Received: 6 May 2018 | Accepted: 26 August 2018
DOI: 10.1002/cae.22068

RESEARCH ARTICLE

WILEY

Lexical based automated teaching evaluation via students' short reviews

Qika Lin[1] | **Yifan Zhu**[1] | **Sifan Zhang**[1] | **Pengfei Shi**[1] | **Qing Guo**[1] | **Zhendong Niu**[1,2]

[1] School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China

[2] School of Computing and Information, University of Pittsburgh, Pittsburgh, Pennsylvania

Correspondence
Zhendong Niu, School of Computer Science and Technology, Beijing Institute of Technology, 5 South Zhongguancun Street, Haidian District, Beijing, China.
Email: zniu@bit.edu.cn

Funding information
National Natural Science Foundation of China, Grant number: 61370137; Ministry of Education-China Mobile Research Foundation Project, Grant number: 2016/2-7; Postgraduate Education Research Project of Beijing Institute of Technology, Grant number: 2017JYYJG-004

Abstract

Student evaluations of teaching (SET) have become a popular approach to assess faculties' teaching. Question-score-based questionnaire is the most common SET measure adopted in universities. However, it fails to cover important facets of teaching process that not mentioned in the predefined questionnaire, which can be substantially obtained from students' short reviews. In this paper, we propose two lexical-based methods, specifically knowledge-based and machine learning-based, to automatically extract opinions from short reviews. Furthermore, the diversity of reviews' themes and styles of same sentiment polarity reviews can be observed from the extracted opinion results. The experimental results show that the proposed methods are able to achieve accuracies of 78.13 and 84.78%, respectively in the task of student review sentiment classification. Further investigation on linguistic features shows that reviews with same sentiment polarity shares similar language patterns. Finally, we present an application scenario in real SET process by utilizing aforementioned methods and discoveries.

KEYWORDS
evaluation system, knowledge-based evaluation method, machine learning, students' review, teaching evaluation

1 | INTRODUCTION

As an intuitive and direct reflect on the performance of classroom instruction, student evaluations of teaching (SET) plays a significant role for faculties' professional competence [26]. For instance, teachers get feedback from SET to improve their class lectures in future while school take SET results into faculties' work performance evaluation. With the development of social network and web service, students often try to get reference from public accessed anonymous SET sites (e.g., RateMyProfessors[1] and MySupervisor[2]) [5]. In many cases of university administrated student evaluations of teaching (UA-SET), students in universities are asked to fill out a questionnaire of the course they select at the end of each semester. The questionnaire usually consists of a series of questions about quality factors during the whole class with options or blanks. Students answer each question by filling a score into the blank or selecting an option corresponding to a specific score. Most questions come from experience of previous teaching problems or experts' suggestions. However, the "Predefined question + Score" model often has problems such as different standards in students scoring and cannot cover all aspects of the teaching process. In other words, traditional questionnaire method is not free enough for students to explain their opinions toward courses they have taken. In addition, non-English-based SET reviews like Chinese still lack of systemic investigation.

[1] https://www.ratemyprofessors.com

[2] https://www.mysupervisor.org

Therefore, using a commentary paragraph as a supplement of the questionnaire can effectively alleviate the aforementioned problems. With the development of opinion mining and sentiment analysis technology, emotion-oriented computing has become a breakthrough to the rating subject in intelligent education. There has been some research in the evaluation of students' learning status [36], student attention [19], and teacher performance [3] through sentiment analysis. Automatic calculating the emotional polarity of student assessment text to teachers makes teacher evaluation module more objective and reduce the process workload. In practical applications, the calculation of emotion often depends on the universal emotion model [1]. Furthermore, to summarize, there is currently neither a widely accepted sentiment analysis approach, nor customized practical applications on the subject of student review to the best of our knowledge.

In this paper, two approaches are constructed to calculate the emotion orientation of the students' short review texts, which are sentiment dictionary-based in this domain and machine learning-based models using lexical features. After that, we try to further analyze the potential linguistic features when different students give same sentiment reviews. We found that negative evaluations are more evidence-based to support their viewpoints than positive ones, which help to further teacher modeling researches. Compared with the previous research, the main contribution of this paper are described as follows: (1) In knowledge-based approach, we used the word co-occurrence method and word-embedding similarity method to automatically build and expand the universal sentiment dictionary. Accordingly, the accuracy of the sentiment analysis at the domain of teacher evaluation is improved; (2) A variety of machine learning models and multi-lexical features were used, and finally obtain a machine learning model with 12.11% higher accuracy than using general sentiment lexicon; and (3) The textual characteristics of student reviews are analyzed. And then we find out the differences in the content styles when students give positive or negative ratings.

The main organization of this paper comprises six sections. The remaining parts are follows: section 2 analyzes the related literature of emotion calculation and education-related text emotion computation. Section 3 describes the methods proposed in this paper to calculate the sentiment polarity. Section 4 introduces the data set we used and conduct experiments by two different methods. The main contents are the expansion of the sentiment lexicon, the sentiment calculation based on lexicon, the extraction, and selection of lexical features and the sentiment value fitting by some machine learning models. Section 5 discusses the student evaluation style and teacher evaluation system design. Finally, conclusion and outlook are proposed in section 6.

2 | RELATED WORK

This section mainly discusses the related research and development in the domain of emotion analysis and teacher evaluation in recent years. It has been 20 years since Hatzivassiloglou et al. [17] first conducted an in-depth study of the semantic orientation on adjectives in 1997. Emotional analysis field has made tremendous progress. Different from the field of English emotional analysis, many Chinese scholars have made different studies on Chinese field because of the particularity of its expression.

