Enhancing the Accuracy of Predicting Students Grades in Open-Ended Questions through Adjustments to Attention Weights

Masaki Koike
Chubu University
masa1357@mprg.cs.chubu.ac.jp

Hirokazu Kohama
Chubu University
tuna0724@mprg.cs.chubu.ac.jp

Tsubasa Hirakawa
Chubu University
hirakawa@mprg.cs.chubu.ac.jp

Takayoshi Yamashita
Chubu University
takayoshi@isc.chubu.ac.jp

Hironobu Fujiyoshi
Chubu University
fujiyoshi@isc.chubu.ac.jp

ABSTRACT

With the digitalization of the educational environment, educational support is anticipated by predicting student performance from the operation log data of digital teaching materials. However, these methods require the construction of large-scale systems and have to collect extensive long-term log data. Therefore, we focus on the response sentences from lecture questionnaires, which have a simple recording system. We collected the response sentences from lectures given at Japanese universities, we will classify students' grades using a Transformer Encoder. In particular, utilizing Term Frequency-Inverse Document Frequency (TF-IDF) to analyze written responses, we identify words indicative of each students' grade. Then, to emphasize the identified words during the inference phase of the Transformer Encoder model for grade prediction, we aim to improve the accuracy of the predictions. In the evaluation experiment using the proposed method, the accuracy of grade prediction improved by 2.5 pt and the f1-score improved by 1.2 pt, compared to the baseline.

Keywords

student surveys, student performance, text mining, TF-IDF, NLP

1. INTRODUCTION

Efforts have been made to analyze lecture comprehension using learning log data to identify students with low grades early and improve their learning behaviors [10]. Previous studies on grade prediction used data such as digital teaching material usage, past grades, attendance, and homework, and predicted grades using models such as decision trees, neural networks, and support vector machines (SVMs) [4,6]. Additionally, Stephen et al. combined multiple linear regression (MLR) and principal component analysis (PCA) to predict grades using data from students video viewing behaviors, exercises, assignment responses, and quiz grades for digital materials [9]. However, these methods require a large-scale data collection system and time to accumulate the necessary learning behavior data for predicting grades. For a simpler and more easily collected approach, we focused on the response sentences from lecture open-ended questionnaires. Questionnaires are easier to collect than log data on student behavior and can be conveniently collected using existing questionnaire applications. Additionally, these methods enable early predictions of student performance from the first lecture onwards, allowing for faster student support. Furthermore, open-ended questionnaires are more closely related to the students' understanding of the lecture than multiple-choice questionnaires and allow for a more concrete analysis of the ideas held by the students. However, applying existing approaches to questionnaire responses is challenging due to their free-answer format. Therefore, to analyze questionnaire responses, we identified unique expressions for each grade using the term frequency-inverse document frequency (TF-IDF) method, a word frequency analysis method. The identified words are then emphasized in the inference process of the Transformer Encoder [7] to improve the accuracy of grade predictions.

2. RELATED WORKS

Studies have investigated factors associated with students' grades in education. Stephanie et al. [2] conducted three open-ended questionnaires for undergraduate engineering students to explore any correlation between the responses and GPA. The study included an analysis of word frequency, t-tests, z-tests, and the length and number of words in the responses. The analysis revealed a distinct difference in the vocabulary used by students with high and low grades. Particularly in response to the question "In your own words, what do engineers do?", words like "why" and "test" (as a verb) were found to be associate with the students' grades. One of the methods to measure the importance of words within a text is TF-IDF [5]. TF-IDF is a metric that signifies the significance of a word in the documents of a corpus. TF (Term Frequency) is the frequency of a specific word's appearance in a document, whereas IDF (Inverse Document Frequency) is the reciprocal of the total number of documents divided by the number of documents con-

M. Koike, H. Kohama, T. Hirakawa, T. Yamashita, and H. Fujiyoshi. Enhancing the accuracy of predicting students grades in open-ended questions through adjustments to attention weights. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 872–876, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12729979

taining corpus. These two metrics allow for the analysis of highly important words that are frequently found in specific documents. Dadgar et al. [1] proposed a method for news classification tasks, one that uses the TF-IDF to compute word importance and that feeds the computed results into an SVM. Their method showed that the use of the TF-IDF led to higher accuracy compared with other classification techniques on both the BBC dataset and the 20Newsgroup dataset.

