2019 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)

Detecting Interest-factor Influenced Abnormal Evaluation of Teaching via Multimodal Embedding and Priori Knowledge based Neural Network

Yu Mao[1,†], *Student Member, IEEE,* Yifan Zhu[1,†], *Student Member, IEEE,* Sifan Zhang[1], Dexiu Zhang [2], Fuquan Zhang[3,*], Xiaozhong Fan[1]

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

[2]*School of Computer Science, Minnan Normal University, Zhangzhou, China*

[3]*Fujian Provincial Key Laboratory of Information Processing and Intelligent Control, Minjiang University, Fuzhou, China*

maoyu_bit@163.com, {zhuyifan, zhangsifan}@bit.edu.cn, zhdxmnnu@163.com, 8528750@qq.com, fxz@bit.edu.cn

***Abstract*—Student evaluations of teaching (SET) helps to quantitatively measure the teaching performance of classes. However, concerns arouse when the result of SET and students final grades are highly positively correlated, which means the students may make the evaluation based on the final grades they take. For the purpose of avoiding retaliation and mutual reward rating, some methods based on statistics or rules are used to detect the presence of anomaly evaluation. Whereas, the problem is still challenging since the characteristics of this abnormal evaluation are implicit and multimodal. Besides, additional information such as the domain of subject, instructors teaching style, and the reputation spread by students also requires corresponding prior knowledge to supplement the abnormal evaluation detection model. Therefore, in this paper, we proposed a multimodal embedding and prior knowledge based neural network to detect potential so called Interest Factor Influenced Abnormal Evaluation (IFAE). The method proposed in this paper uses SDNE and PV-DM to embed the evaluation network between students and teachers and the comment text of students. Taking into account the continuity of mutual evaluation between students and teachers during the teaching cycles and the different emphasis of students' comment texts on teachers, the features of students and teachers are comprehensively constructed. Then, the attention mechanism is used in the comment text to perform final prediction jointly with above features. The experiment result shows that our proposed model outperforms other state-of-art models which based on single type of features on F1 score by 9.25%.**

***Index Terms*—Interest factor influenced abnormal evaluation, student evaluation of teaching, fusion of multimodal feature, priori knowledge, abnormal evaluation detection.**

I. Introduction

Student evaluations of teaching (SET) plays a significant role in evaluating the performance of classroom instruction, thereby assessing faculties professional competence [1]. Coupled with the existing statistical and intelligent analysis techniques, students can anonymously feedback the feelings and opinions in the course to the school quality control department and teachers themselves as a reference for future improved teaching [2], [3]. However, this evaluation method also experiences this challenge while the teaching quality control department can qualitatively and quantitatively analyze the classroom performance of the classroom teachers [4]. The conflict of interest exists when students evaluate their teachers considering the final score of the class. In fact, cases have been demonstrated that the evaluation can be used as weapons of revenge to teacher who give the bad score or benefits of reward to the one give the score beyond expectation [5]. Although we believe these extreme cases are not common in all classes, such stimulus may still drive teachers change their teaching style and plan once their faculties professional competence are associated with the result of SET [6]. Therefore, a motivation to be a kinder, softer and easier to give high score is generated to teachers, which perturb the original object of both university teaching and SET. Compared with university administrated

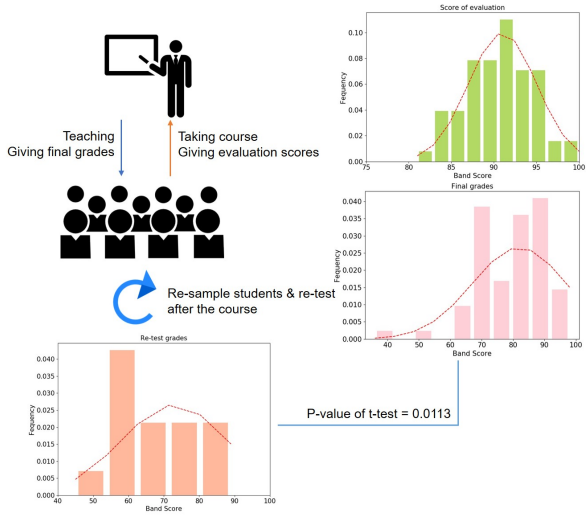

Fig. 1. An illustration of an example of suspected abnormal evaluation: students and the teacher highly scored each other in a course but a re-test shows that the students' mastery of knowledge is not as high as the level their final grades reflected (Data is collected from SET system in Minnan Normal University).

[†]The authors who share the same contribution to this work; [*]Corresponding author.

978-1-7281-4328-6/19/$31.00 ©2019 IEEE
DOI 10.1109/ISPA-BDCloud-SustainCom-SocialCom48970.2019.00171

SET (UA-SET), the situation gets worse at public anonymous SET systems. Boasting from friends even self and criticism without evidence damage the reliability to a large extent [4]. A vivid example is presented in Fig 1, where in a Software Engineering class, sixty-seven students and one course teacher scored highly of each other, but an after-course re-test by randomly sampling fifteen students in this class suggests their mastering of knowledge is obviously poorer than it thought to be (paper of examination in the re-test is replaced with the same course from another university with similar school-running level.).

In practice, several measures have been taken to reduce the such problems among universities. For instance, evaluation of pass rate can be used to limit the extremely score distribution. However, this data-control method only covers up the problems and introduces more technical even ethical problems. Another method is that the college organize a unified standard examination to measure the performance of teaching. However, the drawbacks are also obvious: it can only be applied to large-scale multi-classes teaching and the basis may be conveyed to teachers that they are not trustable, making the teachers bear the pressure that should have been borne by the students. In other words, the isolation between students score and their reviews makes it difficult to remove the influence of conflicts of interest from SET. It needs a method to identify such abnormal evaluation before we simple refer to students SET reviews. To summarize, the recessiveness and the diversity of feature types make it difficult to quantify these potential teaching evaluations.

