Automated Multi-Dimensional Analysis of Peer Feedback in Middle School Mathematics

Jiayi Zhang, University of Pennsylvania, joycez@upenn.edu
Ryan S. Baker, University of Pennsylvania, ryanshaunbaker@gmail.com
J. M. Alexandra Andres, University of Pennsylvania, aandres@gse.upenn.edu
Stephen Hutt, University of Denver, stephen.hutt@du.edu
Sheela Sethuraman, CueThink, sheela@cuethink.com

Abstract: Peer review, a commonly-used pedagogy in contemporary education has been found to positively influence student learning, benefitting both feedback provider and recipient. However, the quality of the feedback may vary, and lower-quality feedback (e.g., lacking specificity), is less likely to be implemented by the recipient, leading to suboptimal outcomes. Although recent work has used criteria to scaffold feedback to ensure quality, it is often difficult to monitor whether students follow these criteria. In this study, we develop models that automatically detect the attributes of student feedback, reflecting the presence of three pedagogically relevant constructs: 1) commenting on the process, 2) commenting on the answer, and 3) relating to self. We find models employing sentence embeddings produce the best results, with AUC ROCs ranging from .90-.96, and are robust to algorithmic bias.

Background

Peer review, an important pedagogy in contemporary education, has been commonly and increasingly adopted in classrooms to facilitate collaborative learning, improving communication, domain knowledge, and cognitive engagement. Positive learning effects have been found for both the students who receive feedback and the students who give feedback (Li et al., 2010).

For students who receive feedback, feedback provides an evaluation of current performance, strengths and weaknesses of their work, and provides possible suggestions for improvements. Students are expected to learn from implementing the feedback and revising their work accordingly (Shute, 2008). Compared to teacher's feedback, peer review enables students to receive a greater quantity of feedback that is richer, more diverse, and perceived as more open to negotiation (Topping, 2009). Under the right circumstances, peer feedback can be just as effective as teacher's feedback at improving students' math proficiency (Patchan et al., 2022).

Providing feedback has also been shown to lead to better learning outcomes. Cho & Cho (2011) found that the students who provided feedback had higher quality revisions, while limited effect was found for those who received feedback. Roscoe & Chi (2008) use the term reflective knowledge-building to refer to this effect, in which assessors monitor their own understanding and repair misconceptions as they evaluate peers' works.

However, the effectiveness of peer review can be influenced by its quality (Shute, 2008), and not being able to ensure quality has been the main concern limiting the use of this pedagogical method by instructors. For this reason, studies have examined what components are key to peer feedback and analyzed how these components influence the likelihood of feedback implementation by the receivers and the quality of revision. In general, feedback appears to work best when it is specific, actionable, descriptive, and focuses on the task (Shute, 2008).

Hattie and Timperley (2007) proposed a theoretical model that categorizes feedback into four levels. The first level of feedback evaluates the product of a task, such as the correctness of the answer. The second level of feedback comments on the process or strategies used to complete a task, such as what strategies are used to solve a problem. The third level focuses on promoting self-regulation, where the comment intends to help students (feedback recipients) to understand their current progress in relation to their goals. Lastly, the fourth level refers to feedback that is not directly related to the learning tasks. Empirical evidence indicates that feedback should focus on more than one aspect of learning, including components that evaluate the product, process, and progress of the learning (Shute, 2008). Thus, evaluative criteria are often introduced in peer review activities to scaffold peer feedback (Liu & Carless, 2006), specifying the critical components that feedback should include. However, it is often difficult to monitor whether students follow these criteria, to ensure quality peer feedback.

To address this issue, a few studies have employed text analytics to analyze the quality of feedback based on metrics such as problem localization, feedback type, and the relevance of feedback (Nguyen & Litman, 2015). In recent work, Darvishi et al., (2022) created three automated functions to examine and scaffold peer feedback in an AI-assisted learning platform. Compared to the control group, students who used the functions were more likely to provide comments that were constructive, specific, and better aligned to the rubric.

In line with previous work that utilizes text analytics to evaluate the quality of peer feedback, in this study, we explored the possibility of automatically detecting key attributes of peer feedback. We collected peer comments from a digital learning platform for middle school math, and identified and operationalized three constructs that are pedagogically relevant to detect. These constructs are 1) commenting on the process, 2) commenting on the answer, and 3) relating to self. We distilled key features of each example of peer feedback using a variety of natural language processing (NLP) tools as well as content-based feature engineering, and then input the features into a neural network model. We compared the model performance across feature sets and evaluated the fairness of the best performing models across student demographic groups.

Methods

Learning platform & data collection

In this study, the digital learning application CueThink was implemented at a middle school in the southwest of the US. The platform structures a math problem into a *Thinklet*, which scaffolds the problem-solving process into four phases. In each phase, students are prompted to answer various questions and to complete different tasks. (See Zhang et al., (2022) for a full description of the platform).