3. PROPOSED METHOD

In this section, we propose a method using TF-IDF to estimate words of high importance for each grade, with these estimated words are strongly reflected in grade predictions through the Transformer Encoder.

3.1 Estimation of important words by grade using the TF-IDF

Latent Dirichlet Allocation (LDA) and TF-IDF are well-known methods for analyzing text data. LDA is particularly effective in classifying multiple topics, and it is less suitable for questionnaires concerning a single topic. Therefore, we use TF-IDF to evaluate the importance of words in the responses and to emphasize the high-importance words for each grade in the response sentences. The TF-IDF is a metric that represents word significance as a score based on its frequency in a text. When the set of sentences G_i corresponds to the grade i within all sentences, the TF-IDF score TF-IDF for each word t in each sentence $g \in G_i$ is determined using the Equation (1):

$$\text{TF-IDF}(t, g, G_i) = \log(1 + c(t, g)) \cdot \log\left(\frac{|G_i|}{df(t)}\right), \quad (1)$$

where $c(t, g)$ is the number of occurrences of the word t in the sentence g, $|G_i|$ is the number of sentences in grade i, and $df(t)$ is the number of sentences containing the word t in G_i. The average TF-IDF score for each word is calculated using Equation (2) to obtain the final word importance:

$$S(t, G_i) = \frac{\sum_{g \in G_i} \text{TF-IDF}(t, g, G_i)}{df(t)}. \quad (2)$$

The importance of each word is compared to its maximum importance across other grades. Words whose importance exceeds twice their maximum importance across other grades are defined as important.

3.2 Emphasis on attention weights for important words

The attention mechanism in the Transformer model treats input tokens as queries and calculates attention weights through the inner product with keys. To emphasize important words, the model compares the queries with identified important words and adds bias to the attention weights where matches occur. This process enhances the representation of important words in the model's outputs. Figure 1 illustrates the procedure for reflecting important words in inference using our method. First, we calculate the bias to add to the attention weights, as shown in Equation (3):

$$\text{bias} = \begin{cases} 0.3 & \text{if } S(\mathbf{Q}, G_i) \geq 2 \cdot \max(S(\mathbf{Q}, G_j)) \\ 0 & \text{otherwise,} \end{cases} \quad (3)$$

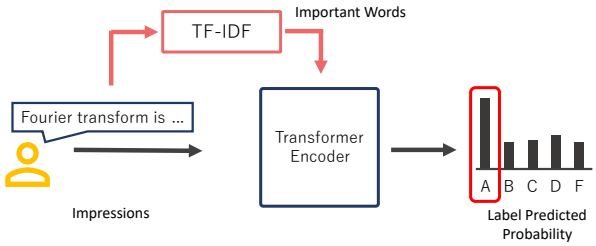


Figure 1: Procedure for reflecting important words in inference using our method.

where G_j represents the sentences in group G_j, excluding those in G_i. Second, the formula to calculate the attention weight A, considering the bias for the input token $\mathbf{Q}$, is presented in Equation (4):

$$A(\mathbf{Q}, \mathbf{K}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} + \text{bias}), \quad (4)$$

where $\mathbf{K}$ is the key, d_k is the dimensionality of both $\mathbf{Q}$ and $\mathbf{K}$.

4. EXPERIMENTS

We evaluate the effectiveness of our method, which involves modifying Attention weights using TF-IDF, by comparing its grade prediction accuracy against that of existing models.

4.1 DATASETS

Our dataset was derived from the "Information Science" course conducted at Kyushu University, under the approval of an Ethics Committee. This course covers a span of 14 weeks and has a final examination. After each lecture, we posed five reflective questions to assess the course material.

Q1: Please explain today's content in your own words.

Q2: Write down what you understood and what you were able to do from today's content.

Q3: Write down what you did not understand or were not able to do from today's content.

Q4: If you have any questions, please write them down.

