Predicting Student Performance Using Teacher Observation Reports

Menna Fateen
Kyushu University
menna.fateen@m.ait.kyushu-u.ac.jp

Tsunenori Mine
Kyushu University
mine@ait.kyushu-u.ac.jp

ABSTRACT

Studying for entrance examinations can be a distressing period for numerous students. Consequently, many students decide to attend cram schools to assist them in preparing for these exams. For such schools and for all educational institutes, it is necessary to obtain the best tools to provide the highest quality of learning and guidance. Performance prediction is one tool that can serve as a resource for insights that are valuable to all educational stakeholders. With accurate predictions of their grades, students can be further guided and fostered in order to achieve their optimal learning goals. In this regard, we target middle school students to be able to guide them on their educational journey as early as possible. We propose a method to predict the students' performance in entrance examinations using the comments that cram school teachers made throughout the lessons. Teachers in cram schools observe their student's behavior closely and give reports on the efforts taken in their subject material. We show that the teachers' comments are qualified to construct a tool that is capable of predicting students' grades efficiently. This is a new method because previous studies focus on predicting grades mainly using student data such as their reflection comments or earlier scores. Experimental results show that using readily available feedback from teachers can remarkably contribute to the accuracy of student performance prediction.

Keywords

text mining, student grade prediction, teacher observation reports, machine learning

1. INTRODUCTION

"If you could reinvent higher education for the twenty-first century, what would it look like?". A question like this one invites many observations about the advantages and issues that the current state of higher education has in the world. As a matter of fact, this question has been addressed specifically by the founders of the Minerva Schools at KGI [1] in the United States. At such innovative universities and schools, active learning and student engagement with the material are highly encouraged [2, 3, 4, 5, 6]. Additionally, the student/teacher ratio is expected to be lower than in traditional schools for higher teacher effectiveness [7]. Students are assessed and observed closely by their teachers and they can receive written feedback from their teachers daily. These reports clarify any confusion, reinforce strong points and give more specific advice and guidance [8, 9]. Besides, since teachers frequently engage with students, research has proven that these teachers, especially those with professional development, can accurately judge and forecast their students' computational skills [10].

In this paper, we propose a novel method for predicting students' performance or final grades. We show that we can use reports carefully written by teachers that closely observe the students, to construct a grade prediction model. If these predictions can be made accurately, it would be an invaluable resource to help the teachers better regulate their students' learning. Future performance prediction is considered a powerful means that can provide all educational stakeholders with insights that are beneficial to them. Many grade prediction models have been proposed by researchers in the last decade [11, 12, 13], but no model has used teacher reports as far as we know. The teacher reports we use are provided by a cram school in Japan. Cram schools are specialized in providing extra and more attentive education for students who want to achieve certain goals, particularly studying for high school or university entrance exams [14]. To capture the meanings of the teacher reports, we obtain vector representations by applying the term-frequency inverse document-frequency (TF-IDF) method and extracting BERT embeddings. Our model uses these vectorized reports as the explanatory variables for a Gradient Boosting regressor. The regressor then predicts the students' scores. Our experiment results show that when adding teachers' reports to the regular student exam scores, we can predict their letter grade with an accuracy up to 62%. To sum up, our contributions can be outlined as follows:

- We propose a new performance prediction method using teacher observation reports represented using TF-IDF and BERT.
- We conducted 2 main different models of prediction and compared the experiment results to show that using teacher reports has the potential to contribute to an increase in accuracy of grade prediction models.

All in all, to the best of our knowledge, this is the first study to use NLP to mine teacher observation comments

to predict student grades. Our research and experimental results demonstrate the potential that these unstructured teacher observation comments have in predicting students' total scores and final letter grades.

2. RELATED WORK

The utilization of data mining and machine learning or deep learning tools to construct predictive models are increasingly being adopted in many different fields [15, 16]. Needless to say, the educational field has not been an exception. Topics in educational data mining vary widely from course recommendation systems [17] to automatic assessment [18]. More specifically, an extensive amount of studies have been dedicated to prediction modeling whether it be predicting student grades or performance such as next-term grade prediction [19] or student dropout. These prediction models are essential since they underlie applications to important educational AI-based decision-making systems [20]. With accurate predictions, the performance of students can be monitored using these systems, and students that have difficulties in their studies can easily be detected and given further guidance early on.

