Improving the Quality of Students' Written Reflections Using Natural Language Processing: Model Design and Classroom Evaluation

Ahmed Magooda[1](✉), Diane Litman[1](✉), Ahmed Ashraf[2], and Muhsin Menekse[2]

[1] University of Pittsburgh, Pittsburgh, USA
{aem132,dlitman}@pitt.edu
[2] Purdue University, West Lafayette, USA
{butt5,menekse}@purdue.edu

Abstract. Having students write reflections has been shown to help teachers improve their instruction and students improve their learning outcomes. With the aid of Natural Language Processing (NLP), real-time educational applications that can assess and provide feedback on reflection quality can be deployed. In this work, we first evaluate various NLP approaches for developing a reflection quality prediction model, aiming to find a configuration that balances model simplicity and generalizability across courses. Second, using the model that best balances runtime performance and predictive accuracy, we evaluate the impact of using this model to trigger real-time feedback regarding reflection quality in a mobile application currently being deployed in multiple courses across universities. Analysis of students' long-term (semester-level) and short-term (reflection writing level) changes in reflection quality across multiple classes demonstrate the utility of the deployed model in encouraging students to submit reflections with higher quality.

Keywords: Reflections · NLP · Quality prediction · Feedback

1 Introduction

Enabling students to write *free-text responses* to *reflection prompts* has been shown to improve learning gain and teaching quality [10]. Prior computational work has largely focused on reflection quality assessment [3,7,9,12,13], but has typically considered data from only single course domains and evaluated models for accuracy without regard to runtime performance. Moreover, while reflection quality modeling has been used to understand learning outcomes, its potential

The research reported here was supported by a grant from the Institute of Education Sciences (R305A180477). The opinions expressed do not represent the views of the U.S. Department of Education.

M. M. Rodrigo et al. (Eds.): AIED 2022, LNCS 13355, pp. 519–525, 2022.
https://doi.org/10.1007/978-3-031-11644-5_43

for adaptive reflection scaffolding largely remains an area for future research [2] or only studied in the lab [5]. Expanding on this prior literature, we perform research in two stages to provide students with real-time reflection quality assessment and feedback. **In the first stage**, we design new *quality prediction models* using recent transformer-based NLP techniques, and investigate model performance along two dimensions: 1) accuracy within and across conditions common to classroom use cases (e.g., differing courses), and 2) run-time (e.g., to determine which models can be integrated into a real-time application). **In the second stage**, we incorporate the best model into the *CourseMIRROR* mobile app, with the goal of providing students with *real-time feedback*. An in-the-wild evaluation of the technology deployment across multiple college classes demonstrates how providing feedback improves the quality of submitted reflections.

2 First Stage: Reflection Quality Prediction

Data for Model Development. We use the publicly available *CourseMIRROR (CM)* corpus[1] [5,9] which contains student reflections collected from 4 undergraduate classes (Chemistry (Chm), Statistics (ST), and Material Science (MSG1, MSG2)) at the end of each lecture. Each reflection was scored for quality in terms of specificity by trained raters on a 4 point scale, according to the guidelines in [9]. Table 1 summarizes the reflection distributions in the corpus.

Model Design. Following prior work, we implement a *feature extractor module* to encode reflections, followed by a *prediction module* to predict the reflection quality. While early models used handcrafted predictive features [7,9,12], recent research used neural network (NN) encoders to automatically extract features [3,6,13]. We similarly use a NN to extract all features, but unlike prior work, we use recent BERT-based transformer encoders as they have achieved better performance on many downstream tasks compared to earlier NN encoders (e.g., word2vec, GloVe, etc.) [1]. We integrate and compare two sentence encoders within our model. **DistilBERT** [11] is a distilled version of the BERT transformer-based encoder [4]. **RoBERTa** [8] is an optimized version of the original BERT, where model hyperparameters were tuned to achieve better performance compared to BERT. We predict quality using **classification** (following CourseMIRROR [5,9]) and report results using support vector machines (SVM).

Model and Data Configurations. We experimented with three different configurations of RoBERTa to observe the impact of model encoder size: RoBERTa (base, and large), and DistilRoBERTa. The encoder parameters are kept fixed during model training. For evaluation, we use **leave-one-out** to split the data, where reflections in each testing fold come from a held-out course not used during model training. This corresponds to the use case for the second stage of our

[1] https://engineering.purdue.edu/coursemirror/download/reflections-quality-data/.

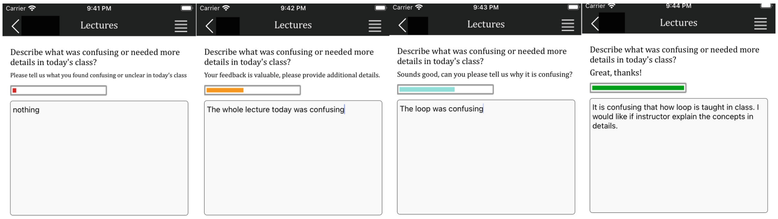
research, where the model trained at the conclusion of stage one will be used to predict the quality of reflections from new courses in stage two.