Q5: Write down your thoughts or reflections on today's lesson.

We gathered a total of 70 open-text responses from each student, corresponding to five questions per week for 14 weeks. In addition to these responses, we obtained each students' final grades, which were classified as A, B, C, D, or F. In this research, we collected responses from the same course across three academic terms: 2021-1, 2021-2, and 2022-1. Table 1 shows the number of students enrolled in the course for each term, along with the distribution of final grades in our dataset. We obtained 17,660 usable responses after excluding instances with no response and unclassifiable responses such as "nothing in particular". We used responses from 80% (298 students) of the dataset as training data and responses from the remaining 20% (75 students) as evaluation data.

Table 1: Distribution of student grades in our dataset

Grade	A	B	C	D	F	Total
2021-Course-1	9	53	32	7	6	107
2021-Course-2	15	88	37	9	25	174
2022-Course-1	17	37	34	4	4	96
Total	41	178	103	20	35	377

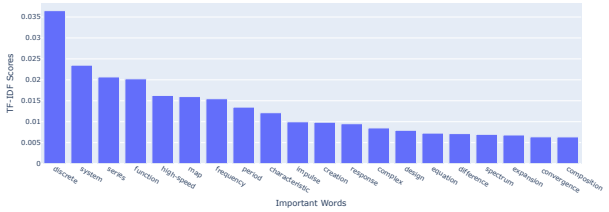


Figure 2: Top 20 important words for Grade A.

4.2 Obtaining important words with the TF-IDF

We applied TF-IDF to the dataset to identify unique words for each grade, which were then designated as important words. Figure 2 shows the words obtained from the grade A responses and their importance scores as an example. The word importance was compared with the maximum word importance of the other four grades. Words with importance more than twice the maximum in the other grades were defined as important. We selected up to three important words for each grade and used them to train the model. Table 2 shows the important words for each grade.

4.3 Experimental Setup

We use a Transformer Encoder model to predict students' understanding from their responses to the questionnaire.

4.3.1 BERT

BERT is composed of several Transformer Encoders and can make predictions by considering the relationships between all words in a sentence through self-attention. Moreover, BERT can understand bidirectional relationships between words and their context by performing pre-training tasks 'Masked Language Modeling (MLM)' and 'Next Sentence Prediction (NSP)' on a large-scale unlabeled text dataset. After acquiring the pre-trained language model, it can be fine-tuned with relatively small numbers of labeled data to handle various tasks like classification and inference. In this study, we use the Japanese pre-trained model 'cl-tohoku/bert-japanese-base' released by Tohoku University.

Table 2: Calculated important words

Grade	important words
A	'discrete', 'system', 'series'
B	'copyright', 'security', 'legal'
C	'concept', 'indicate', 'match'
D	'gain', 'network', 'neural'
F	'cryptography', 'investigation', 'key'

Table 3: Comparison of the accuracy[%] and F1-score[%] with and without our method

	Baseline		Ours	
	Accuracy	F1-score	Accuracy	F1-score
SVM			61.3	45.4
BERT	34.5	**51.3**	**62.5**	48.1
RoBERTa	45.7	**53.1**	**63.7**	51.8
LUKE	70.1	55.1	**72.6**	**56.3**

4.3.2 RoBERTa

A robustly optimized BERT pretraining approach (RoBERTa) [3] is a model based on BERT that conducts its pretraining solely with MLM. RoBERTa has shown superior results compared to BERT due to several changes, including increased batch sizes, dynamic masking, and the elimination of NSP. In this study, we use the Japanese pre-trained model 'nlp-waseda/roberta-base-japanese' released by Waseda University.

4.3.3 LUKE

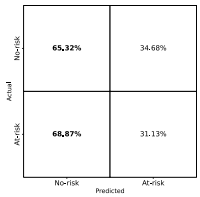
LUKE [8] is a model built on RoBERTa that demonstrates superior results by incorporating entity representations into the attention mechanism. LUKE defines entities as linguistic representations of objects or concepts within a text and treats words and entities as independent tokens, enabling predictions that consider proper nouns. Additionally, LUKE is designed with varying queries for different word combinations during Attention calculations, thereby strongly recognizing the relationship between text and entities. In this study, we used the Japanese pre-trained model 'studio-ousia/luke-japanese-base'.