Over the past years, several methods have been developed to predict student's performance using Natural Language Processing (NLP) techniques. It has been proven that mining unstructured text using NLP has the capacity to contribute to accurately predicting students' success over the information obtained from usual fixed-response items [21]. Luo et. al [13] proposed a method to predict student grades based on their free-style reflection comments collected after each lesson. The comments were collected according to the PCN method [22] that categorizes the students' comments. To represent the students' reflection comments, Word2Vec embeddings were adopted followed by an artificial neural network. Their experiments show a correct rate of 80%. Teacher or advisor notes have been used by Jayaraman, not to predict student grades, but to detect students that are at risk of dropping out of college [23]. In their study, they use sentiment analysis to extract the positive and negative sentiment from the advisors' notes and use those as features to train a model. The model achieves 73% accuracy at identifying at-risk students.

3. DATA DESCRIPTION

The dataset obtained and used for our model was provided by a cram school in Fukuoka, Japan. To ensure confidentiality, no student names or other identifying data were presented. Reports were obtained monthly and sent as CSV files. Since our model is focused on predicting the performance of students in their entrance examinations, we focused on those students in their final year of middle school. The final dataset after preprocessing composed of 11,960 reports over the period from May to October for 159 students.

3.1 Monthly Reports

In addition to the student ID and the class date, each report also consisted of the subject code, the teacher's comments, understanding, attitude and homework scores. More data in the reports were also provided but were unstructured and considered redundant for the prediction model. The features that were extracted from the reports and used in the study are discussed in more detail in Section 4.1. However our main explanatory variable used in the study is the teachers' observation comments written in Japanese. The average length of these comments is 96 characters. In addition, by analyzing comments, it was observed that teachers tend to encourage and energize their students by using words such as "better" and "work on". Moreover, the words used in the comments depend on the context or class subject to some degree. For example, the expression "calculation problem" is likely to be used in math lessons.

Table 1: Number of Reports in Each Subject

	Japanese	Math	Science	Social	English
Number of Reports	1157 (9.7%)	3547 (29.6%)	2428 (20.3%)	1669 (14%)	3159 (26.4%)

In the cram school, students take different lessons for each subject. These lessons fall under the 5 main subjects: Japanese, Mathematics, Science, Social Studies and English. Since the main objective of our model is to predict a student's total score, reports in all 5 subjects are required. Therefore, testing the model was only possible for those students who attended classes for all subjects. The number of reports that fall under each subject are shown in Table 1. The values in the table show that the most taken lessons and therefore the most reports provided were in the subject of Mathematics followed directly by English. The number of total reports for each student varied depending on the classes attended. The average number of total reports recorded for each student was 82 reports with a maximum of 206 and a minimum of 24 reports.

3.2 Test Scores

Students attending the cram school were naturally registered in many different schools. The results of their regularly taken examinations at school were recorded and provided. These scores were what we considered student data and would be traditionally used as the main feature to predict their performance in the entrance exam. To teach the model to perform these predictions, we adopted the supervised learning method. In supervised learning, training data needs to be labeled with the required outputs for each input. This enables the model to train its learning function by altering it based on the correct result so that the function can then be applied to new inputs. In our study, we used the students' results in their cram school simulation exams as the labels for the model since their actual performance in the entrance exam was unattainable.

The simulation scores for the 159 students were recorded for all subjects and also provided as the total score. To visualize the distribution of the students' scores, histograms were plotted as shown in Figure 1. The shape of the graph for the subject scores distribution and total score distribution is approximately bell-shaped and seems symmetric about the mean, so it is assumed that the scores follow the normal distribution. The standard deviation, σ, for all scores are displayed in Table 2 to show how dispersed the values are.

