Education and Information Technologies (2023) 28:797–814
https://doi.org/10.1007/s10639-022-11151-z

Automatic scoring of student feedback for teaching evaluation based on aspect-level sentiment analysis

Ping Ren[1] · Liu Yang[1] · Fang Luo[2]

Received: 22 September 2021 / Accepted: 2 June 2022 / Published online: 7 July 2022

Abstract
Student feedback is crucial for evaluating the performance of teachers and the quality of teaching. Free-form text comments obtained from open-ended questions are seldom analyzed comprehensively since it is difficult to interpret and score compared to standardized rating scales. To solve this problem, the present study employed aspect-level sentiment analysis using deep learning and dictionary-based approaches to automatically calculate the emotion orientation of text-based feedback. The results showed that the model using the topic dictionary as input and the attention mechanism had the strongest prediction effect in student review sentiment classification, with a precision rate of 80%, a recall rate of 79% and an F1 value of 79%. The findings identified issues that were not otherwise apparent from analyses of purely quantitative data, providing a deeper and more constructive understanding of curriculum and teaching performance.

Keywords Student evaluations of teaching · Sentiment analysis · Aspect level · Dictionary-based approach · Deep learning

1 Introduction

The quality of education is a hot topic around the world, and it intrinsically depends on teaching efficiency. Thus, accurate and efficient evaluation of teaching has become

✉ Fang Luo
luof@bnu.edu.cn

[1] Collaborative Innovation Center of Assessment for Basic Education Quality, Beijing Normal University, Beijing, China

[2] Beijing Key Laboratory of Applied Experimental Psychology, School of Psychology, Beijing Normal University, Beijing, China

an important academic research field. In recent years, the value of "student-oriented" education has been widely publicized and recognized, with greater attention placed on the opinions and voices of students (Annan et al., 2013). As intuitive and direct reflections of classroom instruction, student evaluations of teaching (SETs) are considered reliable, effective and valuable assessments (Ory, 2000; Rajput et al., 2016). Historically, student feedback on courses and teachers has been integral to accountability in education (Annan et al., 2013). In practice, student evaluations are not only linked to administrative decisions about teachers' promotions and merit pay raises (Beran & Rokosh, 2009; Emerson & Records, 2007), but they also provide teachers with an overall picture of their teaching performance and help improve levels of instruction (Hoon et al., 2014; Smith, 2008). Such evaluations can strengthen the communication between teachers and students to promote students' growth and development (Kulik, 2001; Nasser & Fresko, 2002).

2 Literature review

2.1 Forms of SETs

Although SETs have been proved to have many benefits, a problem that persists in academia concerns the instruments used for such evaluations, as most student evaluation tools rely on rating scales (Denson et al., 2010) and rarely use open-ended items (Onwuegbuzie et al., 2009). This design has the following disadvantages. First, rating scores obtained from Likert-format scales are usually biased. The "predefined question + score" model is often problematic (i.e., using different standards for scoring), while attitudes obtained from written comments are likely to offer more stable feedback (Emerson & Records, 2007; Serdyukova et al., 2010). Second, SETs instruments differ considerably in the quality of items, the way the teaching-effectiveness construct is operationalized, and particular dimensions that are included. In practice, many instruments are based on a mixture of logical and pragmatic considerations and are rarely assessed psychometrically (Annan et al., 2013; Donnon et al., 2010). Poorly worded or inappropriate items will not provide useful information. Third, students who evaluate teaching using rating instruments are more likely to purposely falsify answers influenced by the "halo effect". Clayson & Haley (2011) found that up to 31% of their respondents admitted to recording false information on evaluation forms; in comparison, only 19.4% claimed to falsify written comments. In conclusion, data collected from scale measures alone is not sufficient to comprehensively evaluate teachers (Beran et al., 2007), and student ratings used for summative purposes may not be optimal for helping teachers improve. Researchers even caution that the results of such measures do not necessarily improve faculty teaching performance without other forms of supportive feedback, such as commentary text.

Different from the feedback provided on highly structured rating scales, written comments represent an often-overlooked method for identifying a broad range of variables—including strengths and weaknesses—in teaching, as open-ended items allow students to express themselves freely with their own words and to focus on what they perceive as most important (Stupans et al., 2016). Besides, this format allows students to offer suggestions and opinions about various teaching aspects not covered by quantitative metrics (Hammond et al., 2003; Hodges & Stanton, 2007), and this can make up for any deficiency in the rating scales (Alhija & Fresko, 2009). However, free-form text comments obtained from open-ended questions are only infrequently analyzed comprehensively, as the manual analysis of written feedback from large classes is extremely tedious and time consuming. The result of this is a loss of rich insight into student perceptions of their teachers' teaching (Brockx et al., 2012; Rajput et al., 2016). Thus, there is a need to automate this process to analyze students' feedback to obtain the desired outcome more conveniently.

