It had 42,872 positive comments and 14,290 negative ones. The text emotional polarity label is represented by 1 and 0, respectively. 1 represents positive emotional comment and 0 represents negative emotional comment.

4.2 Analysis of experiment

4.2.1 Experimental environment.

The experimental environment configuration is as follows: The operating system is Windows 10, the CPU is Intel(R) Core (TM) i5-10210u CPU @ 1.60GHz 2.11 GHz, the memory is 16GB, the program language is python3.7, and the deep learning model framework is PyTorch1.9.0.

4.2.2 Topics extraction.

LDA is used for topic extraction of online comments. Firstly, several feature words are extracted from the corpus, and corresponding sets are combined to generate the word frequency vector of any two or two topics between these sets, and the average cosine similarity of each topic number is calculated, as shown in Figure 5, from which the optimal topic number is selected.

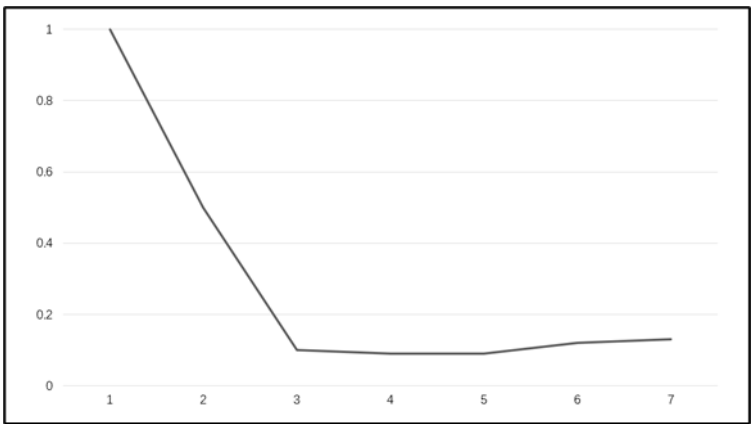

Figure 5. Average cosine similarity graph between topics

Figure 6 shows that for the comment data, when the number of topics is 4 or 5, the average cosine similarity reaches the minimum value, so the number of topics is 5. The gensim module of Python was used to build the LDA topic extraction model, extract the attribute words, and cluster the attribute words into 5 topics named A1-A5.

4.3 Data Preprocessing

The comment corpus was input into the AP-Bert model for emotion analysis, and the emotion distribution of each comment text was calculated to obtain the emotion score. For example, the topic emotion score of Test1 was 0.8080; Test2's theme sentiment score was 0.0323; Test3 has a thematic sentiment score of 0.5652. It can be seen that each emotion output score is in the range of 0~1. The closer it is to 1, the more positive the emotion is, and the closer it is to 0. 0-1 is divided into five ranges, 0-0.2 is very dissatisfied, 0.2-0.4 is dissatisfied, 0.4-0.6 is neutral, 0.6-0.8 is satisfied, and 0.8-1.0 is very satisfied. Based on this division, on the basis of LDA topic extraction results and AP-Bert model sentiment analysis results, combined with PLTSs, the overall satisfaction ratio of each topic is obtained, as shown in Table1.

Table 1. The satisfaction-topics proportion matrix.

satisfaction	The overall satisfaction proportion for attributes				
	A_1	A_2	A_3	A_4	A_5
Highly Satisfactory	0.7163	0.7130	0.4194	0.2032	0.3030
Satisfactory	0.0091	0.0104	0.0047	0.0111	0.0127
ordinary	0.0040	0.0089	0.0086	0.0083	0.3094
Dissatisfied	0.0046	0.0084	0.00989	0.3163	0.0129
Very Dissatisfied	0.2660	0.2593	0.5574	0.4610	0.3621

4.4 Data Preprocessing

In order to further confirm the effectiveness of the AP-Bert model in this paper, we selected common models for sentiment analysis, including support vector Machine (SVM), convolutional Neural network (CNN), LSTM and original BERT model as the benchmark algorithm. Precision, Recall and F1 were used as evaluation indexes to comprehensively compare the performance of each model in sentiment analysis. The evaluation indexes were evaluated by the formula.