Facing the challenges and issues discussed above, in this paper, we name this kind of abnormal evaluation as Interest Factor Influenced Abnormal Evaluation (IFAE). We propose a neural network structure to predict such potential IFAE based on three types feature: graphical bidirectional scoring feature, textual comment feature and numerical prior scoring benchmark feature. In detail, we first extract corresponding embedded representation of graphical and textual features, and then aggregate them with introducing the priori scoring benchmark for each course and student in a unified network. The contribution of this paper is two folded: First, to the best of our knowledge, it is the first time to present a predicting model to detect potential IFAE according to the multimodal SET data; Second, we have integrated and initially explored the benefits of different types of evaluation information to identify IFAE.

The rest of this paper organizes as follow: in Section 2 we present a brief literature review on previous efforts in finding abnormal evaluations. Section 3 shows the methods and structure of predicting model we utilized to identify IFAE. Experimental results are presented in Section 4. We discuss our results and note the limitations in Section 5. Section 6 gives the conclusion remarks.

II. Related work

Over the years, with the deepening of teaching theory research and the expansion of practical fields, the role of teaching evaluation is becoming more and more important. Educational researchers have investigated many factors considered to affect student learning. There are continuing researches about how much the extant teacher effectiveness literature can be trusted to identify characteristics of effective teachers [7]–[9], and additional researches about the effect of psychological perception on student achievement [5], [10], [11], thereby measuring the quality of teaching from the aspects of students' psychological emotions when evaluating teachers, the effectiveness of evaluating teaching content, and the selection bias in evaluating courses. In addition, [12] took the scores students obtained as an important indicator of teaching evaluation, and study intends to explore the relationship among students perceptions of course and teacher and achievement at higher level. Mori *et al.* and Hou *et al.* discussed the influence of some artificial and social factors on teaching evaluation from a deeper level of teacher personality and disciplinary mechanism, and investigated whether the relations between instructional and personality ratings and the general course evaluation varied by major [13], [14]. Although these studies have put forward some new evaluation opinions, they only measured the quality of the evaluation system from the perspective of evaluation indicators, and did not propose how to identify abnormal evaluation from previous evaluation data.

At present, a few of the students sentiment texts analyzing research have been concentrated on teaching evaluation, which focused most of their attention on e-learning environments [15], [16]. In addition, some researchers constructed the level of opinion result to determine opinion result of teachers. Aung *et al.* and Binali *et al.* respectively analyzed the students' text feedback automatically using lexicon based approach to predict the level of teaching performance [17], [18]. A database of English sentiment words was created as a lexical source to get the polarity of words, and by analyzing the sentiment information including intensifier words extracting from students' feedback to describing the level of positive or negative opinions. Other researchers process text evaluation data by extracting new features. [19], [20] selected different types of features to improve sentiment classification performance. They extracted the sentiment-words, substrings, substring-groups, and key-substring-groups as features in sentiment classification area, and then compared and analyzed the features. Moreover, it is also a research hotspot to process the evaluation data by analyzing the semantic relationship between words in the evaluation data. Zhang *et al.* proposed a method for sentiment classification based on word2vec and SVM to get the semantic features [21]. Esparza *et al.* applied Support Vector Machines algorithm with three kernels: linear, radial and polynomial, to predict a classification of comments in positive, negative or neutral [22]. [23] and [24] have also achieved good performance by combining sentiment analyzing methods with other features to mine and analyze text evaluation data.

Although these studies above have performed sentiment analysis on the comment text data to some extent, they only get the results of students evaluation to teacher by using comment text, and can not identify abnormal evaluation. To

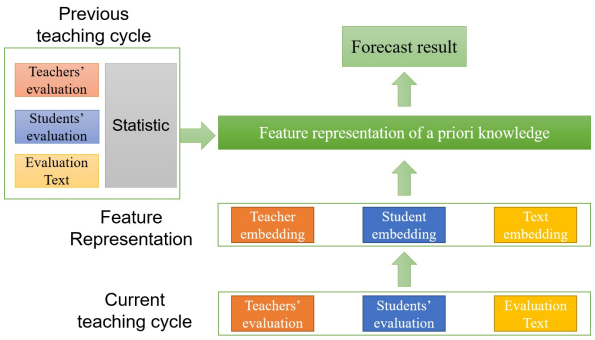

Fig. 2. The pipeline of the proposed model.

summarize, the current research on teacher evaluation based on student reviews has not been investigated enough in abnormal evaluation field yet.

III. Methods

The architecture of the proposed method is shown in Fig. 2. Among them, the teachers evaluation, students evaluation, and students comment text are the data sources.

Due to the periodicity of teaching activities, students often have a certain understanding of the corresponding teachers from the previous students before class, and they will have corresponding expectations, so the evaluation to teachers will be affected by this. Similarly, a teacher's teaching style and evaluation model are relatively fixed, and it will hardly produce significant changes within two teaching cycles. Therefore, the evaluation of students will have certain similarities with the evaluations in the previous cycle.

In other words, in the last teaching cycle, the teacher's evaluation to the students will be partially extended to this cycle, and evaluation from students to teachers will also have a certain impact in this cycle. Therefore, in predicting the results of this teaching cycle, we must consider the results of mutual evaluation between teachers and students in the previous cycle.

As shown in Fig. 2, this paper firstly constructs the graph network according to the mutual evaluation of teachers and students in this cycle, and then embeds the nodes in the network by using the graph embedding algorithm. At the same time, considering that the student's comment text of the teacher can reflect the students' attitude towards the teacher, the comment text is also included as part of the model input. Moreover, the mutual evaluation of teachers and students in the last teaching cycle was comprehensively considered. Finally, the above materials are comprehensively constructed as classification features and classified by deep network.