4. METHODOLOGY

4.1 Feature Selection

Figure 1: Distribution of Simulation Test Scores

Table 2: Standard deviation of subject scores

	Japanese	Math	Science	Social	English	Total
σ	11.85	16.62	20.33	16.93	18.47	70.01

For our experimental settings, we adopt 3 main feature sets for the sake of comparison. The first feature set, FS_1, consists of using teachers' report contents as the main explanatory variables. A teacher's report in one lesson evaluating the student consists of 1-Comments 2- Understanding Score 3- Attitude Score and 4-Homework Score. We use all of these attributes except for the homework score. This is mainly because more than 36% of the reports did not include homework scores since not all lessons necessarily require homework. After each lesson, the teacher writes some comments based on their observations, assesses the student on their understanding giving them a score of either (0-30-60-80-100) and an attitude score of either (1-2-3-4). The second feature set, FS_2, consists of student-related data only, specifically their gender and the score of their regularly scheduled exam at school. Since we predict each subject score separately, the regular score corresponds to the subject score. As for the students' gender, the Pearson correlation coefficient between it and the score is 0.12 while the correlation coefficient between the regular score and the simulation score is 0.80 which suggests that the important factor in FS_2 is essentially the student regular score and not the gender. Finally, we investigate using both teachers' reports and the regular student scores to verify whether adding teachers' reports contributes to the accuracy of the prediction model or not. The third feature set, FS_3, is essentially a concatenation of FS_1 and FS_2. A sample of FS_1 is shown in Table 3.

4.2 Natural Language Processing

There are numerous ways to represent text data for a machine learning model to convey the original meanings of the text and prevent information loss. In our experiments, we chose to represent the teachers' comments using two techniques. We used the traditional TFIDF vectorization method and compared it with BERT embeddings.

4.2.1 TF-IDF

The first essential step in transforming text into a numerical representation is preprocessing the text. This step begins with tokenization or splitting the sentences into words. Tokenization in languages such as English can be done by splitting the sentence strings at each space. However, for Japanese, this step is merged with the next, which is morphological analysis, since there are no spaces in Japanese sentences. We use the fugashi [24] parser for this step, which is essentially a wrapper for Mecab[1], a Japanese tokenizer and morphological analysis tool. Our parser extracts from each report the following parts of speech: nouns, verbs, auxiliary verbs, adjectives and adverbs. We use the corresponding terms to these extracted parts of speech to build a bag-of-words vector with weights given by the TF-IDF method implemented by sklearn [25]. Since the teachers' comments are given in Japanese, we provide the mentioned parser to the tokenizer parameter. We also give a list of predefined Japanese stop words to the vectorizer.

4.2.2 BERT

BERT or Bidirectional Encoder Representations from Transformers is a new method of pre-training language representations presented by Google [26]. BERT obtains state-of-the-art results on many NLP tasks. It is a Transformer Encoder stack that pre-trains language representations. A pre-trained BERT model is basically a general purpose language understanding model trained on a large corpus which can then be used for downstream tasks. The BERT model we used for the comments was pretrained by Inui Laboratory, Tohoku University[2]. The corpus they used for pretraining was Japanese Wikipedia and the model was trained with the same configuration as the original BERT. In the experiments shown in this paper, we used the BERT [CLS] token embeddings as our BERT embeddings.

4.3 Evaluation Metrics

To evaluate our experiments, we use the Mean Absolute Error (MAE) metric. The MAE is calculated using the following formula :

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |\text{score}_{pred,i} - \text{score}_{true,i}| \quad (1)$$

where $\text{score}_{true,i}$ is the actual score that student i obtained. The predicted score ($\text{score}_{pred,i}$) is calculated differently for subject scores and total score. For a specific subject $s \in$ S, where S = {Japanese, Math, Science, Social Studies, English}, a student i can attend a variable number t of lessons. Therefore, to predict the subject score ($\text{SubjectScore}_{pred,i,s}$) of student i we use each of their reports as independent inputs to the model and obtain an ordered list $X_{i,s,t}$ of pre-

[1] `https://taku910.github.io/mecab/#parse`
[2] `https://github.com/cl-tohoku/bert-japanese`

Table 3: A sample of FS_1: teachers' reports (comments originally in Japanese)

