Received: 8 June 2017 | Revised: 20 November 2017 | Accepted: 22 January 2018

DOI: 10.1111/jcal.12247

ORIGINAL ARTICLE

WILEY Journal of Computer Assisted Learning

Improving early prediction of academic failure using sentiment analysis on self-evaluated comments

L.C. Yu[1] | C.W. Lee[1] | H.I. Pan[1] | C.Y. Chou[2] | P.Y. Chao[3] | Z.H. Chen[4] | S.F. Tseng[1] | C.L. Chan[1] | K.R. Lai[2]

[1] Department of Information Management, Yuan Ze University, Taoyuan City, Taiwan

[2] Department of Computer Science and Engineering, Yuan Ze University, Taoyuan City, Taiwan

[3] Department of Information Communication, Yuan Ze University, Taoyuan City, Taiwan

[4] Graduate Institute of Information and Computer Education, National Taiwan Normal University, Taipei City, Taiwan

Correspondence
Liang-Chih Yu, Department of Information Management, Yuan Ze University, No. 135, Yuandong Road, Zhongli District, Taoyuan City 320, Taiwan.
Email: lcyu@saturn.yzu.edu.tw

Funding information
Ministry of Science and Technology, Grant/Award Numbers: MOST 105-2218-E-006-028 and MOST 105-2221-E-155-059-MY2

Abstract

This study presents a model for the early identification of students who are likely to fail in an academic course. To enhance predictive accuracy, sentiment analysis is used to identify affective information from text-based self-evaluated comments written by students. Experimental results demonstrated that adding extracted sentiment information from student self-evaluations yields a significant improvement in early-stage prediction quality. The results also indicate the limited early-stage predictive value of structured data, such as homework completion, attendance, and exam grades, due to data sparseness at the beginning of the course. Thus, applying sentiment analysis to unstructured data (e.g., self-evaluation comments) can play an important role in improving the accuracy of early-stage predictions. The findings present educators with an opportunity to provide students with real-time feedback and support to help students become self-regulated learners. Using the exploring results for improvement in teaching and learning initiatives is important to maintain students' performances and the effectiveness of the learning process.

KEYWORDS

early prediction, learning analytics, sentiment analysis, unstructured data

1 | INTRODUCTION

Academic performance estimation is important not only for helping students take control of their own learning and become self-regulated learners but also for allowing educators to identify students at risk and intervene early to reduce the risk of failure. Formative and summative assessments are generally adopted for academic performance estimation. Data used for such assessments contain valuable information, such as trends and patterns related to the learning process in educational practices, which can be used to understand learning status. However, such assessments are complex and time-consuming to implement, and important details may be overlooked. An alternative approach is to apply automatic computational techniques to analyse learner data and to facilitate and improve the assessment process.

With the emergence of big data analysis techniques and advances in computation, learning analytics has emerged as a growing research focus in the education domain. Learning analytics seeks to identify learning-related patterns in educational contexts and thereby improve understanding of educational processes, assess learning outcomes, and make performance predictions (Jayaprakash, Moody, Lauría, Regan, & Baron, 2014). Many useful learning analytics algorithms and methods have been developed including clustering, classification, decision trees, association rule mining, regression, and visualization, and these approaches have been widely applied in the development of models for predicting academic failure. Bayer, Bydzovská, Géryk, Obsivac, and Popelinsky (2012) focused on predicting dropouts and school failures using data based on students' social behaviours, such as student-to-student and student-to-teacher relationships and interactions identified by social network analysis. Such behaviours were then processed by various machine learning methods (e.g., SVM, naïve Bayes, and J48 decision tree) to improve prediction accuracy. Shahiri, Husain, and Rashid (2015) provided an overview on predicting final academic grades using various analytical techniques. Smith, Lange, and Huston (2012) developed predictive models using naïve Bayes classification to identify at-risk students based on dozens of key variables. These studies demonstrated effective predictive ability and identified key variables to accurately predict failure. However, these studies are not concerned with two critical aspects: (a) the importance of timing and (b) the relationship of student emotional state and academic success.

Clearly, early intervention is a key factor in preventing academic failure by at-risk students (Jayaprakash et al., 2014), and Santana, Costa, Fonseca, Rego, and de Araújo (2017) stressed the importance of predicting failure early enough to allow for pedagogical interventions to prevent academic failure. To achieve the early identification of at-risk students, Santana et al. (2017) sought to improve data quality using a sequence of data preprocessing steps such as data extraction, cleaning and transformation, and enhancing model effectiveness using an algorithm to select attributes that best fit the model. However, Pritchard and Wilson (2003) noted that academic success has traditionally been predicted using structured data such as demographic and academic variables, but these variables clearly do not account for all of the variation in academic success. To move beyond traditional predictors, valuable unstructured data sources need to be included to achieve early identification of at-risk students.

