PILOT STUDY ON APPLYING NATURAL LANGUAGE PROCESSING TECHNIQUES TO CLASSIFY STUDENT RESPONSES TO OPEN-ENDED PROBLEMS TO IMPROVE PEER REVIEW ASSIGNMENTS

Matthew A. Verleger
Engineering Fundamentals
Embry-Riddle Aeronautical University

Abstract

As an educational tool, peer review can be a valuable way to provide students feedback without a significant increase in instructor workload. Despite all that is currently known about our students, the best mechanism for assigning reviewers to reviewees in a peer review of artifacts is still considered to be blind, random assignment. The underlying conjecture of this research project is that "there has to be a better way". This paper represents a follow-up to earlier work by the author [1]. That study presented the results of an attempt to develop a classification schema using a large *archival* database of student work. This paper takes the resulting algorithms produced from that archival dataset and applies them to new student work, identifying how well the archival-based classification works on a new data set. The implications of that application on future algorithm design will be discussed as well as the next steps for the research.

Introduction

The use of peer review is an essential part of the engineering design process. The American Society of Civil Engineers maintains an official policy, formally supporting the use of peer review in engineering [2]. As an educational tool, peer review can be a valuable way to provide students feedback without a significant increase in instructor workload. Despite all that is currently known about our students, the best mechanism for assigning reviewers to reviewees in a peer review of artifacts is still considered to be blind, random assignment. The underlying conjecture of this research project is that "there has to be a better way". Specifically, if a mechanism can be identified to accurately predict reviewer quality and reviewee need, complementing matches can be made – assigning high quality reviewers to high need reviewees and vice-versa.

In 2008, as part of his PhD dissertation work, the author explored this topic using exclusively quantitative approaches in the context of Model Eliciting Activities (MEAs) and found that the classification of reviewer quality and reviewee need was subject to numerous hidden assumptions [3]. To combat those assumptions, textual analysis would be necessary and future work has centered around Natural Language Processing (NLP) techniques for addressing and mitigating those assumptions. By analyzing the text students and teams produce, the goal is to develop a more accurate classification model for making reviewer-reviewee matches.

Background

Peer Review

Peer review has been common in scientific achievement since the mid-1950's [4], but has struggled to become widely adopted as a pedagogical tool. The challenge is in getting students to accept that their peers are valid sources of feedback. This problem has only been exacerbated by frequent news articles regarding the prevalence of pseudo-scientific journals publishing clearly unscientific work that was supposedly peer reviewed [5], calling into question the larger value of peer review. When Felisa Wolfe-Simon's team published a paper in the journal *Science* claiming to have

discovered "a bacterium that can grow by using arsenic instead of phosphorous" [6], she was met with great backlash and scorn from the broader scientific community. Taking to Twitter, she defended the paper, stating "We went through a solid peer-review and made responses and revision in response to them." [7] Clearly, the peer review process has serious flaws, even at a prestigious journal like *Science*.

Pedagogically, the largest hindrance toward broader adoption is getting students to accept that it is a viable source for feedback and assessment. In his survey of peer review literature, Topping[8] found that students rate the value of feedback from instructors more highly than that of peers, even when the feedback is purely meant to be formative. In a survey given by Ballantyne, et al [9] to 1,654 students who participated in peer review, 734 gave a response to a question regarding the worst aspect of the peer review process. 31% of those 734 responders (14% of the total) mentioned concerns about the competency of either themselves or their peers. One underlying problem becomes the "one-bad-apple" phenomenon, where a single poorly formed comment results in a wider dismissal of the review.

Conducting peer review has become a relatively trivial matter with it being built in to most modern learning management systems [10-13] and numerous custom platforms specifically built to include peer review [14-21]. While these systems all streamline the process, reducing the time-intensive overhead associated with conducting peer review, one element which all are lacking is an informed mapping system for assigning reviewers to reviewees. Most systems rely on some form of random or instructor-based assignment, not leveraging the benefits of data analytics to make more beneficial matches.

