Rodrigo, M. M. T. et al. (Eds.) (2021). Proceedings of the 29th International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

An Improved Model to Predict Student Performance using Teacher Observation Reports

Menna FATEEN*, Kyouhei UENO & Tsunenori MINE
Kyushu University, Fukuoka, Japan
*menna.fateen@m.ait.kyushu-u.ac.jp

Abstract: Predicting students' performance is a highly discussed problem in educational data mining. A tool that can accurately give such predictions would serve as a valuable resource to teachers, students, and all educational stakeholders as it would provide essential insights. Students can be further guided and fostered to achieve their optimal learning goals. In this paper, we propose an improved method to predict students' performance in entrance examinations using comments that their cram school teachers took throughout lessons. Teachers in these cram schools observe their students' behavior closely and give reports on the efforts taken in their subject material. We compare our previous model with a new and improved one to show that teachers' comments are qualified to construct a reliable tool capable of predicting students' grades efficiently. These methods are new since studies previously focused on predicting grades mainly using student data such as their reflection comments or earlier scores. Our improved experimental results show that using this readily available feedback from teachers can predict students' letter grades with an accuracy of 68%.

Keywords: Text mining, student grade prediction, teacher observation reports, machine learning

1. Introduction

Grade prediction is one of the most prominent challenges in the educational data mining community with a wide range of diverse solutions (Bretana et al., 2020; Sweeney et al., 2015). A tool that can accurately predict students' future grades is considered a powerful means that can provide valuable and beneficial insights to all educational stakeholders. These insights include early identification of at-risk students, factors that affect student performance and more. Grade prediction tools usually adopt either historical student data such as their previous grades or require that students write comments after class (Luo et al., 2015).

On the other hand, in educational systems, a trend can also be observed where educational programs evolve towards more active learning. Teachers spend more time observing their students and designing class material that encourage engagement. In innovative universities such as the Minerva Schools at KGI in the United States, the student/teacher ratio is low, and students receive written feedback from their teachers daily that clarifies any confusion, reinforces strong points, and gives more specific advice and guidance (Kerrey, 2018; Han & Xu, 2020). This trend produces a large amount of unstructured data generated by the teachers in form of reports and comments.

In this paper, we introduce an improved grade prediction model that can adequately exploit unstructured data to provide essential insights to students and teachers alike. We improve upon our previous model (Fateen & Mine, 2021) that employs teacher reports provided to us 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. The cram school sector in Japan, or 'juku', is a very large and influential one, bringing in billions of yen in profit each year. Research in the area of cram school education is needed and encouraged (Lowe, 2015).

Our two main developed grade prediction models utilize both classic and state-of-the-art techniques in Natural Language Processing (NLP) to capture the meanings of the teachers' comments. We use these vectorized reports as explanatory variables for our machine learning regression models. The experimental results of our improved model proposed in this paper show that when adding teachers' reports to the regular student exam scores, we can correctly predict their letter grade with an accuracy up to 68%. To sum up, our contributions can be outlined as:

- We propose an improved grade prediction method and compare it with our previous method. Both models use teacher observation reports represented using classic (bag of words, term-frequency inverse-document-frequency) and state-of-the-art (BERT) methods.
- We conduct extensive experiments on real data sets so as to prove the capability of teachers' reports for building accurate grade prediction models.

Finally, to the best of our knowledge, this is the first study to mine teacher observation reports to predict student grades. Our research and experimental results demonstrate the potential that these teacher observation comments have in predicting students' total scores and final letter grades.

2. Related Work

Data mining techniques are being increasingly adapted in many different fields from engineering to healthcare (Chen et al., 2017; Rushdi et al., 2020). Needless to say, many diverse solutions to significant problems in the educational community have been introduced that utilize machine learning and artificial intelligence. Educational data mining solutions vary widely from course recommendation systems (Ma et al., 2017) to automatic feedback models (Makhlouf & Mine, 2020). More specifically, many studies have been dedicated to build precise grade prediction models using various techniques such as predicting next term grades using cumulative knowledge bases (Morsy & Karypis, 2017). Bydžovská (2016) used two approaches to predict students' final grades. The first approach utilized students' social behavior while the second approach used a collaborative filtering technique where the final grade was predicted based on previous achievements of similar students. Both approaches had similar average results and the paper described each approach's advantages and disadvantages.

