An AI-BERT-Bi-GRU-LDA algorithm for negative sentiment analysis on Bilibili comments

Ziyu Liang and Jun Chen

School of Education, Guizhou Normal University, Guiyang, China

ABSTRACT

The number of online self-learning users has been increasing due to the promotion of various lifelong learning programs. Unstructured commentary text related to their real learning experience regarding the learning process is generally published by users to show their opinions and complaints. The article aims to utilize the dataset of real text comments of 10 high school mathematics courses participated by high school students in the Bilibili platform and construct a hybrid algorithm called the Artificial Intelligence-Bidirectional Encoder Representations from Transformers (BERT) + Bidirectional Gated Recurrent Unit (BiGRU) and linear discriminant analysis (LDA) to crunch data and extract their sentiments. A series of tests regarding algorithm comparison were conducted on the educational review datasets. Comparative analysis found that the proposed algorithm achieves higher precision and lower loss rates than other alternative algorithms. Meanwhile, the loss ratio of the proposed algorithm was kept at a low level. At the topic mining level, the topic clustering of negative comments found that the barrage content was very messy and the complexity of the course content was generally reported by the students. Some problems related to videos were also mentioned. The outcomes are promising that the fundamental issues underlined by the students can be effectively resolved to improve curriculum and teaching quality.

INTRODUCTION

In the era of digital learning, the learning methods utilized by students grow more diverse, not only for students in higher education but also for those in basic educational levels. So, the analysis method developed for online course platforms can be applied to MOOCs and other similar courses. For example, Chinese students often use a video-sharing website called Bilibili to learn and entertain. Many excellent first-line teachers and course developers have shared high-quality online course resources on the Bilibili platform. These online courses are free, open, and easy to use anytime. Since the online learning method requires users to be familiar with the technology to carry out regular educational and teaching activities, it is prone to new problems due to social reasons such as insufficient interaction between teachers and students, the lack of online learners' emotional participation, and other issues.

Corresponding author
Ziyu Liang, lzy15926@163.com

On the other hand, online learners often communicate and interact in online discussion forums. These comments are complex and diverse regarding the content, cover a wide range of areas, and become the primary way to share and communicate among learners (*Bai et al., 2019*). In addition, the comments, for example, on the Bilibili platform are also very different from other types of comments, not only because the sentences are shorter but also because online learners will record their real experiences and feelings during the online learning in the comment area, which is related to the quality of courses, teachers' performances and the service quality of the platform, which provides some inspiration and reference to improve the quality of courses.

When users' sentiments are a concern, positive comments may be temporary, and the content of the comments may be monotonous. It may not contain substantial information that could not help improve the course content. At the same time, negative comments usually point out the shortcomings of courses and the service quality of the platform. Thus, teachers must pay close attention to negative comments (*Li et al., 2019*). However, teachers and course developers cannot read every negative comment pertinent to each course individually.

Moreover, micro texts are shorter and have a more diversified emotional vocabulary than short texts such as news, *etc.* Therefore, performing sentiment analysis on micro texts such as educational review texts is more challenging (*Gui et al., 2017*). The research employs hybrid sentiment classification algorithms to extract the emotional polarity of comment texts on online courses, primarily to derive negative texts with negative comments and carry out topic clustering, which helps assess online courses' quality and provides helpful information for teachers. Hence, valuable feedback will promote the quality improvement of online courses.

The remainder of the article is organized as follows: Section 'Related Works' provides a literature review of sentiment analysis and its applications in education. Section 'The Proposed Algorithm: ABGAL' introduces the Al-BERT-BiGRU+LDA algorithm based on the attention mechanism. Section 'Experimentation' presents the experimental dataset and experimental results. Section 'Conclusion' concludes the research and presents a direction for future research.

RELATED WORKS

Pre-trained model

Most researchers (*Szegedy et al., 2017*; *Wu, Shen & Van DenHengel, 2019*; *Singh, Hoiem & Forsyth, 2016*) have confirmed through experimental studies that using pretraining technology can significantly improve the effect of downstream tasks. The early pretraining technique was a static encoding technique. In 2003, Bengio incorporated the idea of deep learning into the language model and proposed the NNLM (*Bengio, Ducharme & Vincent, 2000*) for a pretraining model.

In 2013, Word2Vec borrowed the idea from NNLM and employed a language model to construct word vectors. This was followed by Global Vectors (GloVe) (*Pennington, Socher & Manning, 2014*) and FastText (*Joulin et al., 2016*). Word2Vec and GloVe are the most

widely used language models in natural language processing (NLP). Many studies have shown (*Al-Amin, Islam & Uzzal, 2017*; *Tang et al., 2015*; *Grover & Leskovec, 2016*) that the Word2Vec and GloVe models can significantly improve text vectorization's representation effect in NLP tasks. Therefore, the vectorized representation of a text is fixed, and the model cannot effectively combine the text context information of the context globally. To resolve the problem of polysemy and word ambiguity and capture more useful deep semantic information, In 2018, ELMo (*Sarzynska-Wawer et al., 2021*) was proposed as a text representation method that could combine context. Since then, the GPT model (*Radford et al., 2018*), BERT algorithm (*Devlin et al., 2018*), *etc.*, have been suggested.

