International Society of the Learning Sciences

Analyzing Students' Written Arguments by Combining Qualitative and Computational Approaches

Julia Gouvea, Ruijie Jiang, and Shuchin Aeron
julia.gouvea@tufts.edu, ruijie.jiang@tufts.edu, shuchin@ece.tufts.edu
Tufts University

Abstract: Education researchers have proposed that qualitative and computational machine learning (ML) approaches can be productively combined to advance analyses of student-generated artifacts for evidence of engagement in scientific practices. We applied such a combined approach to written arguments excerpted from university students' biology laboratory reports. These texts are lengthy and contain multiple different features. We present two outcomes that illustrate the possible affordances of combined workflows: 1) Comparing ML and human-generated scores allowed us to identify and reanalyze mismatches, increasing our overall confidence in the coding; and 2) ML-identified word clusters allowed us to interpret the overlap in meaning between the original coding scheme and the ML predicted scores, providing insight into which features of students' writing can be used to differentiate rote from more meaningful engagement in scientific argumentation.

Introduction

An emerging line of education research positions computational approaches as "partners" of, rather than replacements for, qualitative analyses by human coders (Rosenberg & Krist, 2021; Sherin, 2013; Sherin et al., 2018). This line of inquiry aims broadly towards a "bootstrapping" program through which each approach can be improved through interaction with the other (Sherin, 2013). Each form of coding, human and machine, has its strengths and limitations. Qualitative analyses identify patterns that are meaningful for researchers but can be time-consuming and biased by what the coders are primed or able to notice. Computational approaches can find empirical patterns and statistical relationships in the data that human analysts might miss, but these patterns may not necessarily reflect meaningful constructs. When the two approaches are combined, the power to find patterns and the ability to interpret them can potentially be enhanced.

We add to this line of work by combining qualitative analyses with statistical machine learning (ML) modeling of university students' lab reports. In a prior study, our purpose in analyzing these reports was to distinguish between different types of engagement with argumentation in students' writing. Some reports suggested rote engagement in argumentation, while others contained evidence that students were more authentically engaged in the practice of scientific argumentation. In the prior work, qualitative analysts modified and applied a multi-dimensional coding scheme to categorize these differences in argumentation (Gouvea et al., under revision). In this work, we describe insights gained from a collaborative workflow in which ML approaches were applied to the dataset from this original study followed by additional qualitative analyses. After establishing an initial baseline of agreement between the two approaches, we highlight two main outcomes of this interaction. First, we compared human scores to computational predictions to identify and reanalyze mismatches between the two approaches. In this process, the ML approach functioned as a check on the consistency in human coders, flagging potentially mis-coded or borderline cases for additional review. Second, we developed a computational approach for identifying "discriminatory" words that capture differences between higher- and lower-scoring arguments. Qualitative analysts then checked these words against themes from the original coding scheme. This approach lends interpretability to the ML, a common limitation of such approaches (Lee et al., 2019).

Scientific Argumentation in Lab Reports

The original study, and context for this collaboration, is a university-level introductory biology laboratory course in which students write lab reports. Research on these reports was motivated by patterns suggesting rote engagement in argumentation. For example, many students organized their arguments around claims provided in the lab manual, even when the evidence suggested more complicated interpretations. In addition, some features of lab reports (e.g., describing sources of "error") seemed included to meet expectations rather than functioning to alter or support claims. We interpreted these features as evidence of "pseudoargumentation" that can occur when students see their role as complying with the expectations of scientific writing rather than using writing as an activity in which to explore and express their own reasoning (Berland & Hammer, 2012).

To shift how students engaged with argumentation, we undertook a redesign of the laboratory course. Our approach was to design a context motivating a "need" to engage in argumentation rather than to provide

explicit scaffolding (Berland & Reiser, 2011; Manz, 2015). We did so in two ways: (1) We increased uncertainty by designing experiments and computer simulations that allowed students to encounter more complex and potentially conflicting patterns in the results they observed, and (2) We also introduced class discussions and re-framed the purpose of lab and lab reports to emphasize the value of students' reasoning over canonical conclusions and adherence to formatting guidelines. Our analyses of lab reports detected a shift in the quality of arguments that we believe reflected a shift in how students understood and participated in the activity of writing a lab report (Gouvea et al., under revision).