2.1 | Calculation of sentiment polarity

In general, there are two methods for Chinese sentiment analysis: knowledge-based approach and machine learning-based approach [28]. Knowledge-based approaches are often unsupervised learning methods. The main idea is to construct a sentiment dictionary through a large corpus. HowNet [12] and NTU Sentiment Dictionary (NTUSD) [20] are the most commonly and widely used ones in the general Chinese sentiment analysis field. As a traditional method, it has been very popular in the past, but one of its drawback is that the sentiment analysis results depend heavily on the sentiment dictionary. Thus, specific sentiment lexicon is needed to constructed to obtain high performance in a specific domain, or in specific corpora. Cao et al. [7] constructed an emotional dictionary of comments related to the field of transportation, with the best accuracy of 82.51%. Huang et al. [18] proposed a novel automatic construction strategy of domain-specific sentiment lexicon based on constrained label propagation. Similar researches also can be found in disaster management [30], sports players' evaluation [27], climate and meteorology study [16,22], etc. With the gradual and extensive development of machine learning and deep learning, some machine learning methods are also applied to the analysis of emotions in recent years. The main steps of this supervised approach are as follows [28]: extract features from the text, train, and test on dataset using a machine learning model. Many scholars have done lots of research on feature extraction and model design in the field of Chinese sentiment analysis [15,33,39,40]. Tan and Zhang [33] presents an empirical study of sentiment categorization on Chinese documents. Four feature selection methods (MI, IG, CHI, and DF) and five learning methods (centroid classifier, K-nearest neighbor, winnow classifier, Naive Bayes, and SVM) are investigated on a Chinese sentiment corpus with a size of 1,021 documents. The result showed that the IG features and SVM model provided the best performances for sentiment categorization coupled with domain or topic dependent classifiers. Zhai et al. [39] extracted features from sentiment-words, substrings, substring-groups, and key substring-groups. They found that different types of features

possess different discriminative capabilities in Chinese classification and substring-group features have greater potential to improve the performance. Zhang et al. [40] proposed a method for sentiment classification based on word2vec and SVM^{pref} and achieves encouraging performance. However, when using different algorithms of machine learning, researchers seldom consider the textual theme or emotion words as features.

2.2 | Sentiment analysis in education domain

At present, a few of the students' sentiment texts analyzing research concentrated on online education, which focused most of their attention on system design [4,31]. Adinolfi et al. [1] conducted emotional analysis on the online platform. They focused on proposing an architecture of sentiment analysis for higher education purposes and showing how it is employed to monitor student satisfaction on different online platforms. However, they only added the module of calculating emotional polarity through universal emotional dictionary, and did not explore the issue of emotional analysis, nor the accuracy or recall statistics. Aung et al. [2] constructed a sentiment lexicon with the size of 745 words manually to analyze sentiment of students' comment. Unfortunately, their experiments did not consider the complexity of construction in big data environments and similar problem can also be found in References [29,41]. Esparza et al. [13] presented a model called Social Mining using a corpus of real comments in Spanish about teacher performance assessment. They used different SVM models and acquired a relative high performance with accuracy of 81.49%. Tian et al. [36] conducted extensive experiments with a large amount of data on BBS and QQ which their results showed great significance in psychological problems prevention. However, they ignored the bias of the web platform and did not suggest a good solution to extract features from comments. To summarize, the current research on teacher evaluation based on student reviews has not been investigated enough in Chinese field yet.

3 | METHODOLOGY

In this section we show how to use sentiment lexicon and machine learning to analyze student review texts. The overall flow chart of sentiment polarity calculation in this paper is shown in Figure 1. For Chinese, before using the text, the word segmentation processing must be conducted. In this paper, commonly used Chinese word segmentation tool, jieba[3], is utilized to handle this work, and entire processes are described as follows. The text is first processed by the pre-processing and segmentation steps. Two methods are conducted to calculate the emotion value and evaluate the performance of the model.

[3]https://github.com/fxsjy/jieba

3.1 | Knowledge-based approach

To avoid duplication of effort, we surveyed and adopted the currently existing sentiment dictionaries in the Chinese field as base dictionary for sentiment analysis. Finally, HowNet and NTUSD are exploited. HowNet dictionary is a common knowledge repository based on inter-conceptual relations and inter-attribute relations related on both Chinese and English. It contains 4,370 negative terms and 4,566 positive ones [12]. The NTUSD dictionary is derived from the Chinese emotional polarity dictionary of National Taiwan University's Natural Language Processing Laboratory. The simplified Chinese version contains 8,276 negative terms and 2,810 positive terms. In the following subsections we present the method of computation, construction and expand on these two base dictionaries.

3.1.1 | Calculation of emotion polarity

The main steps to calculate the emotional polarity through the sentiment lexicon are as follows:

1. Emotional intensity calculation of words
2. Processing modified words
3. Calculation of sentence emotion value

First, we obtain the emotional word polarity degree value through the emotional dictionary. In Chinese expressions, we can observe the phenomenon that when people read a new word they do not know, they can guess the meaning of the word according to the Chinese characters that make up the word. Based on this, we can make the assumption that the emotional polarity value of a Chinese word is a function of the polarity (emotional propensity) of its all morphemes. Using the method proposed by Ku et al. [21], we can calculate the emotional polarity of each morpheme based on the different appearing frequencies in positive and negative dictionaries.

$$\text{Weight } P_{c_i} = \frac{fp_{c_i}/\sum_1^n fp_{c_i}}{fp_{c_i}/\sum_1^n fp_{c_i} + fn_{c_i}/\sum_1^m fn_{c_i}} \quad (1)$$

$$\text{Weight } N_{c_i} = \frac{fn_{c_i}/\sum_1^m fn_{c_i}}{fp_{c_i}/\sum_1^n fp_{c_i} + fn_{c_i}/\sum_1^m fn_{c_i}} \quad (2)$$

As shown in Equations (1) and (2), fp_{ci} and fn_{ci} denote the occurrence frequency of a Chinese morpheme(c_i) in each set of complimentary and derogatory terms, and n and m denote the numbers of different Chinese characters in each set, respectively. Based on the above calculated result, the

