Topic Extraction of Online Curriculum Reviews

Cheng Wang¹, Ya Zhou¹*

¹Guangxi Key Lab of Trusted Software, Guilin University of Electronic Technology, Guangxi, Guilin, 541004, China
*Corresponding author’s e-mail: yzhou@guet.edu.cn

Abstract. With the advent of the era of intelligent education, online learning has become a new way of learning. Online learning will produce a lot of review data, this paper focuses on how to correctly reflect learners' emotional attitude towards the curriculum, this paper studies the mining of text topic emotion in MOOC course review data set. Firstly, the balanced algorithm is used for sample equalization, secondly, we use Albert + biLSTM model to classify course reviews, finally, the LDA model is used to extract the corresponding topics. Compared with the experimental model, the proposed model has higher accuracy and stability. Machine learning method has some advantages, but the model is more complex, and it can be simplified in the future.

1. Introduction

With the influence of Internet lifestyle and the birth of information-based learning style, we have also opened up a new way of Education - internet teaching. MOOC is an interactive Internet IT skills free learning website, because of the convenience and openness of its learning, many teachers and learners are attracted to carry out corresponding teaching activities on this platform, and learners can comment on the course they study, it has become one of the important ways of multiple evaluation in pedagogy. As a main objective way to show the quality of a course, curriculum review is real, timely and extensive, it is the main channel for people to understand the quality of curriculum. However, in the face of massive course reviews. How can people quickly know the quality of a course, whether it is what you need to learn.

At present, scholars at home and abroad have done some research on topic sentence extraction. There are many ways to mine the thematic emotion of text, but the mainstream method is topic mining based on topic probability model. Wang et al.¹ extract topics based on LDA model, Gibbs sampling method is used to reflect the subject information from multiple aspects, the smoothness of topic probability distribution is used to calculate the credibility, get the final topic sentence, after a series of experiments, good results have been obtained. Sun et al.² established utsu model on the basis of sampling emotion tags for sentences and topic tags for words, its sampling method is more in line with the characteristics of emotional expression in language, the performance of emotion classification is better than JST and asum. In 2018, Ma et al.³ used LDA model to extract literature topics and build a "content method" two-dimensional network model, to explore the relationship between research content and research methods, discover the hidden knowledge in the subject field. Although there are many researches on topic sentence extraction in text, however, there are few researches on topic sentence extraction of MOOC course reviews. As a special text, MOOCS curriculum review has its own characteristics, there are important differences between its research and other texts. At present, the research on the extraction of MOOCS reviews mainly focuses on the extraction of keywords and phrases in reviews.
The emergence of MOOC platform is relatively short, therefore, the research results of MOOC curriculum review are relatively few[4]. Research on topic extraction analysis of MOOC course review data, it’s only in recent years.

This paper is based on the reference of previous people, this paper analyzes the characteristics of the text itself, the rules are formulated according to the distribution of theme sentences in comments and the reference function of the theme, finally, the feature weighting method is combined with LDA topic extraction model. It mainly focuses on how to mine explicit theme emotion in explicit opinions, but in some sentences, although there is no clear subject word and opinion word as a guide, but we can express subjective emotions through objective statements, these are called implicit themes and emotions. It can effectively improve the accuracy of emotion analysis by extracting the double implied theme emotion.

2. Model overview
As a medium of curriculum quality reference, curriculum review is a kind of media. A special kind of text, it has its own characteristics. General course reviews express learners' emotional tendencies[5], positive emotion or negative emotion. In order to improve the accuracy and accuracy of topic extraction, we use Albert + bilstm to classify course reviews, it is divided into positive emotion and negative emotion. The LDA model is used for topic extraction.

Albert is a lightweight Bert. At present, it is favored by industry and academia, this is mainly because Albert uses factorization, there are three methods: cross layer parameter sharing and inter sentence coherence, to a certain extent, it solves the shortcomings of Bert, a large number of parameters and huge resources. Albert uses transformer compiler with self-attention mechanism in the whole pre training process, can capture the two-way relationship in the statement more thoroughly, it can realize the two-way learning of language representation in all layers, it is a multitasking model, the task consists of two self-monitoring tasks, they are masked language model (MLM) and next sentence prediction (NSP). It solves the limitation that other models can't combine with context information. Bilstm is a bidirectional LSTM, LSTM is a variant of RNN, the RNN gradient vanishing problem is solved to some extent, but LSTM can only process data from forward sequence, it is also very important to process data in reverse sequence in text sentiment classification. The basic component of bilstm is LSTM, it consists of forward LSTM and backward LSTM, two independent hidden layers are used to process the data simultaneously from forward and backward directions to get complete semantic information.