2.2 Automatic scoring of teaching evaluations via student short reviews based on sentiment analysis

Fortunately, with the rapid development of natural language processing (NLP) —an area of research and application that explores how computational techniques can be employed to learn and understand human language content — it becomes feasible to automatically mine valuable information from student text feedback (Chong et al., 2020). There are many subfields in NLP such as text summarization, question answering and sentiment analysis, and each subfield has their own individual progress, development and applications. This work will focus mostly on sentiment analysis. Sentiment analysis (SA) is a common textual data quantification method, which aims at identifying feeling tendencies (i.e., attitudes, emotions and opinions) by certain people through the analysis of given subjective text (Bing, 2012). In short, it is the process of analyzing, processing, inducing and inferring the emotional tones of subjective text. The most important step of this analysis is the classification of the polarity of the text as either positive or negative (Medhat et al., 2014). For example, a sentence such as "she is a kind-hearted teacher" is classified as positive due to the presence of the positive affect word "kind-hearted", while "she is always impatient with students" is classified as negative due to the presence of negative affect word "impatient". SA has been applied to a wide range of practical problems (Srinvas & Hanumanthappa, 2017), including public opinion monitoring, consumer preference analysis and so on. In a similar fashion, the present work aims to identify the polarity of a student's feedback. However, most prior research has focused on determining the polarity on document level (Rajput et al., 2016; Tseng et al., 2018) or sentence level (Shaikh & Doudpotta, 2019; Sindhu et al., 2019). Document-level sentiment analysis regards the whole document as a basic information unit, whereas sentence-level sentiment analysis aims to determine whether a particular sentence expresses a positive or negative opinion (Chauhan et al., 2018, pp. 260). The reality is that, within free-form text, students may describe various aspects of a teacher like teaching method and technical knowledge. Though this classification of texts at the document or sentence level is useful for the educational system, it does not provide the in-depth mining of valuable information and details needed to improve the teaching/learning process (Sindhu et al., 2019). Separate consideration for the different aspects of student feedback provides better sentiment analysis. For instance, the comment "She is a good teacher with a gentle personality, but her teaching method is boring" contains two opposing sentiments for two different aspects of the teacher — teacher quality and teaching method. Therefore, the present study used aspect-level SA to identify opinion orientation with respect to a particular teaching aspect in the overall text, combined with context information to indicate specific teaching areas in need of improvement.

There are two main methods for traditional sentiment analysis: lexicon based and machine learning based (Li et al., 2020). The main idea of the lexicon based method is to construct a sentiment dictionary through a large corpus and match the opinion words in the dictionary with data to determine polarity, which usually requires the combination of a "sentiment dictionary+manual judgment" (Aung & Myo, 2017; Hong & Li, 2019). This approach is very popular, but one of its main drawbacks is that the SA results rely heavily on the sentiment dictionary. In recent years, with the explosion of network data and the continuous updating of language expressions, emotional vocabularies within a specific domain or corpora have been expected to achieve good predictive results. The creation of these vocabularies involves a significant amount of manual tagging, which can impact both cost and accuracy (Lin et al., 2019). Accordingly, this method has been unable to solve problems associated with a large number of unknown words and complex ambiguous words. However, the approach has the advantage of high accuracy for small amounts of text, generating results with a certain authority. For example, Rajput et al. (2016) calculated sentiment scores to classify feedback as positive, negative or neutral by constructing a sentiment dictionary; the proposed approach achieved an accuracy of 91.2%. Thus, for the present research, we considered using this approach in combination with other methods.

Based on our consideration of the abovementioned problems, we proposed the method of deep learning. Deep learning is a powerful machine learning algorithm due to its feature of self-learning (Li et al., 2020; Tenzin et al., 2020). Generally, deep learning models are based on neural network. Currently, many networks for deep learning have been developed like convolutional neural network (CNN), recurrent neural network (RNN), and so on. Especially, in the field of natural language processing, long short-term memory (LSTM), a representative method of RNN, is the most common network module to capture the sentence semantic. Meanwhile, the existing studies have shown that the effect of deep learning is more obvious than that of traditional methods (Zhang et al., 2019). For example, Tseng et al., (2018) used classifiers to judge and classify the emotional information of a set of college students' teaching evaluations. The results showed that the attention LSTM classifier—with a positive sentiment recognition rate of up to 97% and a negative sentiment recognition rate of up to 87%—was the best text emotion classifier. Although deep learning has strong discrimination and self-learning ability, the emotional evaluation can easily deviate due to the diversity and complexity of language texts, especially within the Chinese language (Li et al., 2020). Therefore, in order to overcome these shortcomings, the present study conducted SA based on deep learning, using the artificially screened dictionary as an input to improve prediction.

3 The present research

Following the wide application of the student evaluation of teaching in the education system, handling the qualitative opinions of students efficiently while automatic report generation emerges as a challenging task that needs to be solved urgently around the world. However, there is little research has been conducted on how to more effectively score students' teaching evaluation texts automatically, especially in China. To extract information effectively, the present study aimed at using SA to conduct a fine-grained analysis. A topic dictionary and sentiment dictionary were constructed according to the textual comments, and then introduced to the deep learning model as valuable reference sets for prediction. Further, the indices of different models were compared. Finally, the prediction results were converted into two indicators—favor-

Table 1 Demographic background of study participants

	n	%	*M*	*SD*
Students' characteristics				
Age			13.36	0.93
Gender				
Boy	2,511	56.01		
Girl	1,972	43.99		
Teachers' characteristics				
Age			28.44	2.17
Gender				
Male	12	12.12		
Female	87	87.88		
School subjects				
Mathematics	33	33.33		
English	34	34.34		
Chinese	32	32.33		
Education level				
Junior high school education	1	1.01		
Senior high school education	0	0		
Secondary specialized school education	0	0		
Associate's degree	0	0		
Bachelor degree	51	51.52		
Master degree or above	47	47.47		
Professional titles				
No title	0	0		
The third-class	2	2.02		
The second-level	68	68.69		
The first-level	29	29.29		
The senior level	0	0		
Homeroom teacher				
Yes	57	57.58		
No	42	42.42		

Notes: The teacher's professional title representing teacher's professional ability is a kind of academic and technical level classified into four grades able rating and average score—to realize the automatic scoring of written teaching evaluation comments. The resulting scores should significantly facilitate schools and regional teaching managers in understanding the performance of their teachers from the student perspective, and improving their teachers' education and skills.

4 Methods

4.1 Participants

The data came from the teaching evaluation project of a Chinese Education Bureau, which included the written comments of 4,483 junior school students (grades 7–9) about 99 teachers. The demographic background characteristics are shown in Table 1.