Moreover, it outperforms other pre-trained language algorithms in many typical tasks in NLP, which becomes an essential milestone in NLP. Our work uses the Al-BERT algorithm as the basic pretraining model. It obtains rich dynamic encoded semantic information through the Al-BERT algorithm to pave the way for deep text sentiment classification to extract text features better.

Sentiment classification approaches
Traditional approaches for sentiment classification

First, the sentiment lexicon has been broadly implemented as an essential approach in sentiment classification tasks. *Wilson, Wiebe & Hoffmann (2005)* employed manual annotators to manually annotate a set of words to build a sentiment dictionary The algorithm can automatically detect the polarity of sentiment expressions in subsets and achieves excellent results. *Tang et al. (2014)* proposed an algorithm for constructing a large-scale emotional vocabulary based on representation learning methods. They introduced the Urban Dictionary to expand several emotional seed words to obtain more training data for constructing phrase-level emotional classifiers (*Tang et al., 2014*). *Khoo & Johnkhan (2018)* introduced a new universal sentiment dictionary, the WKWSCI sentiment dictionary, and conducted a comparative experiment with five existing sentiment dictionaries (HuLiu Opinion Lexicon, *etc.*) and found that Hu-Liu Opinion Lexicon performed well. It is suitable for sentiment mining of commentary texts, while the WKWSCI sentiment dictionary is suitable for non-commentary texts. Researchers have also tried combining sentiment lexicons with deep learning algorithms to improve classification performance in recent years. *Yang et al. (2020)* suggested a hybrid algorithm based on a sentiment lexicon combined with CNN and Bi-GRU based on an attention mechanism. Experimental results show that the algorithm can effectively improve the text performance of sentiment classification. However, due to the inability of the method used to construct the sentiment dictionary to unify contextual semantic features well, the classification precision of the algorithm is low, and there are issues such as an extended model construction period and slow update speed.

As classic supervised algorithms, machine learning algorithms usually implement sentiment classification through artificial feature engineering and achieve reliable classification results (*Bao et al., 2021*). *Joachims (2005)* suggested a new method to extract the emotional polarity of the text, that is, first determine whether a sentence expression is neutral or emotionally polarized and then distinguish the sentence expression with

emotional polarity. Through this method, the emotional polarity of many sentence texts can be automatically identified. Classifying texts with negative reviews and implementing machine learning models also achieved good outcomes (*Liu & Yang, 2020*). Machine learning is representative of supervised learning. Model training must depend on much training data with manual annotations. Experts need to formulate rules and perform manual annotations in advance. Nevertheless, the process is difficult, and there are problems of poor portability and time-consuming and labor-intensive problems. Thus, a semi-supervised learning algorithm is an effective solution (*Tan et al., 2021*).

Sentiment classification based on deep learning approach

The last decade has witnessed deep learning gradually replacing machine learning algorithms, especially in NLP tasks, and its effects have been praised (*Sailunaz & Alhajj, 2019*; *Xu et al., 2020*; *Kastrati, Imran & Kurti, 2020*; *Hassan & Mahmood, 2017*; *Chen et al., 2017*). As a compelling model, CNN has made significant image and video processing breakthroughs due to its excellent local feature extraction ability. However, the learning ability for sequential data, especially text data, has been severely challenged (*Liu & Guo, 2019*). With the advent of recurrent neural networks (RNNs), Possibilities were seen for sequence data processing and prediction. RNN realizes the natural processing of time series data by storing and recalling information. However, in actual operation, we found that when RNN processed long text sequences, it was hard to converge during model training, impacting the training effect and model stability (*Pascanu, Mikolov & Bengio, 2013*). To resolve the issue, researchers proposed the LSTM algorithm (*Hochreiter & Schmidhuber, 1997*). LSTM preserved long-term dependencies on text content by introducing forget and update gates to manipulate, update, and ignore information in the hidden state. However, LSTM must deal with a long training time, numerous parameters, and high computational complexity. This complexity is the price of performance improvement, but it also becomes an obstacle to its use. Based on these issues, researchers combined linear variables with nonlinear gating structures to propose a more efficient GRU algorithm (*Chung et al., 2014*). In the design of GRU, by combining the update gate and the forget gate into one update gate, the network structure is simplified, parameters are reduced, calculations are saved, and better convergence is obtained. In addition, to get the front and rear information of the text sequence simultaneously, Bi-GRU integrates the forward and backward hidden layer information to attain richer contextual text information. This algorithm simplifies the model by unifying the update gate and the forget gate, which resolves the problem that RNNs face.

