ASEE 2019 Annual Conference
CHARGED UP FOR THE NEXT 125 Years
126TH ANNUAL CONFERENCE & EXPOSITION
ASEE

Paper ID #25230

Board 39: Designing Intelligent Review Forms for Peer Assessment: A Data-driven Approach

Zachariah J Beasley, University of South Florida

Zachariah J. Beasley is a Ph.D. candidate, teaching assistant, and instructor at the University of South Florida in the Department of Computer Science and Engineering. He received his M.S. in Computer Science from USF in May 2017 and a B.A. in Computer Science and a B.A. in Applied Mathematics from Franklin College in May 2015. His teaching and research interests include Data Mining, Natural Language Processing (sentiment analysis, text processing), Crowd sourcing, Cyberlearning, Software Engineering, and Data Structures. He is a musician (guitar and bass) and was a collegiate athlete (soccer and tennis).

Prof. Les A Piegl, University of South Florida

Les A. Piegl is a professor of computer science at the University of South Florida.

Paul Rosen, University of South Florida

Paul Rosen is an Assistant Professor at the University of South Florida in the Department of Computer Science and Engineering. He received his PhD from the Computer Science Department of Purdue University. His research interests include data visualization, topological data analysis, and computer science education. Along with his collaborators, he has received awards for best paper at PacificVis 2016, IVAPP 2016, PacificVis 2014, and SIBGRAPI 2013 and honorable mentions at the VAST Challenge 2017 and CG&A 2011 best paper.

Designing Intelligent Review Forms for Peer Assessment: A Data-driven Approach

Abstract

This evidence-based practice paper employs a data-driven, explainable, and scalable approach to the development and application of an online peer review system in computer science and engineering courses. Crowd-sourced grading through peer review is an effective evaluation methodology that 1) allows the use of meaningful assignments in large or online classes (e.g. assignments other than true/false, multiple choice, or short answer), 2) fosters learning and critical thinking in a student evaluating another's work, and 3) provides a defendable and non-biased score through the wisdom of the crowd. Although peer review is widely utilized, to the authors' best knowledge, the form and associated grading process have never been subjected to data-driven analysis and design. We present a novel, iterative approach by first gathering the most appropriate review form questions through intelligent data mining of past student reviews. During this process, key words and ideas are gathered for positive and negative sentiment dictionaries, a flag word dictionary, and a negate word dictionary. Next, we revise our grading algorithm using simulations and perturbation to determine robustness (measured by standard deviation within a section). Using the dictionaries, we leverage sentiment gathered from review comments as a quality assurance mechanism to generate a crowd comment "grade". This grade supplements the weighted average of other review form sections. The result of this semi-automated, innovative process is a peer assessment package (intelligently-designed review form and robust grading algorithm leveraging crowd sentiment) based on actual student work that can be used by an educator to confidently assign and grade meaningful open-ended assignments in any size class.

Keywords: optimal review form, automated assessment, peer review, sentiment analysis, data mining, opinion mining

1 Introduction

In our combined experience of teaching computer science and engineering courses, we have noticed a lack of engineering principles applied to assessment. Too often assessments are chosen not for how they would benefit the student, but for ease of grading. In today's increasingly large (and online) classrooms, an assessment solution must be provided that maximizes the students' ability to communicate what they know, allows them to express their creativity and independence, encourages critical thinking, and finally, is easy to grade. These constraints seem conflicting, but they are not necessarily so.

Assessment of essays in large, online, or massively open online courses (MOOCs) has increasingly turned to one of two areas: automatic essay scorers (AES) or peer review. There are many limitations of automated text grading: missing semantic meaning by focusing solely on textual features, inappropriate (short) content length, susceptibility to gaming, and evaluating factual claims [1], [2]. There are also limitations of peer grading, including grading time burden and seven common human rater errors, according to Zhang: severity/leniency, scale shrinkage, inconsistency, halo effect, stereotyping, perception difference, rater drift [2]. Zhang postulates that lack of full understanding of the rater's process could bleed into an AES process that is inaccurate. He offers three conditions for fully automated grading: 1) the internal mechanism for grading must be sufficiently transparent 2) enough evidence must be collected to validate fairness and 3) a quality-control mechanism must be available to correct poor results. For a mixture of human and automated assessment, he proposes two options: weighting both or using automated as a validation (not contributing to the rater score).