In Verleger et al. [3], individuals were given a reviewer quality score based on their performance on a calibration exercise. Teams were given a reviewee need score based on a TA's evaluation of an earlier draft of their work. Matches were then made based on those two scores in a largely complementary manner and peer review was conducted. The author reevaluated all three drafts of the work of 147 teams to quantify the changes between consecutive drafts. The primary finding from that research was a better understanding about how sensitive the algorithmic assignment was to the driving assumptions and the need for a qualitatively aware approach to rating reviewers and reviewees.

MEAs

It should be noted that large portions of this section have been published elsewhere by the author and his colleagues. The focus of this section is not to present any new or a novel synthesis of literature, but to provide the reader with a basic understanding of the essential components of the context in which this research took place without simply directing them to another source.

This research is exploring the matching process for peer review in the context of Model-Eliciting Activities to provide a proof-of-concept to expand to other contexts. Model-Eliciting Activities (MEA) are realistic, client-driven, open-ended problems that are designed to be both model-eliciting and thought-revealing [22]. They require students to mathematize (e.g., quantify, organize, dimensionalize) information in context. The solution to an MEA requires the development of one or more mathematical, scientific, or engineering concepts that are unspecified by the problem – students must grapple with their existing knowledge to develop a generalizable mathematical model to solve the problem. The point is for students to be involved in the creation of the initial ideas underlying the concept or system, thus establishing the need and motivation to go through cycles of expressing their initial ideas, testing, and refining them. The product teams produce when working on an MEA is a written document describing a generalized solution to the problem. Formative and summative

evaluation of student work in this research focuses on three overarching dimensions (mathematical model, re-usability/modifiability, and share-ability) that align to the MEA design principles [23]. These three dimensions were designed to specifically assess issues which practicing engineers valued [24]. (Later iterations of the rubric have since separated re-usability/modifiability into separate dimensions, but they were combined in this study to maintain parity with archival data). The numeric items of the MEA Rubric can be seen in Table 1.

The mathematical model dimension encompasses the assessment of (1) the quality of the solution in terms of how well it addresses the complexity of the problem and accounts for all data provided, and (2) the use of rationales to support the solution method. The root of this dimension is assessing how good the procedure is at providing a solution to the specific problem being given. Does the procedure do what it is explicitly required to do? This dimension of the MEA Rubric contains 3 items and an overall dimension score

The re-usability/modifiability dimension consists of two inter-related but subtly different concepts. The re-usability aspect focuses on the quality of the solution in terms of how easily it can be used by the client in new but similar situations. A re-usable procedure (1) identifies who the direct user is and what the direct user needs in terms of the product, criteria for success, and constraints, (2) provides an overarching description of the procedure, and (3) clarifies assumptions and limitations concerning the use of the procedure. The underlying idea is that engineers rarely develop a procedure specifically to solve a single problem, but often design solutions around a class of problems. Part of that development involves explicitly defining that class in such a way as to make it clear what problems can and cannot be solved with the given procedure.

The modifiability aspect of the re-usability/modifiability dimension assesses how well the procedure can be modified by the direct user for use in different situations. A modifiable procedure (1) contains acceptable rationales for critical steps in the procedure and (2) clearly states assumptions associated with individual procedural steps. Unlike the re-usability dimension, which defines the larger context in which a procedure can and cannot be used, the modifiability dimension is concerned with how difficult it is to modify each step in order to adapt the procedure while maintaining the team's intentions. For example, if a specific value is selected as a threshold value (e.g., "remove the top 10% of the data"), modifiability seeks to measure if the reasoning behind "10%" is made clear. The re-usability/modifiability section of the MEA Rubric item consists of a single item, which is then also used as an overall dimension score.

The share-ability dimension is used to evaluate the quality of the solution in terms of (1) how well the client can understand the procedure, and (2) how accurately the client can replicate results given in the procedure for the provided data set. A portion of this includes responding to all of the client's requests for results. An underlying component of this dimension is not only clarity, but also brevity and avoiding extraneous and unnecessary information. The share-ability dimension of the rubric consists of 3 items and an overall dimension score. The three dimension scores are then combined into an overall procedure score.