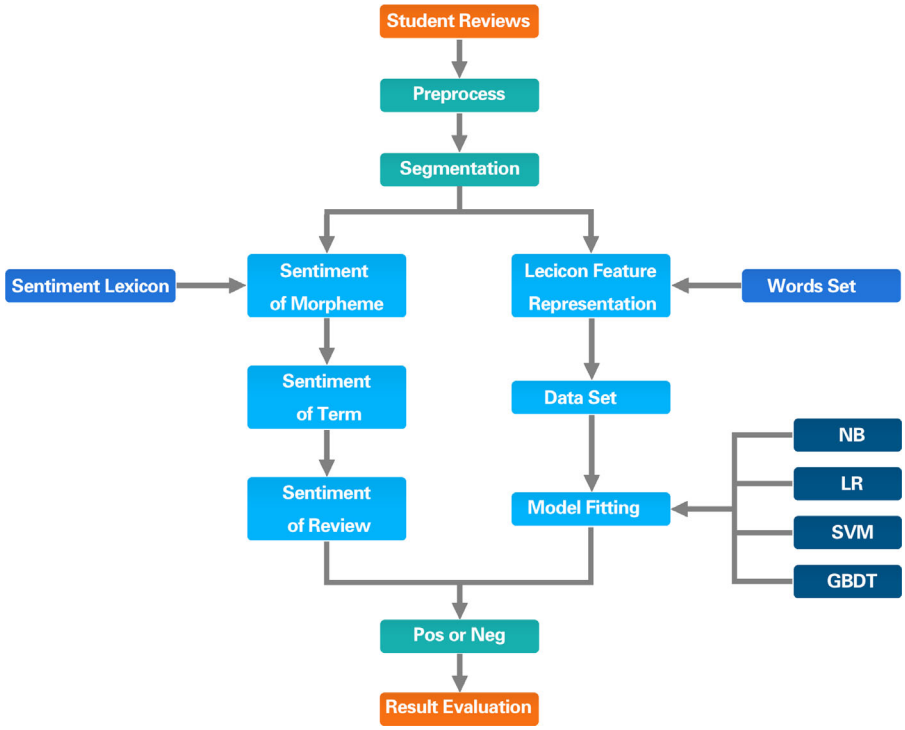


FIGURE 1 The overall process of sentiment polarity computing for text reviews in SET

emotional polarity value of each Chinese morpheme(c_i) is calculated as shown in the following Equation (3).

$$S_{c_i} = \text{Weight } P_{c_i} - \text{Weight } N_{c_i} \tag{3}$$

The value calculated in this approach can largely represent the emotional polarity of Chinese morpheme. The word is considered emotional positive when S_{ci} is mathematical positive. Otherwise, the situation is reversed, and the absolute value can indicate the intensity of emotional polarity. Assume that a Chinese term w consists m Chinese characters, that is, $c_1, c_2, ..., c_m$, so the emotional tendency S_w of w can be calculated by Equation (4).

$$S_w = \frac{1}{m}\sum_1^m S_{c_i} \tag{4}$$

In the case of Chinese usage, especially when it comes to the evaluation with emotional polarity, there are usually a large number of adverbs to modify the emotional terms. Correctly handling these adverbs is very important for understanding the expressed emotion. Referenced to the degree level word set of HowNet dictionary, we set the corresponding degree value of the following various different vocabulary according to the artificial priori knowledge and experimental tests, as shown in Table 1.

The degree level words in HowNet dictionary are divided into six levels, as shown in Table 1. In fact, there are no negation words in the dictionary, that is, there is only "超(over)" in the last category. However, the function of negation is reversing the meaning of emotion expression (e.g., "nice" indicates positive emotion tendency, while "not nice" represents a negative emotion) which works similarly to the "超(over)" category. Thus, in this paper, we merged negations into last category in Table 1.

Supposed that a comment text contains k terms, its emotional value calculation steps include: First, calculate the expressed emotional value of a single term(w_i). Then the overall emotional value can be derived by averaging the emotional value of all terms. The specific methods are formulized in Equations (5) and (6). Note that each emotional term(w_i) with two adverbs (adv_{i1} and adv_{i2}) ahead are extracted in each sentence. If adv_{i1} and adv_{i2} appear in the list of modified words as we introduce in Table 1, the corresponding weight (represent by M) is set equal to degree value, otherwise M equals to 1.

$$O_{w_i} = M_{\text{adv}_{i1}} \times M_{\text{adv}_{i2}} \times S_{w_i} \tag{5}$$

$$O_{\text{sent}} = \frac{1}{k}\sum_1^k O_{w_i} \tag{6}$$

3.1.2 | Expansion of the sentiment lexicon

Universal sentiment lexicon can meet the needs of many emotional analysis tasks. However, its performance may not be sufficient enough in a certain specific domain. To

TABLE 1 Modification degree value of each adverb category

Modification words	Degree value	Introduction
极其|extreme 最|most	2.3	The strongest tone, usually expresses most, extreme, absolute or one hundred percent meaning
很|very	1.8	A strong tone usually expresses very, highly or greatly meaning
较|more	1.2	More intense tone, usually expresses more, relatively or fairly meaning
稍|-ish	0.8	Weak tone, usually expresses slightly or a little meaning
欠|insufficiently	0.5	Very weak tone, usually expresses mild, not very or slightest meaning
超|over 不是|negation	−1	More than the normal range of tone, usually expresses excessive or indiscriminate meaning. Somehow, these words are similar with negations

improve the accuracy of emotion analysis in such scenario, it is necessary to develop a domain based sentiment dictionary. In this paper, we construct a SET reviews based sentiment lexicon via two methods: word co-occurrence method and word-embedding similarity method.

In word co-occurrence method, pointwise mutual information (PMI) [9] is a basic concept in information theory and natural language processing (NLP). It is often used to measure the independence between two terms, as shown in Equation (7).

$$\mathrm{pmi}(x,y) = \log \frac{p(x,y)}{p(x)p(y)} \tag{7}$$

where $p(x,y)$ denotes the probability that the word x and y appear together. $p(x)$ denotes the probability of the word x appears, so does $p(y)$ to word y. $Pmi(x,y)$ represents the degree of coexistence of word x and y. The larger the value is, the more frequent the co-occurrence is. PMI can only be used to determine the co-occurrence of two words, but not enough to determine the polarity of a word. Turney [37] utilized the closeness between positive and negative reference words to determine the sentiment orientation (SO) of a word, which is calculated in Equation (8).

$$SO(w) = \frac{1}{p}\sum_1^p \mathrm{pmi}\left(w,w_i^+\right) - \frac{1}{q}\sum_1^q \mathrm{pmi}\left(w,w_i^-\right) \tag{8}$$

Note that w is the emotional word of the polarity to be determined. w^+ and w^- are the positive and negative seed words, meanwhile p and q represent their numbers, respectively. If the value of SO is greater than the threshold, the word is considered closer with positive words than negative words. Thus the probability whether is positive is also greater and vice versa.