A. Graph Embedding

At the end of this teaching cycle, students and teachers evaluation to each other, constitutes a huge student-teacher evaluation network. Then the graph embedding method is used in order to represent teachers and students entities.

Graph embedding [25] aims to map the graph data into a low-dimensional latent space, where each vertex is represented as a low-dimensional vector and the network computing can be directly realized. Generally, learning graph representations faces the following great challenges: High non-linearity, Structure-preserving and Sparsity [26]. Although there are many graph representation methods such as Deepwalk [25], node2vec [27], and TransR [28], they do not solve the above three problems well.

Structural Deep Network Embedding(SDNE) [26] is able to perform graph embedding and fix above questions by using first-order proximity and second-order proximity.

First, SDNE is performed according to the following definitions:

1) Graph: $G = (V, E)$ where $V = \{v_1, , v_n\}$ represents n vertexes and $E = \{e_{i,j}\}_{i,j=1}^{n}$ represents the edges, each edge is associated with a weight $s_(i, j) \leq 0$.
2) First-Order Proximity: The first-order proximity describes the pairwise proximity between vertexes. For any pair of vertexes, if $s_{i,j} > 0$, there exists positive first-order proximity between v_i and v_j, otherwise, the first-order proximity between v_i and v_j is 0.
3) Second-Order Proximity: The second-order proximity between a pair of vertexes describes the proximity of the pairs neighborhood structure. Let $\mathcal{N}_u = \{s_{u,1}, , s_{u,|V|}\}$ denote the first-order proximity between v_u and other vertexes. Then, second-order proximity is determined by the similarity of $\mathcal{N}_u$ and $\mathcal{N}_v$.
4) Graph Embedding: Given a graph denoted as $G = (V, E)$, graph embedding aims to learn a mapping function $f : v_i \rightarrow y_i \in R^d$, where $d \ll |V|$. The objective of the function is to make the similarity between y_i and y_j explicitly preserve the first-order and second-order proximity of v_i and v_j.

Then, A semi-supervised deep model is used to perform graph embedding. For each vertex, it is able to find its neighborhood. The supervised component is designed to exploit the first-order proximity as the supervised information to refine the representations in the latent space.

The SDNE model is composed of two parts: the encoder and decoder. Given the input x_i, the hidden representations for each layer are shown as follows:

$$y_i^{(1)} = \sigma(W^{(1)} x_i + b^{(1)}) \tag{1}$$

$$y_i^{(k)} = \sigma(W^{(k)} y_i^{(k-1)} + b^{(k)}), k = 2, , K \tag{2}$$

And the objective function is shown in Eq. 2:

$$\mathcal{L} = L_{2nd} + \alpha \mathcal{L}_{1st} + \nu \mathcal{L}_{reg} \tag{3}$$

Each part in Eq.2 can be expanded as Eq.4, Eq.5 and Eq.6:

$$\mathcal{L}_{2nd} = \|(\widehat{X} - X) \odot B\|_F^2 \tag{4}$$

For $\mathcal{L}_{2nd}$, $\odot$ means the Hadamard product, and it is proposed to add more penalty to the reconstruction error of the non-zero elements than that of zero elements.

$$\mathcal{L}_{1st} = \sum_{i,j=1}^{n} s_{i,j} \|y_i - y_j\|_2^2 \tag{5}$$

For $\mathcal{L}_{1st}$, it is not only necessary to preserve the global network structure, but also essential to capture the local structure.So the supervised component to exploit the first-order proximity is used.

$$\mathcal{L}_{reg} = \frac{1}{2} \sum_{k=1}^{K} (\|W^{(k)}\|_F^2 + \|\widehat{W}^{(k)}\|_F^2) \tag{6}$$

Finally, $\mathcal{L}_{reg}$ is an $\mathcal{L}2 - norm$ regularizer term to prevent overfitting.

B. Text Embedding

Our previous studies have shown that students' evaluation to teachers will have many well-defined words with extreme emotional polarity, these words can directly reflect the students' attitude towards teachers [3]. Therefore, analyzing the semantic expression and emotional polarity of the teacher's comment text can assist the prediction.

In this paper, the PV-DM method [29] is used to model the comment text at the paragraph level, and the corresponding word vector is obtained. Specifically, similar to word2vec [30], each paragraph and each word are mapped to a unique vector, represented by a column in the matrix D and W, respectively. Given a sequence of words, the model is to maximize the average log probability:

$$\frac{1}{T} \sum_{t=k}^{T-k} log\, p(w_t | w_{t-k}, ..., w_{t+k}, d_i) \tag{7}$$

The prediction task is typically performed via a multi-class classifier, such as softmax. So, we have

$$p(w_t | w_{t-k}, ..., w_{t+k}, d_i) = \frac{e^{y_{wt}}}{\sum_i e^{y_i}} \tag{8}$$

Each of y_i is an normalized log-probability for each output word i, which is computed as:

$$y = b + Uh(w_{t-k}, , w_{t+k}, d_i; W, D) \tag{9}$$

where U and b are softmax parameters. H is constructed by concatenating or averaging the word vectors extracted from W and the paragraph vector extracted from D.

In this paper, in addition to obtaining semantic information, the two words with the highest positive or negative emotional polarity in the comment text are selected according to the work [3] and provided together with the semantic information.

C. Detection of IFAE

In each teaching cycle, students and teachers will play different roles in the process of mutual evaluation (scoring). When a student evaluates a teacher, the student plays the role to evaluate others, and the teacher plays the role to be evaluated; and when the teacher evaluates the student, the opposite is true. Therefore, in the network of students' evaluation of teachers and teachers' evaluation of students, this paper constructs these four different entities.

As shown in Fig.3, based on the network of teacher-student mutual evaluation, four independent entities are constructed. S_rating represents students who evaluated teachers and correspondingly S_rated represents students who are evaluated by teachers. They are built by two different networks. T-rating and T-rated represent the same two entities of the teacher.