While AES solves most of the common human rater errors, it does so at the expense of failing to deeply understand the text. It is also confined to the essay domain. Scaling the number of human raters to take advantage of the wisdom of the crowd averages out individual rater errors and is highly adaptable to other assessment types (e.g. projects). Thus, we posit a third option to mixing human and automated grading, leveraging the best of both worlds in a unique way: we utilize the intelligence of peer reviewers to capture content that an AES cannot (and potentially never will): humor, irony, passion, usability, beauty. We then utilize linguistic and natural language processing (NLP) techniques in areas in which they excel—concept recognition and sentiment analysis (determining positive, negative, or neutral feedback in written text)—to assign a comment "grade".

Our research also stems from a desire to understand how students learn and what information they retain so we can tailor delivery and provide specific help and resources. It is well-known that every student is unique, that every student learns differently, and that engaged students have the best chance at success in a course. Thus, we want to customize the classroom experience, including assessment, to their needs. Our model of peer review is formative, not summative — we desire the student to learn by reviewing, not review for assessment's sake so we can assign a grade at the end of the semester. Additionally, while we acknowledge editorial review (revising and resubmitting work upon receiving feedback) is advantageous to the student [3], our process utilizes post-publication peer review so a student only reviews and learns from a final submission.

2 Previous Work

Davis, et al. note that there is no consensus even in a single discipline (biomedical journals) and a narrow focus (research article) for an optimal review form [4]. Of the fourteen general surgery journals they selected, only two questions were shared among all: overall recommendation and comments to the author. They recommend that a set of guidelines should be created to mitigate "potential gaps [that] exist in the review process." In this search for quality reviews/form, the solution is usually pursued in two areas: the grading algorithm itself (fairness) and the quality and interpretation of comments (helpfulness). Most study the helpfulness of peer reviews with the understanding that more helpful reviews contribute to more feedback being implemented in future revisions (of an essay, for example). Rather than research the review form itself, most work attempts to teach the student how to review, tweaks the software to fix comments, or assigns a best-fit reviewer to provide quality feedback.

2.1 Determining helpfulness by content

Xiong, Litmaan, and Schunn observe two review characteristics ("localization information" and "concrete solutions") that promote helpful reviews [5]. However, rather than adjust the review form to request the specific characteristics, they propose an NLP technique that examines the students' reviews and prompts them to correct/improve it. Cho notes the difficulty of an educator monitoring all peer comments as class size grows larger [6]. His system thus classifies review tagged comments from three areas of the paper (intro/theory/experimental setup, data analysis/result, and abstract/conclusion) as “helpful” or “non-helpful” based on specificity and praise. The purpose of this study was to monitor comments to filter poor or inappropriate comments before passing them to the student. Our paradigm views the purpose of reviews fundamentally differently than these and other works. We define helpfulness as clarity of review and perception of the work for the educator, not the student, especially since research has shown that it is better to give peer review than receive it [7]. Thus, we seek to understand what the student is saying, rather than to fit their review into specific characteristics.

2.2 Improving helpfulness by matching

Giannoukos, et al. focus on peer-matching to improve feedback [8]. Their process involves assigning three to five reviewers based on criteria like proficiency, strictness, usefulness, and willingness to review. Our approach differs in a key way: instead of searching for a few key reviewers (which incidentally, should have the profile of an educator), we seek as many reviewers as possible to get insights from diversity, rather than conformity. This leaves out no student (who is reviewed by someone with a poor usefulness/willingness/strictness rating) and conversely, prioritizes no student.

2.3 Creating helpfulness by good questions

Pechenizkiy et al. note the difficulties of choosing questions for their online assessment that are not too closely related in their small-scale study on data mining student data from a 73-student online exam [9]. While they do not focus on helpfulness per se, they do focus on form revision and employ clustering to determine if answering one correctly influenced answering another. This analysis could potentially be useful in our iterative process after we collect and revise our review form to prune questions. However, it is sometimes desirable for question overlap, and answering two questions correctly cannot signify causality (perhaps it simply indicates a student’s understanding of the related material).