In word-embedding similarity method, the vectorization of words has always been a matter of concern. The most intuitive and commonly used word-embedding method is one-hot representation, which expresses each word as a long vector. The dimension of this vector is the size of the vocabulary, where most of the elements are 0 and only one dimension has a value of 1, indicating the word's position in the vocabulary set. However, this approach has two disadvantages: (1) it is easily plagued by the dimensionality disaster, especially when it is applied to deep learning algorithms and (2) the similarity between words cannot be well characterized (known as "word gap"). Another kind of word vectorization method is distributed representation, which was first proposed by Rumelhart [32] to overcome the disadvantage of one-hot representation. The basic idea is to map each word in a language to a short, fixed length vector by training (of course the "short" here is relative to the "long" one-hot representation). Put all these vectors together to form a vector space of words, and each vector is a point in this space. By calculating "distance" in this space, we can judge the similarity (lexical, semantic) between them. In this paper we select word2vec [24,25] to train the corpus and finally express the words as vectors of 300-dimensional length, so that we can observe the similarity between them and then calculate the sentiment polarity of the words, which is described in Equations (9) and (10).

$$\mathrm{Sim}_{\cos}(a,b) = \frac{\sum_1^n (A_i \times B_i)}{\sqrt{\sum_1^n A_i^2} \times \sqrt{\sum_1^n B_i^2}} \tag{9}$$

$$SO(w) = \frac{1}{p}\sum_1^p \mathrm{Sim}_{\cos}\left(w,w_i^+\right) - \frac{1}{q}\sum_1^q \mathrm{Sim}_{\cos}\left(w,w_i^-\right) \tag{10}$$

3.2 | Machine learning-based approach

In recent years, the widespread application of machine learning for emotional analysis indicates the efficiency and accuracy. Features are extracted from data and then used as classification basis by computer. Usually, features are extracted at three levels: lexical, syntactic, and semantic

features. In this paper, we conduct our experiment using all words, sentence keywords and emotion words as lexical features. Keywords are extracted by using the TextRank [23] method, which is a graph-based ranking model for text processing and keyword extraction. The theme of a person's comment has a very strong causal relationship with the emotional polarity that he wants to express. Therefore, we assume that the keyword of text has high correlation with the emotional tendency it expresses. Emotion words are undoubtedly closely related to the emotion expressed in the text.

Then, machine learning models based on above-mentioned lexical features are selected for evaluating teaching performance. In this paper, students' short review texts are classified into two basic sentiment categories, that is, positive and negative one. we choose widely used models like Naive Bayes (NB) [11], Logistic Regression (LR) [10], Support Vector Machine (SVM) [6], and Gradient Boost Decision Tree (GBDT) [14] to conduct experiment. Explanation of these models are as follows:

- Naive Bayes (NB): It is a classic classification method based on Bayes' theorem and the conditional independence assumption of features. For a given training data set, the joint probability distribution of input and output is learned based on the conditional independence assumption of features first. Then, based on this model, Bayes' theorem is used to find the maximal output of the posterior probability for a given input x. This method is simple to implement, and the efficiency of learning and prediction is very high.
- Logistic Regression (LR): LR is a kind of generalized linear model, which is a supervised binary classification model based on Sigmoid function. After a linear change and a Sigmoid function, the output value is converted to a floating point number between 0 and 1. If it is greater than 0.5, it is classified into a positive class, and vice versa.
- Support Vector Machine (SVM): SVM is a binary classification model. Its basic model is a linear classifier with the largest interval defined in the feature space. SVM also includes several kernel techniques, which makes it essentially a non-linear classifier. It has many unique advantages in solving small sample, nonlinear and high-dimensional pattern recognition problem, and can be applied to other machine learning problems.
- Gradient Boost Decision Tree (GBDT): GBDT is an algorithm that classifies data by using an additive model (i.e., a linear combination of base functions) and continuously reducing the residuals produced by the training process. The core of GBDT is that each tree learns the residuals of all previous trees' conclusions. The residuals are the sum of the actual values after adding the predicted values.

4 | EXPERIMENT AND RESULT

4.1 | Dataset and evaluation measurement

We collected 3,926 available SET records from postgraduate evaluation system of Beijing Institute of Technology (BIT), which contains short review texts and grades of the students' teachers. By observing the emotional orientation of the text and the statistical analysis of the scoring, we scored and evaluated the text synthetically to allow experts to separate the positive and negative texts. In this way, we obtained altogether 1,323 negative emotional texts and 1,699 positive emotional texts (3,022 in total).

We use the standard precision, recall, and F1 score as the primary performance indicators to evaluate emotional classifications. In addition, the overall accuracy is also calculated to evaluate the overall performance of the classification. There are four cases of binary classification results: TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative). Our performance indicators' calculations are as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{11}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{12}$$

$$F1 = \frac{2*\text{Precision}*\text{Recall}}{(\text{Precision} + \text{Recall})} \tag{13}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{14}$$

Note that the above calculations of precision and recall rates are all for positive samples, which correspond to negative items.

4.2 | Experiments

4.2.1 | Knowledge-based approach

First, we use the two general sentiment lexicons of NTUSD and HowNet to analyze reviews' sentiment using the methods mentioned in the previous chapter, and the results are shown in Table 2. It can be seen that the sentiment values calculated by these two kinds of sentiment lexicons have achieved a certain degree of accuracy. The performance of NTUSD is better than the HowNet lexicon. However, despite the dominant NTUSD dictionary, its accuracy rate only reaches 72.67%, and the average F1 value is 71.42%. The performance in negative words shows a evident shortcoming of universal sentiment dictionary. These phenomena demonstrate that the universal sentiment dictionary does not storage enough information about STE reviews in educational domain.

TABLE 2 Experiment result of knowledge-based approach

	Positive			Negative			
Sentiment dictionary	**Precision(%)**	**Recall(%)**	**F1(%)**	**Precision(%)**	**Recall(%)**	**F1(%)**	**Accuracy(%)**
NTUSD	72.40	83.05	77.51	73.16	59.33	65.32	72.67
HowNet	62.26	**88.64**	73.93	67.99	30.99	41.63	63.40
SRSDwco	**80.03**	81.40	79.73	75.58	**73.92**	**75.97**	**78.13**
SRSDwes	74.58	86.52	**80.83**	**78.21**	62.13	68.31	75.84

Corresponding to each indicator, each bold value in the table express the best performance value among several sentiment dictionaries.