The above-mentioned S-rating and S-rated alone are not complete for students, because the students' evaluation on teachers will be affected by the evaluation they received, and T-rating and T-rated have the same problems. The teachers and students in the previous teaching cycle have already conducted mutual evaluations, leaving their respective evaluation scores, and the work [3] has indicated that past cycles evaluation will have an impact on that in this cycle. Therefore, in order to fully express the students and teachers, this paper integrates the previous teaching cycles evaluation in current evaluation network of students and teachers.

The average value of the evaluation scores from teachers to their students in last teaching cycle is denoted as $t\_rating_i, i = 1, N$, and the average value of the evaluation scores teachers received is denoted as $t\_rated_i, i = 1, N$ where, i represents the i-th teacher. These two values can also be regarded as the average values of the score that students received and the score students evaluated. Although the students are not the same group of students in last cycle, it is mentioned above that due to the periodicity of teaching activities, these students will be influenced by the previous students. This paper uses these two evaluation networks to fully characterize students and teachers, details are as follows:

$$weight(t\_rating_i) = \frac{e^{t\_rating_i}}{e^{t\_rating_i} + e^{t\_rated_i}} \tag{10}$$

$$weight(t\_rated_i) = \frac{e^{t\_rated_i}}{e^{t\_rating_i} + e^{t\_rated_i}} \tag{11}$$

$$\begin{aligned} Student = & S\_rating \times weight(t\_rated_i) + \\ & S\_rated \times weight(t\_rating_i) \end{aligned} \tag{12}$$

$$\begin{aligned} Teacher = & T\_rating \times weight(t\_rating_i) + \\ & T\_rated \times weight(t\_rated_i) \end{aligned} \tag{13}$$

The student's comment texts for the teacher are also considered. According to the previous work [3], for each student's evaluation comment, after using the PV-DM to perform the embedding representation, the two words with the highest

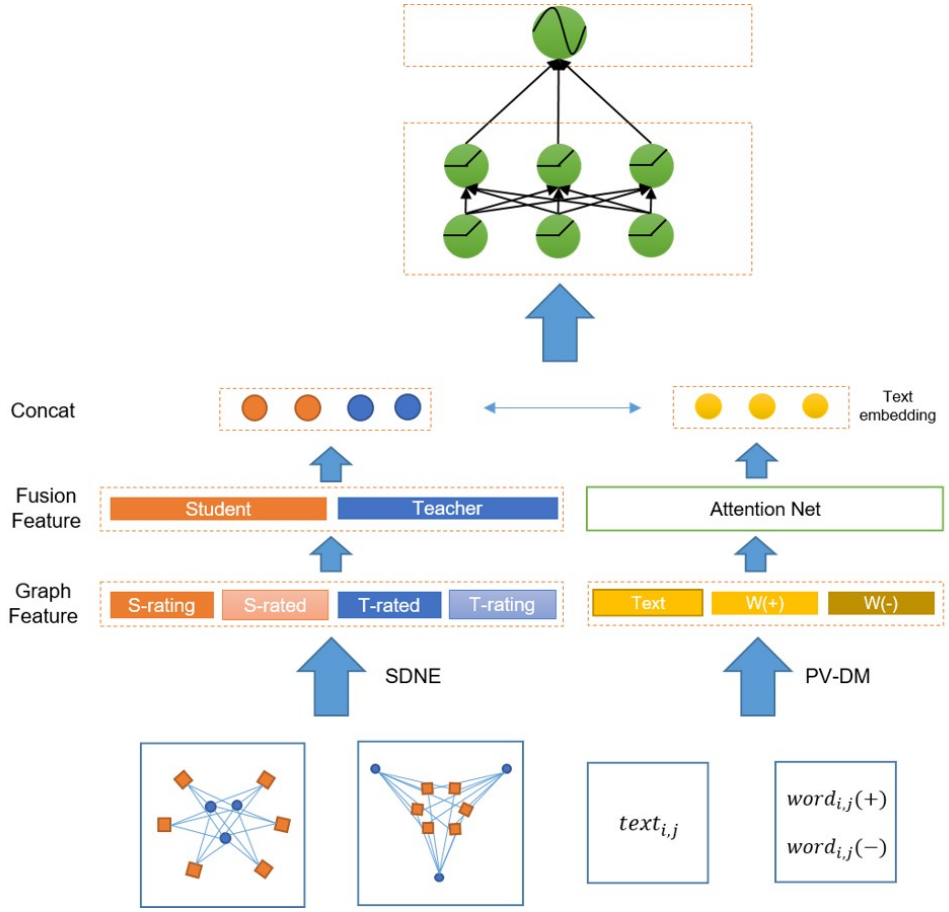

Fig. 3. The architecture of the proposed IFAE detection model.

positive polarity and negative polarity are extracted as features. In that, $text_{i,j}$ represents the evaluation comment of i-th teachers j-th student, $word_{i,j}(+)$ and $word_{i,j}(-)$ represents the positive and negative word with the highest polarity in the text respectively. Taking into account the different emphasis of different students on the evaluation of teachers, this paper uses the attention mechanism to dynamically select the most important features. The output of the attention network is formulated as Eq. 14:

$$\begin{aligned} text\_embedding =& \alpha_1 \times Text + \\ & \alpha_2 \times w(+) + \\ & \alpha_3 \times w(-) \end{aligned} \tag{14}$$

where $_i, i = 1, 2, 3$ is the attention value, which is calculated by a double-layer feedforward neural network. The input of the attention network can be expressed as:

$$\begin{aligned} Input =& W_{text} \times Text + b_{Text} \bowtie \\ & W_{w(+)} \times w(+) + b_{w(+)} \bowtie \\ & W_{w(-)} \times w(-) + b_{w(-)} \end{aligned} \tag{15}$$

where $Input$ is a concatenation of the output of a single layer of neurons with three different biases. b_{Text} refers to the average scores of all the comment texts in the last teaching cycle. $b_{w(+)}$ and $b_{w(-)}$ refers to the average scores of all highest positive word and negative word in the comment text.