Understanding	Attitude	Comments
80	4	We are trying applied problems of resolution into factors. You look like making many mistakes carelessly, but know formulas very well.
80	4	We are trying applied problems of resolution into factors. You look like making many mistakes carelessly, but know formulas very well.
100	4	He took notes while watching the commentary and focused on the problem. If you keep going at this rate, you will be able to meet the target, the 5th time. So, let's do our best!

dicted scores for student$_i$. The estimated score for the subject is then decided using:

$$\text{SubjectScore}_{pred,i,s} = Med(X_{i,s,t}) = \begin{cases} X_i[\frac{t}{2}], & \text{if t is even} \\ \frac{1}{2}(X_i[\frac{t-1}{2}] + X_i[\frac{t+1}{2}]), & \text{if t is odd} \end{cases} \quad (2)$$

To measure the central tendency, we used the median rather than the mean as it is robust to skewness and outliers. Nevertheless, if the estimations follow a normal distribution, the median would be close to the mean. The total predicted score ($\text{TotalScore}_{pred,i}$) can then be estimated by:

$$\text{TotalScore}_{pred,i} = \sum_{s \in S} \text{SubjectScore}_{pred,i,s} \quad (3)$$

Finally, since students receive letter grades for their total score, we map the estimated total score to its closest corresponding letter grade according to the percentages shown in Table 4 [27]. We then compute the percentage of grades that are x ticks away from their actual grades. A tick, as specified by [28], is defined as the difference between two successive letter grades. We name this metric percentage by tick accuracy or PTA. PTA_0 stands for the Percentage by 0 Tick Accuracy which means the model successfully predicted the letter grade with no error while PTA_1 is the percentage of incorrectly predicted grades but are 1 tick away from the true letter grade (e.g. A vs B). A similar metric was used in previous studies regarding grade prediction models [11, 28].

Table 4: Letter grades and their corresponding percentages

Grade	S	A	B	C	D	F
%	90-100	80-89	70-79	60-69	50-59	0-49

5. EXPERIMENTS

5.1 Model Overview

In our experiments, we adopt gradient boosting, a composite machine learning algorithm. We employed its sklearn implementation, `GradientBoostingRegressor` [25] to predict the continuous value of the students' scores in each subject. Since there is no prior research on the effect of using teacher observation reports in predicting students' grades, we use the following method as the baseline in our experiment. At first, subject codes were unavailable for each teacher observation record. Therefore, we constructed a model that used all of each student's reports, regardless of the subject, to directly predict and estimate the total score according to Equation 2. We call this model, the 'Direct' model. Subject codes then became accessible and we were able to map each report to its corresponding subject. Leveraging that, we created a separate regression model for each subject's reports and estimated the total score as shown in Equation 3. This model is called the 'Subjects' model.

5.2 Experimental Results

All experiments in the study were evaluated using group 10-fold cross-validation. The advantage of group k-fold cross validation method is that all data are used for both training and testing, and each instance is used for testing once. This is especially useful in situations where data is limited. Since the dataset comprises reports for 159 students, we used 143 students' reports for each fold as the training set and 16 as the testing set. The number of reports or instances for each subject model, therefore, varied depending on how many lessons each student had attended. The average MAE, which is calculated as in Equation 1, of all ten folds was computed and used as the main evaluation metric. We ran the baseline Direct model with the 3 feature sets described in Section 4.1. Teachers' comments were represented using BERT embeddings. The performance results are shown Using all 3 feature sets, the Subjects model consistently outperforms the Direct baseline model. Specifically, predicting the total score using the Subjects model with FS_3, which uses both teachers' reports and student data, resulted in a decrease in MAE of 5.62. Using teachers' reports alone (FS_1) resulted in a comparatively higher MAE in both models. However, adding teachers' reports to student data (FS_3) showed a smaller value in MAE than using student data only (FS_2) which suggests that teachers' reports as features can contribute to the accuracy of the grade prediction model.

