International Society of the Learning Sciences

# Comparative analysis of the rule-based and machine learning approach for assessing student reflections

Ahmed Ashraf Butt, Purdue University, butt5@purdue.edu
Saira Anwar, Texas A&M University, sairaanwar@tamu.edu
Ahmed Magooda, University of Pittsburgh, aem132@pitt.edu
Muhsin Menekse, Purdue University, menekse@purdue.edu

**Abstract:** This study describes Natural Language Processing (NLP) and rule-based algorithms used to assess the students' written reflection quality. We used the CourseMIRROR application to gather student reflections from 120 engineering students and converted the reflections into an equivalent quality score using human coders and algorithms. We used Cohen's kappa statistics to explore the agreement between the algorithm's reflection quality and the human coding. The results revealed a strong agreement between the NLP algorithm and human coding. Furthermore, we used Spearman correlation statistics to evaluate the relationship between the predicted quality scores and the human coding. The results showed a strong correlation between reflection quality evaluated by the NLP algorithm and human coders. Overall results revealed that the NLP algorithm embedded in the CourseMIRROR application is a good fit for assessing students' reflections and can be used to guide them during the reflection writing process.

**Keywords:** Reflection quality, scaffolding, mobile application, learning environment, reflection specificity, NLP algorithms, rule-based algorithm

## Introduction

Fostering reflective thinking is an important instructional practice (e.g., Bannert, 2006; Stark & Krause, 2009). It involves students in a meaning-making process that helps them make connections between their prior knowledge and learning experiences, deliberately systematizing their thinking with an attitude of valuing themselves and others in a community (Dewey, 1933). Studies have shown that the reflection writing practice can promote students' engagement (Menekse et al., 2020), help them identify their misconceptions (Tawde et al., 2017), and improve their learning outcomes (Miller & Jensen, 2016). Fan et al. (2015) developed a mobile application (i.e., CourseMIRROR) that gathers students' reflections by prompting them to reflect on their experiences after each lecture throughout the semester to facilitate the integration of reflective practices in the classroom. Also, it provides real-time feedback by assessing the reflection quality during the reflection writing process and guiding users towards writing specific rather than generic reflections. Here, the reflection quality indicates the extent to which the reflection is relevant to the reflection prompt and the course.

This study describes two quality prediction algorithms (based on Natural Language Processing (NLP) and a rule-based approach) that can be used to improve the application's ability to evaluate the students' reflection quality. For the study, we collected students' reflections using the CourseMIRROR application and converted the reflections into the equivalent reflection quality scores coded by two human coders. Then, we explored whether the reflection quality assessed by the NLP algorithm and the rule-based algorithm is similar to the reflection evaluation conducted by human coders. Specifically, this study is guided by the following research question: how does the reflection quality assessed by two quality prediction algorithms relate to human coding?

## Related work

Based on the reflection literature, the commonly used method to evaluate reflection writing is through content analysis (Ullmann, 2019). A typical content analysis process includes 1) reading students' reflections, 2) manually labeling them with self-described categories, and 3) converting these categories into quality scores of the reflection. Furthermore, the focus of prior studies using content analysis was to identify the level of reflection quality (i.e., shallow or deep reflection) by analyzing students' reflection journals and essays (Kovanović et al., 2018). Hence, researchers are exploring ways to automate this reflection process and facilitate the adoption of reflective practice.

Researchers are utilizing computational advancements to design automatic text analytics for assessing the students' reflection texts among the ongoing efforts. According to Ullmann (2019), these text analytics are being designed by broadly using three approaches, i.e., dictionary-based approach (find the occurrence of predefined words in the text), rule-based approach (using the rule to get better inference from the text), and the machine learning (ML) approach (using ML algorithms). Based on the previous approaches, this study utilized two algorithms 1) Rule-based algorithm where the expert has defined the pre-set pattern and their associated

 

weightage, and 2) NLP algorithm trained on different educational reflection sets. Further, this study investigated the relative efficacy of these algorithms in the reflection quality.

## Participants and dataset

We recruited 120 first-year engineering (FYE) students enrolled in a required engineering course from a large midwestern university located in the United States for the study. The topics covered in the course were introductory computer programming concepts, development of mathematical models, data visualization, and designing solutions for engineering problems. Of the participants, 83% were male students, and 17% were female students. Also, 61.7% were White, 21.7% were International Students, and 16.6% were People of Color (POC). Furthermore, the dataset used for this study consists of 3452 individual reflections from 28 lectures.