Then, through two layers of network, the attention scores are calculated and normalized through the softmax function:

$$\alpha_i^{'} = h^T \, ReLU(W \times Input + b) \tag{16}$$

$$\alpha_i = \frac{e^{\alpha_i^{'}}}{\sum_{i=1}^{3} e^{\alpha_i^{'}}} \tag{17}$$

Afterwords, the mixed features of students, teachers and the output of attention network are cascaded, then input into the feedforward neural network to obtain the predicted values:

$$\hat{y} = Sigmoid(ReLu((W2 \times ReLU(W1 \times feature + b1) + b2)) \tag{18}$$

and the loss function uses log loss and adds L2 regularization to prevent over-fitting that may occur:

$$loss = -y \times log(\hat{y}) - (1 - y) \times log(1 - \hat{y}) + \alpha \sum \omega \tag{19}$$

Where y refers to the true value and $\hat{y}$ refers to the predicted value.

Finally, the arithmetically averages of the predicted value of the teacher with all of his students are calculated to obtain the final predicted result.

$$\hat{y}_i = \frac{1}{M} \sum_{j=1}^{M} \hat{y}_{i,j} \tag{20}$$

M refers to that the i-th teacher (i.e.) has M students.

IV. Results

A. Data collection

To validate the feasibility of our proposed method, a real-world SET dataset is constructed by collecting teaching evaluation data from Minnan Normal University, Zhangzhou city, Fujian province, China in the school year of 2017-2018. In particular, we collect the records containing both teaching evaluation and final grade for each student and course. The evaluation and grade are submitted according to the aforementioned "static game scenario": Students and teachers can read the result of evaluation and grades only after both sides of them have submitted their scores. The prior knowledge of each course is determined by the teacher's score and student evaluation in the previous academic year (2016-2017 school year). Additionally, the labels of the dataset (i.e. IFAE suspected course) are set by courses which have guest students: their opinions and comments towards a course are thought to be more objective because they have less such conflicts of interests discussed in this paper. Thus, we filtered these to keep courses which have both evaluations from normal and guest students and grades from teachers. The labels are determined according to the statistical significance of evaluation between the normal student group and guest student group, In other words, the label is set to true (i.e. suspected IFAE) if the p-value of t-test on evaluation scores between the two groups is less than 0.01.

Finally, a 3,817-record dataset including 412 students and 108 courses from College of Computer and College of Physics and Information Engineering are established. Among these records, about a quarter of the evaluations were labeled as suspected IFAE samples (961). We further randomly select 380 records as the test set and the rest is regarded as training set.

B. Performance metrics and baseline models

As discussed before, the IFAE detection is regarded as a binary classification. Thus, the metrics of IFAE detection is same to other classification tasks. In this paper, we use F1 score and ROC (Receiver Operating Characteristic) curve to assess the performance of different models.

We also select the following models as baselines for comparison:

- LSTM [31]: First we chose to use the subjective text comments of all students as features and identify the IFAE using the LSTM model. Specifically, each student's comment text uses a word embedding model (word2vec) to obtain a corresponding vector sequence after the word segmentation. These vector sequences are then trained and classified through a bidirectional Gated Recurrent Unit (GRU) layer before a fully connected layer as well as softmax layer.
- GCN [32]: Based on the evaluation and grading network, a Graph Convolutional Network (GCN) can be applied to further learn the graphical relationship, thereby establishing classification driven by the network structure. In particular, the node feature matrix (two types pf nodes: teachers and students) and the adjacency matrix are convolved through two GCN layers which activated by ReLu function, and then are propagated through a dense layer and a softmax layer.
- SDNE [26]: As a graph embedding approach, SDNE transfers nodes in a network into lower dimensional numerical representation. In the experiment, we use SNDE to extract the corresponding vectors of each evaluating/grading pair between student and teacher, and then use dense layer as well as a softmax layer to establish the classification. In other words, this implementation is similar to the right half of our proposed model which shown in Fig.2.
- SRSwco [3]: Instead of machine learning approach, knowledge based approaches also received recent focus in the academic community. In this paper, we also compare the result from a teaching evaluation domain based sentiment lexicon named SRSwco. In particular, the standard deviation of sentiment scores which each course receives is measured and a threshold is utilized to judge whether IFAE was happened in this course.

C. Experiment and results

The implementation of our model was based on Tensorflow (version 1.5) and Keras (version 2.2.4) and the implementation of SDNE was originated from OpenNE. Spektral is utilized to implement the GCN and other graph based neural network models.

Shown in Table I, our proposed model outperforms other baselines in both precision and recall measurement, thereby achieving the best F1 score. In these models, text or content-based approaches (LSTM, SRSwco) generally performs worse than graph-based approaches (GCN, SDNE). This result reflects the fact that the comment text brings a lot of noise while providing the effectiveness of the evaluation analysis. Sparse and small amount of short comments for teaching evaluation can be difficult to achieve accurate IFAE detection alone without the aid of prior knowledge. But on the other hand, when the model reintroduces text features and prior knowledge on the basis of the basic features of the graph structure, these text features can benefit the original IFAE detection.