Evaluation Results. With respect to *predictive performance*, Table 1 shows that while all transformer models perform very closely, the best QWK (Quadratic Weighted Kappa) score is achieved by DistilBERT. As predicted, all four BERT-based transformer encoders outperform a baseline model using a GloVe NN encoder (which was used to encode reflections in [6]). With respect to *runtime*, Table 1 shows that RoBERTa large takes on average 6 times and 3 times the time to embed compared to DistilBERT and DistilRoBERTa, respectively. We decided to choose the DistilBERT model for our real-time deployment as it is slightly faster than DistilRoBERTa and achieves best predictive performance.

Table 1. CourseMIRROR (CM) corpus reflections distribution and model performance (QWK) and runtime (in seconds) results (best in **bold**).

CM data distribution				Model	QWK	Reflection embedding time		
Courses		Scores				Max	Avg	Min
ST	1769	1	1354	GloVe (Baseline)	0.66	NA		
MSG1	395	2	2035	RoBERTa base	0.77	0.24	0.13	0.11
Chm	1034	3	2377	RoBERTa large	0.78	0.9	0.35	0.26
MSG2	3626	4	1058	DistilBERT	**0.79**	**0.13**	**0.06**	**0.05**
Total = 6824				DistilRoBERTa	0.77	0.16	0.10	0.09

3 Second Stage: Improving Reflection Quality



[redacted]

Describe what was confusing or needed more details in today's class?

Please tell us what you found confusing or unclear in today's class

nothing

[redacted]

Describe what was confusing or needed more details in today's class?

Your feedback is valuable, please provide additional details.

The whole lecture today was confusing

[redacted]

Describe what was confusing or needed more details in today's class?

Sounds good, can you please tell us why it is confusing?

The loop was confusing

[redacted]

Describe what was confusing or needed more details in today's class?

Great, thanks!

It is confusing that how loop is taught in class. I would like if instructor explain the concepts in details.

Fig. 1. Real-time quality feedback as a reflection is being written.

We now turn to presenting real-time feedback to students while writing reflections in a mobile application. First, we hosted the DistilBERT model from Sect. 2 on a server and provided an API to communicate with the hosted quality prediction model. Second, we integrated communication into the *CourseMIRROR* mobile application that students used to write and submit reflections, to enable

CourseMIRROR to provide a real-time indicator of the predicted reflection quality while students are actively writing before submission. Figure 1 shows the interface of the reflection submission mobile application. To avoid flooding the server with requests, we decided not to call the API for each student's change within the typing session as we didn't expect the quality score to change with every character change. Instead, we performed API calls whenever the number of words became odd, e.g., number of words is 1, 3, 5, 3, 5, 7, etc.

Table 2. Data used for feedback evaluation: across semester analysis and within session analysis. ARPL refers to average number of reflections per lecture.

Across semester analysis					Within session analysis			
	Course	# Lecs.	# Students	ARPL	Course	# Lecs.	# Students	ARPL
+ Feedback	PHYS1	32	143	72	CS1	26	33	14.19
	PHYS2	18	123	47	CS2	28	30	12.14
	PHYS3	9	92	17	CS3	26	46	12.34
	ENGR1	20	90	53	IS1	16	54	20.9
No feedback	ENGR2	26	124	64	CS4	27	19	4.5
					Avg # logs per reflection			9.06

Table 3. *Percentage of students* with last lecture's reflection score less than first lecture's reflection score and vice versa (left), and *average reflection quality score* for submitted reflections for the first and last lecture of the semester (right).

		Last < First	Last >= First	First Lec mean	Last Lec mean
Avg. of with feedback		27.5%	**72.5%**	2.89	2.85
No feedback	ENGR2	**65%**	35%	3.2	2.5

3.1 Experiment 1: Does Real-Time Quality feedback result in Better Reflections Through a Semester, Compared to No Feedback?

Reflection Data. We collected the reflections summarized in Table 2 (left) using the mobile application in 5 different college-level courses across two universities. Reflections from four courses were collected after integrating the real-time feedback algorithm during the Spring 2021 semester. Reflections from the remaining course were used as a control group,[2] as they were collected before the feedback algorithm was integrated into the mobile application. ***We performed human annotation of reflection quality for data from all courses and carried out the analysis using these human scores***. Three annotators evaluated the reflections based on the annotation guidelines [9].

[2] We didn't randomly assign students to feedback and control groups, as the data collection happened in two different semesters.

Evaluation Results. Table 3 compares average reflection quality change for the courses with real-time quality feedback versus the course without. At the *reflection level*, the last two columns show the average score of submitted reflections for the first and last lecture of the semester.[3] For the average of the four courses with feedback, score of the first and last lectures are very close, with a 0.04 difference. For the course with no feedback, the scores show a 0.7 degradation, which is around a seventeen times larger difference than the feedback course difference. At the *student level*, the first two columns show the percentage of students where their last lecture's reflection score was less than their first lecture's score and vice-versa. With feedback, the percentage of students submitting equal or higher-quality reflections for the last lecture is greater (bolded) than those who submit lower-quality reflections. When no feedback is presented, the majority of students (bolded) tend to submit lower-quality reflections at the semester's end. In sum, our results support feedback utility.