Several methods that utilize natural language processing have been developed to predict student performance especially since the field of NLP has seen many significant breakthroughs over the past few years. These NLP methods have been proven to have the capacity to contribute to accurately predicting students' success or failure over the information usually obtained from fixed-response items (Robinson et al., 2016). Luo et al. (2015) proposed a method that predicts students' grades based on free-style comments taken according to the PCN categorical method (Goda & Mine, 2011). Word2Vec embeddings were adopted to reflect the meanings of the students' comments. This was followed by an artificial neural network to predict the student grades. Their results showed a correct rate of 80%.

Teacher comments have been used by Jayaraman (2020) not to predict students' grade, but to detect those students who are at risk of dropping out of university. Sentiment analysis was used in their study to first extract the positive and negative words from the advisors' notes and then fed into a model as features. Their method achieved a 73% accuracy at predicting dropout.

3. Data Description

The data used in our experiments were obtained from a cram school in Fukuoka, Japan. No student names or any other identifiers were included to ensure confidentiality. We obtained the teachers' reports monthly as CSV files in addition to the students' scores of their exams taken at their regular schools. The final dataset compromised 11,960 reports for 167 students over the period from May to October 2020.

3.1 Monthly Reports

In addition to the teachers' comments, each report also contained the class date, subject code, understanding, attitude and homework scores. The feature sets adopted in our experiments are discussed in detail in Section 4.2. The main explanatory variable, however, is the teachers' comments. The average length of these comments was 96 characters and a word cloud of the most used words in the teacher's comments is shown in Figure 1.

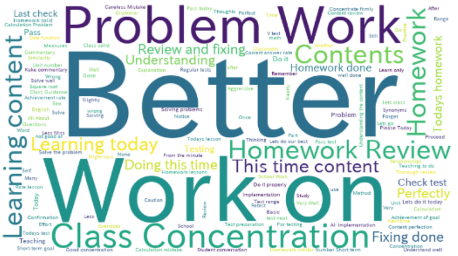

Figure 1. Word Cloud of Translated Comment Words.

The comments are originally written in Japanese and the words were translated using Google Translate. From the cloud, it can be perceived 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 frequently in math lessons.

Students in the cram school take different lessons for each subject. The lessons fall under the five main subjects: Japanese, Mathematics, Science, Social Studies and English. Since the main goal of our model is to predict a student’s total score, reports in all the five subjects were required. In our first model we thus eliminate from our experiments the students that did not attend lessons for all the subjects. However, in the second model, we also conduct experiments with a compensation model to try to estimate the missing subject scores. This will be discussed in Section 5.3.

Table 1 shows the number of reports that fall under each subject. The number of total reports for each student varied depending on the number of 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. A histogram that visualizes the distribution of the number of reports for students is shown in Figure 2.

Table 1. *The Number of Reports in Each Subject along with the Standard Deviation and Mean of the Students' Scores in Each Subject*

	Japanese	Math	Science	Social Studies	English
Number of Reports	1157 (9.7%)	3547 (29.6%)	2428 (20.3%)	1669 (14%)	3159 (26.4%)
Stand. Dev.	11.85	16.62	20.33	16.93	18.47
Mean	54.56	49.41	48.10	44.74	47.12

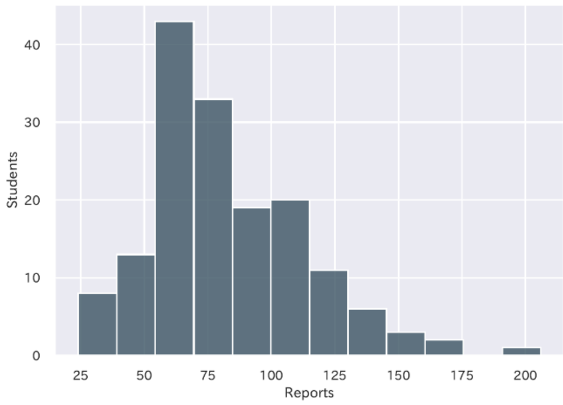

Figure 2. Histogram of the Number of Reports Distributed among Students.