Second, when constructing an emotional dictionary through word co-occurrence, we set a threshold at 20. A total of 1,664 new positive words and 2,519 negative words were confirmed and merged into dictionary to become the students' reviews sentiment dictionary with word co-occurrence method (SRSDwco). As to word-embedding similarity method, we initialize the parameter of *min_count* to 5 when using word2vec, that is, we only consider words that appear at least five times. Then similarity threshold with cosine distance was set to 0.02 and finally got 301 positive words and 285 negative words to expand the emotional dictionary as the students' reviews sentiment dictionary with word-embedding similarity method (SRSDwes). The final evaluation results are shown in Table 2. Contrasted with the previous experiment, both the word co-occurrence method and the word-embedding similarity method improve the total performance, especially the former one. The positive and negative F1 values increase by 2.22 and 10.65%, respectively. It shows that on the original basis, this method has made great progress in negative corpus evaluation compared with the positive ones. Furthermore, the accuracy of SRSDwco has also increased by 5.46%. As for the SRSDwes, both positive corpus and negative corpus are about 2–3% of this range. Examples of positive terms newly found by aforementioned methods include "低调(modest)," "为人正直(fair-minded)," "亦师亦友(be both friend and tutor)," "深入浅出(easy to understand)," etc. Negative words such as "侮辱(indignity)," "噩梦(nightmare)," "臭名远扬(notorious)," "高高在上(arrogant)" were excavated as well.

4.2.2 | Machine learning-based approach

Before using machine learning methods, vectorization should be conducted to represent the meaning of sentences in the corpus. In this experiment, we use lexical features as comparison. The selection of model is very important for effectively fitting. In this section, we classify the record in training set into two categories (Positive and Negative) by using naive Bayes(NB), logistic regression(LR), support vector machine (SVM), and gradient boost decision tree (GBDT) algorithms. Our data set consists of 3,022 student reviews and 5-fold cross-validation is utilized for measurement of accuracy. Also, we compare three theme words sets in the experiment: All words in the corpus excluding the stop words (All Words for short 7,259 in total), theme words generated by textrank algorithm (Theme Words for short, 5,819 in total), all sentiment words in SRSDwco (Sentiment Words for short, 4,183 in total). For each review, all extracted words are fused as input features for machine learning-based classification models and the review's sentimental polarity is regarded as its classification label.

TABLE 3 Experiment result of machine learning-based approach

		Positive			Negative			
Feature word selection	**Model**	**Precision(%)**	**Recall(%)**	**F1(%)**	**Precision(%)**	**Recall(%)**	**F1(%)**	**Accuracy(%)**
All words	NB	80.78	88.08	84.47	81.76	71.83	76.21	81.15
	LR	79.89	91.67	86.64	86.35	69.57	75.33	82.14
	SVM	82.67	85.77	83.86	81.15	77.31	79.59	82.03
	GBDT	84.65	84.31	83.66	81.07	81.46	82.23	83.02
Theme words	NB	75.46	89.69	83.47	83.12	63.52	70.05	78.06
	LR	83.03	88.87	86.08	84.12	76.46	79.81	83.46
	SVM	85.63	86.13	85.06	81.89	81.27	82.62	84.01
	GBDT	**87.69**	80.87	81.99	78.38	**85.93**	**84.51**	83.13
Sentiment words	NB	73.83	90.8	82.85	81.87	56.36	64.79	76.19
	LR	79.39	89.68	85.50	85.39	72.15	76.63	81.69
	SVM	83.89	**89.98**	**87.24**	85.87	77.89	81.14	84.67
	GBDT	83.14	89.38	87.02	**86.96**	79.63	82.12	**84.78**

Corresponding to each indicator, each bold value in the table express the best performance value among several sentiment dictionaries.

As depicted in Table 3, compared with only knowledge-based approach, the prediction of sentiment orientation using machine learning has made great progress both in accuracy and recall. The bold values play the same role as in Table 2. The effect of the classifier generally increases gradually when using lexical features among All Words, Theme Words and Sentiment Words. It means that many words play a role of noise to a large extent, resulting in poor performance via All Words. It is also worth mentioning that the Theme Words set and Sentiment Words set share some words, that is, $\{ThemeWords\} \bigcap \{SentimentWords\} \neq \emptyset$. SVM and GBDT are the two most popular classifiers nowadays and have achieved excellent results in this experiment. They performed very well on both positive and negative data with little relative deviation. Finally, using GBDT model and Sentiment Words set as a vectorized dictionary we can get the best classifier, with an accuracy rate of 84.78%.

To obtain a more intuitive understanding of these methods, Figure 2 shows the performance of precision, recall, and F1 score on positive and negative classification tasks and overall accuracy. Note that we select one model with the best accuracy or method of each approach, that is, NTUSD, SRSDwco, and Sentiment Words based GBDT model as general lexicon, customized lexicon and machine learning method, respectively. It is obviously that the best accuracy rates can reach at 72.67, 78.13, and 84.78%. Our proposed methods over perform 5.46 and 12.11% than the universal one. Also, other evaluation indicators also support our methodology.

5 | UNDERSTANDABLE LINGUISTIC FEATURE AND ITS APPLICATION SCENARIOS

5.1 | Reviews' style

After automated classification is conducted, it is possible to acquire human understanding of linguistic features by tracing back every review text of same classification category. We