Once a student has completed a math problem, the *Thinklet* that consists of the student's final answer, textual responses, and a screencast which students record to explain the problem-solving process, is shared with peers to review. While reviewing the responses and the screencast, reviewers are encouraged to annotate the work by providing a text-based comment describing if they agree or disagree with the work and explain what they like. These comments are then sent back to the *Thinklet* author for possible revision. To understand how students annotate each other's work, we collected 229 comments from 63 students (reviewers), annotating on 170 *Thinklets* created by 117 students (authors). On average, each reviewer provided 3.6 comments, and each *Thinklet* received 1.3 comments.

Text labeling

Using grounded theory, we identified three constructs that are salient in the data and pedagogically relevant. These constructs are 1) commenting on the process, 2) commenting on the answer, and 3) relating to self.

As shown in Table 1, commenting on the process and commenting on the answer are two constructs that identify a difference in the focus of a comment (process vs. outcome), described as the first two levels in Hattie and Timperley's (2017) four level model. While comments on the process involves the reviewer focusing on evaluating how a math problem is solved, commenting on the answer reflects reviewers' focus on the outcome of the problem. These constructs can both occur in the same review. When labeling commenting on the process, we only included comments that are specific and pertinent to solving the math problem, excluding comments on the overall process of the problem-solving, or comments irrelevant to solving the math problem. This focus was chosen based on evidence that feedback tends to be more effective if it is specific to the task (Shute, 2008)

With the goal of capturing the reflective knowledge-building described in Roscoe & Chi (2008), we operationalized and coded the construct relating to self. This construct identifies behaviors when learners judge a peer's work by relating it to their own work, reflecting on how peer's work is similar or different from themselves, either on the problem-solving process or on the results.

Table 1

Constructs, Definition, and Examples

Construct	Working definition	Examples
Commenting on the process (CP)	When evaluating peers' works, learners comment on the process of the work	"I like how you had 20 then subtracted 12 and got 8"
Commenting on the answer (CA)	When evaluating peers' works, learners comment on the outcome of the work	"You got it correct good job" "I like how you organized it but I respectfully disagree with your answer."
Relating to self (RS)	Learners judge another person's work by relating to his/her own work.	"You are right, because I got the same answer." "My strategy is like yours because I also did 3x2"

Coding the data

Two coders worked on coding the data, labeling the presence of the three constructs in each comment. To check reliability, the two coders labeled the same set of comments (N = 30) separately and compared their coding. An acceptable kappa was reached for all three constructs (κ_{CP}=0.859; κ_{CA}=0.783; κ_{RS}=1.0). Once the reliability was established, the coders proceeded to code the remainder of the comments. We labeled all 229 comments collected, obtaining 25 positive labels for CP, 65 positive labels for CA, and 51 positive labels for RS.

Building the detector

We trained four models that differed in the complexity and interpretability of their feature set, and compared their performance. For each of the feature sets, we fit a model using neural network with one hidden-layer, using the TensorFlow library in Python. We describe the process of developing and training these models in the following subsections and uploaded the Python script to a Github repository (https://bit.ly/CSCL23feedback).

Bag of words models. Bag of words (BOW), as a simple NLP approach, examines the frequency of words included in an utterance and uses the frequencies to understand the sentiment or semantic meaning of a sentence. In our study, we constructed a unigram and bigram bag-of-words models using the Natural Language Toolkit in Python. We first stemmed and removed common English stop words from each comment. With the preprocessed data, we created a set of features using the unigram and bigram methods for each of the three constructs. The unigram method treats each unique word as a vector and uses these vectors to create a frequency table, while the bigram treats any two adjacent words as a word pair, which constitutes a vector in this case. By vectorizing a comment, a frequency table is created, denoting the number of times a word or a word pair is used in the comment. With these features, we constructed a ReLu neural network model with one hidden layer to fit the model.

Part-of-speech (POS)/semantic/sentiment taggers. POS/semantic/sentiment tagging, another commonly used NLP technique, categorizes and counts words in a corpus based on the part-of-speech (grammatical classification of nouns, verbs, etc.), semantic, and sentiment groupings. Using LIWC, we extracted a set of features that are designed to reflect the content, linguistic characteristics, and sentiment in students' comments. In total, 22 features were extracted (see Github repository). With these features, we trained models using the same neural network algorithm to predict the presence of the constructs.

Sentence embedding model. A caveat of the BOW approach is its inability to make connections between words that are not adjacent, which impedes the ability to examine a sentence comprehensively. Sentence embedding addresses this issue by encoding sentences into a high-dimensional vector space, using the relationships of words learned from previously encountered sentences. We used the sentence encoder from TensorFlow's Universal Sentence Encoder large v5 (Cer et al., 2018), which generates a 512-dimension embedding based on each word in a sentence and the words surrounding it, converting the text-based input into a numerical representation as output. Using these numerical outputs, we trained a model with the same neural network.

Results

Model performance

To evaluate the model, each model was trained with 5-fold student-level cross-validation. We computed the Area Under the Receiver Operating Characteristic curve (AUC ROC) for each of the five testing folds, averaged the AUC across folds, and calculated the standard deviations (SD) to demonstrate the variability.

Overall, models that use sentence embedding outperform the other approaches on all three constructs (see Table 2). With this approach, the average AUC ROC is 0.899 for commenting on the process, 0.963 for commenting on the answer, and 0.96 for relating to self. In the following sections, we narrow the discussion to the results from the best performing models (highlighted in Table 2).