3.2 Test Scores

Students attending the cram school are normally registered in many different schools. The results of their examinations regularly taken at school were also provided to us. These scores were what we considered as student data and would be traditionally used as the main feature to predict their performance in the entrance exam.

In our proposed models, we adopted a supervised machine learning approach where training data needs to be labeled with the required outputs for each input. This is required so that the model can alter its learning function based on the correct results. Since the students' actual performance in their entrance examinations is unattainable to us, we used the students' results in their simulation exams as the model labels. Histograms for each subject scores were plotted to visualize the distribution as shown in Figure 3. We can observe that the distributions are approximately bell-shaped and seem symmetric about the mean so we can assume that they follow a normal distribution. To show how dispersed the values are, we display the standard deviation for subject scores in Table 1.

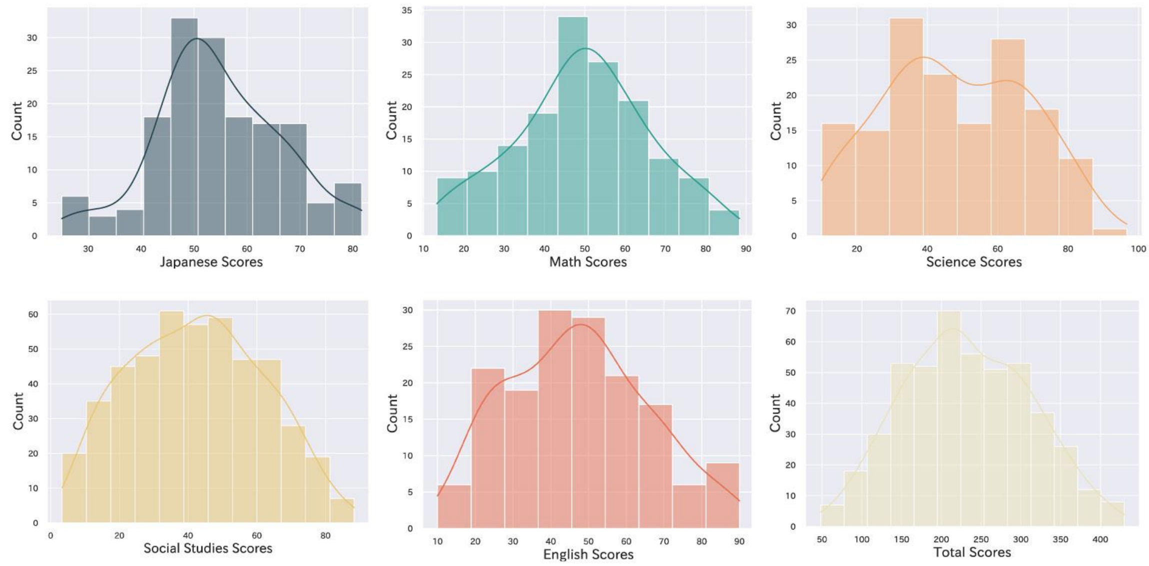

Figure 3. Distribution of Simulation Test Scores.

4. Methodology

4.1 Proposed Models: Reports Model and Students Model

Two main approaches were taken in our experiments to predict students' final grades. In our previous model called "Reports Model," each report is treated as an independent instance. 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 and therefore have *t* reports. To predict the subject score of student *i*, each report *t* is fed into a regression model separately and an ordered list $X_{i,s,t}$ of predicted scores for student *i* is obtained. Finally, to determine the estimated predicted score of a subject, we use $SubjectScore_{pred,i,s} = Median(X_{i,s,t})$. The total predicted score ($TotalScore_{pred,i}$) can then be estimated by $TotalScore_{pred,i} = \sum SubjectScore_{pred,i,s}$ for $s \in S$. These steps are illustrated in detail in Figure 4.

We introduce another model called "Students Model" as an improved one in this paper. In the model, a separate regression model for each subject judges a student's performance based on all their reports combined in the specified subject. In this model, we aim to encapsulate a student's performance in one vector since the performance can naturally vary from class to class due to different factors. The elements in each vector of the Students Model depend on the feature set used, which is discussed in the next section.