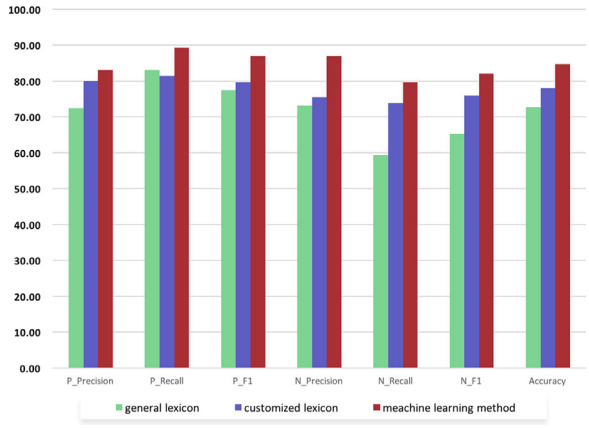


FIGURE 2 The comparison of different methods performance

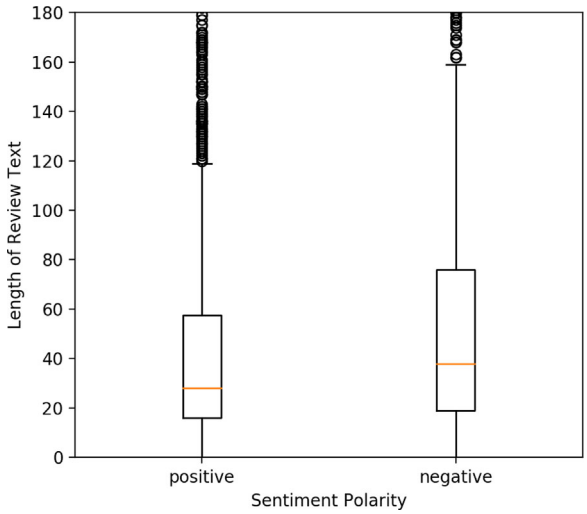


FIGURE 3 Box plots of review's length of positive and negative reviews

notice a very interesting phenomenon when manually checking the classification results: most positive reviews are described in simpler terms and more abstract while negative ones are more evidential. For example, brief or general words often appear in positive reviews, like "好(nice)," "有耐心的(patient)," "知识渊博(have a wide range of knowledge)," etc. Correspondingly in the negative ones, besides using some negative emotion words, students often use evidential texts to explain why they believe their teacher is not good. These evidential stories include the complaint about lecturing performance, teacher's rejection to answer questions asked by student and other specific and detailed teachers' behavior.

To examine this conclusion, we conduct statistics analysis in two ways. First, we record the amount of Chinese characters except punctuation in each evaluation text, and compare the difference between the positive reviews and the negative ones

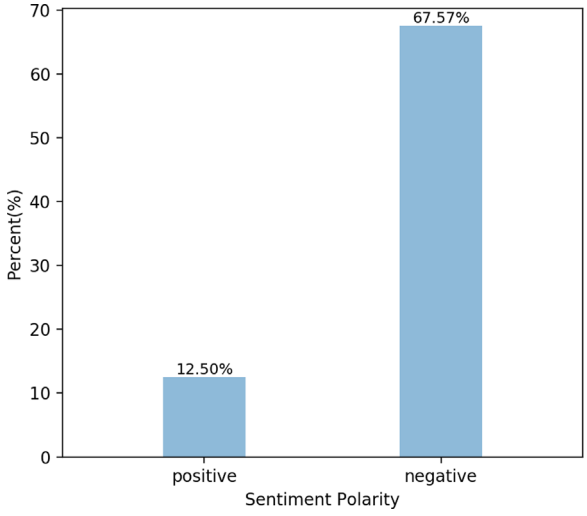


FIGURE 4 Proportion of evidential words of positive and negative reviews

TABLE 4 Examples of student reviews

Sentiment polarity	No.	Review text
Positive	1	很好的老师, 对学生开放平等尊重。(Very good teacher, very liberal and respectful to students).
	2	好老师, 对学生很好, 提倡因材施教。(Good teacher, very good for students, advocates individualized teaching).
	3	非常好的老师, 一生受益匪浅 (Very good teacher, I will benefit a lifetime).
Negative	4	整天打击 压榨学生, 给学生没有一点指导, 还整天挖苦学生 打击你的自信。(Criticize and squeeze students all day, give students no guidance, but also torment students and combat your self-confidence).
	5	上课不好好教学, 基本都在吹嘘自己, 学生基本靠自学。(The teacher doesn't seriously teach in class and always boasts himself. Students rely on their own self-study).
	6	很多问题没有详细解释清楚, 让学生自己看书, 前去答疑也是不好好讲, 不知道是能力问题还是态度问题。(Many problems are not explained in detail, let the students read their books. When asked questions, He always does not seriously explain. I do not know if this is due to his ability or attitude).

on this parameter. As shown in Figure 3, the box charts for negative evaluations were significantly higher than the positive ones, with their lower quartile, median, upper quartile, and mean are respectively locating at 16, 28, 57.5, 48.51 and 19, 38, 76, 71.36. Negative reviews have more 10, 22.85 Chinese characters in median and mean value, respectively.

Second, as described above, we define the "evidential words" as words or stories that describe a specific teaching activity or section. Then we manually label the records whether contains such evidential words. Figure 4 presents the percentage of evidential words in each category. It is obvious that there is a big gap between positive and negative category (12.50 and 67.56%). These two results demonstrate our original suppose that "Students tend to use more words and provide more evidences to explain the reasons they give negative evaluation reviews." In addition, we list some examples of student reviews in Table 4 to obtain a more intuitive understanding.

5.2 | An application scenario: BIT-UASET

University administrated student evaluations of teaching system (UASET) is an apposite application scenario because its regularity and convenience for management. We apply our sentiment analysis tools to the postgraduate class oriented UASET system in Beijing Institute of Technology (BIT-UASET) which score students' reviews automatically as a supplementary reference for teaching quality control. Then the process of BIT-UASET is modified as Figure 5. The bottom layer of the system is the data acquisition layer. Students input the evaluation information for the teacher into the system through text and questionnaire selection. The

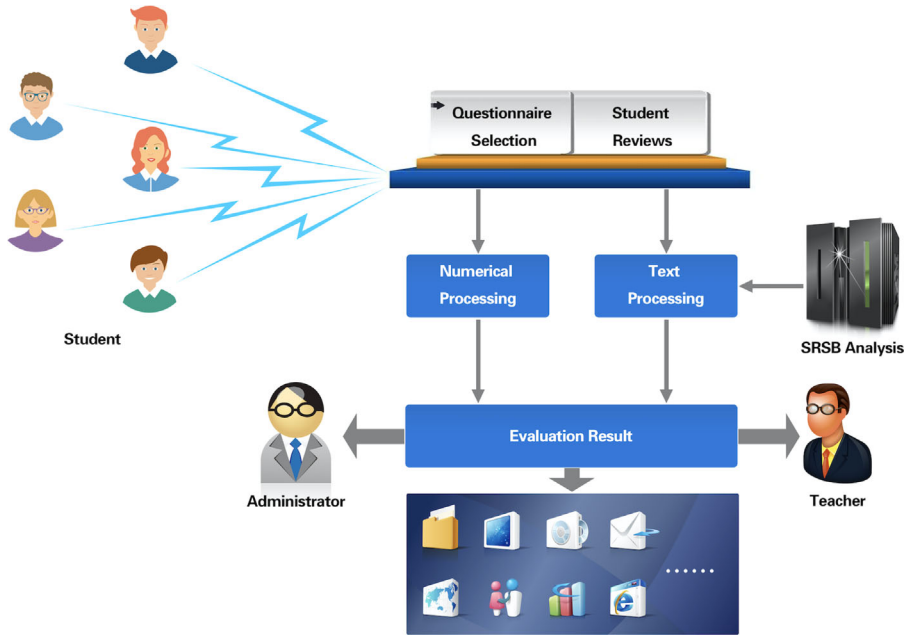


FIGURE 5 Modified evaluation process after the application of our sentiment analysis tools