Duers’ paper on the learner as co-creator is perhaps the closest to our idea that students should contribute to the creation of their assessment forms [10]. Their new form, built specifically by twenty-five nursing students, for nursing students, was well received by most but not all students, was condensed, and contained mostly language assessing human qualities like how “polite”, “professional” and “responsive” a student was. The study was designed to prevent nurses from feeling “torn to shreds” during peer evaluations. Unlike Duers’ study, our review form process is designed to be applicable to any field of study where students deliver information, and is based on a corpora of almost five thousand examples of what students *actually* said. Since it is anonymous peer evaluation, it is not imperative that students hold back or soften their opinion — they can express their true perception of their peers’ presentations. This work is perhaps the closest to ours in ideology, but differs widely in its breadth and implementation.

2.4 Balancing review burden and fairness

It is assumed, based on the principle of the crowd, that the more reviewers per work, the fairer evaluation will be. However, Shah, et al. proposed that peer review alone does not scale since there is a predictable proportion of incorrect peer review scores [11]. In their approach, 3-5 students review anothers' work on a pass/fail basis with the educator's grade as the ground truth. Raising the number of reviews per student can become burdensome, so they propose two methods as a form of dimensionality reduction: grouping like submissions which all receive the same grade and grouping like parts of submissions (a method also employed by Wei and Wu [12]). In either case, it is difficult to define and assess the clustering algorithm (similarity threshold, max/min number of clusters, etc.) and it seems unfair for a student's work to receive a score without actually being viewed by a peer. Their educational model differs from ours, which employs the project not simply as a universal assignment, but as a group-specific research project that requires the student to teach their reviewing peers. In addition, we create group assignments, but require students to review individually, which increases our number of available reviewers.

Kulkarni et al. combine an automated grader and peer review in two pertinent ways: 1) assigning one to three reviewers to a work depending on the confidence of the automated grader or 2) assigning one to three identifiers and one to three verifiers to annotate answer features [1]. The grade is determined by the median of the machine and human (verification) grades. TAs added attributes to the educator-provided rubric based on the subset of works they grade to determine the ground truth answers.

Identify-verify consumed the same effort as peer-median grading, for 92% of the accuracy in questions with non-binary answers. Fewer students reported liking identify-verify, and students reported low confidence in its grading accuracy, with one verifier citing a concern that other students were not reviewing properly. Roughly 34% of their submissions had high enough confidence for fewer than three reviewers. They estimated that less than 3% of students (n = 41) who should have passed the course did not due to their grading accuracy (67% to 82%).

The authors only use short answer questions that can be partitioned into components, which are evaluated for presence or absence. This reduces peer reviewing to the monotony of pattern-matching. Additionally, the authors admit to sacrificing grade accuracy (which was highest in peer-median grading) to ease review burden. However, this is not a sacrifice we are willing to make. We desire our students to have confidence in our system. Two aspects of our system lower the reviewing burden of a student: 1) group work (reduces the number of submissions) and 2) a reviewing to learn model with open-ended assignments that introduces variety and requires analysis/fosters learning.

3 Educational Principles

3.1 Learning environment

In order to accomplish our two general goals of innovation in teaching/assessment and tailoring a course to student needs, we employ a project-based class with brainstorming, teamwork, and co-creativity. Students learn the material organically, as they will once they graduate, and practice communicating information to peers clearly and concisely. This is a valuable skill in today's marketplace of ideas. The internet is a digital knowledge base that students must learn to wield effectively. Quality information must be found through careful mining, analysis, and critical thinking, then narrowed down and summarized to be understood by one's peers (see Cummings' architecture of internet-enabled learning [13] for a more detailed description). This approach (Figure 1) accommodates all learning styles and allows creativity and learning at one's own pace. It also accommodates students who begin with different levels of knowledge, allowing gaps in minimum basic knowledge to be filled as needed. The overall emphasis is on problem solving and communication, not cramming facts and regurgitating them on a piece of paper. These kinds of projects are enormously beneficial to students, but are hard to grade in a timely and objective manner in a typical large course. In short, while grading of numerical assessments can scale easily, grading higher-level open-ended assessments cannot. For example, between July 2016 and June 2017 almost five-hundred and sixty thousand students took the GRE revised general test, which requires responses to two essay prompts, for a total of over one million essays to grade [14]. Grading by hand requires an enormous amount of time and cost. Similarly, UC San Diego Professor Scott Klemmer has taught online courses in Human Computer Interaction that have garnered over 3,600 students [15]. For an individual or small team to grade an open-ended assignment (e.g. "design a website") in such a course would be effectively impossible. To solve such a scalability problem while still keeping the project-based architecture, we employ the widely-used evaluation methodology of peer assessment.