4.2 Measures

4.2.1 Open-ended item to evaluate teaching

Students were asked to answer a single open-ended question: "Some students think he/she is a good teacher, but some students don't think so. What do you think of him/her? Please elaborate on his/her performance in various aspects." Due to teachers usually enjoy high authority in the education system, especially in the Chinese cultural background, students may feel great pressure when evaluating their teachers and may be reluctant to report their true attitudes and views. In order to eliminate the possible sensitivity of the question, students were presented with two completely opposite sentiment orientations before asking their own views, so as to make students clear that any answer is acceptable. In addition, in the process of testing, the principle of anonymity was emphasized to students, which can also improve the authenticity of students' answers.

4.2.2 Teaching evaluation rating scale for students

Students subjectively evaluated their teacher using a 59-item *questionnaire.* Questionnaire items were designed to elicit students' ratings of their teachers across five dimensions: morality, teaching content, teaching attitude, teaching ability and teaching effectiveness. Each item was answered on a Likert 5-point scale ranging from 1 ("very inconsistent") to 5 ("very consistent"). Higher scores corresponded to higher student evaluations. Cronbach's alphas for each dimension were 0.81, 0.75, 0.89, 0.88, and 0.94, respectively, indicating high reliability and good validity (Table 2).

Table 2 Structural validity test of the teaching evaluation questionnaire

GFI	AGFI	RMSEA	CFI	NFI	TLI	IFI
0.87	0.84	0.065	0.87	0.83	0.85	0.87

4.3 Data preprocessing

First, 127 short reviews were deleted due to incompletion, content that was entirely composed of punctuation marks or content that was obviously quoted and/or meaningless. Following this, pretreatment was carried out to (a) manually filter out “stop words,” such as some modal particles and function words; (b) replace and delete typos/words; and (c) exclude data written in English. As a result of this process, 4,483 valid written comments were obtained.

4.4 Building the topic and sentiment dictionaries

The topic and sentiment dictionaries were constructed using manual screening. First, for the Chinese language, word segmentation processing had to be conducted before the text could be used. For this purpose, the commonly used Chinese word segmentation tool Jieba (Sun, 2012) was utilized to remove repeat words and extract the usable set of nouns, verbs and adjectives. Second, in the student evaluation texts, most of the topic words consisted of verbs and nouns (e.g., “class” [上课], “correction” [批改], “effect” [效果]), while the emotional words consisted of mainly adjectives (e.g., “gentle” [温柔], “novel” [新颖]). Thus, the topic words were selected from the noun and verb sets, the emotion words were selected from the adjective set and words expressing both topic and emotion were selected from all word sets. For example, although “junzi” [君子] is a noun, it also expresses positive feelings towards the teacher; likewise, although “punish” [体罚] is a verb, it also expresses negative feelings towards the teacher. In addition, although the implicit topic keywords could not directly reflect the teaching evaluation topic in the way that explicit topic keywords could, they were also included in the topic dictionary. For example, although there is no explicit topic keyword “teaching attitude” [教学态度] in the sentence “The teacher is partial to good students,” “partial” [偏心] actually expresses a negative evaluation of the teacher’s attitude. Finally, after reading the short teaching evaluations and correcting keywords that were not classified accurately, topic and emotional keywords were added directly to the dictionaries.

4.5 Coding of the teaching evaluation texts

NVivo 11 was used to code the teaching evaluation texts. Referring to Wang’s (2018) nine dimensions of teaching evaluation text analysis and the topic dictionary constructed in this study, annotation nodes were determined as follows: teacher quality, teacher image, teaching method, teaching content, teaching ability, teaching attitude, teaching effect, teacher-student relationship and classroom atmosphere. Specifically, teacher quality referred to teachers’ stable personal characteristics, including personality, temper and morality; teacher image referred to teachers’ physical characteristics, including dress and posture; teaching method referred to teachers’ methods of instruction; teaching content referred to teachers’ subject knowledge or life experience that was passed on to students; teaching attitude referred to teachers’ attitudes towards teaching and students; teaching effectiveness referred to students’ achievement, interest and self-confidence as a result of the teaching; teaching ability referred to teachers' knowledge level and classroom management; teacher-student relationship referred to the relationship between teachers and students formed in the process of education, as well as their attitudes towards one another; and classroom atmosphere referred to the overall attitudes and emotions within the teaching environment.

Then, a total of 18 tags—combining the abovementioned nine aspects with two kinds of emotional tendencies (positive and negative)—were used to code the teaching evaluation texts, whereby "-1" represented text with no emotional tendency, "0" represented text with a negative emotional tendency, "1" represented text with a positive emotional tendency and "2" represented text with both positive and negative emotional tendencies. Finally, the aspect-level of the teaching evaluation texts was labeled. All coding work was independently completed by three team members who subsequently met to discuss discrepancies. Together, they coded 80 copies, discussing the results and reaching a unified coding rule. After completing all the coding work, the team members checked the coding results again to ensure the consistency of the coding.

4.6 Aspect-level SA model

First, 4,483 written comments were randomly disrupted. To test the criterion-related validity of the machine prediction on the test set, the training set and the test set were divided according to teachers (as all students' teaching evaluation texts in the test set corresponded to approved teachers). Finally, 8 students were randomly selected from the three subjects of Chinese, mathematics and English, and the teaching evaluation texts for 24 teachers (1,042) were used as the test set and those for the remaining 75 teachers were used as the training set. The training set was used for model training and the test set was used for verification.

4.6.1 Problem formalization

As abovementioned, the teaching evaluation via text can be formalized as a task of aspect-level sentiment analysis. In this task, the basic input is a sentence and an aspect word. Aspect-level SA model will mine the interaction between the both to predict the sentiment polarity on the given aspect. For example, a sentence "… Also, he welcomes students to ask him questions after class. He will answer them enthusiastically and won't be bored." […而且, 他也欢迎同学们下课去问他题目, 他都会很热情的解答, 不会感到厌烦。] and an aspect word "teaching attitude" [教学态度] are the original input. The output will be 1, that is, positive. Generally, when building a deep learning model in the field of natural language processing, textual input data will be regarded as a word sequence. In this task, the sentence is expressed as$X = [x_1, x_2, \ldots, x_n]$ and the aspect was represented as $A = [a_1, a_2, \ldots, a_m]$. The key issue is to establish a model $f(X, A)$, which outputs the probability that the sentiment polarity is positive. Especially, we take sentiment dictionary as priori, which consists of several words, denoted as $AW = [aw_1, aw_2, \ldots, aw_l]$. Further, the task is to establish a new model $g(X, A, AW)$ which further improves the performance on teaching evaluation.