system processes both of these two kinds of messages separately. For the text, our proposed method is utilized to calculate the sentiment score of each review. After the questionnaire selection data are processed by statistical analysis, the teacher's evaluation result is finally generated from these two aspects. In addition, the system has been deployed for teaching evaluation on postgraduate courses in BIT, which will receive ~4,000 postgraduates' evaluation on ~800 teachers every semester. Therefore, the deployment of our system has high potentiality for further and deeper educational mining research.

6 | CONCLUSION AND DISCUSSION

In this paper, we use the traditional knowledge-based approach to analyze the emotion of student reviews and build a sentiment lexicon (1664 positive words and 2,519 negative words). Based on this, we also use four machine learning methods (NB, LR, SVM, and GBDT) through lexical features to calculate the emotion polarity of the evaluation texts. The best accuracy rates of using general dictionaries, customized dictionaries, and machine learning methods were 72.67, 78.13, and 84.78% respectively, with the latter two methods increasing by 5.46% and 12.11%. Finally, we discussed the human understanding of linguistic features among each category of reviews and presented its application to BIT-UASET.

Our article focuses on the teacher evaluation based on students' short reviews. It is somehow different from the related literature as said in section 2.2, resulting in that the comparison between them is impractical. We look forward to collecting more data and comparing with neural networks in the future.

Further investigation can be conducted: First, sentence- and paragraph-level sentiment analysis methods are also need for teaching evaluation work. Second, multimodal fusion from aspects to obtain accurate teacher modeling can relay on aggregating students' assessment (including multiple-choice answers and evaluation texts), educational background, academic social network, and research abilities. Third, we have already done some research on e-learning recommendations [8,34,35,38], and we intend to integrate our BIT-UASET and these e-learning recommender technologies to improve teacher modeling.

ACKNOWLEDGMENTS

The authors would like to thank Graduate School of Beijing Institute of Technology for the access to data. This work is supported by the National Natural Science Foundation of China (No. 61370137), the Ministry of Education-China Mobile Research Foundation Project (No. 2016/2-7), and the Postgraduate Education Research Project of Beijing Institute of Technology (No. 2017JYYJG-004).

ORCID

Qika Lin http://orcid.org/0000-0001-5650-0600

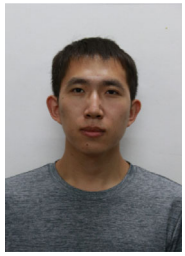
REFERENCES

1. P. Adinolfi et al., *Sentiment analysis to evaluate teaching performance*, Int. J. Knowledge Soc. Res. **7** (2016), 86–107.
2. K. Z. Aung and N. N. Myo, *Sentiment analysis of students' comment using lexicon based approach*, Ieee/acis International Conference on Computer and Information Science, 2017, pp. 149–154.
3. R. Barrett, *Multi-aspect and multi-class based document sentiment analysis of educational data catering accreditation process*, International Conference on Cloud & Ubiquitous Computing & Emerging Technologies, 2014, pp. 188–192.
4. H. H. Binali, C. Wu, and V. Potdar, *A new significant area: Emotion detection in e-learning using opinion mining techniques*, IEEE International Conference on Digital Ecosystems and Technologies, 2009, pp. 259–264.
5. S. S. Boswell, *Ratemyprofessors is hogwash (but i care): Effects of ratemyprofessors and university-administered teaching evaluations on professors*, Comput. Human Behav. **56** (2016) 155–162.
6. C. J. C. Burges, *A tutorial on support vector machines for pattern recognition*, Data Min. Knowl. Discov. **2** (1998), 121–167.
7. J. Cao et al., *Web-based traffic sentiment analysis: Methods and applications*, IEEE trans. Intell. Transp. Syst. **15** (2014), 844–853.
8. W. Chen et al., *A hybrid recommendation algorithm adapted in e-learning environments*, World Wide Web. **17** (2014), 271–284.
9. K. W. Church and P. Hanks, *Word association norms, mutual information, and lexicography*, Comput. Ling. **16** (1990), 22–29.
10. A. Cucchiara, *Applied logistic regression*, Technometrics **34** (1989), 358–359.
11. P. M. Domingos and M. J. Pazzani, *On the optimality of the simple bayesian classifier under zero-one loss*, Mach. Learn. **29** (1997), 103–130.
12. Z. Dong and Q. Dong, *Hownet – a hybrid language and knowledge resource*, International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings, 2003, pp. 820–824.
13. G. G. Esparza et al., *A sentiment analysis model to analyze students reviews of teacher performance using support vector machines*, International Symposium on Distributed Computing and Artificial Intelligence, 2017, pp. 157–164.
14. J. H. Friedman, *Greedy function approximation: A gradient boosting machine*, Ann. Stat. **29** (2001), 1189–1232.
15. A. García-Pablos, M. Cuadros, and G. Rigau, *W2vlda: Almost unsupervised system for aspect based sentiment analysis*, Expert Syst. Appl. **91** (2017), 127–137.
16. S. Gul et al., *Twitter sentiments related to natural calamities: Analysing tweets related to the jammu and kashmir floods of 2014*, Electronic Library **36** (2018), 38–54.
17. V. Hatzivassiloglou and K. R. Mckeown, *Predicting the semantic orientation of adjectives*, Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (1997), 174–181.
18. S. Huang, Z. Niu, and C. Shi, *Automatic construction of domain-specific sentiment lexicon based on constrained label propagation*, Knowl. Based Syst. **56** (2014), 191–200.