With the continuous advancement of deep learning algorithms, the attention mechanism has become widely implemented in various deep learning fields. Some researchers have applied it to text extraction and summary generation tasks. *Rush, Chopra & Weston (2015)* proposed an entirely data-driven approach to text summarization by employing a local attention model. Compared with other algorithms, the correct rate of summaries generated by the proposed method is greatly improved. *Parikh et al. (2016)* proposed a simple reasoning framework for NLP with the help of an attention mechanism, which can

decompose complex text language into several simple subtexts. It dramatically reduces the complexity when a text is processed.

To sum up, LSTM, GRU, and Bi-GRU resolve a series of complex challenges that conventional algorithms face, especially in processing complex and extended sequences of text information. They show excellent performance and scalability, opened a new chapter in good model universality, and contributed significantly to improving and optimizing algorithms and effects in NLP. Therefore, this study combines the Al-BERT dynamic pretraining model with the Bi-GRU-Attention deep learning algorithm and the LDA model to obtain the deep emotional features of the learners' comment texts and perform topic clustering and analysis on the negative comment texts.

THE PROPOSED ALGORITHM: ABGAL

This section introduces the algorithm architecture ABGAL. The model is constructed with two modules: the sentiment classification and the topic clustering. The sentiment classification module is divided into three layers: the Senti-Al-BERT layer, the deep Senti-extraction layer, and the softmax layer, respectively. After the sentiment classification module is run, the topic clustering module is built to realize the text topic clustering of negative comment texts to capture high-value information in negative comments. The diagram of the proposed algorithm is shown in Fig. 1.

Al-BERT layer

The BERT model was suggested by Google in 2018 and achieved good results by stacking the Transformer substructure as a feature extractor and implementing the self-attention mechanism to encode words directly. However, when the model is expanded to a particular scale, the model training will take up excellent memory usage, the training time will be so long, and the model will degenerate and cause overfitting problems. To resolve these issues, scholars such as *Lan et al. (2019)* proposed a lightweight version of the BERT model, namely, Al-BERT (A Lite BERT), by employing the factorized embedding parameterization method and the cross-layer parameter sharing approach to cause the amount of the model's parameter significantly reduced. It performs better on NLP tasks while reducing the number of parameters. Therefore, this study uses Al-BERT as a pretraining model. In the vectorized representation of text generated by Al-BERT, the vector corresponding to input in the algorithm has three superimposed vectors. It is called the Token Embedding Segment.

Embedding and Position Embedding. By converting each word into matrix–vector forms of Q, K, and V through different linear transformations and calculating independently of each other, a unique vector encoding for each term is obtained so that it can have a deeper perception of the context (*Liu et al., 2020*). Equation (1) presents the self-attention calculation.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V \tag{1}$$

Q, K, and V denote the query, key, and value matrix, respectively. The research implements the trained Al-BERT algorithm, which consists of 12 layers of Transformer. The embedding
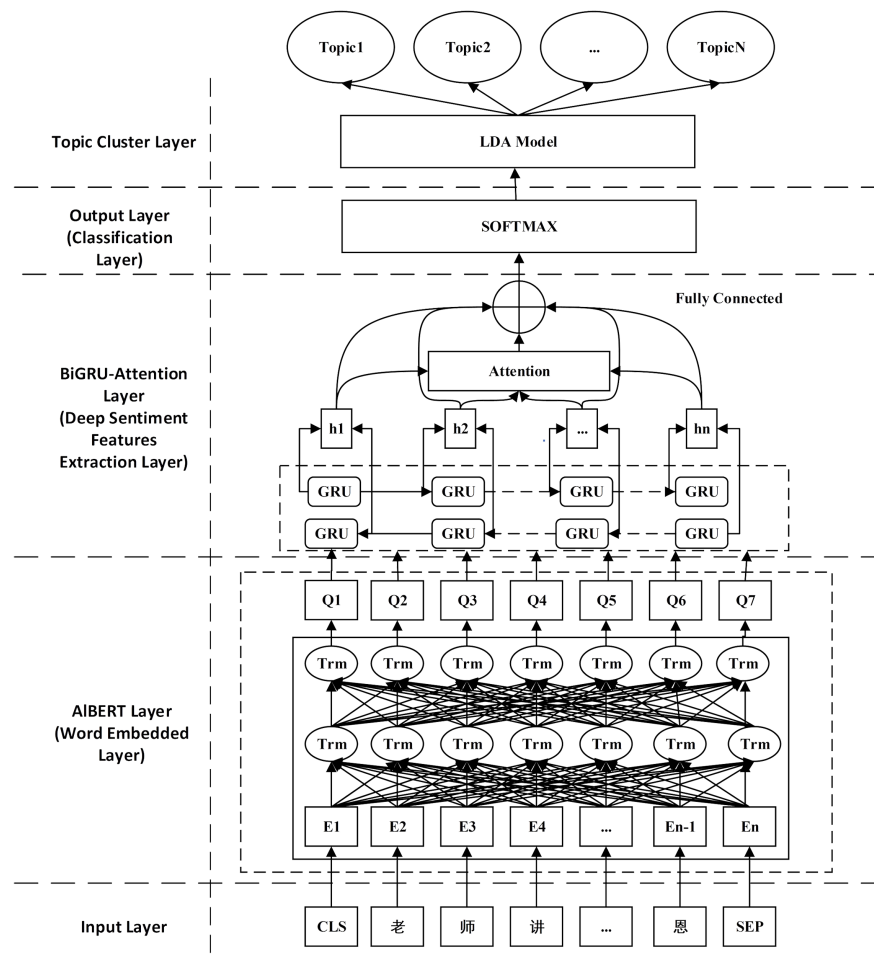
Figure 1   Structural diagram of ABGAL model.