4.6.2 Presentation of the proposed models

Three models were applied to the aspect-level SA of the teaching evaluation texts. Model 1 and Model 2 are following $f\left(X,A\right)$, and Model 2 is following $g\left(X,A,AW\right)$. Specifically, Model 1 used aspect-based Bi-LSTM SA, including a word embedding layer, an average pooling layer, a Bi-LSTM layer, and an output layer (see Fig. 1, red box). Model 2 used aspect-level SA of the self-attention mechanism, including a word embedding layer, an average pooling layer, a Bi-LSTM layer, an attention layer and an output layer (see Fig. 1, yellow box). Model 3 used aspect-level SA based on the topic dictionary, including a word embedding layer, an average pooling layer, a Bi-LSTM layer, an attention layer and an output layer (see Fig. 1, blue box). Model 3 differed from Model 2 in the realization of an attention layer. Model 2 constructed the attention mechanism based on aspect input, while Model 3 constructed the attention mechanism based on aspect dictionary input. Commonly used indicators of model classification effectiveness include precision, recall and F1 value. The value range of the above indicators is 0 ~ 1. The closer to 1, the better the precision effect of the model. However, it is difficult to achieve this ideal value in practice. Generally, each situation must be judged independently, according to research needs.

(1) Model input.

The input for Model 1 and Model 2 consisted of text and aspect. Input text was expressed as$X=\left[{x}_{1},{x}_{2},...,{x}_{n}\right]$, with n representing the length of the input text and ${x}_{i}$representing the i th word in the text. Input aspect was represented as $A=\left[{a}_{1},{a}_{2},...,{a}_{m}\right]$, with m representing the length of the input aspect text and ${a}_{i}$ representing the i th word in the aspect text. In addition, the dictionary correspond-

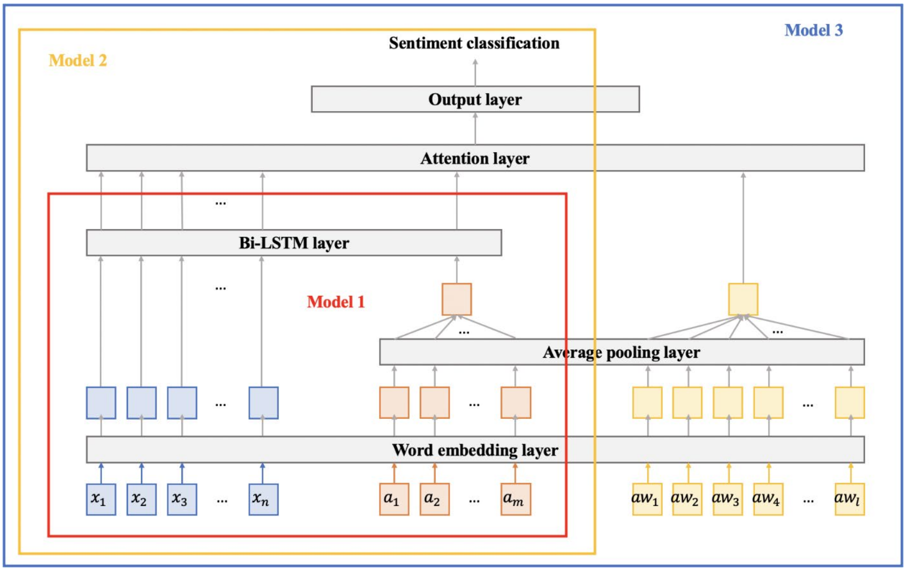


Fig. 1 Overall flow chart of the SA model

ing to aspect was added in Model 3, expressed as $AW = [aw_1, aw_2, \ldots, aw_l]$, with l representing the number of words in the dictionary and aw_i representing the i th word in the dictionary. Concretely, the word sequence is obtained by Jieba segmentation. When the sequence was input into the model, each word is transformed into a unique ID and further corresponds to a one-hot vector.

(2) Word embedding layer.

The word embedding layer encoded each word in the model input as a distributional vector of specified length k, with the aim of modelling word-level semantic information. Each word has a unique deep representation in word embedding layer. A unified word embedding layer was constructed for the text input, aspect input and dictionary input corresponding to aspect, recorded as $E(\bullet)$. Through the word embedding layer, input$X$, A and AW were coded as $E\left(X\right) \in \mathbb{R}^{n\times k}$、$E\left(A\right) \in \mathbb{R}^{m\times k}$ and $E\left(AW\right) \in \mathbb{R}^{l\times k}$, respectively.

(3) Average pooling layer.

The average pooling layer was denoted as$P(\bullet)$, representing the input aspect text or aspect dictionary as a vector. The layer is to capture one whole semantic for the input content. Specifically, the embedded aspect input was represented as$E\left(A\right) \in \mathbb{R}^{m\times k}$. After pooling, it was represented as $P\left(A\right) = \frac{1}{m}\left[E\left(a_1\right) + E\left(a_2\right) + \cdots + E\left(a_m\right)\right] \in \mathbb{R}^{1\times k}$. The embedded aspect dictionary input was represented as $P\left(AW\right) = \frac{1}{l}[E\left(aw_1\right) + E\left(aw_2\right) + \cdots + E(aw_l\,)] \in \mathbb{R}^{1\times k}$ 。

(4) Bi-LSTM layer.