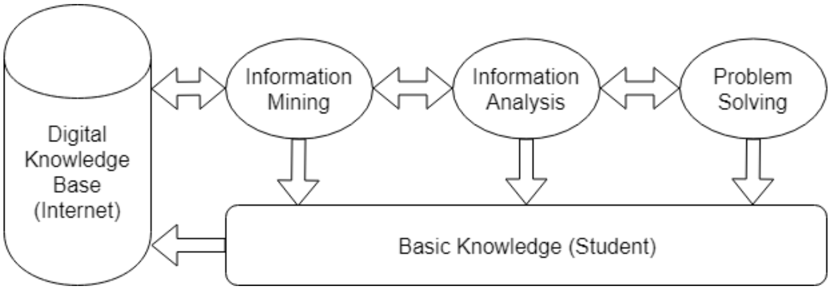


Figure 1: Learning Environment

One useful by-product of a peer review crowd is an unbaised grade. Bias exists among graders for multiple reasons: a grader may dislike (or favor) a particular student, be experiencing a particularly frustrating day, or succumb to fatigue. Any of these factors may result in a biased score. But a peer-reviewing crowd as a whole will not have these limitations. Indeed, there may be a few poor reviewers, but their marks will be averaged out by the principle of the wisdom of the crowd.

3.2 Peer Review: How to choose the questions?

Assuming that course projects are chosen well, aligning not only with the desired learning objectives, but with students' background, interests, and abilities, we then face a challenging task: how do we craft our peer review form? Sadly enough, even in the discipline of engineering, our experience (and to our best knowledge, the only practice) is for the educator to semi-randomly choose questions, assign weights to each question, and create a scoring algorithm. We use the term "semi-randomly", because there are tips and tricks to crafting a rubric (qualitative vs. quantitative, 5-point

Likert scale, analytic vs. holistic, etc.). In reality, though this rubric may weight items according to the educator's desires, it may not sample the knowledge field effectively (Figure 2) and it may not be *fair*. That is, it may not accurately capture what the students are saying, and thus its feedback is not reliable. The overarching desire is to probe students' knowledge as it relates to course content, using appropriately placed questions that adequately cover the domain (i.e. adhering to construct validity in [16]). In doing so, we force the students to study as they review. To approach asking the right questions on our review form, we asked the following questions ourselves:

- How many questions should we ask?
- How broad should their coverage be?
- Where should we place the questions?
- How should we group the questions?
- How do we weight the questions?
- How should we aggregate the answers?

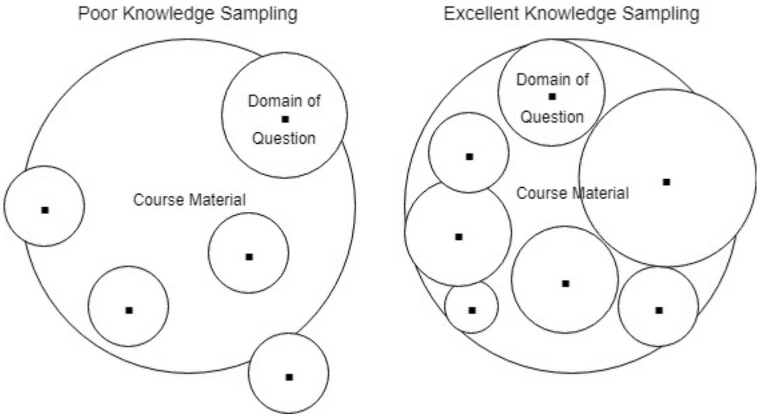


Figure 2: Knowledge Sampling

In addition to the questions, we have a number of considerations to avoid:

- Form fatigue (if the review is too long or boring)
- Too many answers per question (adds confusion and incorrect responses)
- A form that is difficult to complete (must be accessible and intuitive)

4 Review Form Design

Below we detail the most important aspects of our process, at a very high level. We do not have space to list in detail every experiment or analysis task, thus we focus on our pedagogical approach while still seeking to provide enough specific evidence to support our claims.