LDA model comes from semantic analysis, in 2003, BLEI et al. Proposed a three-layer Bayesian probability topic model LDA, the method is based on singular value decomposition, the high dimensional document vector is approximately mapped to a low dimensional latent semantic space, in order to reduce the dimension of documents and eliminate the synonymy and polysemy of words. In the topic model, topic represents a concept, an aspect, and a series of Related words, which is the conditional probability of these words. Figuratively speaking, the theme is a bucket, which contains words with high probability of occurrence. These words have a strong correlation with the theme.

2.1. Emotion classification task definition
When analyzing the sentiment tendency of course reviews, we define all the review data sets as a set H. Set H contains a lot of comments, we define every comment sentence as an X_t, each t represents the position of the comment sentence set, H = \{X_1, X_2, X_3, X_4, X_5 ......\}. Each word in the comment sentence is defined as X_i. Where i is the position index of each sentence, S = \{X_1, X_2, X_3, X_4, X_5 ......\}. The result set of emotion classification is E = \{Y, N\}. Y represents positive emotion, which is the praise of learners. N represents negative emotion, which is the poor evaluation of learners.
2.2. Emotion analysis model of course review

Data pretraining, using Albert pretraining model, map each Chinese character in the MOOCs course comment sentence to a low-dimensional continuous vector space \( w_i \in \mathbb{R}^{d_w} \), where \( i \) is the word vector dimension in which \( d_w \) was the dimension of the word vector. The output of the word-embedding layer was the contextual vector of the comment: \( \{w_1, w_2, \ldots, w_n\} \in \mathbb{R}^{n \times d_w} \).

The BiLSTM can extract the context semantic information effectively by the forward sequence and the latter sequence, and then input the word vector into the model from the two aspects. From formula (1) to get the above emotional tendency feature, from formula (2) to get the following emotional tendency feature, and finally from formula (3) to get the global feature containing context information.

\[
\begin{align*}
\bar{h} &= f(\overline{wF}_i + \overline{uF}_{i-1}) \\
\hat{h} &= f(\overline{wB}_i + \overline{uB}_{i+1}) \\
\overline{h} &= [\overline{h}, \overline{\hat{h}}]
\end{align*}
\]

Emotion analysis results output, the emotion feature vectors generated by semantic extraction layer are input into softmax classifier. Finally, the emotion classification result predicted by the model is obtained, as shown in (4)

\[
p = \text{softmax}(w \cdot r + b)
\]

2.3. Topic extraction model of course review

LDA theme model is a three-layer Bayesian probability theme model proposed by BLEI et al. in 2003. The empty points in the graph represent the hidden variables, the solid point represents the observable value, and the matrix represents the repeated process. The large matrix represents the document set (a
total of \( Y \) positive emotions or \( n \) negative emotions). Small rectangles represent words that are repeatedly sampled from the topic distribution to produce documents.

\[
\begin{align*}
p(\text{word} | \text{document}) &= \sum_{\text{theme}} p(\text{word} | \text{theme}) \times p(\text{word} | \text{document}) \\
p(\beta, \gamma, w | \alpha, \theta) &= p(\beta | \alpha) \prod_{n=1}^{n} p(Z_n | \beta) p(w_n | Z_n, \beta)
\end{align*}
\]

3. Data set and evaluation index

We crawled 80,000 real Chinese comment data sets of 50 national excellent courses on the MOOC online learning platform through crawler technology. In order to improve the accuracy of topic extraction, we manually delete invalid comments and spam comments. Two graduate students majoring in natural language processing annotated topic sentences respectively. Finally, 7480 comments were selected from the two students, which were labeled as the experimental materials. In the experiment, 80% of the corpus is randomly selected as training data, and the remaining 20% as test data.

In order to eliminate the influence of different course review materials on the experimental results, cross validation is used. In order to verify the feasibility of the proposed model, two experiments are carried out. The first experiment is to verify the emotion classification accuracy of Albert + bilstm and other methods on the same dataset. The second experiment is to verify the topic extraction effect of this model and other models.

Experiment 1: in order to verify the superiority of Albert + bilstm. This paper uses different methods to carry out experiments on the same corpus, and compares the sentiment classification of three kinds of course reviews which are processed by bilstm algorithm, Bert + LSTM algorithm and algorithm. The accuracy of their sentiment classification is shown in the figure 3 below.
It can be seen that Albert + bilstm model is the best, this is because Albert uses sense order prediction (SOP), which focuses on the coherence between sentences rather than the matching between sentences, while bert + LSTM uses NSP task, which is too simple to learn from MLM task. In contrast, Bi LSTM considers capturing the global information of context in reverse order, but it also ignores the local information of comments, resulting in poor performance of the model.