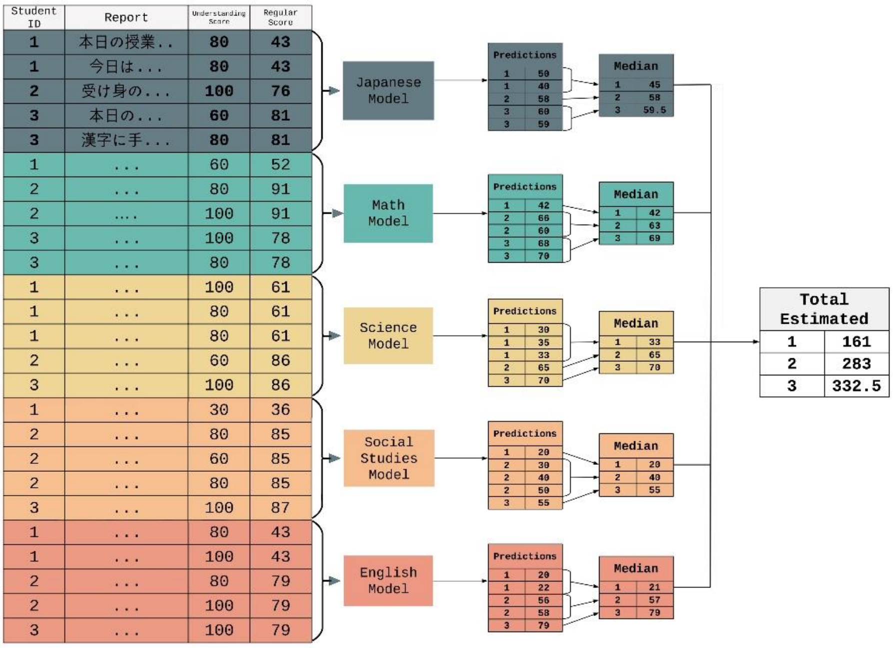

Figure 4. Reports Model Architecture. Each subject model takes a report instance which as input which includes the teachers' comments, an assessment of their understanding given by a score of either (0-30-60-80-100) and an attitude score of (1-2-3-4).

4.2 Feature Selection

For the sake of comparison, we adopt three primary feature sets in our experimental settings. The first feature set, FS_1 consists of teachers' report contents only as the main explanatory variables. After each lesson, the teacher provides a report for each student consisting of written comments based on their observations, an assessment of their understanding given by a score of either (0-30-60-80-100) and an attitude score of (1-2-3-4). In the Reports Model, FS_1 used the complete report as attributes except for the homework score since the score was not included in more than 36% of the reports. Conversely, since the Students Model aims to encapsulate the meaning of all provided reports, a more statistical approach had to be taken. Therefore, in the Students Model, FS_1 consists of the number of classes taken by the student, the first, second and third quartile of the understanding score and finally the teachers' comments.

The second feature set, FS_2, consists of student-related data only. In the Reports Model, this set consisted of their gender and the score of their regularly scheduled exam at school, which we call the students' regular scores or the regular score. Since we predict each subject score separately, the regular score corresponds to the subject score. The gender attribute was omitted for the Students Model since the Pearson correlation coefficient between the gender and the regular score is 0.12, which shows no statistical significance while the correlation coefficient between the regular score and the simulation score is 0.80. Moreover, in the Students Model, all 413 students' regular scores were taken into consideration since this feature set does not depend on subject reports or classes. Finally, we investigate using both teachers' reports and the regular scores to verify whether adding teachers' reports contributes to the accuracy of the prediction model or not. FS_3 therefore is a concatenation between each model's FS_1 and FS_2. Figure 5 shows the vector elements when using FS_3 in the Students Model.

4.3 Natural Language Processing

Various ways exist to represent texts to convey the original meanings and prevent information loss. We chose to represent the teachers' comments using three main NLP techniques. In the Reports Model, a TFIDF (term frequency-inverse document frequency) vectorization method is adopted and compared with BERT embeddings. In the Students Model, we use the classic bag of words method along with BERT.