19. W. Y. Hwang et al., *Exploring effects of discussion on visual attention, learning performance, and perceptions of students learning with str-support*, Comput. Educ. **116** (2018), 225–236.
20. L. W. Ku and H. H. Chen, *Mining opinions from the web: Beyond relevance retrieval*, J. Am. Soc. Inf. Sci. Tech. **58** (2007), 1838–1850.
21. L. W. Ku, Y. T. Liang, and H. H. Chen, *Opinion extraction, summarization and tracking in news and blog corpora*, Proceedings of AAAI (2006), 100–107.
22. T. M. Lee et al., *Predictors of public climate change awareness and risk perception around the world*, Nat. Clim. Chang. **5** (2015), 1014–1023.
23. R. Mihalcea and P. Tarau, *Textrank: Bringing order into texts*, Conference on Empirical Methods in Natural Language Processing, EMNLP 2004, A Meeting of Sigdat, A Special Interest Group of the Acl, Held in Conjunction with ACL 2004, 25–26 July 2004, Barcelona, Spain, 2004, pp. 404–411.
24. T. Mikolov et al., *Efficient estimation of word representations in vector space*, arXiv preprint arXiv **1301.3781** (2013). available at: https://arxiv.org/pdf/1301.3781.pdf
25. T. Mikolov et al., *Distributed representations of words and phrases and their compositionality*, Advances in Neural Information Processing Systems (2013), 3111–3119.
26. J. C. Ory, *Teaching evaluation: Past, present, and future*, New Directions Teach Learn. **2000** (2010), 13–18.
27. Y. J. Park et al., *A deep learning-based sports player evaluation model based on game statistics and news articles*, Knowl. Based Syst. **138** (2017), 15–26.
28. H. Peng, E. Cambria, and A. Hussain, *A review of sentiment analysis research in chinese language*, Cognit. Comput. **9** (2017), 1–13.
29. C. Ponginwong and W. S. Rungworawut, *Teaching senti-lexicon for automated sentiment polarity definition in teaching evaluation*, International Conference on Semantics, Knowledge and Grids, 2014, pp. 84–91.
30. J. R. Ragini, P. M. R. Anand, and V. Bhaskar, *Mining crisis information: A strategic approach for detection of people at risk through social media analysis*, Int. J. Disaster Risk Reduct. **27** (2017), 556–566.
31. P. Rodriguez, A. Ortigosa, and R. M. Carro, *Extracting emotions from texts in e-learning environments*, Sixth International Conference on Complex, Intelligent and Software Intensive Systems, 2012, pp. 887–892.
32. D. E. Rumelhart, *Learning internal representations by back-propagating errors*, Nature **323** (1986), 533–536.
33. S. Tan and J. Zhang, *An empirical study of sentiment analysis for chinese documents*, Expert Syst. Appl. **34** (2008), 2622–2629.
34. J. K. Tarus, Z. Niu, and D. Kalui, *A hybrid recommender system for e-learning based on context awareness and sequential pattern mining*, Soft Comput. **6** (2017), 1–13.
35. J. K. Tarus, Z. Niu, and A. Yousif, *A hybrid knowledge-based recommender system for e-learning based on ontology and sequential pattern mining*, Future Gener. Comput. Syst. **72** (2017), 37–48.
36. F. Tian et al., *Recognizing and regulating e-learners' emotions based on interactive chinese texts in e-learning systems*, Knowl. Based Syst. **55** (2014), 148–164.
37. P. D. Turney, *Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews*, Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (2002), 417–424.
38. S. Wan and Z. Niu, *A learner oriented learning recommendation approach based on mixed concept mapping and immune algorithm*, Knowl. Based Syst. **103** (2016), 28–40.
39. Z. Zhai et al., *Exploiting effective features for chinese sentiment classification*, Expert Syst. Appl. **38** (2011), 9139–9146.
40. D. Zhang et al., *Chinese comments sentiment classification based on word2vec and svm perf*, Expert Syst. Appl. **42** (2015), 1857–1863.
41. H. Zhao et al., *A teaching evaluation method based on sentiment classification*, Intern. J. Comput. Sci. Math. **7** (2016), 54–62.

Q. LIN received his BE degree in Chemical Engineering and Technology from Beijing Institute of Technology, Beijing, China in 2016. He is currently working toward the MS degree at School of Computer Science & Technology, Beijing Institute of Technology, Beijing, China. His research interests include data mining, recommendation system, and natural language processing.

Y. ZHU received his BE degree in Computer Science from Beijing Information Science & Technology University, Beijing, China in 2016. He is currently working toward the PhD degree at School of Computer Science & Technology, Beijing Institute of Technology, Beijing, China. His research interests include opinion mining, user profiling, and social computing.



S. ZHANG received his BE degree in Packaging Engineering from Jinan University, Zhuhai, China in 2016. He is currently working toward the MS degree at School of Computer Science & Technology, Beijing Institute of Technology, Beijing, China. His research interests include data mining, recommendation system, and natural language processing.

P. SHI is currently working toward the MS degree at School of Computer Science & Technology, Beijing Institute of Technology, Beijing, China. His research interests include information security, data mining, and user profiling.

Q. GUO is a senior engineer in Dawning Information Industry Co., Ltd. and a PhD student in Beijing Institute of Technology. His research interests are distributed systems, big data storage, and processing.

Z. NIU received his PhD degree in Computer Science from the Beijing Institute of Technology, Beijing, China, in 1995. He was a post-doctoral researcher with the University of Pittsburgh, Pittsburgh, PA, USA, from 1996 to 1998, a researcher/adjunct faculty member with Carnegie Mellon University, Pittsburgh, from 1999 to 2004, and a joint professor with School of Computing and Information of University of Pittsburgh from 2006. He is a professor and the deputy dean with the School of Computer Science and Technology, Beijing Institute of Technology. His current research interests include informational retrieval, software architecture, digital libraries, and Web-based learning techniques. Prof. Niu is a recipient of the IBM Faculty Innovation Award in 2005 and the new century excellent talents in University of Ministry of Education of China in 2006.

How to cite this article: Lin Q, Zhu Y, Zhang S, Shi P, Guo Q, Niu Z. Lexical based automated teaching evaluation via students' short reviews. *Comput Appl Eng Educ*. 2019;27:194–205. https://doi.org/10.1002/cae.22068