Self-evaluations written by students are a useful unstructured data source, containing rich emotional information that can help illuminate the emotional states of students. Sorour, Goda, and Mine (2015) emphasized the potential for student comments to reflect student learning attitudes, subject matter comprehension, and lesson difficulty, helping teachers adjust and refine student learning activities. Pekrun, Goetz, Titz, and Perry (2002) suggested that student emotions while attending class, studying, and taking tests are important variables in determining academic success. Several studies have confirmed the impact of emotional states on academic performance (D'Mello et al., 2008; Luo, Sorour, Goda, & Mine, 2015; Sorour, Mine, Goda, & Hirokawa, 2015). Blikstein and Worsley (2016) further indicated that affective state analysis (or sentiment analysis) is a novel multimodal learning analytics technique that can identify students' emotional states in studying and learning.

This study presents a model for the early identification of students who are likely to fail in an academic course. To enhance early predictive accuracy, sentiment analysis is used to identify affective information within text-based self-evaluations written by students. This affective information is then combined with structured data (e.g., attendance, homework completion, and in-class participation) as input to build a prediction model. Machine learning algorithms are used to determine whether the model can improve the prediction performance beyond using structured data alone. A total of 181 undergraduate students participated in experiments wherein they were asked to produce written self-evaluations after each lesson, yielding 2,476 comments. Each comment was manually annotated with a discrete score called a sentiment intensity or strength between 1 and 9, based on the Self-Assessment Manikin rating scale (Lang, 1980). A score of 1 denotes an extremely negative comment, whereas 9 denotes an extremely positive comment, and 5 denotes a neutral sentiment. As shown in Figure 1, a line chart is used to represent the student's emotions over time, allowing teachers to provide real-time feedback to students with fluctuating or consistently low moods and emotions. For example, a student comment mentioned that the student felt confused by excessive new terms used during the third week of class. This comment is given an emotional intensity score of 4, which indicates it is slightly negative. In reaction to this, the teacher could ask a teaching assistant to provide supplemental explanation before the next lesson.

To highlight useful information and support decision making, we developed a self-regulated system called the dynamic diagnostic and self-regulated (DDS) system using graphical techniques to achieve information visualization. In educational contexts, visual representations of course activities and usage records can provide educators and students with a general view of learning status (Romero & Ventura, 2010). The DDS system contains both structured data (e.g., attendance, homework completion, and in-class participation) and unstructured data (self-evaluated comments) collected from the university's proprietary virtual learning platform, which can provide real-time suggestions according to the learner learning status to promote learner autonomy.

The rest of this study is structured as follows. Section 2 presents prior work related to our proposed approach. Section 3 presents data sources and research methods used in this study. Section 3.2.1 summarizes the experimental results and introduces the proposed self-regulated system. Findings and conclusions are finally presented in Section 4.1.

2 | LITERATURE REVIEW

2.1 | Student emotions and learning

Students' emotions can profoundly affect their thoughts, motivations, and actions. D'Mello et al. (2008) emphasized that the detection of learner emotions is important in promoting effective learning due to the inextricable linkages between cognition and emotion. Positive emotions may increase student interest and motivation in learning. Pekrun et al. (2002) defined learner emotions as the emotional

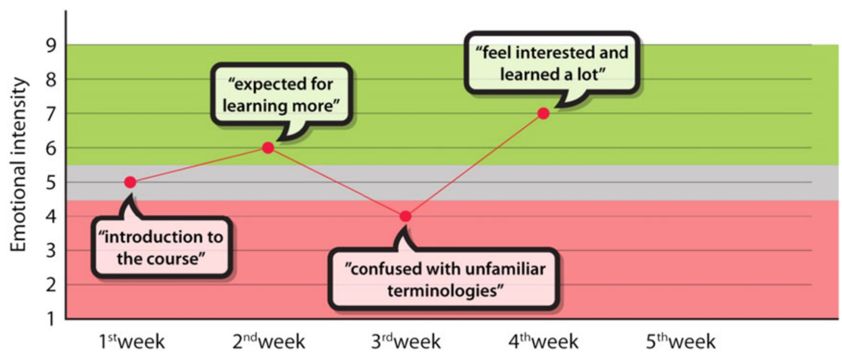

FIGURE 1 Student emotional intensity extracted from self-evaluated comment [Colour figure can be viewed at wileyonlinelibrary.com]

experience students have in academic settings when attending class, studying, and taking tests. Students' emotional states are significantly related to their motivation and academic achievement. For example, test anxiety is negatively related to learning and achievement. High levels of test anxiety stem from comparing self-performance to peers, considering the consequences of failure, and feeling unprepared for tests, all of which are associated with significantly lower test scores (Cassady & Johnson, 2002; Hembree, 1988). Pritchard and Wilson (2003) examined the relationship between emotional factors and student performance, finding that students who are emotionally healthy are more likely to succeed in college. Altrabsheh, Cocea, and Fallahkhair (2015) used several machine learning techniques to predict emotion types relevant for learning from students' textual feedback. Various pedagogical strategies can be used to keep students engaged in response to different emotion types. For example, bored learners should be given more challenging tasks, whereas confused learners require more examples to increase understanding.