The hyperparameters used for each model are a batch size of 16 and a learning rate of 5e-5 for 5 epochs. The bias value added by the proposed method significantly impacts the predictions of the model, so it was tested with multiple values. As a result, 0.3 was chosen as the most effective value. Each model was trained five times using cross-entropy loss, and the average values of the results were taken as the final evaluation metrics. As a comparison, we present the results from an SVM trained using features derived from TF-IDF scores for each word. Both accuracy and f1-score are used as evaluation metrics.

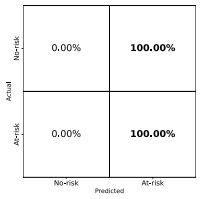
4.4 Experimental Results

We conduct a two-class classification evaluation to predict the students' understanding of the lectures. Students with grades A and B are defined as those "no-risk student", while those with grades C, D, and F are defined as "at-risk students". Learning and evaluation are performed using multiple models to examine the changes in accuracy with and without the proposed method. Table 3 presents a comparison of the prediction accuracy in lecture understanding for each model, with(Ours) and without(Baseline) the proposed method. Table 3 shows that our method improved the accuracy of all models, achieving higher scores than the SVM model. Moreover, the LUKE model improved the accuracy by 2.5 pts and the F1-score by 1.2 pts with our method, achieving the highest precision in prediction. The confu-

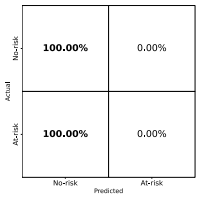
sion matrices for each model compare the baseline with our method, as shown in Figure 3.



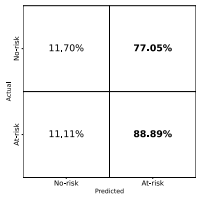
(a) Baseline
confusion matrix in the SVM model



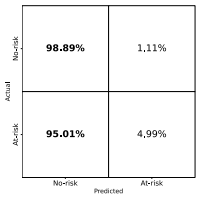
(b) Baseline



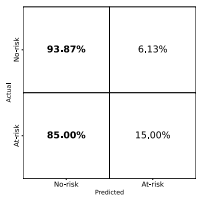
(c) Ours
Comparison of the prediction using the confusion matrix in the BERT model



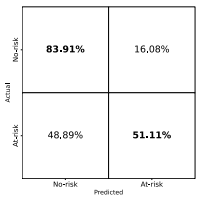
(d) Baseline



(e) Ours
Comparison of the prediction using the confusion matrix in the RoBERTa model



(f) Baseline



(g) Ours
Comparison of the prediction using the confusion matrix in the LUKE model

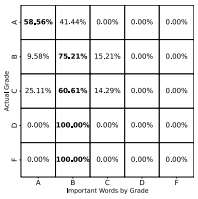
Figure 3: Comparison of the prediction using the confusion matrix in each model.

Figure 3 shows that while the predictions of the SVM, BERT, and RoBERTa models were biased, the results of the LUKE model were corrected by our method, improving the prediction accuracy for the students risk. These observations demonstrate that our method effectively improved the accuracy of predicting students' grades. Next, we evaluated the models for five classes: grades A, B, C, D, and F. Table 4 shows the accuracy of each model for predicting student grades, with and without our method. Table 4 shows that our method improved the accuracy. Figure 4 shows the

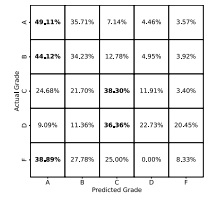
Table 4: Comparison of the accuracy[%] and F1-score[%] with and without our method for five-5class problems

	Baseline		Ours	
	Accuracy	F1-score	Accuracy	F1-score
SVM			51.1	45.7
BERT	25.7	10.5	**62.5**	48.0
RoBERTa	7.5	5.3	**50.5**	33.9
LUKE	37.5	39.0	**52.1**	42.8