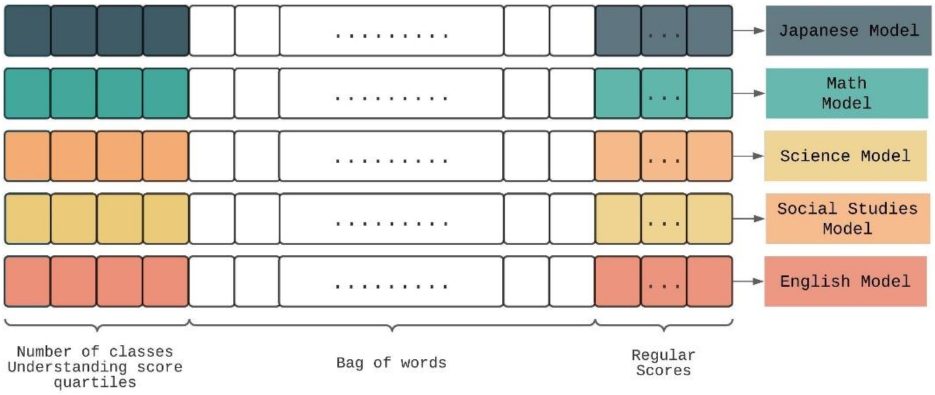

Figure 5. Vectors using FS_3 in the Students Model. Each vector in this case consists of (the number of classes taken by the student, the first, second and third quartile of the understanding score, the teachers' comments and the students' regular scores)

4.3.1 Bag of Words

The first essential step in transforming a text into a numerical representation is preprocessing the text. For English sentences, this step begins with splitting the sentences into words or tokenization. Tokenization in languages such as English can be simply 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 Mecab (Kudo, 2006), a Japanese tokenizer and morphological analysis tool to extract the nouns and verbs from the teachers' comments. We then use these extracted words to build the corpus of vocabulary for Bag of Words (BOW). In a BOW vector, each element corresponds to a word in the corpus and represents its frequency in the sentence. In the Students Model, we employed sklearn's CountVectorizer to build the BOW vector (Pedregosa & Karypis, 2011). The length of the BOW vector is 6033 words. In the Reports Model, we experimented with a variant of BOW, TFIDF, where high weights are given for those terms that occur often in a particular sentence or document but rarely in other documents.

4.3.2 BERT

BERT (Bidirectional Encoder Representations from Transformers) is a novel technique of pre-training language representations presented by Google (Devlin et al., 2018). On many NLP tasks, BERT obtains state-of-the-art results. A pre-trained BERT model is essentially a general-purpose language understanding model trained on a large corpus like Wikipedia. This pretrained model can then be utilized for downstream NLP tasks. When building both the Reports and Students Models, we used the BERT model pretrained by Inui Laboratory at Tohoku University (cl-tohoku, 2019). The model was trained with the same configuration as the original BERT and the pretraining corpus used was Japanese Wikipedia. We used the [CLS] token embeddings as our BERT embeddings to transform the teachers' comments into constant length vectors.

4.4 Evaluation Metrics

To evaluate our experiments, we adopt two main evaluation metrics: MAE (Mean Absolute Error) and PTA (Percentage by Tick Accuracy). MAE is calculated using the following formula:

$$MAE = \frac{1}{n}\sum |score_{pred,i} - score_{true,i}|$$

, where $score_{true,i}$ is the actual score that student i obtained. In the Reports Model, $score_{pred,i}$ for subjects is calculated by taking the median, as explained in Section 4.1, while in the Students Model the subject score is predicted directly by the regression model. The total score is the summation of all subject scores predicted by either model.

PTA has been used in previous studies to evaluate grade prediction models. Since students receive letter grades for their total score, we map the estimated total score to its corresponding letter grade according to the percentages as shown in Table 2. A tick, as specified by (Polyzou & Karypis, 2016) is the difference between two successive letter grades. In the experiments we employ PTA_0 and PTA_1, which mean the model successfully predicted the exact letter grade or predicted it incorrectly but with 1 tick away from the true grade (e.g., B vs C), respectively.

Table 2. *Letter Grades and their Corresponding Percentages*

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

5. Experiments

5.1 Overview

Gradient boosting, a composite machine learning algorithm was the regressor adopted in the Reports Model, or the baseline, since it showed to have yielded the best results. However, in the Students Model, XGBoost was employed instead. XGBoost is a more efficient and scalable implementation of Gradient Boosting machines (Chen et al., 2015). A brief summary of the main differences between the highest performing Reports and Students Models is displayed in Table 3.