Table 6 shows the MAE, PTA_0 and PTA_1 of each subject's score prediction model. We ran the subject model with all 3 feature sets. For FS_1 and FS_3, we compared the performance of the two text representations, TF-IDF vectors and BERT embeddings. Values in bold indicate the leading scores for each metric in all subjects. In terms of MAE, using FS_3 consistently outperforms the other feature sets. It can also be seen that BERT embeddings tend to have better overall

Table 5: Average MAE of total score prediction with Direct model vs Subject model using the 3 feature sets: FS_2: student data, FS_1: teacher reports, FS_3: FS_1 + FS_2

	FS_2	FS_1	FS_3
Direct	42.73	53.81	38.91
Subjects	**36.83**	**52.02**	**33.29**

Table 6: Evaluation metric scores in all subjects using the 3 feature sets and comparing between using TFIDF for text representation vs using BERT embeddings. Values in bold indicate the best metric value in a specific subject.

		Japanese			Math			Science			Social Studies			English			Total		
		MAE	PTA_0	PTA_1	MAE	PTA_0	PTA_1	MAE	PTA_0	PTA_1	MAE	PTA_0	PTA_1	MAE	PTA_0	PTA_1	MAE	PTA_0	PTA_1
	FS_2	10.32	0.37	0.20	10.96	**0.53**	0.12	15.02	**0.49**	0.12	13.43	0.51	0.087	12.48	0.58	0.12	36.83	0.58	0.15
TFIDF	FS_1	9.79	0.36	0.22	12.53	0.47	0.10	17.25	0.37	0.09	13.57	**0.60**	0.01	14.93	0.52	0.00	54.81	0.47	0.07
	FS_3	**9.16**	**0.38**	0.20	10.37	0.50	0.16	14.07	0.44	0.16	12.08	0.56	0.086	12.10	0.58	**0.13**	35.19	0.621	0.14
BERT	FS_1	9.47	0.27	**0.23**	12.36	0.45	0.07	16.66	0.40	0.11	13.92	0.55	0.02	14.51	0.52	0.02	52.02	0.49	0.07
	FS_3	9.32	0.37	0.22	**10.12**	0.52	**0.18**	**13.31**	0.43	**0.18**	**12.00**	0.53	**0.095**	**10.99**	**0.62**	0.11	**33.29**	**0.622**	**0.17**

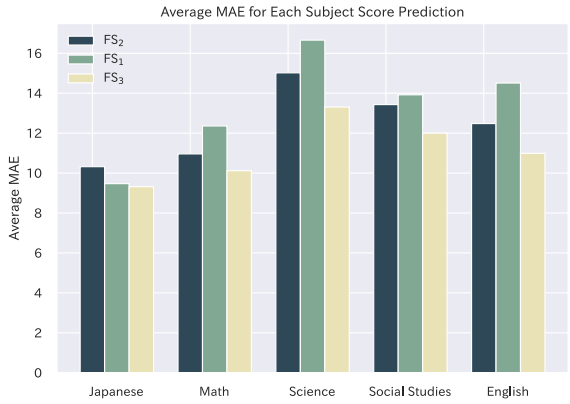

Figure 2: Average MAE of subject scores across all FS

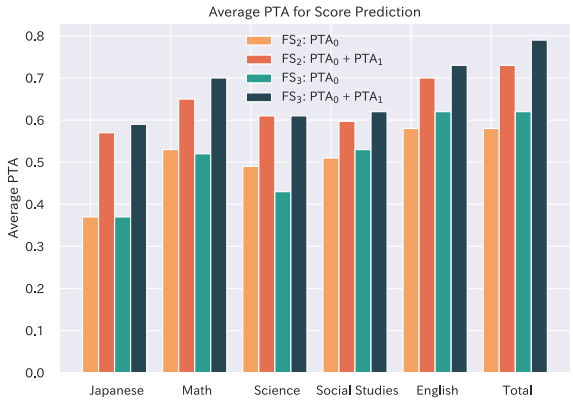

Figure 3: A comparison of PTA metric evaluated when using FS_2 and FS_3 across all subject scores and total score

performance than the TF-IDF vectors. Moreover, running the Subjects model with FS_1 using BERT resulted in lower MAE than when using TF-IDF. Finally, when predicting the total score, using FS_3 with BERT held the top scores across all evaluation metrics.