## Instrument

We used CourseMIRROR educational application to collect students' reflections. This application prompts students with two open-ended questions to reflect on the confusing or interesting concepts at the end of each lecture. Additionally, it uses Natural Language Processing (NLP) algorithms to create students' reflections summaries by combining the reflections based on common themes (Luo et al., 2015). For the study, students voluntarily participated and submitted 3452 reflections (i.e.,1726 reflections for each question) in 28 lectures throughout the semester. Furthermore, these reflections were analyzed by two human coders based on a rubric (Heo et al., 2018; Menekse et al., 2011) and then assigned a quality score. We calculated the agreement between the two raters, and they showed good agreement, as $\kappa$ (MP) = .617 and $\kappa$ (POI) = .652 (Altman, 1990).

## Reflectivity quality models

Following are the two approaches used to design the quality prediction algorithms that are used to assess the reflection quality:

### Machine learning approach

In this approach, we utilize recent SOTA models in NLP to automatically produce a quality score. Our work uses a basic model consisting of a feature generation module followed by a classification module for score generation. In order to generate features, we used a transformer-based bidirectional deep contextual language model to automatically generate features. The module operates on the raw input text and automatically converts the text into numerical features. We used DistilBERT (Sanh et al., 2019) model for feature generation. The model is a distilled version of the original BERT (Devlin et al., 2018) transformer-based encoder. DistilBERT reduces the number of parameters to around 60% of the original BERT (110M to 66M). This, in turn, allows the model to be faster and more suited to real-time quality prediction. We used a logistic regression classifier for the classification module that operates on the generated features and produces a score of 1, 2, 3, or 4. We only train the logistic regression classification module and keep the DistilBERT parameters fixed. This can reduce the load required for training the model and similarly allows us to easily fine-tune the model with new samples acquired through time.

### Rule-based approach

In this approach, a team of educational researchers defined a set of patterns and their associated weightage to score the reflection quality. Based on these rules, we designed an algorithm used to predict the reflection quality. Table 1 describes the few rules and their associated score used to design this algorithm:

**Table 1**
Rules with associated weightage used for rule-based algorithm

| Number | Pattern | Quality Score |
|---|---|---|
| 1 | "how\|what\|when\|where\|which" | 0.25 |
| 2 | Keywords (extracted from the syllabus, and learning objective of the class) | 1 |
| 3 | "compare\|comparison\|difference\|differentiating\|determining\|determination \|relationship\|relation\|between" | 0.29 |

 

| | | |
|---|---|---|
| 6 | "^pretty clear\|straightforward" | -0.6 |
| 7 | "^everything\|^nothing\|^none\|^nope\|^null\|^no$\|^no\\s" | -0.6 |
| 10 | “understood everything” | -0.6 |
| 11 | "I\|explanation\|example" | -0.3 |
| 12 | "this lecture\|the lecture\|all\|completely" | -0.3 |

To evaluate the reflection quality, this algorithm parses the students’ reflections and uses regexes to count the occurrence of each pattern in the string. After that, it calculates the reflection quality by rounding off the sum of the associated weightage of each pattern based on their occurrence. Furthermore, we modify the predicted quality score to either 1 or 4, if the values go less than score 1 or more than score 4, respectively. This way, the range of the quality score is between 1 to 4 points.

## Results

We divided the reflection into two sets: 1) reflections discussing the interesting aspects of the lecture and 2) reflections discussing the confusing aspects of the lecture. For the remaining paper, we will refer to the first set of reflections as “Reflections 1.0” and the second as “Reflections 2.0,” respectively. Then, we converted the reflections into the equivalent quality scores evaluated by NLP and rule-based algorithms. The quality scores range from 1 to 4 points, where a score of 1 indicates poor relevancy, and a score of 4 indicates high relevancy with the course/question. For analysis, we calculated Cohen’s kappa to evaluate the agreement between the human, NLP, and rule-based algorithm’s reflection quality scores. The result of the analysis is shown in table 2:

**Table 2**
*Cohen’s κ co-efficient among the reflection quality scores*

| | Reflections 1.0 | | | Reflection 2.0 | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 1 | 2 | 3 |
| 1. Human coding | - | - | - | - | - | - |
| 2. Rule-based algorithm | 0.042 | - | - | 0.068 | - | - |
| 3. NLP algorithm | 0.775 | 0.033 | - | 0.773 | 0.067 | - |