By letting the predicting model output the probability of each sample in the test set, we plot the Receiver Operating Characteristic (ROC) curves for the different models (Fig 4). The area under ROC curve (AUC) also validates the

TABLE I
PERFORMANCE COMPARISON OF DIFFERENT MODELS ON TEST SET (N=380)

Model	Precision	Recall	F1	Accuracy
LSTM	0.3815	0.6875	0.4907	0.6394
GCN	0.5556	0.6770	0.6103	0.7815
SDNE	0.6115	0.7708	0.6820	0.8184
SRSwco	0.5473	0.5416	0.5445	0.7710
Ours	0.7314	0.8229	0.7745	0.8789

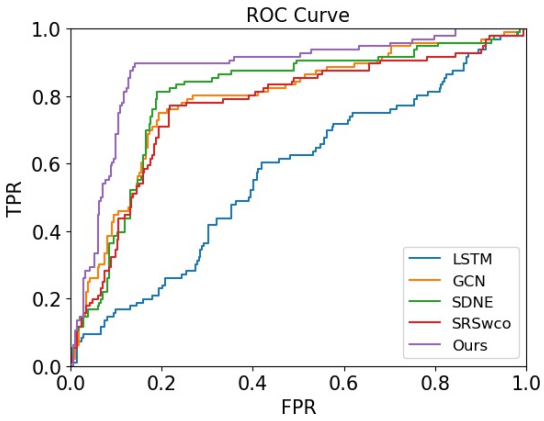

Fig. 4. ROC Curve of different models.

aforementioned explanation and indicates our proposed model reached the best predicting performance.

V. DISCUSSION

Evaluation of teaching is a long-term and important research area. However, the bias seems unavoidable due to not only the different evaluation criteria but also the potential conflict of interest. The nature of education is to guide people acquire more knowledge instead of playing number of tricks. Therefore, reducing the influence caused by the human factors of interest is helpful to the completeness of teacher professional assessment. In this paper we presented an unsupervised framework to identify non-teaching related abnormal evaluation of teachers. The results show that our method has the ability to detect IFAE initially.

Based on the results of the experiment in this paper, the following issues should be noted.

First, to measure the effectiveness of graph embedding, we additionally perform an extra experiment to examine the influence of prediction by varying the size of dimension of textual and graphic embedding. Shown in Fig 5, our proposed model derive the best performance of F1 when setting the graphical dimension to 150 and textual dimension to 200.

Based on the results, the contribution of different types of feature are different. In addition to previous discussion on the benefit of using textual feature to assist graphic feature, we exhibit a more intuitional understanding in Fig. 6. The long-tail distribution of words in the comments (Fig. 6 (a)) and overall short length of comments (Fig. 6 (b)) indicates that students are tend to give a brief straight but narrow expression on their taken courses. Therefore, although we require that each student provides a short comment on each course to be evaluated, the length as well as diversity of words are still limited which makes the text-based models overfit during the training process. On the other hand, the introduction of prior knowledge of each course not only improves the performance of the learning model, but also reflects the course selection in reality: Students do not choose courses randomly and independently of each other, and they will refer to the prior reputation and related requirements of the courses before choosing them.

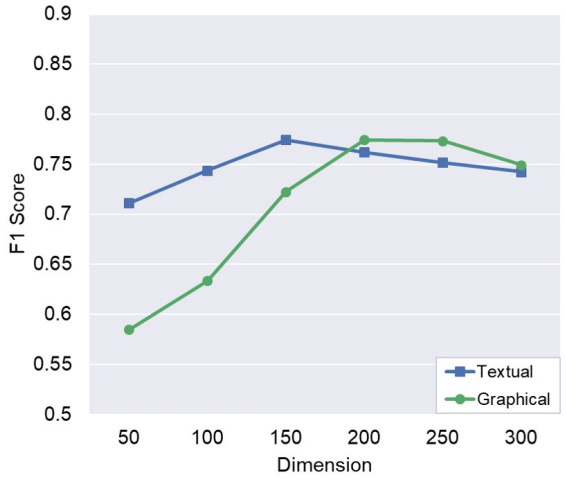

Fig. 5. The variation of F1 score by changing the dimension of text embedding or graph embedding.

From the aspect of IFAE detection itself, this paper only focuses on detecting abnormal evaluation in a holistic view. In other words, we took an assumption that only large-scale abnormal evaluation records from students towards a course can be thought as an overall abnormal evaluation on this course. According to the subsequent investigation on the courses which have large differences on evaluation between normal students and guest students, we believe the IFAE are often caused in an unaware way in both sides of the teacher and students. Therefore, the purpose of our proposed IFAE detection is to remind course teachers with suggestions on improving teaching styles and approaches, rather than catching misconducts.

There are several areas may benefit from this study. From the view of scholar evaluation, this work may help to bring a novel perspective towards a more comprehensive profiling of scholars such as academic ability prediction [33], academic resource recommendation [34], [35], educational fine-grained sentiment analysis [36], [37], etc. Additionally, students can also benefit from course recommendation to overcome information overload problems filtered by our model [38], [39] and a better understanding of students colonial preferences [40].