4.1 Iterative process

Our review form has followed a data-driven, iterative process (Figure 3). The first iteration of the form was created using the "typical" approach as outlined above: we created the form from scratch, using our intuition to choose questions and assign weights. In addition to the numerical (analytic) section, we provided a detailed comments (holistic) section where students were encouraged to write an evaluation of their peer's work. In each following iteration of the form, potential future questions were created by intelligent data combing: a process of selecting information-rich key words and phrases, through human intelligence, for the purpose of correctly analyzing and summarizing student observations. In this way, we can capture content that other automated graders cannot (and perhaps never will): humor, irony, creativity, perceived preparedness, etc. The form questions were updated to probe for the most common student feedback, and the process is repeated until we reach a steady state of questions.

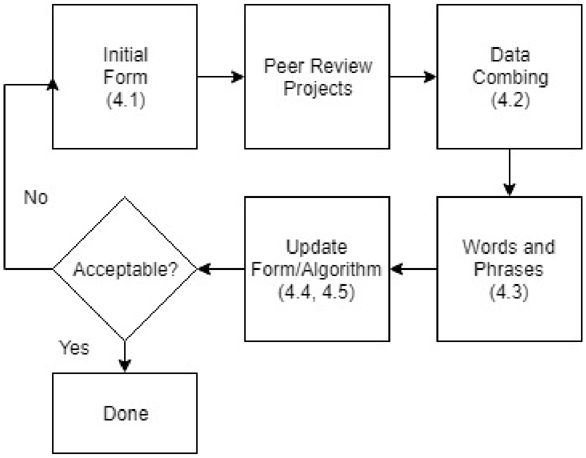


Figure 3: Seed-Growing Algorithm

4.2 Gathering questions: Intelligent data combing

We noticed student feedback separated naturally into three main clusters, thus in every iteration of the review form analytic questions were placed into: Overall score, Technical score, and Personalization score. Table 1 displays the way our review form and scoring algorithm were updated after each iteration.

	Review Form Questions				Grading Algorithm
Iteration	Total	Overall	Technical	Personalization	Description
1	11	1	4	6	Sum within section, weighted by minimum std[1]
2	11	1	4	6	Sum within section, weighted by intuition
3	22	1	16	5	Sum within section, weighted by intuition
4	22	1	16	5	Average within section, weighted by experiment

[1] standard deviation

Table 1: Review Process Iterations

The questions were chosen based on key ideas, words, and phrases gathered by hand from student detailed comments the previous semester. Human intelligence was required for this task — we

Positive	Negative	Negate	Flag
teamwork	lag	not	copy
conclusion	slow	cant	paste
preparation	heavy	no	infringement
in-depth	improve	none	cheat
coordination	lost	nothing	cheater

Table 2: Sample of Dictionary Words

could not simply select the most common feedback (e.g. "good" or "great work") because it did not add meaningful information. Instead, we cut through the noise by selecting unique words and phrases that provided rich meaning but were used frequently enough to be matched. Table 2 shows some sample dictionary key words. Questions and answers were created from the selected words and phrases and grouped based on the category under which they best fit (Section 4.3). Every question has three answers with the exception of overall score, which has eight. Answers were chosen to provide the maximum possible semantic distance between choices. For the third iteration of the review algorithm, answers were chosen to reflect the question weight of 0 (worst), 3.01 (average: a 'B'), and 4.3 (excellent: an 'A+').

The process has resulted in two different review forms to date. The seed form was designed based on teaching style, the students and their preparation, and the courses. The second iteration removed two questions and added thirteen leveraging our process for capturing sentiment. Since these factors vary widely, our particular form may not be appropriate for other courses. Although the questions are not solely limited to the field of engineering, they do reflect feedback from students in our discipline. We do not believe there is a one-size-fits-all review form — it is a mistake to use one tuned to a specific course/discipline without going through the process of iteratively mining student comments and updating the review form.

Students filled out a review within a week of viewing a group's in-class presentation and/or viewing their posted power point slides and youtube presentation. They also completed a review for multiple team essays and term projects. To date, we have over four thousand eight hundred reviews with student detailed comments.