dimension of the word vector is assigned to 768. To ensure that the vector input to the Al-BERT algorithm is easy to operate, this research sets the length of the input sequence to 512. Only the part within the maximum range is reserved for sentences exceeding the maximum length, and < padding > is used to fill in the insufficient length sequence.

Bi-GRU layer

The GRU maintains the same effect as the LSTM and combines the input gate and the forget gate in LSTM into one gate, called the update gate, so the algorithm only consists of the reset gate and the update gate. Figure 2 shows the GRU structure. The parameters are more simplified and have better convergence. Equation (2) presents the update gate.

$$Z_t = \sigma \left(W_z S_t + U_z H_{t-1} + B_z \right) \tag{2}$$
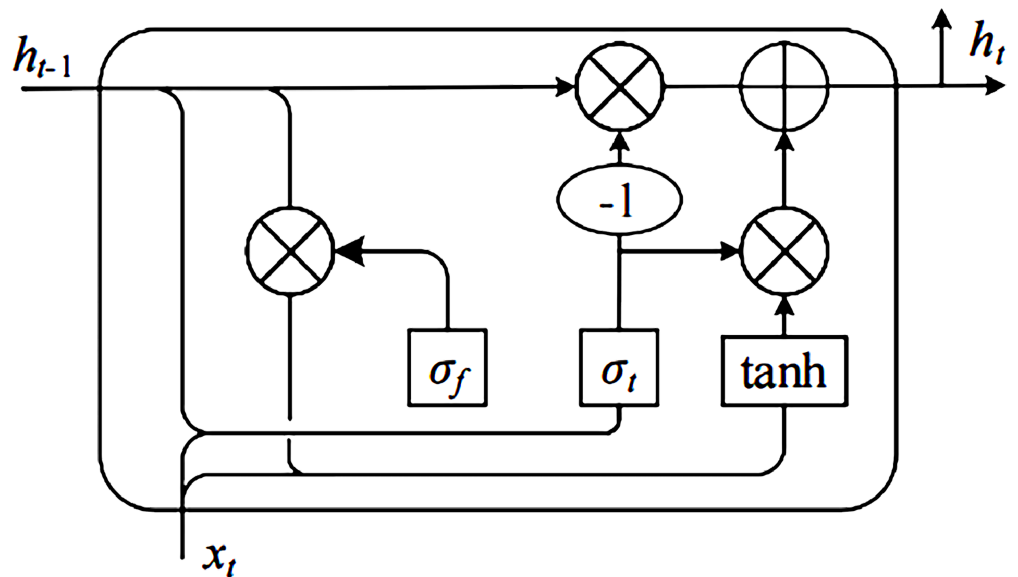
Figure 2 Fig Structural diagram of GRU.

where W_z and U_z denote weights, and H_{t-1} represents the input and B_z denotes the bias. Then, GRU calculates the candidate content that needs to be retained. W and U denote weighted, and B represents the bias:

$$\hat{H}_t = \tanh(WS_t + UR_t H_{t-1} + B). \tag{3}$$

Finally, the GRU calculates the final output based on the above-obtained results:

$$H_t = (1 - Z_t)\hat{H}_t + Z_t H_{t-1}. \tag{4}$$

In the Bi-GRU layer, the emotional features are extracted in the forward and backward GRU networks, which can capture the global features of the context. The two reverse GRUs jointly determine the output of the Bi-GRU network.