The Bi-LSTM layer acted on the connection matrix of text representation$E\left(X\right)$ and aspect representation $P\left(A\right)$, aimed at capturing the semantic information of the entire sentence and the interaction between X and A, recorded as $BiLSTM(X, A) \in \mathbb{R}^{(n+1)\times k}$. In Model 1, the $BiLSTM(X, A)_{n+1} \in \mathbb{R}^{1\times k}$, representing the sentence, was directly connected to the output layer to obtain the final prediction result. In Models 2 and 3, the matrix $BiLSTM(X, A)$ was input to the attention layer to further optimize the sentence representation.

(5) Attention layer.

For different aspect inputs, different words in the input text contributed differently to sentence semantics. The attention layer was designed to model words' contribution weight to optimize the sentence-level semantic representation. The aspect obtained by the average pooling layer was represented as the query vector in the attention mechanism, and the contribution weight of each word was expressed as follows:

$$\alpha_i = \frac{\exp\left(E(\,x_i) \bullet P(A)^T\right)}{\sum_j \exp\left(E(\,x_j) \bullet P(A)^T\right)} or \frac{\exp\left(E(\,x_i) \bullet P(AW)^T\right)}{\sum_j \exp\left(E(\,x_j) \bullet P(AW)^T\right)}$$

Model 2 used$P\left(A\right)$ as aspect representation, and Model 3 used$P\left(AW\right)$as aspect representation. Then, the text sentences were expressed as follows:

$$S\left(X, A\right) = \sum_{i=1}^{n} \alpha_i E\left(x_i\right) \in \mathbb{R}^{1\times k}$$

(6) Output layer.

Based on the sentence representation $S(X, A)$ obtained from the Bi-LSTM or attention layer, the output layer obtained the final probability distribution of sentiment classification through a layer of linear transformation and softmax activation. The specific operation was as follows:

$$p = softmax\left(S(X, A) \bullet W + b\right)$$

$W \in \mathbb{R}^{k\times cn}$and $b \in \mathbb{R}^{1\times cn}$ was the trainable parameter and cn was the number of predefined emotion categories.

4.6.3 Hyper-parameters of the models

To ensure a uniform size of the vector matrix of input words in the network, the maximum length n of text input was set to 300. The dimension k of the word embedding layer was set to 300 and the hidden layer dimension of the Bi-LSTM layer was set to 128. The learning rate of the model was set to 0.001, and RMSprop was selected by the optimizer. The optimization goal was to minimize the cross-entropy loss function, and the batch size was 128.

4.7 Automatic scoring of teaching evaluation texts

According to the automatic output results, student IDs could be matched with teacher IDs, and the following indicators could be calculated to evaluate teachers.

(1) Praise rate referred to the proportion of students with a positive emotional tendency to the total number of students in the class with respect to a certain aspect (i.e. the proportion of output results of “1”); rate of poor evaluation referred to the proportion of negative evaluations in the class (with reference to output results of “0”); non-mention rate referred to the proportion of evaluations marked as “-1”; and the contradiction rate referred to the proportion of output results labeled “2.”

(2) Average score. The output results of the sentiment classification were recoded and scored with “-1” as 0, “2” as 0, “0” as – 1, and “1” as 1. The average score of all students in a class with respect to the nine aspects could be calculated to obtain the teacher’s score on each aspect. When the average score was less than 0, the students considered the teacher inadequate in this aspect; when the average score was more than 0, the students considered the teacher good in this aspect; and when the average value was 0, the students had no obvious emotional reaction to the teacher in this aspect.

5 Result

5.1 Results of the topic and sentiment dictionary construction

A total of 4,053 nouns, verbs and adjectives were obtained through the process of text segmentation. The constructed topic dictionary had 646 words, including 365 explicit keywords and 291 implicit keywords. By induction, these were divided into the following nine themes: teacher quality, teacher image, teaching method, teaching content, teaching attitude, teaching effectiveness, teacher ability, classroom atmosphere and the teacher–student relationship. Ultimately, there were 1,147 words in the sentiment dictionary, including 779 positive words and 368 negative words.

5.2 Evaluation and comparison of the results

The results of the three models are shown in Table 3; Fig. 2. The prediction effect of Model 3 was obviously superior to that of the other two models on each evaluation index.

5.3 Relationship between machine prediction, manual coding sentiment classification and the teaching evaluation rating scale

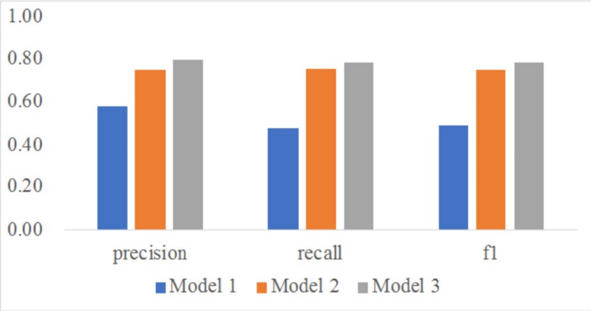
5.3.1 Correlation analysis among machine prediction, manual coding and the criterion questionnaire

Total scores for the machine prediction, manual coding and *student questionnaire* were calculated, followed by a Pearson correlation analysis. The results showed that there was a significant correlation between the total scores predicted by the machine and the questionnaire ($r = 0.17$, $p < 0.001$); a significant correlation between the total scores predicted by manual coding and the questionnaire ($r = 0.18$, $p < 0.001$); and a significant correlation between the results predicted by the machine and manual coding ($r = 0.73$, $p < 0.001$). Although the correlation coefficients between the machine prediction, manual coding and questionnaire predictions were relatively low, they were close to the correlation coefficient of the teaching evaluation questionnaire and

Table 3 Comparison of model results

Model	Precision	Recall	F1
Model 1	0.58	0.48	0.49
Model 2	0.75	0.76	0.75
Model 3	0.80	0.79	0.79

Fig. 2 Comparison of model results



there was a high correlation between the machine and manual coding predictions, indicating that the machine prediction results were consistent with the manual coding results.

