Feedback on Feedback: Automated Detection of Peer Feedback Quality

Hutt, Stephen; Baker, Ryan S.; DePiro, Allison; Wang, Joann; Rhodes, Samuel; Ocumpaugh, Jaclyn; Mills, Caitlin

Introduction

Feedback is essential to successful learning and instruction. Much of the recent feedback literature has focused on instructors providing feedback to students (Van Boekel et al., 2021), however, there is also a growing body of work considering peer feedback, where learners provide feedback to other learners. The effect of peer feedback is two-fold: 1) learners receive feedback from their peers, which they can use much as they would use feedback from an instructor (Misiejuk et al., 2021), 2) learners provide feedback to others, developing both their metacognitive skills in evaluating their own work as well as important communication and teamwork skills (Donia et al., 2022).

Providing peer feedback allows students to develop a better understanding of the assessment process and criteria, and in turn, can improve self-assessment skills (Sadler, 2010). Some studies have shown that giving feedback, is just as effective, or more so, for improving a learner's understanding and communication of a topic than receiving feedback (Hattie & Timperley, 2007). Given the benefits of both giving and receiving feedback, incorporating peer-based feedback into learning technologies could significantly improve learning. However, it is not yet clear how that technology might teach students to deliver "good" feedback.

In this proposal, we consider automated "feedback on feedback" within an existing learning technology that has a peer review element. We hand-coded peer feedback comments from an online mathematics platform, where students provide comments for their peers on their problem-solving approach. We then used natural language processing and supervised machine learning to develop an automated coding process for student comments. To provide actionable insights, we interrogate the model through feature analysis to derive the most important features of successful comments.

Methods

Participants

Data was collected from 116 middle school students as they used the online math problem-solving environment CueThink.

Procedure

CueThink asks students both to solve a math problem and to create a shareable screen-cast video that provides the student's solution and demonstrates their problem-solving process. Once students have completed the problem-solving process and recorded their solution, their video is shared with their class for peer review. In this process, students annotate their peers solutions, often asking the student for their underlying reasoning or why the student picked specific methods. These annotations are then sent back to the video's author for possible revision. These annotations are the focus of the current analyses.

Three coders developed a 5-tiered code for feedback robustness, with tier 1 being the lowest and tier 5 being the highest. The coders initially coded individually and then came together to discuss disagreements and finalize the codebook. A complete set of peer feedback (N=116) was then coded to train supervised machine-learning models. The goal of this process was to develop automated evaluation of the peer feedback comments, using the human codes as ground truth variables. We first tokenized each peer-review comment and then extracted features for each comment using the nltk package in python (Loper & Bird, 2002). We then recorded the length of the comment with and without stop words. These two measures gave an impression of the length of the comment, by removing stop words, we also gain an approximate measure of the number of content words. We next counted the number of "starter" phrases (i.e., sentence scaffolds provided by CueThink) used in the comment. Finally, we generated features using nltk's 32 tags for parts of speech, removing the 7 that were not found in our data. Our final feature set was a total of 28 features.

We used the scikit-learn library (Pedregosa et al., 2011) to implement commonly-used regressors: Bayesian ridge, linear, XGBoost (via the XGBoost library (Chen & Guestrin, 2016)), Huber, and random forest. Hyperparameters were tuned on the training set using the cross-validated grid search where appropriate. We also generated a chance baseline by shuffling the codes and comparing to ground truth

(human codes). This provided a random baseline that preserved the original distribution of codes. All models were trained using 4-fold student-level cross-validation and repeated for ten iterations, each with a new random seed.

Results and Discussion

We compare model accuracy by computing the correlation between the model predictions and ground truth codes (described above in Table 1). We used the Spearman correlation coefficient (i.e., Spearman rho) since the true labels are on an ordinal scale and the model predictions are continuous All results reported are from the test folds. We note that all results are above the chance baseline, with Random Forest Regression providing the best detector.

Table 1. Automated Detection Results – Spearman's rho

Regressor	Rho
Chance	.16
Linear Regression	.85
Huber Regression	.75
Bayesian Ridge Regression	.82
Random Forest Regression	.91
XGBoost Regressor	.89

We interrogated the models to understand how features related to predictions of feedback robustness. SHapley Additive exPlanations (SHAP) values (Lundberg & Lee, 2017) as implemented in the *shap* library in python. Table 2 lists the Shapley values with the largest impact on predictions.

Table 2. High Shapley Value Variables

Predictor	Directionality	Predictor	Directionality
Digits	Positive	Unrecognized Words	Negative
Plural Nouns	Positive	Verb base form	Negative
Length of Annotation	Positive	Adverbs	Negative
Preposition/Subordinating	Positive	Modal	Negative
Determiner	Positive		
Adjective	Positive		
Verb 3rd person	Positive		

Conclusion

Peer feedback has learning benefits for both the student providing feedback, and the student receiving feedback. However, training students to give "good" feedback, remains an open question. We developed an automated detector of peer feedback quality. We use relatively simple descriptive features (i.e., length, parts of speech) to develop an automated model with high classification accuracy. This work serves as a first step towards automated, real-time, feedback-on-feedback, that can help learners develop important communication skills as well as content knowledge.

References

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Donia, M. B. L., Mach, M., O'Neill, T. A., & Brutus, S. (2022). Student satisfaction with use of an online peer feedback system. *Assessment & Evaluation in Higher Education*, *47*(2), 269–283. https://doi.org/10.1080/02602938.2021.1912286

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. *ArXiv Preprint Cs/0205028*.

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 4768–4777.

Misiejuk, K., Wasson, B., & Egelandsdal, K. (2021). Using learning analytics to understand student perceptions of peer feedback. *Computers in Human Behavior*, *117*, 106658. https://doi.org/10.1016/j.chb.2020.106658

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://doi.org/10.1007/s13398-014-0173-7.2

Sadler, D. R. (2010). Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, *35*(5), 535–550. https://doi.org/10.1080/02602930903541015

Van Boekel, M., Weisen, S., & Hufnagle, A. (2021). Feedback in the Wild: Discrepancies Between Academics' and Students' Views on the Intended Purpose and Desired Type of Feedback. *Proceedings of the 15th International Conference of the Learning Sciences-ICLS 2021.*