Attention layer

As one of the most substantial concepts in deep learning, the attention mechanism is a weight allocation method that simulates the allocation mechanism of human brain attention. At a particular moment, the human brain will focus on a specific area that needs attention, reducing or ignoring other regions to grasp more valuable information (*Niu, Zhong & Yu, 2021*). The Attention mechanism assigns more significant weight to crucial details based on probability distribution so that the algorithm can pay more attention to this type of information. In short, an attention mechanism-based text classification can amplify core keywords' impact on text features. As shown in Fig. 1, The result of the weight distribution will affect the final classification outcome (*Liu et al., 2019*). Equations (5) and

(6) present attention allocation.

$$C_t = \sum_{t=1}^{T_h} a_{ij} H_t \tag{5}$$

$$a_{ij} = \mathrm{softmax}(W_{a2}\tanh(W_{a1}H_t)) \tag{6}$$

where a_{ij} denotes attention weight, H_t represents the length of the data, W_{a1} and W_{a2} designate weights.

EXPERIMENTATION

In this section, experiments are conducted to evaluate the performance of the sentiment classification algorithm proposed, and the educational review dataset is implemented in the article.

Experimental procedure

- **Data collection and pre-processing:** First, we utilized crawler technology to collect the comments written by ten high school mathematics courses that high school students participated in on the Bilibili platform. The total number of texts is 20,000 in Chinese. Furthermore, the collected corpus is pre-processed to remove irrelevant comments, meaningless URLs, and non-Chinese comments (including comments with only symbols and English characters). A total of 12,655 usable comments are obtained.
- **Labeling:** The comments with obvious positive adjectives and adverbs are marked as positive, sentences without adjectives are marked as unfavorable, and adjectives or adverbs with negative emotional color and interrogative sentences are marked as negative. Through iterative processing, 6,929 positive text comments, 3,686 neutral comments, and 2,042 negative comments were obtained. Since the number of features has a particular impact on the precision analysis, "more is better" does not guarantee better outcomes. Therefore, the feature selection process must consider both efficiency and precision (*Zheng, Wang & Gao, 2018*). Therefore, 2,500 positive reviews were selected randomly. The distribution of the data set is shown in Table 1.
- **Word representation:** We use the trained Chinese model to vectorize Chinese text comments, and the processed text vector has 768 dimensions.
- **Building deep sentiment classification model:** Al-BERT-Bi-GRU-Att+LDA: The bi-GRU model with attention mechanism based on the proposed Al-BERT pretraining algorithm. In the Bi-GRU layer, the input dimension is consistent with the dimensions of the Al-BERT layer, and the hidden layer dimension is assigned to 256. The input sentence length is assigned to 100. In the attention layer, different weights are assigned to each word through the attention mechanism, and topic clustering is performed on the comments classified as negative.
- **Optimizing model:** During the training stage, the hyperparameters are modified to optimize the final model's performance.
- **Model evaluation:** We evaluate the prediction results with precision, recall, and F1 score.

Table 1	Dataset statistics.			
Dataset	Positive	Neutral	Negative	Total
Train set	4,170	2,288	1,274	7,732
Validation set	1,205	602	345	2,152
Test set	1,471	766	416	2,653

- **Topic clustering** After the model obtains the final classification result; the LDA topic model is used for topic clustering for the comments whose classification results are negative. Thus, the problems that learners are concerned about can be detected and the course developers can present solutions.

Evaluation criteria

Evaluating the model performance is an essential link, and different evaluation indicators comprehensively reflect the quality of the model's performance. The model evaluation indicators used in the article are precision, precision, recall, and the F1-scores, consistent with most studies' indicators. They are presented in Table 2.

$$precision = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{10}$$

Parameter setting

The experiment's data set was 5-fold cross-validated to obtain accurate experimental results, and the samples were randomly divided into five groups. One group was used for training, and the remaining four were used for the validation and test sets. The research compares the performance of LSTM, Bi-LSTM, Bi-LSTM-Att, Text-CNN, and Al-BERT-Bi-GRU-Att algorithms regarding precision and loss value, respectively. The hyperparameters of the proposed algorithm are shown in Table 3:

Baseline methods

To verify the actual performance of the proposed algorithm, baseline text classification models are implemented as controls.

LSTM: The input dimension is assigned to 300, and the number of hidden layer dimensions is 300.

Bi-LSTM: the input dimension is word vector size, and the number of hidden layer dimensions is set to 300

Bi-LSTM-Att: The hidden layer dimension is assigned to 300 and combined with an attention mechanism to achieve accurate weight allocation.

Text-CNN: The filter length is 300, and the widths are set to 3,4 and 5, respectively.

Table 2 Classification results of the model.

True situation	Forecast result	
	Positive	**Negative**
Positive	TP	FN
Negative	FP	TN

Table 3 Parameter list.