2.2 | Sentiment analysis

Learning outcomes, motivations, and attitudes are typically evaluated using exams and by observing student performance in class. However, both approaches can be time and labour intensive, raising the need for a more efficient way to provide continuous and longitudinal assessments of learning effectiveness and attitudes. Luo et al. (2015) noted that comments written by students after each lesson reflect their learning attitudes towards that lesson, understanding of course contents, and perceived difficulty of leaning. They asked students to write comments following each lesson and then extracted information from these comments to continuously track student learning status. Automated text mining techniques can be effectively used to extract and analyse large quantities of textual data. Several studies (Calvo & D'Mello, 2010; Feldman, 2013; Yu et al., 2016) have used sentiment analysis to automatically identify affective information from texts, which can be used to automatically recognize student emotional states. In sentiment analysis, affective states are generally represented using either categorical or dimensional approaches.

The categorical approach represents affective states as several discrete classes such as positive, neutral, negative, and Ekman's (1992) six basic emotions (e.g., anger, happiness, fear, sadness, disgust, and surprise). On the basis of this representation, various practical applications have been developed such as aspect-based sentiment analysis (Pontiki et al., 2014; Schouten & Frasincar, 2016), Twitter sentiment analysis (Rosenthal et al., 2015; Saif, Fernandez, He, & Alani, 2013), deceptive opinion spam detection (Li, Ott, Cardie, & Hovy, 2014), and crosslingual portability (Banea, Mihalcea, & Wiebe, 2013; Xu, Gui, Xu, Lu, & Wong, 2015).

In contrast to the categorical approach, the dimensional approach represents affect states using a continuous numerical value for multiple dimensions. The most commonly used dimensions include the valence-arousal space, a two-dimensional approach where the valence dimension reflects the degree of positive and negative sentiment, and the arousal dimension reflects the degree of calm and excitement. Any sentiment expression can then be represented as a point in the valence-arousal coordinate plane by recognizing its valence-arousal ratings. Applications can use such representations to provide more fine-grained (e.g., real-valued) sentiment analysis (Kiritchenko & Mohammad, 2016). According to Yu et al. (2016), the arousal dimension is more difficult to predict than the valence dimension for both English and Chinese language texts. However, it is still useful because one-dimensional valence analysis focuses on predicting a real-valued score, called sentiment intensity or strength (Kiritchenko, Mohammad, & Salameh, 2016), which can provide both positive and negative directions and intensity levels for use in emotional analysis (Mishne & De Rijke, 2006; Nguyen, Phung, Dao, Venkatesh, & Berk, 2014).

2.3 | Information visualization

Simon (1996) noted that solving a problem simply means representing it so as to make the solution transparent. Information visualization allows users to assess and understand data at a glance. This branch of computer graphics and user interface design is concerned with the effective presentation of interactive or animated digital images (Spence, 2001). These techniques facilitate analysis of large amounts of information by generating a visual representation of the data. Information visualization of analysis results provides a general view of a student's learning achievement and supports effective decision making through highlighting useful information. Romero and Ventura (2010) emphasized the benefit of information visualization of analysed data in educational environments, especially for solving time-consuming interpretation process.

Learner data were collected and analysed to produce information visualizations. Teachers and learners were provided with statistical graphs summarizing student attendance, assignment completion, access to resources, exam scores, and emotional status to offer a general view of learning status.

3 | RESEARCH METHOD

3.1 | Data source

Data were obtained from 181 undergraduate students in two separate Introduction to Computer Science courses during the 19-week summer session at Taiwan's Yuan Ze University. Students were evaluated in terms of attendance, homework completion, in-class participation, and exams in the 5th, 11th, and 18th week of the course. A passing grade for the course is 60. The scores for the two classes were (mean, standard deviation) = (57.2, 15.9) for Class A and (51.8, 15.3) for Class B. Learning outcomes and grade distributions are shown in Table 1.

Following each lesson, the students were asked to write self-evaluations to elaborate on their personal leaning status. A total of 2,476 such comments were collected. Once duplicated or incomprehensible texts had been eliminated, the sample totaled 2,449 texts with a total of 126,318 Chinese characters (without punctuation), for an average of 51.58 characters per text. A total of 12 comments contained fewer than five characters, including “好!” (Good!); “有難度” (It is difficult to understand); “老師風趣” (The professor has a good sense of humor); and “學習到很多” (Learned a lot in this lesson). Despite their

TABLE 1 Learning outcomes and grade distributions

	Num. of students	Learning outcome		Scores	
		Num. of pass (score > 60)	Num. of fail (score < 60)	Mean score	Standard deviation
Class A	95	44	51	57.2	15.9
Class B	86	27	59	51.8	15.3
Total	181	71	110	54.6	15.8

brevity, these comments contained useful information on student learning status. At the other end of the spectrum, a total of 263 texts contained more than 100 characters. All self-evaluation comments indicated the degree of understanding and achievement of each student at each stage of the course.