5.3.2 Examination of the differences between machine prediction and manual coding on the criterion questionnaire

To test for differences between the four classification results of the aspect-level sentiment prediction model and manual coding on the criterion questionnaire, SPSS 22.0 was used to analyze the model results using a one-way ANOVA. For this purpose, 1,042 test sets were used for machine prediction and 4,483 evaluation texts were used for manual coding. Thus, the questionnaire data corresponded to the machine prediction and manual coding data, respectively.

With respect to the nine aspects examined in the teaching evaluation texts, the five aspects of teacher quality, teaching content, teaching attitude, teaching ability and teaching effectiveness were similar to or the same as the dimensions examined by the criterion questionnaire. To test the actual effect of classification, the four classification results (i.e., not mentioned, negative, positive, both positive and negative) of the machine prediction and manual coding on the above-mentioned five aspects of teaching were analyzed according to variance from the questionnaire data.

6 Discussion

6.1 Prediction effect of the aspect-level SA model

The present study attempted to establish an aspect-level sentiment prediction model for teaching evaluation texts. The model used Bi-LSTM, Bi-LSTM and attention methods, and tried to add a topic dictionary as input to improve the prediction effect of the LSTM model. By training and testing three models on the same training set and test set, the results showed that the prediction effect of Model 3 (with the dictionary) was significantly superior to that of Model 1 and Model 2 (with no topic dictionary). This indicates that the topic dictionary constructed in this study had good validity and can be used to train models. Further comparisons between Models 2 and 3 showed that Model 3 had the optimal effect, with a precision rate of 0.80, a recall rate of 0.79 and an F1 of 0.79. Model 3 used the attention mechanism to calculate the attention of each text, by adding the topic dictionary. This made more effective use of the information provided by the topic dictionary. Model 2 used the self-attention mechanism to evaluate word dependence within the teaching evaluation texts and to capture the internal structure of the texts. Therefore, the prediction effect of Model 3 was better than that of Model 2. As expected, the results of this study show that the aspect-level SA for the education domain can successfully be performed by introducing deep learning and lexicon based method at the same time. This just confirms that even though neural networks seem to be taking over most of the NLP tasks, some of the more recent aspect-level SA studies (Bhatnagar et al., 2018; Gupta et al., 2019) show that approaches based on dictionaries can still be very effective. As the result,

Table 4 Descriptive statistical analysis and ANOVA results

	Teacher quality		Teaching attitude		Teaching content		Teaching ability		Teaching effectiveness	
	Machine prediction	Manual coding	Machine prediction	Manual coding	Machine prediction	Manual coding	Machine prediction	Manual coding	Machine prediction	Manual coding
	M±*SD*	*M*±*SD*	*M*±*SD*	*M*±*SD*	*M*±*SD*	*M*±*SD*	*M*±*SD*	*M*±*SD*	*M*±*SD*	*M*±*SD*
Not mentioned	4.451±0.541	4.439±0.572	4.478±0.599	4.489±0.612	4.076±0.582	4.072±0.586	4.444±0.557	4.330±0.545	4.369±0.685	4.385±0.664
Negative	4.118±0.767	3.936±0.750	3.808±0.599	3.380±0.937	3.481±0.554	3.520±0.600	2.833±0.997	3.644±0.641	3.750±0.775	3.690±0.843
Positive	4.518±0.507	4.534±0.500	4.543±0.523	4.579±0.503	4.170±0.571	4.161±0.550	4.448±0.477	4.474±0.488	4.493±0.563	4.525±0.567
Both positive and negative	4.311±0.610	4.039±0.553	4.385±0.613	3.851±0.741	4.212±0.515	3.942±0.550	4.266±0.569	4.117±0.664	4.400±0.586	4.025±0.560
F	16.770**	37.826***	20.229**	73.483***	17.424***	35.340***	13.403***	60.908***	15.646**	51.130***

Legend: $^{***}p<0.001$, $^{**}p<0.01$, $^{*}p<0.05$

the automation of the student opinion mining process will reduce the time and effort needed from human curators. By closely monitoring the aspects that the students are satisfied (or unsatisfied) with, teachers can tailor their performance timely in such a way that the level of service provided to the students is constantly improving.

6.2 Validity analysis of the aspect-level SA

In this study, *students' subjective evaluations of teaching* (i.e., the questionnaires) were used to test the validity of the results of the prediction models. By recoding the classification results to generate a total score, and performing a Pearson product-moment correlation analysis with the questionnaire total score, the results showed a significant but weak ($r=0.17$) correlation. A possible reason for this is that the teaching evaluation texts were obtained from student evaluations of teachers' typical behaviors and characteristics, based on students' actual feelings; the questionnaire, in contrast, was developed by experts according to the general dimensions of teaching evaluation, leading to potential differences in content. However, the teaching aspects referenced in the teaching evaluation texts extended beyond the five aspects examined in the questionnaire. Specifically, the students also conveyed that teacher image, classroom atmosphere, and the teacher-student relationship were also of great value to overall teaching performance. This showed that, although the correlation coefficient between the teaching evaluation texts and the questionnaire was small, the value of the teaching evaluation texts could not be denied, as these reflected students' personal feelings and priorities. Thus, teaching evaluation texts may more accurately reflect students' overall teaching evaluation than questionnaires.

In addition, the questionnaire examined five aspects: teacher morality, teaching content, teaching attitude, teaching ability and teaching effectiveness. These represented only five of the nine teaching evaluation aspects summarized in this study. Therefore, to investigate the effect of manual coding and machine prediction on sentiment classification, a one-way ANOVA was conducted. As shown in Table 4, the differences between the four sentiment tendencies (i.e., not mentioned, positive, negative, both positive and negative) were significant ($p<0.001$) with respect to these five aspects, as a result of either manual coding or Model 3. Accordingly, it can be concluded that students with different types of sentiment tendencies predicted by the teaching evaluation texts obtained different scores on the questionnaire. This may be because the students were more likely to tend toward the middle when answering the questionnaire compared with responding to open-ended items (Elhassan, 2009; Emerson & Records, 2007; Greenwald, 1997). Should this be the case, teaching evaluation texts may reveal students' more explicit and direct attitudes towards teachers. In such teaching evaluation texts, students may be more likely to report on only the teaching aspects they consider most important or prominent (Stupans et al., 2016). Thus, dimensions that may tend toward the middle on a questionnaire may not be significantly reflected in a teaching evaluation text.