Parameter name	Parameter value
Word vector	AlBERT
Vector dimension	768
Loss function	Cross entropy
Optimizer	Adam
Maximum text length	512
Batch size	64
Epoch	150

Findings
Model comparison

The precision and loss values of the algorithm on the dataset are shown in Fig. 3. According to the precision rate curve, the precision rate of LSTM is lower than that of other classification algorithms. By the 10th training cycle, the precision rate has just reached 0.62 in the first training cycle of other algorithms. Text-CNN reached the highest precision rate of 0.86 in the 10th training cycle, and then the precision rate stabilized at 0.86 and fluctuated back and forth. The overall change in the precision of Bi-LSTM and Bi-LSTM-Att is relatively small, reaching the same level in the tenth training cycle. Both are 0.85, which is slightly lower than the Text-CNN model. Compared with the baseline algorithm, the proposed algorithm reaches the highest level of 0.93 in the 10th training cycle. It employs Al-BERT as the pretraining model, which reduces the risk of overfitting and the computational overhead, which is better than the baseline algorithm.

According to the loss rate curve, the loss value of LSTM is the largest among all deep learning algorithms. Compared with the LSTM algorithms, the loss rate of Bi-LSTM decreases significantly, and it has better convergence. When compared with the Bi-LSTM algorithm, the precision and loss rates of Bi-LSTM-Att do not considerably change the precision and loss rate scores. Compared with other types of neural networks, Text-CNN performs well overall because the convolutional layer is good at extracting local features to handle classification tasks better.

The proposed model has undergone 5-fold cross-validation, and the loss value is reduced to about 6% in the final stage. Compared with other experimental algorithms, the precision of the proposed algorithm has increased by at least 5%, and the loss rate dropped by at least 3%. Experiments show that Al-BERT-Bi-GRU-Att-Softmax can enhance feature extraction capability and improve classification precision.
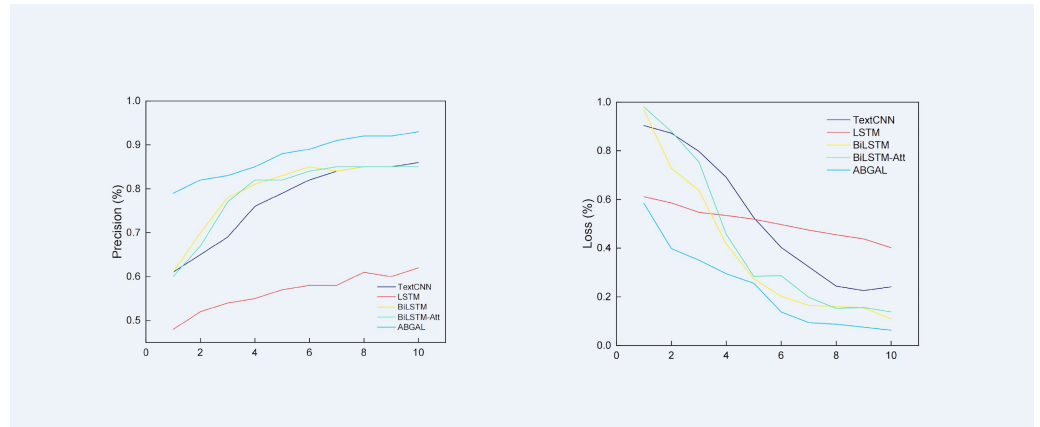
Figure 3 Precision and loss rate of different deep learning models.

Topic clustering

When the model's classification is a concern, we also perform topic clustering on predicted negative reviews. *Hew, Qiao & Tang (2018)* suggested that course reviews were divided into six topics for analysis. The primary process is to divide all negative comments into six topics for statistics after word frequency statistics are performed on the dataset: course content, course difficulty, video quality, resource acquisition, teacher attributes, and online interactions. Also, word frequency statistics for keywords under each topic are performed and quantitative analysis is conducted based on the final statistical results. Table 4 presents the outcomes.

Discussion

According to the statistical results of negative comments under each topic, the following conclusions are attained: First, most online learners generally reflect that the course content is challenging to understand, and the verbal expressions such as "difficult," "complex", and "do not understand" appear the most frequently, generally reflecting that the course content is hard to understand. That is mainly related to the nature of the mathematics itself. Some students also discuss questions about mathematics knowledge in the comment area. Students generally feel anxious about learning mathematical subjects and the exams they take. Secondly, there is a lot of feedback on bullet chatting. Many students complain that the content of bullet chatting is very messy and chaotic, and most of them have nothing to do with the course content, which affects their learning success. Some students say that the video screen suddenly turned black, they could not watch it, and the video sound was low.

Moreover, some students reported that more than the video content was needed and hoped to increase the range. Students have less negative feedback about teachers regarding teacher attributes since high teaching quality is available in online courses. Participants also reported that the order of the courses needed to be corrected. Regarding resource acquisition, a few students noted that courseware and handouts could not be downloaded because links expired.

Table 4 Topic clustering quantitative statistical results.