The data source contains the following information for each student: attendance, homework completion, in-class participation, and student's emotions extracted from the self-evaluation comments. Details are shown in Table 2.

3.2 | Methods for analysis

3.2.1 | Emotion prediction

Sentiment analysis is used to extract student emotion intensity from the self-evaluated comments. Each comment was rated by three volunteer annotators from 1 to 9 in terms of valence based on the Self-Assessment Manikin rating scale (Lang, 1980). The rated values were then averaged to produce the comment's valence and categorized to three different emotion types as Negative (valence rating under 4.5), Neutral (valence rating between 4.5 and 5.5), and Positive (valence rating above 5.5). Table 3 shows the number of comments and mean valence for various emotion types according to each annotator. Comparing the emotion types rated by each annotator shows an averaged Cohen's kappa (Cohen, 1960) of 0.85 for interrater reliability.

In developing sentiment applications, affective lexicons and corpora with valence ratings are useful resources for opinion mining. Affective resources are usually obtained by either self-labelling or manual annotation (Yu et al., 2016). In the self-labelling approach, users proactively offer feelings and opinions in response to browsing web-based content. In the manual annotation method, trained annotators are asked to create affective annotations for specific language resources. We built a Chinese affective lexicon (Yu et al., 2016) with valence ratings for each affective word and used it to automatically extract emotions from self-evaluation comments. The valence rating of each comment was calculated by averaging the valence ratings of the words matched in the lexicon in that comment. For example, Figure 1 shows a student comment from the third week which notes, “I am confused by unfamiliar terminologies.” The valence rating of the affective word “confused” in the lexicon is 4, which means student experienced negative emotions after this lesson. In response to the abovementioned difficulty with new vocabulary, in Week 4, the student was provided with supplementary material, and the learner's subsequent self-evaluation featured the comment: “I feel interested and learned a lot.” The affective word “interested” in the lexicon has a valence rating of 7, indicating the student experienced more positive emotions after attending this lesson. If two or more affective words appear in the self-evaluated comments, the valence rating of the given comment is calculated by averaging the valence ratings of each word. Once the valence rating for each comment is obtained, the rating is converted to an emotion type classified as Negative (valence rating under 4.5), Neutral (valence rating between 4.5 and 5.5), or Positive (valence rating above 5.5).

TABLE 2 Data source

Attribute type		Attribute name	Description
Predictors	Structured data	Homework grade	The grades of homework in average
		Homework click	Number of reading homework instruction
		Attendance	Number of attending class
		Discussion	Number of postdiscussion
	Unstructured data of students' emotions	Positive valence	Number of positive valence
		Negative valence	Number of negative valence
		Neutral valence	Number of neutral valence
		Average of valence	Valence in average
Target		Academic performance	Failed or passed this course

TABLE 3 Results of annotation

Valence		Annotator 1		Annotator 2		Annotator 3		All (average)	
Emotion type	Valence range	Num. of comments (Total: 2,449)	Mean of valence	Num. of comments (Total: 2,449)	Mean of valence	Num. of comments (Total: 2,449)	Mean of valence	Num. of comments (Total: 2,449)	Mean of valence
Negative	<4.5	284	3.965	455	3.440	373	3.440	391	3.835
Neutral	4.5–5.5	1,596	4.995	1,162	5.000	1,331	5.000	1,316	5.037
Positive	>5.5	569	6.037	832	6.380	745	6.600	742	6.145

The performance of the lexicon-based approach to emotion prediction is evaluated using accuracy, defined as

$$\text{Accuracy} = \frac{\text{the number of comments whose emotion types are correctly identified}}{\text{the total number of comments}}. \quad (1)$$

3.2.2 | Academic performance prediction

Data analytic approaches for educational data, support vector machine (SVM), and convolutional neural networks (CNN) were applied to predict student performance. SVM is a binary classification technique based on supervised machine learning in the field of artificial intelligence (Drucker, Burges, Kaufman, Smola, & Vapnik, 1997). Basic SVM takes a set of input data for each given input and predicts which of two possible classes will form the output, making it a nonprobabilistic binary linear classifier (Cortes & Vapnik, 1995). There are many benefits to applying SVM in processing educational data, such as its flexibility with regard to interactions between educational parameters from different sources, short training times to create new models, and a better sensitivity for modeling dependent variables (Ifenthaler & Widanapathirana, 2014). In recent years, deep neural networks such as CNN, which employs convolution and pooling layers to capture useful dependencies between attributes, have achieved remarkable results in various classification tasks such as text classification (Kim, 2014) and image classification (Krizhevsky, Sutskever, & Hinton, 2012). In this study, we trained a simple CNN with one convolution layer based on both structured and unstructured attributes presented in Table 2.

For the early identification of at-risk students, we examine the predictive ability of our model in the fifth (first quarter), seventh, and ninth week (before midterm) of the semester. Class sessions prior to the fifth week were excluded because not enough time had passed to collect sufficient data. Classes following the ninth week were excluded because the resulting feedback would have come too late to help students improve their performance before the course ended.