4.3 Interpreting questions

Students were incentivized with points to provide a quality review comment. The following is one example of a positive review from which we drew concepts that provided questions and answers for our review form:

> "This presentation was very well done. The presenters **understood the material** and that was shown in their delivery. The organization of the content was such that it **promoted engagement and triggered discussion**. It was **technically accurate** and provided **a plethora of resources** to be used in the development of the final project. To me, this presentation marks an important milestone (with regards to the

> information it covers) and I am glad that the presentation enabled a **clear understanding of the material**."

Thus, we see topics that are important to the student — soundness (technically correct), resources, comprehensibility (understanding the material), and engagement. From the topics we crafted questions and answers like the following:

	• Understood at first reading
Comprehensibility	• Several readings required
	• Incomprehensible

We also took observations from negative reviews:

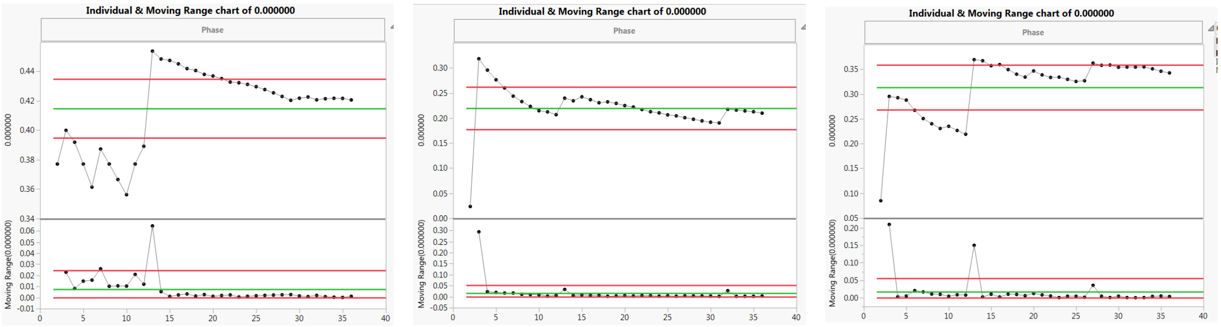
> "The only issue is that the presentation was really long. **Ridiculously long**. After a little more than halfway, **it started feeling like a slog** — well-written and useful, but a slog nonetheless."

From this review (and others), we added questions on length and creativity. The process of intelligently mining comments continued until we found no new topics and key words. We periodically renew the process with subsequent reviews to validate, modify, or add to our existing questions.

4.4 Scoring algorithm

Reviews are scored on a scale of [0, 4.3] corresponding to a letter grade: > 4.2 is an A+, > 3.8 is an A, > 3.5 is an A-, so on and so forth. The first iteration of the scoring algorithm gave a score based on weighted averages, where distance from the mean determined an answer's weight. Similarly, the second and third iterations of the scoring algorithm simply added up the question weights of each section, took the average of each section across all reviews, and took a weighted average of the section means. On the fourth (current) iteration of the scoring algorithm, we adjusted the algorithm to average the question weights (0, 3.01, or 4.3) of each section, take the average of each section across all reviews, and take a weighted average of the section means. Other works (with fewer reviewers) use the median score rather than the mean to mitigate the effect of outliers [1], [11]. Though we have run experiments using median rather than mean, the difference is negligible as our number of reviewers for presentations (35-40) is typically ten times that of other works. Finally, some propose an ordinal method for determining score (e.g. "rate these three students' works from best to worst") [17]. This is not a method we are interested in, as there are many problems with this approach in an educational setting related to fairness. Rather than pressure students to compete against one another, quality work should be recognized and encouraged, regardless of the quality of an individual's peers.

Since there is no ground truth grade for a student's work, we rely heavily on the standard deviation of students' peer reviews to validate our scoring algorithm's reliability (as opposed to educator grade). Using statistical software, we are able to plot the average and moving standard deviation and range (Figure 4). This information is valuable for determining 1) outliers (e.g. 4a and 4c, which are smoothed as additional reviews come in), 2) confidence in our grader (e.g. the Technical average standard deviation is only 0.20, or 5% — suggesting that students mostly agree in their evaluation),



(a) Overall deviation (b) Technical deviation (c) Personalization deviation

Figure 4: One team's standard deviation information

and 3) an ideal number of reviewers. For instance, we notice after about twenty reviewers, our standard deviations remain mostly within the bounding box (much lower in this instance if we only consider Technical score).