As shown in the table, it is quite evident that the NLP algorithm has a strong agreement while evaluating the reflection of both reflection sets (i.e., κ (NLP) = .775 and κ (NLP) = .773) with human coding. The rule-based algorithm has close to no agreement (i.e., κ (NLP) = .042 and κ (NLP) = .0.068) for both reflection sets with human coding. We conducted the Spearman Correlation statistics to further enhance our analysis to explore the correlation between reflection quality evaluated by the quality prediction algorithms and the human coders. The result of the analysis is shown in Table 3:

**Table 3**
*Spearman correlation among the reflection quality scores*

| | Reflections 1.0 | | | Reflection 2.0 | | |
|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 1 | 2 | 3 |
| 1. Human coding | - | - | - | - | - | - |
| 2. Rule-based algorithm | 0.479** | - | - | 0.435** | - | - |

 

| | | | | | | |
|---|---|---|---|---|---|---|
| 3. NLP algorithm | 0.837** | 0.496** | - | 0.859** | 0.485** | - |

**Correlation is significant at the 0.01 level (2-tailed).

The results showed that the correlation between the rating evaluated by the NLP algorithm and the human rating has a strong correlation for both sets (i.e., r = 0.837, 0.8359), with each having 1752 reflections. Even though the correlation between reflection quality scores evaluated by the rule-based algorithm has a strong relationship (i.e., r = 0.479, 0.435; N = 1752) with human coding, it is much lower than the NLP algorithm.

## Conclusion

The current study has described two approaches that can be used to evaluate the students' reflection quality. One approach used the expert's defined pattern, and the other used the NLP algorithm to evaluate the reflection quality. This study investigated the performance of both approaches in terms of their ability to produce similar results to the human coders. Our results showed that the NLP algorithm based on the DistilRoBERTa and DistilBERT with SVM is a promising model to evaluate the reflection quality.

## References


Altman, D. G. (1990). *Practical statistics for medical research*. Chapman & Hall/CRC.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv Preprint. https://arxiv.org/pdf/1810.04805.pdf&usg=ALkJrhhzxlCL6yTht2BRmH9atgvKFxHsxQ

Dewey, J. (1933). *Think we. A Restatement of the relation of reflective thinking to the educative process*. D.C. Heath & Co.

Fan, X., Luo, W., Menekse, M., Litman, D., & Wang, J. (2015). CourseMIRROR: Enhancing Large Classroom instructor-student interactions via mobile interfaces and natural language processing. *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, 1473–1478. https://doi.org/10.1145/2702613.2732853

Heo, D., Anwar, S., & Menekse, M. (2018). The Relationship between engineering students' achievement goals, reflection behaviors, and learning outcomes. *International Journal of Engineering Education*, *34*, 1634–1643.

Kovanović, V., Joksimović, S., Mirriahi, N., Blaine, E., Gašević, D., Siemens, G., & Dawson, S. (2018). Understand students' self-reflections through learning analytics. *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, 389–398.

Luo, W., Fan, X., Menekse, M., Wang, J., & Litman, D. (2015). Enhancing instructor-student and student-student interactions with mobile interfaces and summarization. *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*.16-20. https://doi.org/10.3115/v1/N15-3004

Menekse, M., Anwar, S., & Akdemir, Z. G. (2020). How do different reflection prompts affect engineering students' academic performance and engagement? *The Journal of Experimental Education*, 1–19. https://doi.org/10.1080/00220973.2020.1786346

Menekse, M., Stump, G., Krause, S., & Chi, M. T. H. (2011). The effectiveness of students' daily reflections on learning in engineering context. *Proceedings of ASEE Annual Conference and Exposition*.

Miller, I., & Jensen, K. (2020). Introduction of mindfulness in an online engineering core course during the covid-19 pandemic. *Advances in Engineering Education*, *8*.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv Preprint. https://arxiv.org/pdf/1910.01108.pdf?ref=https://githubhelp.com

Tawde, M., Boccio, D., & Kolack, K. (2017). Resolving misconceptions through student reflections. *Journal of College Science Teaching*, *47*(1), 12.

Ullmann, T. D. (2019). Automated analysis of reflection in writing: Validating machine learning approaches. *International Journal of Artificial Intelligence in Education*, *29*(2), 217–257. https://doi.org/10.1007/s40593-019-00174-2


## Acknowledgments

We are grateful to the support provided by Institute of Education Sciences, U.S. Department of Education under grant award No: R305A180477.