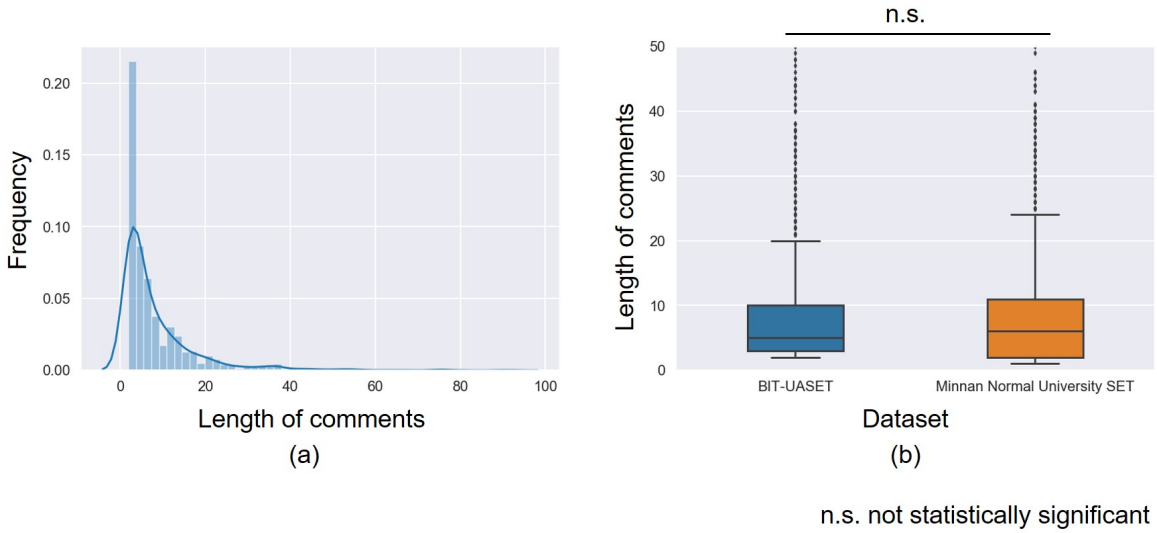

Fig. 6. The distribution characteristics of the texts in the experiment dataset. (a) The frequency of use of the top 100 high frequency words, a long-tail distribution can be clearly observed; (b) Comparison on the length of the evaluation comments between the dataset in this experiment and a similar dataset which used in [3].

Accompanied with these evaluation result, the construction of scholar's knowledge can be completed in a more comprehensive way, thereby forming an more intelligent educational knowledge service [41], [42].

However, several limitations should also be noted. First, although the suspected IFAE was predicted and validated by referring the evaluation score provided by guest students who have less interests related, the final verification still needs further artificial investigation. Therefore, the overall IFAE detection is still time and resource-consuming. In addition, unlike the attendance rate requirement of normal students, the retention rate is a non-negligible issue: if a guest student is unsatisfied with a certain course, he may quit this course or refuse to attend with leaving a bad evaluation on this course. The former action will directly reduce the data collection and the latter one may lead a false positive label because the student only takes the first few lessons. Whereas, if we filtered these evaluations out, a scenario that a course has plenty of bad comments will seldom happen (because few guest students want to continue with a desperate course), which is also harmful to the data complement. How to expand the data sources especially provide more objective IFAE label is still to be solved. Finally, our proposed approach requires multiple operation including preprocessing, network embedding training and learning. How to transfer these procedures into a more direct way or how to propose a end-to-end learning model is a problem to be addressed.

VI. Conclusion

Despite of the limitations, in this paper, we have made an initial effort towards the identification of interest-factor based abnormal evaluation which has grievous damage to the credibility of student evaluation of teaching. In this paper, we extracted the bidirectional scoring feature (i.e. teacher scores students and students evaluate teaching) and linguistic features from the process of evaluation of teaching. Then we utilized both textual features and graphical features by deriving their embedded learning representations and then introduce prior knowledge of each course into an aggregating network to predict suspected IFAE. Finally, experiment shows our method can reach a more effective performance than other state-of-art models.

Many points are valuable for further investigation: First, the feature engineering and feature extraction could be improved by fusing more data sources and modalities to gain a better feature representation. Second, this paper only focuses on a macroscopical IFAE detection, the microscopical one referring to individual student and teacher is still to be discovered. Finally, a wider and deeper data accessibility is suggested to examine the influence by academic hierarchical, cultural, and other social factors.

Acknowledgment

The authors would to thank the academic affairs Office of Minnan Normal University for accessing their student evaluation of teaching database. This work is supported by National Program on Key Basic Research Project (973 Program, Grant No. 2013CB329303, 2012CB720700) and the National Natural Science Foundation of China (Grant No. 61371194).

References

[1] J. C. Ory, "Teaching evaluation: Past, present, and future." *New directions for teaching and learning*, vol. 83, pp. 13–18, 2000.

[2] N. Valakunde and M. Patwardhan, "Multi-aspect and multi-class based document sentiment analysis of educational data catering accreditation process," in *2013 International Conference on Cloud & Ubiquitous Computing & Emerging Technologies*. IEEE, 2013, pp. 188–192.

[3] Q. Lin, Y. Zhu, S. Zhang, P. Shi, Q. Guo, and Z. Niu, "Lexical based automated teaching evaluation via students short reviews," *Computer Applications in Engineering Education*, vol. 27, no. 1, pp. 194–205, 2019.

[4] S. S. Boswell, "Ratemyprofessors is hogwash (but i care): Effects of ratemyprofessors and university-administered teaching evaluations on professors," *Computers in Human Behavior*, vol. 56, pp. 155–162, 2016.

[5] G. A. Boysen, "Revenge and student evaluations of teaching," *Teaching of Psychology*, vol. 35, no. 3, pp. 218–222, 2008.

[6] T. W. Maurer, "Cognitive dissonance or revenge? student grades and course evaluations," *Teaching of Psychology*, vol. 33, no. 3, pp. 176–179, 2006.

[7] W. Stroebe, "Why good teaching evaluations may reward bad teaching: On grade inflation and other unintended consequences of student evaluations," *Perspectives on Psychological Science*, vol. 11, no. 6, pp. 800–816, 2016.

[8] J. Bassett, A. Cleveland, D. Acorn, M. Nix, and T. Snyder, "Are they paying attention? students lack of motivation and attention potentially threaten the utility of course evaluations," *Assessment & Evaluation in Higher Education*, vol. 42, no. 3, pp. 431–442, 2017.

[9] T. M. Tripp, L. Jiang, K. Olson, and M. Graso, "The fair process effect in the classroom: Reducing the influence of grades on student evaluations of teachers," *Journal of Marketing Education*, p. 0273475318772618, 2018.

[10] N. D. Tran *et al.*, "Reconceptualisation of approaches to teaching evaluation in higher education," *Issues in Educational Research*, vol. 25, no. 1, p. 50, 2015.

[11] M. Goos and A. Salomons, "Measuring teaching quality in higher education: assessing selection bias in course evaluations," *Research in Higher Education*, vol. 58, no. 4, pp. 341–364, 2017.