6.3 Limitations and future work

Although the present study demonstrated reasonable performance in the task of sentiment orientation detection (see Table 3; Fig. 2), there were still some shortcomings of the research. First, the correlation between the model classification results and the criterion questionnaire was not high; thus, in a follow-up study, we may divide the prediction model into nine aspects, adding the explicit and implicit keywords for each aspect, to improve the prediction results. Furthermore, in future research, sentiment prediction may be improved by the addition of a sentiment dictionary to the model.

Second, this study conducted SA from a paragraph level. Although one-sided information extraction from a single sentence can be avoided at the paragraph level, the emotional coding may not perfectly map back on to each sentence in the paragraph; thus, the model prediction effect may be inferior at the paragraph level, due to its complexity. Future research may focus on SA at the sentence level, in order to reduce the interference of irrelevant information and improve the classification effect.

Finally, the output of the prediction model was students' personal evaluations. In future research, we may try to use this model to directly predict the overall result of a teacher's evaluations at the level of the whole class. There is a many-to-one relationship between students and teachers, with students in the same class influenced by the same teacher. Thus, there should be a correlation between one student's evaluation of a teacher and his/her classmate's evaluation of the same teacher. When building the prediction model, we may try to obtain the common evaluation of each teacher from all students by iterating each student's evaluation, step by step.

Funding This paper was supported by the National Natural Science Foundation of China (Grant No: U1911201).

Data availability Available on request.

Declarations

Conflicts of interest The authors declare that there is no conflict of interest.

Ethics approval The participants were protected by hiding their personal information during the research process. They knew that their participation was voluntary and they could withdraw from the study at any time.

References

Alhija, F. N. A., & Fresko, B. (2009). Student evaluation of instruction: What can be learned from students' written comments? *Studies in Educational Evaluation*, *35*(1), 37–44. https://doi.org/10.1016/j.stueduc.2009.01.002

Annan, S. L., Tratnack, S., Rubenstein, C., Metzler-Sawin, E., & Hulton, L. (2013). An integrative review of student evaluations of teaching: Implications for evaluation of nursing faculty. *Journal of Professional Nursing*, *29*(5), e10–e24. https://doi.org/10.1016/j.profnurs.2013.06.004

Aung, K. Z., & Myo, N. N. (2017). Sentiment analysis of students' comment using lexicon based approach. In *2017 IEEE/ACIS 16th International Conference on Computer and Information Science* (pp. 149–154)

Beran, T. N., & Rokosh, J. L. (2009). Instructors' perspectives on the utility of student ratings of instruction. *Instructional Science*, *37*(2), 171–184. https://doi.org/10.1007/s11251-007-9045-2

Beran, T., Violato, C., & Kline, D. (2007). What's the "use" of student ratings of instruction for administrators? One university's experience. *Canadian Journal of Higher Education*, *37*(1), 27–43

Bhatnagar, V., Goyal, M., & Hussain, M. A. (2018). A novel aspect based framework for tourism sector with improvised aspect and opinion mining algorithm. *International Journal of Rough Sets and Data Analysis*, *5*(2), 119–130. https://doi.org/10.4018/ijrsda.2018040106

Bing, L. (2012). *Sentiment Analysis and Opinion Mining (Synthesis Lectures on Human Language Technologies)*. Morgan & Claypool Publishers

Brockx, B., Van Roy, K., & Mortelmans, D. (2012). The student as a commentator: Students' comments in student evaluations of teaching. *Procedia-Social and Behavioral Sciences*, *69*, 1122–1133. https://doi.org/10.1016/j.sbspro.2012.12.042

Chauhan, G. S., Agrawal, P., & Meena, Y. K. (2018). Aspect-Based Sentiment Analysis of Students' Feedback to Improve Teaching–Learning Process. *Smart Innovation, Systems and Technologies, 259–266.*https://doi.org/10.1007/978-981-13-1747-7_25

Chong, C., Sheikh, U. U., Samah, N. A., & Ahmad Zuri Sha'ameri. (2020). &. Analysis on Reflective Writing Using Natural Language Processing and Sentiment Analysis. *IOP Conference Series.Materials Science and Engineering, 884*(1), 1–8. https://doi.org/10.1088/1757-899X/884/1/012069

Clayson, D. E., & Haley, D. A. (2011). Are students telling us the truth? A critical look at the student evaluation of teaching. *Marketing Education Review*, *21*(2), 101–112. https://doi.org/10.2753/mer1052-8008210201

Denson, N., Loveday, T., & Dalton, H. (2010). Student evaluation of courses: What predicts satisfaction? *Higher Education Research & Development*, *29*, 339–356. https://doi.org/10.1080/07294360903394466

Donnon, T., Delver, H., & Beran, T. (2010). Student and teaching characteristics related to ratings of instruction in medical sciences graduate programs. *Medical Teacher*, *32*(4), 327–332. https://doi.org/10.3109/01421590903480097

Elhassan, K. (2009). Investigating substantive and consequential validity of student ratings of instruction. *Higher Education Research & Development*, *28*(3), 319–333. https://doi.org/10.1080/07294360902839917

Emerson, R. J., & Records, K. (2007). Design and testing of classroom and clinical teaching evaluation tools for nursing education. *International Journal of Nursing Education Scholarship (IJNES)*, *4*(1), 16. https://doi.org/10.2202/1548-923x.1375

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist*, *52*(11), 1182–1186. https://doi.org/10.1037/0003-066x.52.11.1182