Methodology

MEA Studied

For this study, the Paper Airplane MEA was given to both the archival data group and the new data group. A more detailed description of the MEA can be found in [25], but the underlying problem involved teams developing a procedure to help judges of a paper airplane competition in awarding 4 awards; Most Accurate, Best Floater, Best Boomerang, and Best Overall. Students are given a data file

Table 1. MEA Rubric – Numerical Items.

Dim.	Item Label	Full Item Wording	Points	
Mathematical Model	Mathematical Model Complexity	The procedure fully addresses the complexity of the problem.	4	
		A procedure moderately addresses the complexity of the problem or contains embedded errors.	3	
		A procedure somewhat addresses the complexity of the problem or contains embedded errors.	2	
		Does not achieve the above level.	1	
	Data Usage	The procedure takes into account all types of data provided to generate results OR justifies not using some of the data types provided.	True	4
			False	3
	Rationales	The procedure is supported with rationales for critical steps in the procedure.	True	4
			False	3
Re-Usability/Modifiability	Re-Usability/ Modifiability	The procedure not only works for the data provided but is clearly re-usable and modifiable. Re-usability and modifiability are made clear by well articulated steps and clearly discussed assumptions about the situation and the types of data to which the procedure can be applied.	4	
		The procedure works for the data provided and might be re-usable and modifiable, but it is unclear whether the procedure is re-usable and modifiable because assumptions about the situation and/or the types of data that the procedure can be applied to are not clear or not provided.	3	
		Does not achieve the above level.	2	
Share-ability	Results	Results from applying the procedure to the data provided are presented in the form requested.	True	4
			False	1
	Audience Readability	The procedure is easy for the client to understand and replicate. All steps in the procedure are clearly and completely articulated.	4	
		The procedure is relatively easy for the client to understand and replicate. One or more of the following are needed to improve the procedure: (1) two or more steps must be written more clearly and/or (2) additional description, example calculations using the data provided, or intermediate results from the data provided are needed to clarify the steps.	3	
		Does not achieve the above level.	2	
	Extraneous Information	There is no extraneous information in the response.	True	4
			False	3

containing measurements for each competitor's Distance from Target, Time in Air, and Length of Throw for multiple runs on both a straight path and a boomerang path. Data was purposefully designed to discourage simplistic answers by introducing ties into the most obvious solution paths. The learning objective associated with this MEA is a more practical understanding of basic statistics and the limitations of common statistical tests.

Archival Data

Using a random forest approach, similar to the methodology used in the prior work [1], archival data was used to develop multiple classification trees for each of the 11 rubric items (7 specific items, 3 dimensions, and 1 overall score). For each of the 11 rubric items, 100 classification trees were created. Each iteration of tree creation begins by randomly splitting the full dataset into a training set and a testing set. The random selection of data to build the tree is what produces unique trees. This also means that a subset of the data is not used in the construction of each tree. The tree creation process utilizes a "bag of words" model, where the number of times a particular word or phrase appears in the text, including it not being used, is used to predict a particular outcome. In this case, the prediction was of the expert's evaluation of that rubric item. The results of the process are a set of predictive decision trees and a set of words utilized by those trees for making the predictions. New data can then be presented to the trees and each of the 100 trees provides a predicted value for that sample. The value selected by the majority of trees is then the final rating of the forest.

The trees were generated using archival data - the three drafts of an MEA solution produced by the 147 teams that the author re-evaluated as part of his dissertation work in 2008. This data came from a large, public, mid-west, RU/VH university's offering of an introduction to engineering course.

New Data

The author conducted the same MEA used for the archival data in an offering of the introduction to engineering course at his current institution. Unlike the archival location, the current institution is a medium sized, private, not-for-profit, Business+STEM only, Master's/M university in the southeast. Unlike the large archival data set, only 27 students on 7 teams were in the course being explored. Teams produced 3 drafts, with the first and third draft being evaluated by the author to provide feedback (and a grade) and the 2nd draft being used in the peer review process for feedback. The 2nd draft was subsequently reviewed for this study but the students did not see the ratings as part of their feedback/revision cycle.