Table 4. Score improvement within sessions.

	Final score vs first score (Endpoint category)			Score change direction (Trend category)		
	Improved	Constant	Decreased	Increasing	Constant	Other
Avg of 5 courses	54.1%	39.9%	5.9%	35.6%	39.9%	24.4%

3.2 Experiment 2: Do Students Keep Writing a Reflection Until It Is of High Quality, Each Time They Submit a Reflection?

Reflection Data. For Experiment 2, we logged the changes in reflection quality provided by the deployed model while students were typing the reflections. This can help us observe if the feedback provided helped students improve their submission quality within a writing session. We used data from 5 different college level courses that used the application after within session logging was incorporated into the application. ***Logs were collected from typing sessions and contained scores for each partial reflection***. Table 2 (right) summarizes data size and the average number of logged scores.

Evaluation Results. We first categorize each series of logs using one of three *trend categories*, based on the pattern when considering all logged scores for a given reflection. **Increasing** series are monotonically increasing, **constant** series have constant value, while **other** series are neither monotonically increasing nor constant. We also categorize each series using one of three *endpoint categories*, based on comparing only the starting and ending values. **Improved** series have a final value higher than the starting value, **constant** series have a final value

[3] We performed additional experiments comparing the mean of the first/last quarters of lectures instead of the first/last lecture only, and we observed similar findings.

equal to the starting value, while **decreased** series have a final value less than the starting value. Table 4 shows the distribution of these categories for average of all courses using the application after within-session API logging was implemented (Table 2). For trends (right columns), on average, more than 75% of series are either improving or constant, with around 35% improving. This shows that students often keep improving their reflections until they submit, supporting the utility of real-time feedback. Similarly, comparing the last score to the first score in the series (left columns) shows that in most cases (54%), students end the writing session with higher quality reflections than what they started with. Only around 6% of series end with lower quality reflections than what they started with, suggesting that even when a score drops during a writing session (the "other" trend category), most students recover or improve the quality by the end of the session. In sum, our results again support feedback utility.

4 Summary

Our first stage experiments in model development focused on balancing accuracy and efficiency when predicting reflection quality. Our results suggested DistilBERT as the most promising model for deployment in a real-time application. Our second stage experiments showed that using the model to provide real-time quality feedback did indeed help students submit higher-quality reflections within a reflection writing session and over the semester. For future work, we plan to tackle a few limitations of our current research. First, the feedback generated consists of a color corresponding to an ordinal value and a static message. We would like to explore generating dynamic messages tailored to the reflection content. Additionally, we plan to investigate the utility of generating more personalized feedback that integrates multiple dimensions in addition to specificity.

References

1. Bommasani, R., Davis, K., Cardie, C.: Interpreting pretrained contextualized representations via reductions to static embeddings. In: Proceedings of ACL (2020)
2. Carpenter, D., Cloude, E., Rowe, J., Azevedo, R., Lester, J.: Investigating student reflection during game-based learning in middle grades science. In: LAK21: 11th International Learning Analytics and Knowledge Conference, pp. 280–291 (2021)
3. Carpenter, D., Geden, M., Rowe, J., Azevedo, R., Lester, J.: Automated analysis of middle school students' written reflections during game-based learning. In: Bittencourt, I.I., Cukurova, M., Muldner, K., Luckin, R., Millán, E. (eds.) AIED 2020. LNCS (LNAI), vol. 12163, pp. 67–78. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-52237-7_6
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018)
5. Fan, X., Luo, W., Menekse, M., Litman, D., Wang, J.: Scaling reflection prompts in large classrooms via mobile interfaces and natural language processing. In: Proceedings of 22nd International Conference on Intelligent User Interfaces, pp. 363–374 (2017)

6. Geden, M., Emerson, A., Carpenter, D., Rowe, J., Azevedo, R., Lester, J.: Predictive student modeling in game-based learning environments with word embedding representations of reflection. Int. J. AI Educ. **31**(1), 1–23 (2021)
7. Kovanović, V., et al.: Understand students' self-reflections through learning analytics. In: Proceedings of 8th International Conference on Learning Analytics and Knowledge, pp. 389–398 (2018)
8. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
9. Luo, W., Litman, D.: Determining the quality of a student reflective response. In: The Twenty-Ninth International FLAIRS Conference (2016)
10. Menekse, M., Stump, G., Krause, S., Chi, M.: The effectiveness of students' daily reflections on learning in engineering context. In: ASEE Conference & Exposition (2011)
11. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
12. Ullmann, T.D.: Automated analysis of reflection in writing: validating machine learning approaches. Int. J. AI Educ. **29**(2), 217–257 (2019). https://doi.org/10.1007/s40593-019-00174-2
13. Wulff, P., et al.: Computer-based classification of preservice physics teachers' written reflections. J. Sci. Educ. Technol. **30**(1), 1–15 (2020). https://doi.org/10.1007/s10956-020-09865-1