Experiment 2: the second part mainly verifies the effect of topic extraction. This paper uses different methods, And the experimental comparison was made by using the emotional classification and the non-emotional classification. After emotion classification, the corpus and LDA model, the unclassified corpus and LDA model. The results are as follows:

![Figure 4 Topic extraction results.](image)

It can be seen from the figure 4 that the confusion degree of these two methods decreases with the increase of topics, and then increases with the increase of topics. The effect of topic extraction is a trend from good to bad. It can be seen from the figure 4 that the confusion degree of the unclassified LDA model is relatively high, and the effect of topic extraction is the worst, while the confusion degree of the classified LDA model in this paper is low, and the effect of topic extraction is better. This is because when we use the classification model, we consider the characteristics of the text itself of the MOOC course reviews. The reviews have two classification characteristics, positive emotion and negative emotion. When we carry out classification training, we can integrate the emotional characteristic parameters to pay more attention to different emotional theme tendencies in the topic extraction.

| Theme 1 | Theme 2 | Theme 3 | Theme 4 |
|---------|---------|---------|---------|
| Word | Probability | Word | Probability | Word | Probability | Word | Probability |
| Python | 0.055 | JAVA | 0.037 | C++ | 0.023 | SVM | 0.029 |
| Good | 0.041 | distinct | 0.018 | bored | 0.018 | Chatty | 0.012 |
| Vivid | 0.022 | Interest | 0.16 | Difficulty | 0.009 | Indistinct | 0.019 |
| Quality | 0.011 | Good | 0.15 | Kartun | 0.001 | Understand | 0.020 |

The first two topics shown in table 1 are positive emotions. We can see that Python courses are popular, and the overall evaluation is very good. The second topic is Java. Everyone is interested in this course. The latter two themes express negative emotions. We basically show that these two courses are relatively difficult, and the video quality of these two courses is not very good.

4. Conclusion and Future Work
This paper takes MOOC course review as the research object, and analyzes its text sentiment tendency and theme sentiment. Although certain results have been achieved, there are still many problems, which will be improved from the following aspects in the future.
Although the classification effect of the method proposed in this paper is superior to other deep learning and traditional machine learning methods, the model is more complex and takes a long time to converge. In the future, the model structure can be simplified.

The follow-up work hopes to make the above algorithm into a system and apply it to the online course platform, so that learners and platform managers can more clearly and intuitively understand the quality of the course and learners' concerns.

References

[1] Wang, L., Li, P., Zhu, Q. (2013) A topic sentence extraction method based on LDA model. Computer Engineering and Applications, (2): 160-164.

[2] Sun, Y., Zhou, X., Fu, W. (2013) The sentiment analysis of unsupervised text based on the subject-emotion mixed model. Beijing University Journal: Natural Science Edition, 49 (1): 102-108.

[3] Ma, X., Guo, S., Song, K. (2013) "Content-Method" Co-occurrence Analysis Research Based on LDA Topic Model—Taking Information Science as an Example. Information Science, 36(4):69-74.

[4] Renz, J., Staubitz, T., Willems, C., Klement, H., Meinel, C. (2014) [IEEE 2014 IEEE Global Engineering Education Conference (EDUCon) - Istanbul (2014.4.3-2014.4.5)] 2014 IEEE Global Engineering Education Conference (EDUCon) - Handling Re-Grading of Automatically Graded Assignments in MOOCs, 408-415.

[5] Zhang, K., Song, W., Liu, L. (2019) Bidirectional Long Short-Term Memory for Sentiment Analysis of Chinese Product Reviews. In: 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication. Beijing. pp. 665-668.

[6] Lai, S., Xu, L., Liu, K., Zhao, J. (2015) Recurrent convolutional neural networks for text classification. In: Twenty-ninth AAAI Conference on Artificial Intelligence. Austin. pp. 2267-2273.

[7] Zhao, W., Jiang, J., Yan, H., Li, X.M. (2010) Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Cambridge. pp. 56-65.

[8] Jing, Y., Gou, H., Fu, C., Sun, W. (2019) Sentiment Classification of Online Reviews Based on LDA and Semantic Analysis of Sentimental Words. In: 2019 12th International Symposium on Computational Intelligence and Design (ISCID). Hangzhou. pp. 249-252.

[9] Gu, Y., Xia, T. (2014) Research on Keyword Extraction Based on LDA and TextRank. Modern Library and Information Technology, 2014(21): 41-47.

[10] Chen, Q., Zhang L., Jiang, J. (2019) A review analysis method based on support vector machines and topic models. Journal of Software, 30(5):1547-1560.