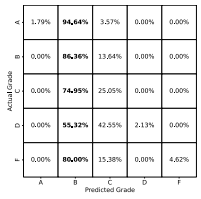
confusion matrices for each model among the five classes, comparing the baseline with our method. Figure 4 shows



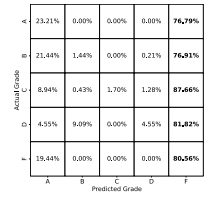
(a) Baseline
confusion matrix in the SVM model



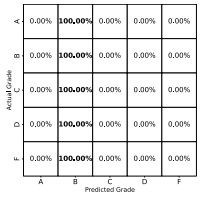
(b) Baseline



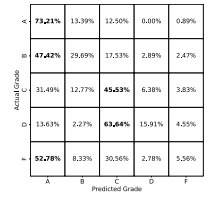
(c) Ours
Comparison of the prediction using the confusion matrix in the BERT model



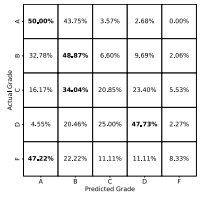
(d) Baseline



(e) Ours
Comparison of the prediction using the confusion matrix in the RoBERTa model



(f) Baseline



(g) Ours
Comparison of the prediction using the confusion matrix in the LUKE model

Figure 4: Comparison of the prediction using the confusion matrix in each model for five-class problems.

that the LUKE model improved with our proposed method, demonstrating increased attention to students with grades D and F. However, BERT and RoBERTa models predominantly predict grade B, which suggests that the proposed method is sensitive and may require optimal parameters for each model.

5. DISCUSSION

As a result of the experiments showed that adding important words calculated by the TF-IDF to the attention weight improved the accuracy of the predicting students' grades in lectures. The LUKE model showed a particularly significant improvement, with a 14.6 pts increase in accuracy and a 3.8 pts increase in F1-score in the five classes. However, the predictions of the other models were biased toward specific grades, leading to inaccurate grade predictions. This bias can be attributed to our method's strong reliance on the students' vocabulary differences, which substantially affects accuracy. As a result, this method may not be suitable for datasets with bias or with repetitive expressions. Therefore, when using this model in a real-world environment, focus to be paid to data bias.

6. CONCLUSION

This study evaluated the effectiveness of a prediction model for student lecture understanding using open-ended questionnaires. The model achieved a maximum accuracy of 70.1% and an F1-score of 55.1%. Furthermore, the accuracy improved by up to 2.5 pt and the F1-score by 1.2 pt after calculating the importance of words for each grade using the TF-IDF and emphasizing important words in the attention mechanism. Grade predictions using questionnaires can be more easily implemented in real-world environments compared to models that predict grades from learning behaviors or that predict students' understanding at an early stage. Our future work will involve designing appropriate support methods for students based on the attention weights of Transformer models.

7. ACKNOWLEDGEMENTS

This work was supported by JST CREST Grant Number JPMJCR22D1, Japan.

8. REFERENCES

[1] S. M. H. Dadgar, M. S. Araghi, and M. M. Farahani. A novel text mining approach based on tf-idf and support vector machine for news classification. In *2016 IEEE International Conference on Engineering and Technology (ICETECH)*, 2016.

[2] S. M. Gratiano and W. J. Palm. Can a five minute, three question survey foretell first-year engineering student performance and retention? *ASEE*, 2016.

[3] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, 2019.

[4] A. Namoun and A. Alshanqiti. Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences*, 11(1), 2021.

[5] G. Salton, E. A. Fox, and H. Wu. Extended boolean information retrieval. *Commun. ACM*, 26(11), 1983.

[6] A. M. Shahiri, W. Husain, and N. A. Rashid. A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, 2015.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

[8] I. Yamada, A. Asai, H. Shindo, H. Takeda, and Y. Matsumoto. LUKE: deep contextualized entity representations with entity-aware self-attention. *CoRR*, 2020.

[9] S. J. Yang, O. H. Lu, A. Y. Huang, J. C. Huang, H. Ogata, and A. J. Lin. Predicting students' academic performance using multiple linear regression and principal component analysis. *Journal of Information Processing*, 26:170–176, 2018.

[10] M. Yağcı. Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 2022.