Table 3. *Comparison between the Characteristics of the Reports Model and the Students Model*

	Reports Model	Students Model
Vector Elements	Report-based	Student-based
FS_1	Teachers' comments in one lesson + understanding score + attitude score	Number of classes + understanding score quartiles + all teachers' comments
FS_2	Regular score + gender	Regular score
Text Representation	BERT	Bag of Words
Regressor	GradientBoosting	XGBoost
Total Score Estimation PTA_0	0.62	0.68

All experiments were evaluated using group 10-fold cross validation. One of the main advantages of group k-fold cross validation is that all data is used for training and testing and each instance is used once for testing. As a result, this validation method is often used especially in situations where data is limited. The average MAE, PTA_0 and PTA_1 of all ten folds were computed for all tests. As shown in Table 3, the Students Model improved the Reports Model on PTA_0 by 6%.

5.2 Results

We ran the model with all the three feature sets as described in Section 4.2. Table 4 shows the results of the Students Model. Bold values indicate the leading scores for each metric in each subject. It can be easily observed that FS3 regularly outperforms the other feature sets. In addition, from the table and as depicted in Figures 6 and 7, we can also realize that experiments using FS_1 achieved lower MAE and higher PTA_0 than FS_2 in all subjects. This strongly suggests that teachers' reports when used to predict student grades can as a matter of fact exceed the performance of a model that uses students' regular scores as the main explanatory variable.

Simultaneously, we conducted experiments for the Students Model using BERT embeddings since they proved to obtain a more accurate Reports Model. Results of the experiment using FS_3 in terms of MAE are shown in Table 5. BERT embeddings were utilized in two main ways: BERT1 and BERT2. In BERT1, the teachers' comments for each class were encoded and then summed up to be combined with the rest of the features. In a different manner, in BERT2, all comments from each class were first concatenated and then represented using BERT embeddings. However, as seen in the table, representing the teachers’ comments using bag-of-words consistently outperformed BERT embeddings in the Students Model.

Table 4. *Evaluation Metric Scores of All Subjects using the 3 Feature Sets in the Student Model. Values in Bold Indicate the Best Metric Value in its Corresponding Subject.*

	Japanese			Math			Science		
	MAE	PTA_0	PTA_1	MAE	PTA_0	PTA_1	MAE	PTA_0	PTA_1
FS_2	11.87	0.48	**0.45**	18.18	0.34	0.47	23.53	0.21	0.47
FS_1	10.62	0.54	0.42	12.42	0.47	0.46	17.32	0.34	0.48
FS_3	**9.42**	**0.55**	0.43	**10.32**	**0.49**	**0.48**	**12.15**	**0.47**	**0.48**
	Social Studies			English			Total		
	MAE	PTA_0	PTA_1	MAE	PTA_0	PTA_1	MAE	PTA_0	PTA_1
FS_2	18.62	0.31	0.48	19.28	0.29	0.49	76.67	0.36	**0.47**
FS_1	14.35	0.36	**0.53**	12.65	0.42	**0.5**	48.02	0.59	0.39
FS_3	**10.92**	**0.54**	0.42	**10.1**	**0.53**	0.42	**31.67**	**0.68**	0.32

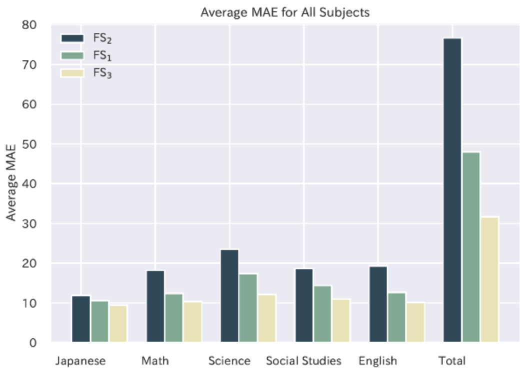

Figure 6. Average MAE for all Subjects using the 3 Feature Sets.