Figure 2 depicts the performance of each subject seperately in terms of MAE across the three feature sets. It can be observed that FS_3 continuously achieves lower MAE than FS_2 and FS_1. In addition, as shown in Figure 3, FS_3 also consistently achieves higher overall PTA. When predicting the total score, FS_3 shows an increase of 6.2% in $PTA_0 + PTA_1$. These results provide evidence and suggest that teachers' reports can in fact add value and contribute to grade prediction models.

6. DISCUSSION

The results presented in the previous section can be summarized into the following main points.

- The highest performance of the grade prediction model can be achieved by using a concatentation of the two feature sets, FS_1 and FS_2.
- When predicting the total score with teachers' reports, using BERT embeddings outperforms TF-IDF.

The success of BERT can be attributed to the fact that the BERT model has been pretrained on huge corpora of Japanese text data. TFIDF vectors, on the other hand, only use the data on hand to produce the representations. However, an important advantage of TFIDF is that the numerical vector representations are computed much faster than extracting BERT embeddings. To further increase the accuracy of the prediction model considering FS_3 and FS_1, we aim to pre-train BERT on each of the 5 subject reports. It has been proven that pretraining BERT on specific domains can lead to a significant increase in performance [29].

7. CONCLUSION

At educational institutes where students are closely observed by their teachers, large amounts of unstructured data exist in the form of reports and comments. In this paper, we attempted to employ and take advantage of these comments to help identify students that may need extra guidance or attention. Our model used teacher observation comments to predict students' total scores. We applied both TF-IDF and BERT embeddings to the observation comments and used the vectors as inputs to a gradient boosting regressor. Three main feature sets were employed in our model, teacher-related features, student-related features, and a concatenation of both. The performance of our model on each set was then demonstrated. Our experimental results showed that the readily available teachers' reports have the potential to create a grade prediction model. Using teachers' reports can increase the accuracy of a grade prediction model that uses only students' previous exam scores by 6.2%. However, there remains room for improvement in our experiments. We believe that with more teachers' comments, the accuracy of our model could increase. We also plan to enhance the text representations by pretraining BERT on the teachers' comments in advance. Additionally, we intend to experiment with another model architecture that would focus on classifying the students' performance first. We hope that with

such well-defined grade prediction models, we can help guide young students and provide a more focused and personalized education to them.

8. ACKNOWLEDGMENTS

This work was supported in part by e-sia corporation and JSPS KAKENHI Grant Numbers: JP21H00907, JP20H01728, JP20H04300, JP19KK0257, and JP18K18656.

9. REFERENCES

[1] R. Kerrey, *Building the intentional university: Minerva and the future of higher education*. MIT Press, 2018.

[2] T. J. Perry and C. Robichaud, "Teaching ethics using simulations: Active learning exercises in political theory," *Journal of Political Science Education*, vol. 16, no. 2, pp. 225–242, 2020.

[3] M. Hernández-de Menéndez, A. V. Guevara, J. C. T. Martínez, D. H. Alcántara, and R. Morales-Menendez, "Active learning in engineering education. a review of fundamentals, best practices and experiences," *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 13, no. 3, pp. 909–922, 2019.

[4] A. Phillipson, A. Riel, and A. B. Leger, "Between knowing and learning: New instructors' experiences in active learning classrooms.," *Canadian Journal for the Scholarship of Teaching and Learning*, vol. 9, no. 1, p. n1, 2018.

[5] J. C. Shin, "University teaching: Redesigning the university as an institution of teaching," 2014.

[6] J. Pirker, M. Riffnaller-Schiefer, and C. Gütl, "Motivational active learning: engaging university students in computer science education," in *Proceedings of the 2014 conference on Innovation & technology in computer science education*, pp. 297–302, 2014.

[7] N. Koc and B. Celik, "The impact of number of students per teacher on student achievement," *Procedia-Social and Behavioral Sciences*, vol. 177, pp. 65–70, 2015.

[8] G. Eyers and M. Hill, "Improving student learning? research evidence about teacher feedback for improvement in new zealand schools," *Waikato Journal of Education*, vol. 10, 2004.