The evaluation metrics, recall (as Equation 2), precision (as Equation 3), and *F*-measure (as Equation 4), are used to evaluate the performance. *F*-measure, as the harmonic mean of precision and recall, is widely used in domains such as information retrieval, machine learning, sentiment analysis, and many others that involve binary classification.

$$\text{Recall} = \frac{\text{number of failed (or passed) student identified by the method correctly}}{\text{total number of failed (or passed) student}}, \quad (2)$$

$$\text{Precision} = \frac{\text{number of failed (or passed) student identified by the method correctly}}{\text{total number of failed (or passed) student identified by the method}}, \quad (3)$$

$$F\text{−measure} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision} + \text{Recall}}. \quad (4)$$

4 | RESULTS AND VISUALIZATION

4.1 | Experimental results

4.1.1 | Emotion prediction

The abovementioned lexicon-based approach was used to determine the emotion types (Positive, Negative, and Neutral) of the 2,449 self-evaluation comments. The ground truth of the emotion types for these comments is presented in Table 3. Comparing the emotion types predicted by the lexicon-based approach and the ground truth generated by human annotators shows an accuracy of 0.76.

4.1.2 | Academic performance prediction

We used machine learning techniques and several meaningful attributes, including structured and unstructured data, to assess early-

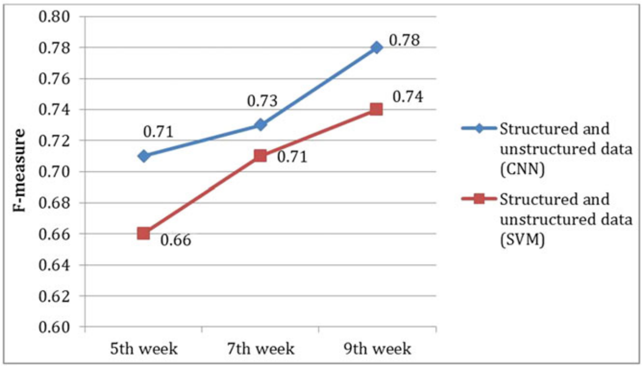

FIGURE 2 Early-stage predictive ability, comparative result. CNN = convolutional neural networks; SVM = support vector machine [Colour figure can be viewed at wileyonlinelibrary.com]

TABLE 4 Confusion matrix of the convolutional neural networks result at Week 9

		Predict	
		Fail	Pass
Actual	Fail	20	7
	Pass	9	37

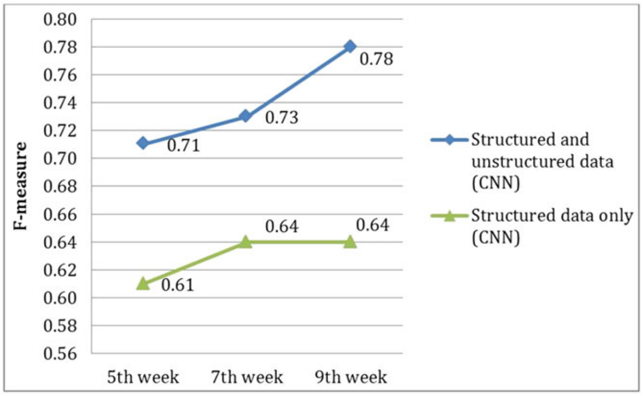

FIGURE 3 Comparative result of adding unstructured data. CNN = convolutional neural networks; SVM = support vector machine [Colour figure can be viewed at wileyonlinelibrary.com]

stage predictive ability of student performance in Weeks 5, 7, and 9 in a 19-week semester. As shown in Table 2, the structured data includes student attendance, in-class participation, and homework completion, whereas the unstructured data are taken from the emotional content of student comments as manually evaluated by annotators. Data for the 181 students were randomly divided into training and test sets at a ratio of 6:4 (108:73 students).

The experimental results show the proposed model provides an effectiveness (represented by *F*-measure) of 0.66 (SVM) and 0.71 (CNN) at Week 5, improving to 0.71/0.73 at Week 7, and 0.74/0.78 at Week 9. The results in Figure 2 show that the proposed model presented the best effectiveness on CNN, achieving an *F*-measure value of 0.78 before the midterm. Table 4 shows the confusion matrix of the CNN at Week 9.

To investigate the effectiveness of sentiment analysis, we examined the differences of *F*-measure with and without unstructured data using the CNN algorithm. The experimental results show the *F*-measure rises from 0.61 at Week 5 to 0.64 at Weeks 7 and 9 using structured data only. Results comparison in Figure 3 shows the improvement of *F*-measure by adding unstructured data and

TABLE 5 Experiment results of CNN and SVM

		CNN results *F*-measure (accuracy)			SVM results *F*-measure (accuracy)		
		5th week	7th week	9th week	5th week	7th week	9th week
Structured data only		0.61 (0.616)	0.64 (0.630)	0.64 (0.630)	0.49 (0.630)	0.66 (0.698)	0.60 (0.643)
Structured and unstructured data		0.71 (0.712)	0.73 (0.739)	0.78 (0.780)	0.66 (0.657)	0.71 (0.712)	0.74 (0.739)
T test	*n*	73	73	73	73	73	73
	t-values	3.604	2.742	1.832	−6.693	−2.917	−2.917
	p (two tailed)	.000	.007	.069	.000	.004	.004

Note. Data for the 181 students were randomly divided into training and test sets at a ratio of 6:4 (108:73 students). CNN = convolutional neural networks; SVM = support vector machine.