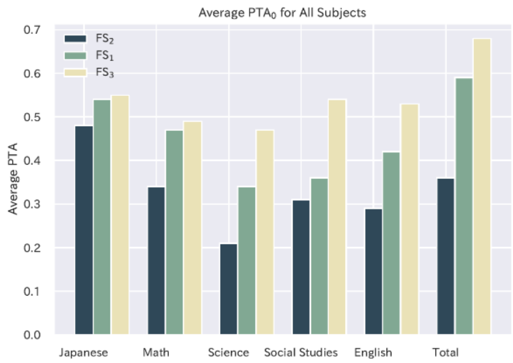

Figure 7. Average PTA_0 for all Subjects using the 3 Feature Sets.

Table 5. *MAE of Students Model with BERT embeddings* ***compared with Bag of*** *Words using FS_3*

	Japanese	Math	Science	Social Studies	English	Total
BERT1	9.52	10.93	12.75	11.13	11.07	34.30
BERT2	9.47	10.73	12.83	11.62	10.87	33.65
BoW	9.42	10.32	12.15	10.92	10.10	31.67

5.3 Compensation Experiments

Supplementary experiments were held to attempt to estimate subject scores for those students that did not attend lessons for all subjects in the cram school and thus reports were unavailable. The estimates were made based on the students’ regular scores of all the subjects. In other words, in cases where FS_3 was unattainable in specific subjects, the predictions were supplemented by using FS_2. When using this compensation for the 413 students, the total score estimation produces an MAE of 38.33 which is an acceptable value compared to the results shown in the previous section.

6. Discussion

An abridged version of the results of the baseline Reports Model is shown in Table 6 where only the BERT results are shown since they outperformed the alternative TFIDF text representation. Predicting the total score using either FS_1 or FS_3 in the Students Model yielded both lower MAE and higher PTA_0

and PTA_1 than in the Reports Model as shown in Table 6. This highlights the fact that student performance cannot be easily judged by separate reports and that an encapsulating statistical approach can accomplish a more improved performance. It is also interesting to highlight the PTA_1 results achieved by the Students Model where it significantly outperforms the baseline in both feature sets. Hence, it can be suggested that the Students Model is a more reliable one since it mostly either correctly identifies the letter grade or misses it by one tick. In the Students Model, $PTA_0 + PTA_1$ reaches an accuracy of 0.98 with FS_1.

On the other hand, the baseline exceeded in performance when using FS_2. However, this can be attributed to the fact that a different strategy or approach was taken. In the baseline, instances were report based and not student based. Thus, an effect similar to oversampling was achieved when using the second feature set.

Table 6. *Students Model Performance compared to the Baseline Model (Reports Model). Best metric values in each subject are shown in boldface, where for pta_1, pta_0+pta_1 is considered.*
7.

		Baseline			Students Model		
		MAE	PTA_0	PTA_1	MAE	PTA_0	PTA_1
FS_1	Japanese	9.47	0.27	0.23	10.62	0.54	0.42
	Math	12.36	0.45	0.07	12.42	0.47	0.46
	Science	16.66	0.40	0.11	17.32	0.34	0.48
	Social Studies	13.92	**0.55**	0.02	14.35	0.36	0.53
	English	14.51	0.52	0.02	12.65	0.42	0.5
	Total	52.02	0.49	0.07	48.02	0.59	0.39
FS_3	Japanese	**9.32**	0.37	0.22	9.42	**0.55**	**0.43**
	Math	**10.12**	**0.52**	0.14	10.32	0.49	**0.48**
	Science	13.31	0.43	0.18	**12.15**	**0.47**	**0.48**
	Social Studies	12.00	0.53	0.095	**10.92**	0.54	**0.42**
	English	10.99	**0.62**	0.11	**10.1**	0.53	**0.42**
	Total	33.29	0.62	0.17	**31.67**	**0.68**	**0.32**

7. Conclusion

At educational institutes where teachers observe their students closely, large amounts of unstructured data are available in the form of comments and reports. In this paper, we proposed an improved model that employs and takes advantage of these comments to produce a reliable and accurate grade prediction model. The model uses teacher observation comments taken after lessons at a cram school in Fukuoka, Japan. The comments in our previous model and the improved model are represented using different natural language processing representations from bag of words to BERT embeddings. We employed three main feature sets: teacher-related features, student-related features, and a concatenation of the two to compare and demonstrate the effect of using each feature set. Our experimental results show that using teachers' comments can not only contribute to the accuracy of existing grade prediction models based on student-related features, but they can also exceed their performance. The improved model introduced in this paper can accurately predict a students' letter grade with an accuracy of 68%, which is an increase of 6% over the previously proposed model.