Table 2

Model Performance Measured by the Average AUC ROC

Construct	Bag of Word (unigram)	Bag of Word (bigram)	POS/Semantic/ Sentiment Tagger	Sentence Embedding
commenting on the process (CP)	0.787 (0.105)	0.798 (0.076)	0.792 (0.077)	0.899 (0.032)
commenting on the answer (CA)	0.859 (0.078)	0.847 (0.06)	0.683 (0.161)	0.963 (0.023)
relating to self (RS)	0.92 (0.059)	0.925 (0.022)	0.667 (0.199)	0.960 (0.023)

Algorithmic bias

To ensure models' fairness, we tested the performance of the sentence embedding models for students in different gender and racial/ethnic groups. However, due to small sample size (fewer than five students), comparisons were not conducted for several groups (e.g., Asian, Native American, gender non-binary students).

We find models that predict the CP and CA perform slightly better for female (AUC_{CP} = .929, AUC_{CA}= .968) than for male students (AUC_{CP} = .907, AUC_{CA}= .944); and the RS model performs slightly better for male (AUC_{RS}= .96) than for female students (AUC_{RS}= 0.937). When examining the model performance across racial/ethnic groups, we find that the CP model performs better for White (AUC_{CP} = 1) and African American (AUC_{CP} = .927), than for Hispanic/Latinx students (AUC_{CP} = .824); the CA model performs slightly better for African American (AUC_{CA}= .963) and Hispanic/Latinx (AUC_{CA}= .962) students, than for White (AUC_{CA}= .93); and the RS model performs better for White (AUC_{RS}= .992), than for Hispanic/Latinx (AUC_{RS}= .981), and for African American (AUC_{RS}= .963) students.

Discussion and conclusions

In this paper, we leveraged NLP and machine learning to automatically detect three key attributes in peer feedback. We find that sentence embedding models can reliably detect these attributes (i.e., commenting on the process, commenting on the answer, and relating to self), and are largely not algorithmically biased.

Two limitations should be addressed in future work. First, larger and more representative samples will need to be collected in order to validate model performance for a broader range of student groups. Second, although we contextualized the three constructs based on past peer review literature, the operationalization of the constructs is situated in one specific learning platform, which may cause the models to be platform-specific. Future work should study the generalizability of these models across platforms and explore how they can be adapted for use in other subjects, learning environments, and for different age groups.

Being able to automatically detect key attributes in peer feedback has several benefits. First, it allows real-time detection to inform scaffolding of peer feedback. These detectors can be used by the learning system to provide automated suggestions, reminding students to include certain components in their comments. Second, these automated detections generate data annotations that enable researchers to conduct scalable analysis on peer review. For example, by correlating these detectors to other data sources such as surveys, it will be possible to examine what motivates students to provide certain types of feedback. It will also be possible to study if contextual features in the learning content influence which types of feedback students provide, a topic difficult to study in the smaller-scale data sets directly available from qualitative coding.

Given these benefits, our future work is focused upon implementing these models in the learning platform. By leveraging the detection, we hope to provide adaptive interventions that scaffold peer feedback to support students in generating more high-quality reviews and comments, increasing the effectiveness of peer review.

References

Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., & Kurzweil, R. (2018). *Universal Sentence Encoder*

Cho, Y., & Cho, K. (2011). Peer reviewers learn from giving comments. *Instructional Science*, *39*(5), 629–643.

Darvishi, A., Khosravi, H., Sadiq, S., & Gašević, D. (2022). Incorporating AI and learning analytics to build trustworthy peer assessment systems. *British Journal of Educational Technology*, *53*(4), 844–875.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112.

Li, L., Liu, X., & Steckelberg, A. L. (2010). Assessor or assessee: How student learning improves by giving and receiving peer feedback. *British Journal of Educational Technology*, *41*(3), 525–536.

Liu, N., & Carless,D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education*, *11*(3), 279–290.

Nguyen, H., & Litman, D. (2015). Extracting Argument and Domain Words for Identifying Argument Components in Texts. *Proceedings of the 2nd Workshop on Argumentation Mining*, 22–28.

Patchan, M., Rambo-Hernandez, K., Deitz, B., & McNeill, J. (2022). Using peer assessment to improve middle school mathematical communication. *The Journal of Educational Research*, *115*(2), 146–160.

Roscoe, R., & Chi, M. (2008). Tutor learning: The role of explaining and responding to questions. *Instructional Science*, *36*(4), 321–350.

Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, *78*(1), 153–189.

Topping, K. J. (2009). Peer Assessment. *Theory Into Practice*, *48*(1), 20–27.

Zhang, J., Andres, J. M. A. L., Hutt, S., Baker, R. S., Ocumpaugh, J., Nasiar, N., Mills, C., Brooks, J., Sethuraman, S., & Young, T. (2022). Using Machine Learning to Detect SMART Model Cognitive Operations in Mathematical Problem-Solving Process. *Journal of Educational Data Mining*, *14*(3), 76-108.