4.5 Weighting each section

In the first three iterations of the form, we used our intuited weights of 0.30 for Overall score, 0.40 for Technical score, and 0.30 for Personalization score. In the fourth (current) iteration of the form, we ran an experiment where we perturbed the weights of each section (including Detailed Comments) from [0, 1] with a step of 0.125. In analyzing the results, we only considered weight groupings that added up to a sum of 1 and gave a letter grade equal to the average letter grade of all 165 different valid combinations of weights. There were a total of 36 different student works in this experiment. Of the fourteen combinations where the letter grades matched, one was significantly closer to the actual average grade than all of the others: 0.25 for Overall score, 0.50 for Technical score, 0.125 for Personalization score, and 0.125 for Detailed Comments score. While this validates our intuitive weights, more data must be collected before we can confidently suggest an ideal weighting of sections.

4.6 Processing detailed comments

Any research on analysis of sentiment in text must choose between two diverging approaches: neural network based architecture and dictionary based architecture. Both approaches have their advantages and disadvantages including performance trade off, accuracy, length of text required, and clarity behind results. For our work, the dictionary based approach was chosen for its intuitiveness, simplicity, and our belief that it would yield clear, meaningful (explainable), and highly accurate results on short segments of text without requiring a large amount of labeled training data. Though detailed comments is one of many factors in a grade, it is important for the educator to understand how and why such a grade was chosen as a quality assurance mechanism.

Thus, our detailed comments are currently processed using standard natural language processing and linguistic methods. In brief, we use a dictionary-based approach with key words gathered through intelligent data combing and weighted by hand. While only the most common concepts were selected for review form questions, many other less common words with a positive or negative sentiment were selected to comprise the dictionaries. We intentionally exclude overused words like "good" and "bad" that provide little quality information. Our positive-word dictionary currently contains 250 words and our negative word dictionary currently contains 187 words. We have an additional dictionary comprised of words that negate sentiment (19 words) and a dictionary for flag words (12 words). The key words matched are aggregated by our algorithm, and if a threshold of key words is matched (suggesting a reliable review), the score is mapped to a range of [1.8, 4.3] — since the algorithm is still in its early stages we do not want to provide a potentially catastrophic score lower than C.

5 Future Work

In the future, we would like to extend our research in two key areas: 1) sentiment analysis and 2) review form modification. In the first area, we would like to validate our comment grading algorithm by obtaining ground truth sentiment per review by a large number of workers representing a diverse population (through Amazon Mechanical Turk or some similar means). We would like to then test our algorithm by comparing it against other open-source sentiment algorithms (both dictionary and neural network). Finally, we would like to then test our dictionary against other publicly-available dictionaries like SentiWordNet [18], MPQA [19], ANEW [20], SlangSD [21], AFINN [22], and Vader [23]. In the second area, we would like to extend our review form to analysis in different courses (i.e. non-engineering) to determine if our dictionary appropriately captures sentiment and provides an accurate score in other academic domains. We are also working on a "topological text reduction", that is, condensing all student review comments into a single, non-overwhelming and understandable visualization so the educator can confirm student sentiment and verify an appropriate score. Finally, we would like to further automate our system so that it will suggest new or signal redundant questions to fine-tune the review form.

6 Conclusion

We do not hold up our review process and form as the golden standard for reviews in every situation and for every class. Rather, we hope to provide an impetus for deeper research and a data-driven approach towards developing effective review forms. The field of educational data mining is ripe with untapped student information, which is of increasing importance and applicability in today's online educational environment. Much more collaboration between data scientists, software engineers, and educators is required for progress. In our experience, formative, post-publication peer review of student work has been well-received by students who are allowed to express their creativity and learning, has produced an incredible aggregate student work rivaling course textbooks, and has proven efficient and reliable to grade. The iterative nature of our model is one that may never reach a final resting state, however, it is accurate, scalable, and defendable, all of which are of utmost importance in assessment today.

References

[1] C. E. Kulkarni, R. Socher, M. S. Bernstein, and S. R. Klemmer, "Scaling short-answer grading by combining peer assessment with algorithmic scoring," in *Proceedings of the first ACM conference on Learning@ scale conference*. ACM, 2014, pp. 99–108.