[9] Y. Han and Y. Xu, "The development of student feedback literacy: the influences of teacher feedback on peer feedback," *Assessment & Evaluation in Higher Education*, vol. 45, no. 5, pp. 680–696, 2020.

[10] K. W. Thiede, J. L. Brendefur, R. D. Osguthorpe, M. B. Carney, A. Bremner, S. Strother, S. Oswalt, J. L. Snow, J. Sutton, and D. Jesse, "Can teachers accurately predict student performance?," *Teaching and Teacher Education*, vol. 49, pp. 36–44, 2015.

[11] Z. Ren, X. Ning, A. S. Lan, and H. Rangwala, "Grade prediction based on cumulative knowledge and co-taken courses.," *International Educational Data Mining Society*, 2019.

[12] Y. Zhao, Q. Xu, M. Chen, and G. M. Weiss, "Predicting student performance in a master of data science program using admissions data.," *International Educational Data Mining Society*, 2020.

[13] J. Luo, S. E. Sorour, K. Goda, and T. Mine, "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*, 2015.

[14] R. J. Lowe, "Cram schools in Japan: The need for research," *The Language Teacher*, vol. 39, no. 1, pp. 26–31, 2015.

[15] M. Chen, Y. Hao, K. Hwang, L. Wang, and L. Wang, "Disease prediction by machine learning over big data from healthcare communities," *Ieee Access*, vol. 5, pp. 8869–8879, 2017.

[16] M. A. Rushdi, A. A. Rushdi, T. N. Dief, A. M. Halawa, S. Yoshida, and R. Schmehl, "Power prediction of airborne wind energy systems using multivariate machine learning," *Energies*, vol. 13, no. 9, p. 2367, 2020.

[17] A. Esteban, A. Zafra, and C. Romero, "A hybrid multi-criteria approach using a genetic algorithm for recommending courses to university students.," *International Educational Data Mining Society*, 2018.

[18] Z. Wang, A. S. Lan, A. E. Waters, P. Grimaldi, and R. G. Baraniuk, "A meta-learning augmented bidirectional transformer model for automatic short answer grading.," in *EDM*, 2019.

[19] S. Morsy and G. Karypis, "Cumulative knowledge-based regression models for next-term grade prediction," in *Proceedings of the 2017 SIAM International Conference on Data Mining*, pp. 552–560, SIAM, 2017.

[20] A. Elbadrawy, A. Polyzou, Z. Ren, M. Sweeney, G. Karypis, and H. Rangwala, "Predicting student performance using personalized analytics," *Computer*, vol. 49, no. 4, pp. 61–69, 2016.

[21] C. Robinson, M. Yeomans, J. Reich, C. Hulleman, and H. Gehlbach, "Forecasting student achievement in moocs with natural language processing," in *Proceedings of the sixth international conference on learning analytics & knowledge*, pp. 383–387, 2016.

[22] K. Goda and T. Mine, "Analysis of students' learning activities through quantifying time-series comments," in *International conference on knowledge-based and intelligent information and engineering systems*, pp. 154–164, Springer, 2011.

[23] J. Jayaraman, "Predicting student dropout by mining advisor notes," in *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, pp. 629–632, 2020.

[24] P. McCann, "fugashi, a tool for tokenizing Japanese in python," in *Proceedings of Second Workshop for NLP Open Source Software (NLP-OSS)*, (Online), pp. 44–51, Association for Computational Linguistics, Nov. 2020.

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[27] C. Vroman, *Japan: a Study of the Educational System of Japan and Guide to the Academic Placement of Students from Japan in United States Educational Institutions: Placement Recommendations by the Council on Evaluation of Foreign Student Credentials, Meeting July 29-30, 1965*. American Association of Collegiate Registrars and Admissions Officers, 1966.

[28] A. Polyzou and G. Karypis, "Grade prediction with models specific to students and courses," *International Journal of Data Science and Analytics*, vol. 2, no. 3, pp. 159–171, 2016.

[29] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, "Don't stop pretraining: Adapt language models to domains and tasks," *arXiv preprint arXiv:2004.10964*, 2020.