Gupta, V., Singh, V. K., Mukhija, P., & Ghose, U. (2019). Aspect-based sentiment analysis of mobile reviews. *Journal of Intelligent and Fuzzy Systems*, *36*(5), 4721–4730. https://doi.org/10.3233/JIFS-179021

Hammond, I., Taylor, J., & McMenamin, P. (2003). Value of a structured participant evaluation questionnaire in the development of a surgical education program. *Australian and New Zealand Journal of Obstetrics and Gynaecology*, *43*(2), 115–118. https://doi.org/10.1046/j.0004-8666.2003.00037.x

Hodges, L. C., & Stanton, K. (2007). Translating comments on student evaluations into the language of learning. *Innovative Higher Education*, *31*(5), 279–286. https://doi.org/10.1007/s10755-006-9027-3

Hong, W., & Li, M. (2019). A review: Text sentiment analysis methods. *Computer Engineering & Science*, *41*(4), 750–757

Hoon, A., Oliver, E., Szpakowska, K., & Newton, P. (2014). Use of the 'stop, start, continue' method is associated with the production of constructive qualitative feedback by students in higher education. *Assessment & Evaluation in Higher Education*, 755–767. https://doi.org/10.1080/02602938.2014.956282

Kulik, J. A. (2001). Student ratings: Validity, utility, and controversy. *New Directions for Institutional Research*, *2001*(109), 9–25. https://doi.org/10.1002/ir.1

Lin, Q., Zhu, Y., Zhang, S., Shi, P., Guo, Q., & Niu, Z. (2019). Lexical based automated teaching evaluation via students' short reviews. *Computer Applications in Engineering Education*, *27*(1), 194–205. https://doi.org/10.1002/cae.22068

Li, W., Jin, B., & Quan, Y. (2020). Review of research on text sentiment analysis based on deep learning. *Open Access Library Journal*, *7*, 1–8. https://doi.org/10.4236/oalib.1106174

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Nasser, F., & Fresko, B. (2002). Faculty views of student evaluation of college teaching. *Assessment & Evaluation in Higher Education*, *27*(2), 187–198. https://doi.org/10.1080/02602930220128751

Onwuegbuzie, A. J., Daniel, L. G., & Collins, K. M. (2009). A meta-validation model for assessing the score-validity of student teaching evaluations. *Quality & Quantity*, *43*(2), 197–209. https://doi.org/10.1007/s11135-007-9112-4

Ory, J. C. (2000). Teaching evaluation: Past, present, and future. *New Directions for Teaching and Learning*, *83*, 13–18. https://doi.org/10.1002/tl.8302

Rajput, Q., Haider, S., & Ghani, S. (2016). Lexicon-based sentiment analysis of teachers' evaluation. *Applied Computational Intelligence and Soft Computing*, 1–12. https://doi.org/10.1155/2016/2385429

Serdyukova, N., Tatum, B. C., & Serdyukova, P. (2010). Student evaluations of courses and teachers. Publication of National University,173

Shaikh, S., & Doudpotta, S. M. (2019). Aspects based opinion mining for teacher and course evaluation. *Sukkur IBA Journal of Computing and Mathematical Sciences*, *3*(1), 34–43

Sindhu, I., Daudpota, S. M., Badar, K., Bakhtyar, M., Baber, J., & Nurunnabi, M. (2019). Aspect-based opinion mining on student's feedback for faculty teaching performance evaluation. *Ieee Access : Practical Innovations, Open Solutions*, *7*, 108729–108741. https://doi.org/10.1109/ACCESS.2019.2928872

Smith, C. (2008). Building effectiveness in teaching through targeted evaluation and response: Connecting evaluation to teaching improvement in higher education. *Assessment & Evaluation in Higher Education*, *33*(5), 517–533. https://doi.org/10.1080/02602930701698942

Srinvas, A., & Hanumanthappa, M. (2017). Viable modern approaches for sentiment analysis: A survey. *International Journal of Advanced Research in Computer Science*, *8*(7), 115–120. https://doi.org/10.26483/ijarcs.v8i7.4095

Stupans, I., McGuren, T., & Babey, A. M. (2016). Student evaluation of teaching: A study exploring student rating instrument free-form text comments. *Innovative Higher Education*, *41*(1), 33–42. https://doi.org/10.1007/s10755-015-9328-5

Sun, J. (2012). Jieba Chinese word segmentation tool. (*2018-01-21)[2018-06-25].* Retrieved from https://github.com/fxsjy/jieba

Tenzin, D., Lemay, D. J., Basnet, R. B., & Bazelais, P. (2020). Predictive analytics in education: a comparison of deep learning frameworks. *Education and Information Technologies*, *25*(3), 1951–1963. https://doi.org/10.1007/s10639-019-10068-4

Tseng, C. W., Chou, J. J., & Tsai, Y. C. (2018). Text mining analysis of teaching evaluation questionnaires for the selection of outstanding teaching faculty members. *Ieee Access : Practical Innovations, Open Solutions*, *6*, 72870–72879. https://doi.org/10.1109/ACCESS.2018.2878478

Wang, H. D. (2018). *Multi-grain sentiment analysis of teaching reviews based on topic* (pp. 25–26). Guang Zhou: South China University of Technology Press

Zhang, J., Chen, F. L., & Zhang, P. Y. (2019). The role and implementation of students' sentiment analysis in curriculum teaching evaluation. *Computer Knowledge and Technology*, *15*(4), 184–188

Ping Ren is an associate professor at Beijing Normal University. Her research interests include the evaluation and promotion of adolescent mental health, regional education quality monitoring and regional education improvement based on monitoring results, etc.

Liu Yang is a doctoral student at Beijing Normal University. Her research focuses on the application of computer technology in the field of education.

Fang Luo is a professor in the school of psychology at Beijing Normal University. Her research agenda includes educational statistics and measurement, as well as the development and application of data mining technology, etc.