In an effort to replicate the conditions of the archival dataset, the same assignment text and data sets were used. The author used the same slides and activities for introducing MEAs as were used in the archival iteration. The only explicitly different component was the institutional choice of the students (large public research vs. small private teaching) and the overall class size.

Results

Data from both the archive and new data sets, for all three drafts were included in the analysis. An error measurement was calculated as the difference between the expert's rubric score and the algorithm's predicted value for each of the 11 rubric items. There are two aspects of interest in this analysis. First is the overall accuracy of the algorithm. Under ideal conditions, the algorithm should exactly match the expert's evaluation. Quantifying the difference between the algorithm and the expert ratings goes to the validity of the algorithm as an accurate predictive tool to quantify a team's needs in the review process. This accuracy is calculated as a percent and is shown in Table 2.

The second aspect of interest is if the distribution of error is the same for both the archival and new offerings. This measurement leads to the generalizability of the algorithms for working in other contexts. This is calculated using χ^2 likelihood ratios between each of the rubric items and the offering. Likelihood ratios were used due to the small n value (21) of the new offering, resulting in each item having more than 20% of the cells having expected counts less than 5. The results of these tests are shown in Table 3. Un-shaded rows represent rubric items where the distribution of the algorithm's ratings are statistically significantly different – indicating that the presentation and context did have an effect on the algorithm's ability to predict the scores.

Table 2 - Algorithm Error Rates (n = 462) – Shaded Cells Unobtainable.

	Algorithm More Generous				**Expert More Generous**		
Difference (Algorithm – Expert)	**3**	**2**	**1**	**0**	**-1**	**-2**	**-3**
Mathematical Model Complexity	0.0%	0.0%	12.8%	69.9%	12.6%	4.5%	0.2%
Data Usage			11.7%	76.8%	11.5%		
Rationales			12.3%	78.1%	9.5%		
Re-Usability/ Modifiability		0.0%	4.1%	80.7%	12.1%	3.0%	
Results	20.6%	0.9%	0.2%	68.6%	0.6%	0.0%	9.1%
Audience Readability		6.5%	31.4%	62.1%	0.0%	0.0%	
Extraneous Information			19.9%	57.6%	22.5%		
Mathematical Model Dimension	0.0%	0.6%	12.6%	59.5%	21.6%	5.2%	0.4%
Re-Usability/Modifiability Dimension		0.0%	4.1%	79.2%	13.4%	3.2%	
Share-ability Dimension	3.9%	21.4%	10.8%	46.8%	9.5%	3.9%	3.7%
Final Score	1.9%	1.1%	6.3%	69.0%	16.5%	4.8%	0.4%

Table 3 - χ^2Likelihood Ratios – 2008 Archival Data versus 2014 New Data.

	Value	**df**	**Asymp. Sig. (2-sided)**
Mathematical Model Complexity	29.82	4	0.000
Data Usage	0.236	2	0.889
Rationales	56.16	4	0.000
Re-Usability/ Modifiability	16.774	3	0.001
Results	57.641	5	0.000
Audience Readability	10.546	2	0.005
Extraneous Information	2.066	2	0.356
Mathematical Model Dimension	44.271	6	0.000
Re-Usability/Modifiability Dimension	18.016	3	0.000
Share-ability Dimension	16.419	6	0.012
Final Score	8.227	6	0.222

Analysis

Analysis of the decision tree method of predicting performance was only somewhat successful and certainly leaves room for improvement. However, with predictive accuracy for each of the items averaging 68%, the algorithmic approach does hold promise. Further, only 2 of the items (Results and the Share-ability Dimension) have fewer than 90% of their predictions within 1 level of being accurate. As the Results item directly contributes to the Share-ability item, poor prediction of the results item would naturally result in a poor prediction of Share-ability. Further, because Results are based on the existence of certain numeric data more so than body text, it is difficult to autonomously identify that the correct results exist, particularly given that the values that should be displayed may be different based on each team's solution process.