[2] M. Zhang, "Contrasting automated and human scoring of essays," *R & D Connections*, vol. 21, no. 2, 2013.

[3] K. N. Ballantyne, G. Edmond, and B. Found, "Peer review in forensic science," *Forensic science international*, vol. 277, pp. 66–76, 2017.

[4] C. H. Davis, B. L. Bass, K. E. Behrns, K. D. Lillemoe, O. J. Garden, M. S. Roh, J. E. Lee, C. M. Balch, and T. A. Aloia, "Reviewing the review: a qualitative assessment of the peer review process in surgical journals," *Research integrity and peer review*, vol. 3, no. 1, p. 4, 2018.

[5] W. Xiong, D. Litmaan, and C. Schunn, "Natural language processing techniques for researching and improving peer feedback," *Journal of Writing Research*, vol. 4, no. 2, pp. 155–176, 2012.

[6] K. Cho, "Machine classification of peer comments in physics," in *Educational Data Mining 2008*, 2008.

[7] K. Lundstrom and W. Baker, "To give is better than to receive: The benefits of peer review to the reviewer's own writing," *Journal of second language writing*, vol. 18, no. 1, pp. 30–43, 2009.

[8] I. Giannoukos, I. Lykourentzou, G. Mpardis, V. Nikolopoulos, V. Loumos, and E. Kayafas, "An adaptive mechanism for author-reviewer matching in online peer assessment," in *Semantics in Adaptive and Personalized Services*. Springer, 2010, pp. 109–126.

[9] M. Pechenizkiy, T. Calders, E. Vasilyeva, and P. De Bra, "Mining the student assessment data: Lessons drawn from a small scale case study," in *Educational Data Mining 2008*, 2008.

[10] L. E. Duers, "The learner as co-creator: A new peer review and self-assessment feedback form created by student nurses," *Nurse education today*, vol. 58, pp. 47–52, 2017.

[11] N. B. Shah, J. Bradley, S. Balakrishnan, A. Parekh, K. Ramchandran, and M. J. Wainwright, "Some scaling laws for mooc assessments," in *KDD Workshop on Data Mining for Educational Assessment and Feedback (ASSESS 2014)*, 2014.

[12] Z. Wei and W. Wu, "A peer grading tool for moocs on programming," in *International Conference of Young Computer Scientists, Engineers and Educators*. Springer, 2015, pp. 378–385.

[13] M. L. Cummings, *Learning in the 21st century: Principles, models, environments*. U-turn Press, 2019.

[14] "Gre worldwide test taker report - july 2012-june 2017," https://www.ets.org/s/gre/pdf/snapshot_test_taker_data_2017.pdf, accessed: 2018-10-26.

[15] C. Piech, J. Huang, Z. chen, C. do, A. Ng, and D. Koller, “Tuned models of peer assessment in moocs,” *arXiv preprint arXiv*, vol. 1307, 2013.

[16] A. E. R. Association, A. P. Association, N. C. on Measurement in Education, J. C. on Standards for Educational, and P. T. (US), *Standards for educational and psychological testing*. Amer Educational Research Assn, 1999.

[17] I. Caragiannis, G. A. Krimpas, and A. A. Voudouris, “Aggregating partial rankings with applications to peer grading in massive online open courses,” in *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 2015, pp. 675–683.

[18] S. Baccianella, A. Esuli, and F. Sebastiani, “Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining.” in *LREC*, vol. 10, no. 2010, 2010, pp. 2200–2204.

[19] L. Deng and J. Wiebe, “Mpqa 3.0: An entity/event-level sentiment corpus,” in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, pp. 1323–1328.

[20] M. M. Bradley and P. J. Lang, “Affective norms for english words (anew): Instruction manual and affective ratings,” Citeseer, Tech. Rep., 1999.

[21] L. Wu, F. Morstatter, and H. Liu, “Slangsd: Building and using a sentiment dictionary of slang words for short-text sentiment classification,” *arXiv preprint arXiv:1608.05129*, 2016.

[22] F. Å. Nielsen, “A new anew: Evaluation of a word list for sentiment analysis in microblogs,” *arXiv preprint arXiv:1103.2903*, 2011.

[23] C. H. E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*, 2014.