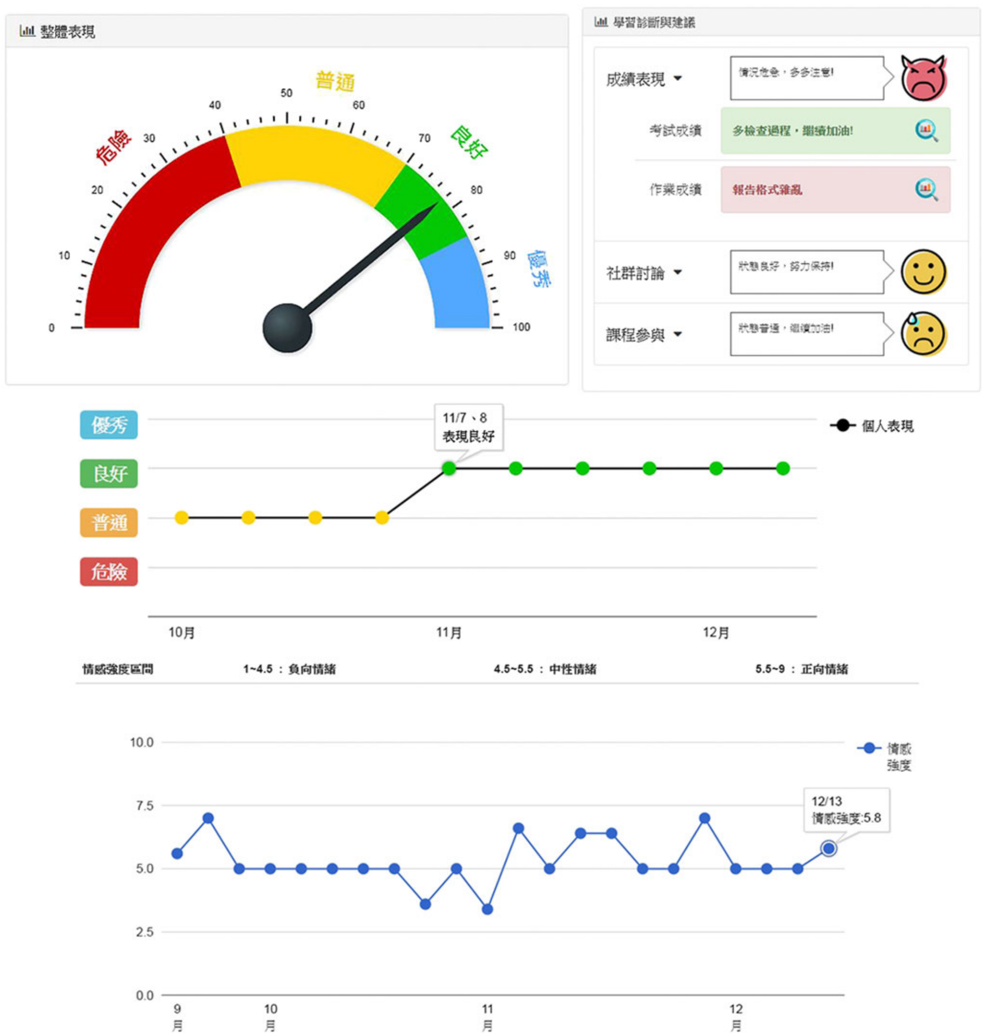

FIGURE 4 Dynamic diagnostic and self-regulated (DDS) system (dashboard on top left; diagnostic and suggestion report on top right; weekly status on the middle; and emotional valence value on the bottom) [Colour figure can be viewed at wileyonlinelibrary.com]

demonstrates the effectiveness of using self-evaluation comments in early prediction.

A paired, two-tailed *t* test was used to compare the predictive results of using structured data alone and with unstructured data. Table 5 lists the *F*-measure, accuracy, and significance level of the *t* test related to CNN and SVM. The results present are statistically significant at the .05 level, except for Week 9 of CNN (p = .069). This finding indicates that the use of the students' emotions extracted from the unstructured self-evaluated comments significantly improved the classifier's predictive ability, especially for earlier stages (e.g., Weeks 5 and 7). One possible reason is that, at such earlier stages, the structured data are relatively too limited to provide sufficient information for prediction; thus, the unstructured comments have a more pronounced effect in improving prediction performance. Conversely, by the later stage (e.g., Week 9), sufficient structured data have been accumulated. Timely recognition of students' emotional states can help educators better understand the learning status of all class members, allowing them to quickly adjust their pedagogical strategies for maximum effectiveness.