Unfortunately, while the predictive process is moderately accurate, the decision trees are only minimally transferrable. While trees can be developed from the results provided by a large course such as the archival one, it is nearly impossible to generate meaningful trees with a small dataset such as the new dataset. This means that trees must be generalizable from the large classroom data to a new small-classroom context. However, only 3 of the 11 scores were considered statistically similar. The remaining 8 were all found to be statistically significantly different. This implies that, while the prediction is moderately acceptable, the trees are not consistent enough to use with new data sets.

Conclusions and Future Directions

This research sought to identify a mechanism for predicting a student teams' need for assistance based on their usage of certain words in their solution to an MEA. While the prediction rates can be somewhat accurately measured based on a common dataset, the use of decision trees is not reasonably transferrable to a new context, even when the new context is purposefully designed to be similar to the old context. One possible source of difference could be the differing populations of students between a large public research institution versus a small private teaching one, though more analysis would be needed to more accurately isolate this. For others interested in developing automated classification techniques for student work, this avenue may offer questionable value for the kind of granularity needed to address this type of problem.

Likewise, for researchers exploring peer review, while decision trees did have moderate success, clearly more work is needed to accurately predict the amount of help that a team needs. Additional options may include other NLP-based approaches such as sentiment or discourse analysis. Another option may be to test having graders quickly scan text and provide a fast "gut feeling" rating for each of the rubric items. While this would not provide detailed feedback, it may be sufficiently accurate for making algorithmic peer review assignments. Both of these concepts will be explored in future research.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No.1264005.

References

1. M. Verleger, "Using Natural Language Processing Tools to Classify Student Responses to Open-Ended Engineering Problems in Large Classes," in *American Society for Engineering Education 2014 National Conference & Exposition*, 2014.

2. "ASCE Policy Statement 351 - Peer Review," 2010. [Online]. Available: http://www.asce.org/issues-and advocacy/public-policy/policy-statement -351---peer-review/. [Accessed: 28-Jan-

2015].

3. M. Verleger, H. Diefes-Dux, M. W. Ohland, M. Besterfield-Sacre, and S. Brophy, “Challenges to Informed Peer Review Matching Algorithms,” *J. Eng. Educ.*, vol. 99, no. 4, pp. 397–408, Oct. 2010.

4. J. C. Burnham, “The Evolution of Editorial Peer Review,” *JAMA J. Am. Med. Assoc.*, vol. 263, no. 10, pp. 1323–1329, 1990.

5. J. Bohannon, “Who’s afraid of peer review?,” *Science*, vol. 342, no. 6154, pp. 60–5, Oct. 2013.

6. F. Wolfe-Simon, J. S. Blum, T. R. Kulp, G. W. Gordon, S. E. Hoeft, J. Pett-Ridge, J. F. Stolz, S. M. Webb, P. K. Weber, P. D. Davies, A. D. Anbar, and R. S. Oremland, “A Bacterium That Can Grow by Using Arsenic Instead of Phosphorus,” *Science (80-.).*, vol. 332, pp. 1163–1166, 2011.

7. F. Wolfe-Simon, “Twitter - We went through a solid peer-review and made responses and revision in response to them.,” 2010. [Online]. Available: https://twitter.com/#!/ironlisa/status/154 96350829383682.

8. K. Topping, “Peer Assessment Between Students in Colleges and Universities,” *Rev. Educ. Res.*, vol. 68, no. 3, pp. 249–276, 1998.

9. R. Ballantyne, K. Hughes, and A. Mylonas, “Developing Procedures for Implementing Peer Assessment in Large Classes Using an Action Research Process,” *Assess. Eval. High. Educ.*, vol. 27, no. 5, pp. 427–441, Sep. 2002.

10. Instructure, “What is a peer review Assignment?,” 2013. [Online]. Available: http://guides.instructure.com/s/2204/m/4152/l/54366-what-is-a-peer-review-assignment. [Accessed: 06-Aug-2013].

11. Blackboard Inc., “Self and Peer Assessment,” 2015. [Online]. Available: https://help.blackboard.com/en-us/Learn/9.1_SP_10_and_SP_11/Instructor/070_Assignments/005_Self_and_Peer_Assessment. [Accessed: 29-Jan-2015].