$$Accuracy = \frac{TP + TN}{(TP + FP) + (TN + FN)} \tag{10}$$

$$Recall = \frac{TP}{TP + FN} \tag{11}$$

$$AUC = \frac{\int_0^1 TP\, dFP}{(TP + FN)(TN + FP)} \tag{12}$$

In order to ensure the accuracy and objectivity of the experimental results, the five models were respectively run for 10 times on the same training and test data set, and the average accuracy, recall rate and AUC values were obtained as the final results of the model, as shown in the table2.

Table 2. The summary table of algorithm results.

Algorithmic model	Evaluation index		
	Accuracy	Recall	AUC
SVM	0.6140	0.8362	0.5811
CNN	0.6371	0.8213	0.6109
LSTM	0.7209	0.8763	0.6128
BERT	0.7358	0.8802	0.6729
AP-BERT	0.7643	0.9189	0.7180

Through experiments, the results are shown in Table 4. It can be seen that compared with other commonly used models, AP-BERT has obviously achieved better results on Acuuracy, Recall and AUC, which indicates that AP-BERT has more advantages in emotion analysis of online comment text.

5. Conclusion

As sentiment analysis models become more and more mature, it is a general trend to propose a better model in the field. Based on this, this paper firstly uses LDA model to extract the topic and mine the attributes of online review text. Secondly, the original BERT model was improved to adapt to this study and accurately measure the emotional propensity score of comments. Finally, LDA theme extraction results and AP-Bert model emotion score are combined with PLTSs to obtain the satisfaction ratio of each subject score. In conclusion, the model and method proposed in this paper can transform the unstructured quantity of online reviews into quantitative information, and achieve better results in terms of accuracy.

References

[1] KIM J, HAN M, LEE Y, et al. Futuristic data-driven scenario building: Incorporating text mining and fuzzy association rule mining into fuzzy cognitive map [J]. Expert systems with applications, 2016, 57: 311-323.

[2] LIANG D C, DAI Z Y, WANG M W, et al. Web celebrity shop assessment and improvement based on online review with probabilistic linguistic term sets by using sentiment analysis and fuzzy cognitive map [J]. Fuzzy optimization and decision making, 2020, 19(4): 561-586.

[3] Taboada M, Brooke J, Tofiloski M, et al. Lexicon-based methods for sentiment analysis [J]. Computational linguistics, 2011, 37(2): 267-307.

[4] Xu GX, Yu ZH, Yao HS, et al. Chinese text sentiment analysis based on extended sentiment dictionary [J]. IEEE Access, 2019, 7: 43749-43762.

[5] Yi Cai, Kai Yang, Dongping Huang, Zikai Zhou, Xue Lei, Haoran Xie, Tak-Lam Wong. A hybrid model for opinion mining based on domain sentiment dictionary[J]. International Journal of Machine Learning and Cybernetics, 2019, 10(8).

[6] Mikolov T, Chen Kai, Corrado G, et al. Efficient est imation of word representations in Vector space. [J]. International Journal of Machine Learning and Cybernetics,2020,7(6).

[7] Pennington J, Socherm R, Manning C. Glove: Global Vectors for Word Representation [C]. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014.

[8] Joulin A, Grave E, Bojanowski P, et al. Bag of Tricks for Efficient Text Classification[J].2016: 427-431.Zhang Yangsen, Jiang Yuru, tong Yixuan. Study of Sentiment Classification for Chinese Microblog Based on Recurrent Neural Network. Chinese Institute of Electronics, 2016, 25(4): 601-607.

[9] Jacob Devlin, Ming-wei Chang, Kenton Lee, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1801.04805v1 [cs. CL], 2018.

[10] Devlin J, Chang M W, Lee K, et al. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding [J]. arXiv preprint arXiv: 1810.04805, 2018.