However, there remains much room for improvement in our experiments. We are working on further improving the performance in hope that with such a well-defined grade prediction model, we can contribute to guiding young students by providing a more focused and personalized education to them.

Acknowledgements

This work was supported in part by e-sia Corporation and by Grant-in-Aid for Scientific Research proposal numbers (JP21H00907, JP20H01728, JP20H04300, JP19KK0257, JP18K18656). We would like to express our deepest gratitude.

References

Bretana, N. A., Robati, M., Rawat, A., Pandey, A., Khatri, S., Kaushal, K., ... & Abadia, R. Predicting Student Success for Programming Courses in a Fully Online Learning Environment. In 28th International Conference on Computers in Education, ICCE 2020 (pp. 47-56). Asia-Pacific Society for Computers in Education.

Sweeney, M., Lester, J., & Rangwala, H. (2015, October). Next-term student grade prediction. In 2015 IEEE International Conference on Big Data (Big Data) (pp. 970-975). IEEE.

Kerrey, R. (2018). Building the intentional university: Minerva and the future of higher education. MIT Press.

Han, Y., & Xu, Y. (2020). The development of student feedback literacy: the influences of teacher feedback on peer feedback. Assessment & Evaluation in Higher Education, 45(5), 680-696.

Fateen, M. & Mine, T. (2021). Predicting Student Performance Using Teacher Observation Reports. In Proceedings of The 14th International Conference on Educational Data Mining (EDM 2021) (pp. 481-486).

Lowe, R. J. (2015). Cram schools in Japan: The need for research. The Language Teacher, 39(1), 26-31.

Chen, M., Hao, Y., Hwang, K., Wang, L., & Wang, L. (2017). Disease prediction by machine learning over big data from healthcare communities. IEEE Access, 5, 8869-8879.

Rushdi, M. A., Rushdi, A. A., Dief, T. N., Halawa, A. M., Yoshida, S., & Schmehl, R. (2020). Power prediction of airborne wind energy systems using multivariate machine learning. Energies, 13(9), 2367.

Makhlouf, J., & Mine, T. (2020, November). Automatic Feedback Models to Students Freely Written Comments. In 28th International Conference on Computers in Education, ICCE 2020 (pp. 336-341). Asia-Pacific Society for Computers in Education.

Ma, H., Wang, X., Hou, J., & Lu, Y. (2017, August). Course recommendation based on semantic similarity analysis. In 2017 3rd IEEE International Conference on Control Science and Systems Engineering (ICCSSE) (pp. 638-641). IEEE.

Morsy, S., & Karypis, G. (2017, June). Cumulative knowledge-based regression models for next-term grade prediction. In Proceedings of the 2017 SIAM International Conference on Data Mining (pp. 552-560). Society for Industrial and Applied Mathematics.

Bydžovská, Hana. "A Comparative Analysis of Techniques for Predicting Student Performance." International Educational Data Mining Society (2016).

Robinson, C., Yeomans, M., Reich, J., Hulleman, C., & Gehlbach, H. (2016, April). Forecasting student achievement in MOOCs with natural language processing. In Proceedings of the sixth international conference on learning analytics & knowledge (pp. 383-387).

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.

Goda, K., & Mine, T. (2011, September). 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, Berlin, Heidelberg.

Jayaraman, J. D. (2020). Predicting Student Dropout by Mining Advisor Notes. In Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020) (pp. 629-632).

Kudo, T. (2006). Mecab: Yet another part-of-speech and morphological analyzer. http://mecab. sourceforge. jp.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

cl-tohoku (2019, November 6) https://github.com/cl-tohoku/bert-japanese/tree/v1.0

Polyzou, A., & Karypis, G. (2016). Grade prediction with models specific to students and courses. International Journal of Data Science and Analytics, 2(3), 159-171.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., & Cho, H. (2015). Xgboost: extreme gradient boosting. R package version 0.4-2, 1(4).