[12] M. Sarwar, M. Dildar, A. A. Shah, and S. Hussain, "Relationship among students academic achievement, students evaluation of teacher and students evaluation of course," *The Dialogue*, vol. 12, no. 1, pp. 49–49, 2017.

[13] S. Mori and Y. Tanabe, "Influence of instructor personality on student evaluation of teaching: A comparison between english majors and non-english majors." *English Language Teaching*, vol. 8, no. 1, pp. 1–10, 2015.

[14] Y.-W. Hou, C.-W. Lee, and M. G. Gunzenhauser, "Student evaluation of teaching as a disciplinary mechanism: A foucauldian analysis," *The Review of Higher Education*, vol. 40, no. 3, pp. 325–352, 2017.

[15] F. Tian, B. An, D. Zheng, J. Qin, Q. Zheng, and Y. Yang, "E-learning oriented emotion regulation mechanism and strategies in interactive text applications," in *Proceedings of the 2011 15th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*. IEEE, 2011, pp. 702–709.

[16] P. Rodriguez, A. Ortigosa, and R. M. Carro, "Extracting emotions from texts in e-learning environments," in *2012 Sixth International Conference on Complex, Intelligent, and Software Intensive Systems*. IEEE, 2012, pp. 887–892.

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

[18] H. H. Binali, C. Wu, and V. Potdar, "A new significant area: Emotion detection in e-learning using opinion mining techniques," in *2009 3rd IEEE International Conference on Digital Ecosystems and Technologies*. IEEE, 2009, pp. 259–264.

[19] A. García-Pablos, M. Cuadros, and G. Rigau, "W2vlda: almost unsupervised system for aspect based sentiment analysis," *Expert Systems with Applications*, vol. 91, pp. 127–137, 2018.

[20] Z. Zhai, H. Xu, B. Kang, and P. Jia, "Exploiting effective features for chinese sentiment classification," *Expert Systems with Applications*, vol. 38, no. 8, pp. 9139–9146, 2011.

[21] D. Zhang, H. Xu, Z. Su, and Y. Xu, "Chinese comments sentiment classification based on word2vec and svmperf," *Expert Systems with Applications*, vol. 42, no. 4, pp. 1857–1863, 2015.

[22] G. G. Esparza, A. de Luna, A. O. Zezzatti, A. Hernandez, J. Ponce, M. Álvarez, E. Cossio, and J. de Jesus Nava, "A sentiment analysis model to analyze students reviews of teacher performance using support vector machines," in *International Symposium on Distributed Computing and Artificial Intelligence*. Springer, 2017, pp. 157–164.

[23] F. F. Balahadia, M. C. G. Fernando, and I. C. Juanatas, "Teacher's performance evaluation tool using opinion mining with sentiment analysis," in *2016 IEEE Region 10 Symposium (TENSYMP)*. IEEE, 2016, pp. 95–98.

[24] B. Dhanalakshmi and A. Chandrasekar, "Analyzing student's performance using efficient opinion mining and ranking method with machine learning techniques," *Journal of Computational and Theoretical Nanoscience*, vol. 15, no. 2, pp. 480–484, 2018.

[25] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 701–710.

[26] D. Wang, P. Cui, and W. Zhu, "Structural deep network embedding," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2016, pp. 1225–1234.

[27] A. Grover and J. Leskovec, "node2vec: Scalable feature learning for networks," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2016, pp. 855–864.

[28] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.

[29] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, 2014, pp. 1188–1196.

[30] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[31] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[32] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *arXiv preprint arXiv:1609.02907*, 2016.

[33] Y. Nie, Y. Zhu, Q. Lin, S. Zhang, P. Shi, and Z. Niu, "Academic rising star prediction via scholars evaluation model and machine learning techniques," *Scientometrics*, pp. 1–16, 2019.

[34] S. Wan and Z. Niu, "An e-learning recommendation approach based on the self-organization of learning resource," *Knowledge-Based Systems*, vol. 160, pp. 71–87, 2018.

[35] J. K. Tarus, Z. Niu, and G. Mustafa, "Knowledge-based recommendation: a review of ontology-based recommender systems for e-learning," *Artificial Intelligence Review*, vol. 50, no. 1, pp. 21–48, 2018.

[36] A. Yousif, Z. Niu, J. Chambua, and Z. Y. Khan, "Multi-task learning model based on recurrent convolutional neural networks for citation sentiment and purpose classification," *Neurocomputing*, vol. 335, pp. 195–205, 2019.

[37] K. Shi, C. Gong, H. Lu, Y. Zhu, and Z. Niu, "Wide-grained capsule network with sentence-level feature to detect meteorological event in social network," *Future Generation Computer Systems*, vol. 102, pp. 323–332, 2020.

[38] M. E. Ibrahim, Y. Yang, D. L. Ndzi, G. Yang, and M. Al-Maliki, "Ontology-based personalized course recommendation framework," *IEEE Access*, vol. 7, pp. 5180–5199, 2018.

[39] J. Zhang, B. Hao, B. Chen, C. Li, H. Chen, and J. Sund, "Hierarchical reinforcement learning for course recommendation in moocs," *Psychology*, vol. 5, no. 4.64, pp. 5–65, 2019.

[40] J. Chambua, Z. Niu, and Y. Zhu, "User preferences prediction approach based on embedded deep summaries," *Expert Systems with Applications*, vol. 132, pp. 87–98, 2019.

[41] P. Groth, "Increasing the productivity of scholarship: The case for knowledge graphs," in *Proceedings of the 24th International Conference on World Wide Web*. ACM, 2015, pp. 993–993.

[42] R. Manrique and O. Mariño, "Knowledge graph-based weighting strategies for a scholarly paper recommendation scenario." in *KaRS@ RecSys*, 2018, pp. 5–8.