4.2 | Information visualization

A DDS system is developed on the basis of the structured and unstructured data used in the experiments. As shown in Figure 4, the DDS system provides a visualized diagnostic report based on fine-grained learning status information. Reports include four components: a dashboard providing a comprehensive assessment overview; diagnostics and performance-based recommendations for activity-specific instruction; weekly status reports; and emotional valence values for student emotional status following each lesson.

The dashboard uses different colours to represent student learning status as follows: blue (*excellent*); green (*good*); yellow (*fair*); and red (*at-risk*). This visual representation provides learners and instructors with personal learning status information at a glance. Diagnostics and recommendations are provided in three categories: grades (including exams and homework), discussion on social media, and in-class participation. Each category is represented by an icon that students can click to receive recommendations and additional performance-related information from teaching assistants or teachers. For example, in Figure 4, the text recommendation for "exams" suggests that the student recheck the computation process for mathematical questions, and that for "homework" suggests that the student improve the messy report format. The recommendations for each category help students understand their performance evaluation and to respond accordingly. Weekly status is presented longitudinally in a line graph that can be used to track changes in student performance over the course of the semester. The emotional valence value chart helps track student learning status trends, with downward shifts in emotional status triggering an alert. Time-relevant self-evaluation comments can be accessed by right-clicking anywhere along the line. This chart can help students reflect on their emotions related to the course in order to improve engagement, homework completion, and performance in tests.

5 | FINDINGS AND CONCLUSIONS

Self-evaluation comments, a form of unstructured data, provide valuable information that can represent student learning status over the duration of a course of instruction. Students' emotions, extracted from such comments, reflect their learning attitudes towards the lesson, understanding of course contents, and difficulties in leaning. This study applies sentiment analysis to self-evaluation comments to better identify at-risk students. Experimental results demonstrated that adding extracted sentiment information from students' self-evaluation comments significantly improves early prediction accuracy. The results also indicate the limited early-stage predictive ability of structured data (e.g., attendance, homework completion, and in-class participation) due to data sparseness at the beginning of the course. Thus, applying sentiment analysis to unstructured data (e.g., self-evaluation comments) can significantly improve the accuracy of early-stage predictions.

The main contributions of this study are twofold. First, the early identification of students who are likely to fail is critical to the success of any pedagogical intervention intended to change the trajectory of an at-risk student. Jayaprakash et al. (2014) suggested that the likelihood of improvement diminishes significantly as time goes on. Applying unstructured data of students' emotions to improve early-stage prediction accuracy can allow the educator and student to address the issue in time to rescue the student's final grade. Second, the proposed DDS system provides valuable support for both educators and students. Pekrun et al. (2002) noted that positive emotions foster self-regulation among students, whereas negative emotions lead to reliance on external guidance. The diagnostic and suggestion reports in DDS system provide useful information for self-regulated learning with positive emotions, and the line chart of the valence value provides instructors with important information for the application of external guidance to students with negative emotions. Our findings allow educators to provide students with real-time feedback and to enhance learner autonomy.

This study suffers from two limitations. First, the study is based on a small sample of students from two sections of a single course at a single university, and the additional work is needed before the findings can be generalized to other groups. Second, the lexicon-based approach used for emotion prediction does not take into account negation (e.g., student is not confused). However, only 5% (115 of 2,449) of all comments in the experimental data set contained negation. Future work will focus on applying the model to larger data sets from multiple universities and course types and improving the effectiveness and usability of the DDS system for both students and educators.

ACKNOWLEDGEMENTS

This work was supported by the Ministry of Science and Technology, Taiwan, ROC, under Grants MOST 105-2221-E-155-059-MY2 and MOST 105-2218-E-006-028. The authors would like to thank the anonymous reviewers and the area chairs for their constructive comments.

ORCID

L.C. Yu http://orcid.org/0000-0003-1443-4347
C.W. Lee http://orcid.org/0000-0001-9098-0660

REFERENCES

Altrabsheh, N., Cocea, M., & Fallahkhair, S. (2015). Predicting learning-related emotions from students' textual classroom feedback via Twitter.

Banea, C., Mihalcea, R., & Wiebe, J. (2013). Porting multilingual subjectivity resources across languages. *IEEE Transactions on Affective Computing*, *4*(2), 211–225.

Bayer, J., Bydzovská, H., Géryk, J., Obsivac, T., & Popelinsky, L. (2012). Predicting drop-out from social behaviour of students. International Educational Data Mining Society.

Blikstein, P., & Worsley, M. (2016). Multimodal learning analytics and education data mining: Using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, *3*(2), 220–238.

Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, *1*(1), 18–37.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety and academic performance. *Contemporary Educational Psychology*, *27*(2), 270–295.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297.

D'Mello, S., Jackson, T., Craig, S., Morgan, B., Chipman, P., White, H., … Picard, R. (2008). AutoTutor detects and responds to learners affective and cognitive states. Paper presented at the Workshop on emotional and cognitive issues at the International Conference on Intelligent Tutoring Systems.