Subject	Keywords	Result
Course difficulty	Difficult, Complex(9)	594
Online interaction	Bullet comment	97
Video quality	Video, Sound(3)	23
Teacher attributes	Voice, Speech rate(5)	23
Resource accessibility	link expired(3)	13

CONCLUSION

The research innovation lies in the quantitative and qualitative analysis of texts when educational reviews generate texts. To do so, the combination of deep sentiment classification and topic models is conducted to derive more insights. The proposed algorithm effectively resolves the ambiguity problems in comment semantics. Experiments show that it generates a better classification result. A more than 5% precision rate is gained when compared with other deep learning algorithms that deal with special sentiment classification tasks, such as educational reviews.

Meanwhile, negative comments are employed to explore the distribution of various topics. The shortcoming of the article is that it only classifies educational review texts whose size is limited. Therefore, more research is needed to examine more extensive texts. Future work will extend this research to larger data sets and try to improve the proposed algorithm, further reviewing its effectiveness and potential impact on performance improvement.

Funding
The authors received no funding for this work.

Competing Interests
The authors declare there are no competing interests.

Author Contributions
- Ziyu Liang conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Jun Chen conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.

Data Availability
The code and raw data are available in the Supplemental Files.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj-cs.2029#supplemental-information.

REFERENCES

Al-Amin M, Islam MS, Uzzal SD. 2017. Sentiment analysis of Bengali comments with word2 vec and sentiment information of words. In: *2017 international conference on electrical, computer and communication engineering (ECCE)*. Piscataway: IEEE, 186–190.

Bai Q, Hu QV, Ge L, He L. 2019. Stories that big danmaku data can tell as a new media. *IEEE Access* **7**:53509–53519 DOI 10.1109/ACCESS.2019.2909054.

Bao T, Ren N, Luo R, Wang B, Shen G, Guo T. 2021. A bert-based hybrid short text classification model incorporating cnn and attention-based bigru. *Journal of Organizational and End User Computing (JOEUC)* **33**:1–21 DOI 10.4018/JOEUC.294580.

Bengio Y, Ducharme R, Vincent P. 2000. A neural probabilistic language model. In: *Advances in neural information processing systems, vol. 13*.

Chen P, Sun Z, Bing L, Yang W. 2017. Recurrent attention network on memory for aspect sentiment analysis. In: *Proceedings of the 2017 conference on empirical methods in natural language processing*, Kerrville: Association for Computational Linguistics, 452–461.

Chung J, Gulcehre C, Cho K, Bengio Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. ArXiv arXiv:1412.3555.

Devlin J, Chang M-W, Lee K, Toutanova K. 2018. Bert: pretraining of deep bidirectional transformers for language understanding. ArXiv arXiv:1810.04805.

Grover A, Leskovec J. 2016. node2vec: scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, New York: Association for Computing Machinery, 855–864.

Gui L, Zhou Y, Xu R, He Y, Lu Q. 2017. Learning representations from the heterogeneous network for sentiment classification of product reviews. *Knowledge-Based Systems* **124**:34–45 DOI 10.1016/j.knosys.2017.02.030.

Hassan A, Mahmood A. 2017. Deep learning for sentence classification, in 2017 IEEE Long Island systems. In: *Applications and technology conference (LISAT)*. Piscataway: IEEE, 1–5.

Hew K, Qiao C, Tang Y. 2018. Understanding student engagement in large-scale open online courses: a machine learning facilitated analysis of student's reflections in 18 highly rated MOOCs. *The International Review of Research in Open and Distributed Learning* **19(3)**:69–93 DOI 10.19173/irrodl.v19i3.3596.

Hochreiter S, Schmidhuber J. 1997. Long short-term memory. *Neural Computation* **9**:1735–1780 DOI 10.1162/neco.1997.9.8.1735.

Joachims T. 2005. Text categorization with support vector machines: learning with many relevant features. In: *Machine learning: ECML98: 10th European conference*

on machine learning Chemnitz, Germany, April (1998) 21–23 Proceedings. Cham: Springer, 137–142.

Joulin A, Grave E, Bojanowski P, Mikolov T. 2016. Bag of tricks for efficient text classification. ArXiv arXiv:1607.01759.

Kastrati Z, Imran AS, Kurti A. 2020. Weakly supervised framework for aspect-based sentiment analysis on students' reviews of MOOCs. *IEEE Access* **8**:106799–106810 DOI 10.1109/ACCESS.2020.3000739.

Khoo CS, Johnkhan SB. 2018. Lexicon-based sentiment analysis: comparative evaluation of six sentiment lexicons. *Journal of Information Science* **44**:491–511 DOI 10.1177/0165551517703514.

Lan Z, Chen M, Goodman S, Gimpel K, Sharma P, Soricut R. 2019. Albert: a lite bert for self-supervised learning of language representations. ArXiv arXiv:1909.11942.

Li X, Zhang H, Ouyang Y, Zhang X, Rong W. 2019. A shallow bert-cnn model for sentiment analysis on MOOCS comments. In: *2019 IEEE international conference on engineering, technology, and education (TALE)*. Piscataway: IEEE, 1–6.