12. “Peer Review Assignment Type - MoodleDocs,” 2015. [Online]. Available: https://docs.moodle.org/19/en/Peer_Review_Assignment_Type. [Accessed: 29-Jan-2015].

13. “Turnitin - Peer Review,” 2015. [Online]. Available: http://turnitin.com/en_us/features/peermark. [Accessed: 29-Jan-2015].

14. O. L. Chapman, “Calibrated Peer Review - The White Paper: A Description of CPR,” 2003. [Online]. Available: http://cpr.molsci.ucla.edu/cpr/resources/documents/misc/CPR_White_Paper.pdf.

15. E. F. Gehringer, “Electronic Peer Review and Peer Grading in Computer-Science Courses,” *Thirty-second SIGCSE technical symposium on Computer Science Education* . ACM, Charlotte, North Carolina, 2001.

16. Eric Zhi-Feng Liu, S. S. J. Lin, Chi-Huang Chiu, and Shyan-Ming Yuan, “Web-based peer review: the learner as both adapter and reviewer,” *IEEE Trans. Educ.*, vol. 44, no. 3, pp. 246–251, 2001.

17. D. A. de Moreira, “No Title,” *Educ. Inf. Technol.*, vol. 8, no. 1, pp. 47–54, 2003.

18. A. H. H. Ngu, J. Shepherd, and D. Magin, “Engineering the ‘Peers’ System: The Development of a Computer-Assisted Approach to Peer Assessment,”

in *Research and Development in Higher Education*, 1995, vol. 18, pp. 582–587.

19. J. Sitthiworachart and M. Joy, "Web-based Peer Assessment in Learning Computer Programming," in *4th Annual Conference of the LTSN Centre for Information and Computer Sciences*, 2003.

20. S. Trahasch, "Towards a Flexible Peer Assessment System," *Fifth International Conference on Information Technology Based Higher Education and Training (ITHET) 2004*. Savannah, GA, 2004.

21. C. Tsai, E. Z. Liu, S. S. J. Lin, and S. Yuan, "A Networked Peer Assessment System Based on a Vee Heuristic," *Innov. Educ. Teach. Int.*, vol. 38, no. 3, pp. 220–230, 2001.

22. R. A. Lesh, M. Hoover, B. Hole, A. Kelly, and T. Post, "Principles for Developing Thought Revealing Activities for Students and Teachers," in *Handbook of Research Design in Mathematics and Science Education*, A. Kelly and R. A. Lesh, Eds. Mahwah, NJ: Lawrence Erlbaum, 2000, pp. 591–645.

23. H. A. Diefes-Dux, M. A. Hjalmarson, T. K. Miller, and R. A. Lesh, "Model-Eliciting Activities for Engineering Education," in *Models and Modeling in Engineering Education: Designing Experiences for All Students*, J. S. Zawojewski, H. Diefes-Dux, and K. Bowman, Eds. Rotterdam, The Netherlands: Sense Publishers, 2008, pp. 17–36.

24. H. A. Diefes-Dux, J. S. Zawojewski, and M. A. Hjalmarson, "Using Educational Research in the Design of Evaluation Tools for Open-Ended Problems," *Int. J. Eng. Educ.*, vol. 26, no. 4, pp. 807–819, 2010.

25. H. A. Diefes-Dux, M. Verleger, J. S. Zawojewski, and M. A. Hjalmarson, "Multi-Dimensional Tool for Assessing Student Team Solutions to Model-Eliciting Activities," *American Society for Engineering Education 2009 National Conference & Exposition*. Austin, TX, 2009.

Biographical Information

Matthew Verleger is an Assistant Professor of Engineering Fundamentals at Embry-Riddle Aeronautical University in Daytona Beach, Florida. Prior to working at Embry-Riddle, he was a faculty member at Utah State University. He holds a PhD in Engineering Education from Purdue University. His research interests are primarily focused on how students learn to develop mathematical models and on developing and using technological tools to improve students' perceptions of the instrumentality and utility value of engineering concepts.