Drucker, H., Burges, C. J., Kaufman, L., Smola, A., & Vapnik, V. (1997). Support vector regression machines. *Advances in Neural Information Processing Systems*, *9*, 155–161.

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3–4), 169–200.

Feldman, R. (2013). Techniques and applications for sentiment analysis. *Communications of the ACM*, *56*(4), 82–89.

Hembree, R. (1988). Correlates, causes, effects, and treatment of test anxiety. *Review of Educational Research*, *58*(1), 47–77.

Ifenthaler, D., & Widanapathirana, C. (2014). Development and validation of a learning analytics framework: Two case studies using support vector machines. *Technology, Knowledge and Learning*, *19*(1–2), 221–240.

Jayaprakash, S. M., Moody, E. W., Lauría, E. J., Regan, J. R., & Baron, J. D. (2014). Early alert of academically at-risk students: An open source analytics initiative. *Journal of Learning Analytics*, *1*(1), 6–47.

Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

Kiritchenko, S., & Mohammad, S. M. (2016). Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling. Paper presented at the Proceedings of NAACL-HLT.

Kiritchenko, S., Mohammad, S. M., & Salameh, M. (2016). SemEval-2016 task 7: Determining sentiment intensity of English and Arabic phrases. *Proceedings of SemEval*, 42–51.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Paper presented at the Advances in neural information processing systems.

Lang, P. J. (1980). Behavioral treatment and bio-behavioral assessment: Computer applications.

Li, J., Ott, M., Cardie, C., & Hovy, E. H. (2014). Towards a general rule for identifying deceptive opinion spam. *Paper presented at the ACL*, *1*.

Luo, J., Sorour, S. E., Goda, K., & Mine, T. (2015). Predicting student grade based on free-style comments using Word2vec and ANN by considering prediction results obtained in consecutive lessons. International Educational Data Mining Society.

Mishne, G., & De Rijke, M. (2006). Capturing global mood levels using blog posts. Paper presented at the AAAI spring symposium: Computational approaches to analyzing weblogs.

Nguyen, T., Phung, D., Dao, B., Venkatesh, S., & Berk, M. (2014). Affective and content analysis of online depression communities. *IEEE Transactions on Affective Computing*, *5*(3), 217–226.

Pekrun, R., Goetz, T., Titz, W., & Perry, R. P. (2002). Academic emotions in students' self-regulated learning and achievement: A program of qualitative and quantitative research. *Educational Psychologist*, *37*(2), 91–105.

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). Semeval-2014 task 4: Aspect based sentiment analysis. Proceedings of SemEval, 27–35.

Pritchard, M. E., & Wilson, G. S. (2003). Using emotional and social factors to predict student success. *Journal of College Student Development*, *44*(1), 18–28.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(6), 601–618.

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S. M., Ritter, A., & Stoyanov, V. (2015). Semeval-2015 task 10: Sentiment analysis in Twitter. Paper presented at the Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015).

Saif, H., Fernandez, M., He, Y., & Alani, H. (2013). Evaluation datasets for Twitter sentiment analysis: A survey and a new dataset, the STS-Gold.

Santana, M. A., Costa, E. B., Fonseca, B., Rego, J., & de Araújo, F. F. (2017). Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2017.01.047

Schouten, K., & Frasincar, F. (2016). Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, *28*(3), 813–830.

Shahiri, A. M., Husain, W., & Rashid, N. a. A. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *72*, 414–422.

Simon, H. A. (1996). *The sciences of the artificial* (3rd edn, orig. publ., 1969 ed.). Cambridge, MA: MIT Press.

Smith, V. C., Lange, A., & Huston, D. R. (2012). Predictive modeling to forecast student outcomes and drive effective interventions in online community college courses. *Journal of Asynchronous Learning Networks*, *16*(3), 51–61.

Sorour, S. E., Goda, K., & Mine, T. (2015). Evaluation of effectiveness of time-series comments by using machine learning techniques. *Journal of Information Processing*, *23*(6), 784–794.

Sorour, S. E., Mine, T., Goda, K., & Hirokawa, S. (2015). A predictive model to evaluate student performance. *Journal of Information Processing*, *23*(2), 192–201.

Spence, R. (2001). *Information visualization (vol. 1)*. Springer.

Xu, R., Gui, L., Xu, J., Lu, Q., & Wong, K.-F. (2015). Cross lingual opinion holder extraction based on multi-kernel SVMs and transfer learning. *World Wide Web*, *18*(2), 299–316.

Yu, L.-C., Lee, L.-H., Hao, S., Wang, J., He, Y., Hu, J., … Zhang, X. (2016). Building Chinese affective resources in valence-arousal dimensions. Paper presented at the Proceedings of NAACL-HLT.

How to cite this article: Yu LC, Lee CW, Pan HI, et al. Improving early prediction of academic failure using sentiment analysis on self-evaluated comments. *J Comput Assist Learn*. 2018;1–8. https://doi.org/10.1111/jcal.12247