Liu G, Guo J. 2019. Bidirectional 1stm with attention mechanism and convolutional layer for text classification. *Neurocomputing* **337**:325–338 DOI 10.1016/j.neucom.2019.01.078.

Liu J, Yang Y, Lv S, Wang J, Chen H. 2019. Attention-based BIGRU-CNN for Chinese question classification. *Journal of Ambient Intelligence and Humanized Computing* **2019**:1–12 DOI 10.1007/s12652-019-01344-9.

Liu S, Yang G. 2020. Negative sentiment analysis of MOOC comments based on machine learning. In: *Advanced graphic communication, printing, and packaging technology: proceedings of 2019 10th China academic conference on printing and packaging*. Springer, 562–567.

Liu Y, Lu J, Yang J, Mao F. 2020. Sentiment analysis for e-commerce product reviews by deep learning model of bert-big-softmax. *Mathematical Biosciences and Engineering* **17**:7819–7837 DOI 10.3934/mbe.2020398.

Niu Z, Zhong G, Yu H. 2021. A review on the attention mechanism of deep learning. *Neurocomputing* **452**:48–62 DOI 10.1016/j.neucom.2021.03.091.

Parikh AP, Täckström O, Das D, Uszkoreit J. 2016. A decomposable attention model for natural language inference. ArXiv arXiv:1606.01933.

Pascanu R, Mikolov T, Bengio Y. 2013. On the difficulty of training recurrent neural networks. In: *International conference on machine learning, PMLR*, 1310–1318.

Pennington J, Socher R, Manning CD. 2014. Glove: Global vectors for word representation. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, Kerrville: Association for Computational Linguistics, 1532–1543.

Radford A, Karthik N, Tim S, Ilya S. 2018. Improving language understanding by generative pre-training.

Rush AM, Chopra S, Weston J. 2015. A neural attention model for abstractive sentence summarization. ArXiv arXiv:1509.00685.

Sailunaz K, Alhajj R. 2019. Emotion and sentiment analysis from Twitter text. *Journal of Computational Science* **36**:101003 DOI 10.1016/j.jocs.2019.05.009.

Sarzynska-Wawer J, Wawer A, Pawlak A, Szymanowska J, Stefaniak I, Jarkiewicz M, Okruszek L. 2021. Detecting formal thought disorder by deep contextualized word representations. *Psychiatry Research* **304**:114135 DOI 10.1016/j.psychres.2021.114135.

Singh S, Hoiem D, Forsyth D. 2016. Swapout: Learning an ensemble of deep archi-tectures. In: *Advances in neural information processing systems* New York: Curran Associates, Inc,.

Szegedy C, Ioffe S, Vanhoucke V, Alemi A. 2017. Inception-v4, inception-Resnet and the impact of residual connections on learning. In: *Proceedings of the AAAI conference on artificial intelligence. Vol. 31*. New York: Association for Computing Machinery, 4278–4284.

Tan X, Zhuang M, Lu X, Mao T. 2021. An analysis of the emotional evolution of large-scale internet public opinion events based on the Bert-LDA hybrid model. *IEEE Access* **9**:15860–15871 DOI 10.1109/ACCESS.2021.3052566.

Tang D, Wei F, Qin B, Zhou M, Liu T. 2014. Building large-scale twitter specific sentiment lexicon: a representation learning approach. In: *Proceedings of Coling 2014, the 25th international conference on computational linguistics: technical papers*, Kerrville: Association for Computational Linguistics, 172–182.

Tang J, Qu M, Wang M, Zhang M, Yan J, Mei Q. 2015. Line: Largescale information network embedding. In: *Proceedings of the 24th international conference on world wide web*, New York: Association for Computing Machinery, 1067–1077.

Wilson T, Wiebe J, Hoffmann P. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In: *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, Kerrville: Association for Computational Linguistics, 347–354.

Wu Z, Shen C, Van DenHengel A. 2019. Wider or deeper: revisiting the resnet model for visual recognition. *Pattern Recognition* **90**:119–133 DOI 10.1016/j.patcog.2019.01.006.

Xu D, Tian Z, Lai R, Kong X, Tan Z, Shi W. 2020. Deep learning-based emotion analysis of microblog texts. *Information Fusion* **64**:1–11 DOI 10.1016/j.inffus.2020.06.002.

Yang L, Li Y, Wang J, Sherratt RS. 2020. Sentiment analysis for e-commerce product reviews in Chinese based on sentiment lexicon and deep learning. *IEEE Access* **8**:23522–23530 DOI 10.1109/ACCESS.2020.2969854.

Zheng L, Wang H, Gao S. 2018. Sentimental feature selection for sentiment analysis of Chinese online reviews. *International Journal of Machine Learning and Cybernetics* **9**:75–84 DOI 10.1007/s13042-015-0347-4.