In the next section, we present an overview of our original method of characterizing argumentation in the lab reports as well as a description of the computational approaches we developed to explore patterns in this dataset. We then describe how these approaches, when combined with additional qualitative analyses, led to additional understandings of the data and confidence in our analysis.

Methods

Qualitative Analysis of Students' Argumentation in Lab Reports

To identify features of students' writing that suggested meaningful engagement in argumentation, we used a modified version of the Structure of Observed Learning Outcomes (SOLO) Taxonomy (Biggs & Collis, 1982) (see Table 1). The first dimension, *argument structure,* captures the number of claims and the degree of integration among them. Arguments that make a single claim are scored at a lower level than those that identify and interrelate multiple claims. The second dimension, *scope of evidence*, accounts for the sources of evidence used to support claims. Arguments that align only a subset of evidence with a single claim score lower than those integrating multiple sources of evidence, including, at the highest level, ideas from outside the context of the lab (e.g., analogies, thought experiments, prior knowledge). Finally, the *consistency and closure* dimension, considers the certainty of conclusions. Lower-level arguments "close" prematurely, claiming to have "proved" claims with certainty even when inconsistencies remain. Higher-level arguments include expressions of uncertainty and appropriately qualify or specify conditions of applicability for claims. Taken together, these dimensions capture a range of approaches to argumentation in lab reports with higher levels reflecting more meaningful engagement, defined by increasing claim complexity, integration of evidence, and appropriate attention to uncertainty.

Table 1
Description of scoring categories for three dimensions

Dimension	Level 1	Level 2	Level 3	Level 4
Argument Structure	Tautological / Simple	One-sided, Additive	Two sided, Relational	Compound / Conditional
Scope of Evidence	Relies on given information	Aligns subset of data with claim	Relates data patterns	Relates data and includes additional information
Consistency and Closure	Closed without rationale	Closed with some rationale	Closed with inconsistencies acknowledged	Appropriately qualified or open-ended

We analyzed 146 Discussion sections of reports from three implementations of the laboratory course: the original lab and two iterations of redesign. Text length averaged about 500 words (ranging from 100 to almost 1500 words). Analysts assigned scores for each of the three dimensions independently and then determined a best fit score. Because the categories are ordinal, we calculated inter-rater reliability of initial pre-consensus coding of the full data set using Cronbach's Kappa with linear weighting (Cohen, 1968). The weighted Kappa was 0.74 (S.E. = 0.03), corresponding to a "good" level of agreement.

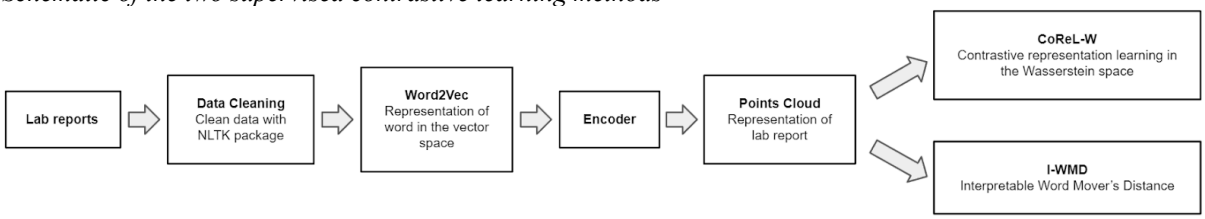
Comparing consensus scores across the three years revealed an increase in argumentation scores (χ^2=79.7; *df*=2; p<.0001). In the original labs (Y1), the median and mode score was 1.5; in the first design iteration (Y2), the median and mode shifted to 2.5, and finally, in the second design iteration (Y3), the median shifted to 3.5, while the mode increased to 4. Thus, the original analysis by human coders was able to identify a shift in students' engagement with argumentation corresponding with an increase in the uncertainty in the lab data as well as an emphasis on reasoning over compliance and correctness (Gouvea et al., under revision).

Overview of computational approaches

Other researchers have used computational machine learning approaches to characterize students' arguments. Notably, Lee et al., (2019) used the c-rater method to automate scoring of students' explanations for claims as well as their articulations of certainty. The c-rater approach detects key words and sentences and represents them as vectors. This method can successfully score short answers (50 words) that correspond to a single construct.

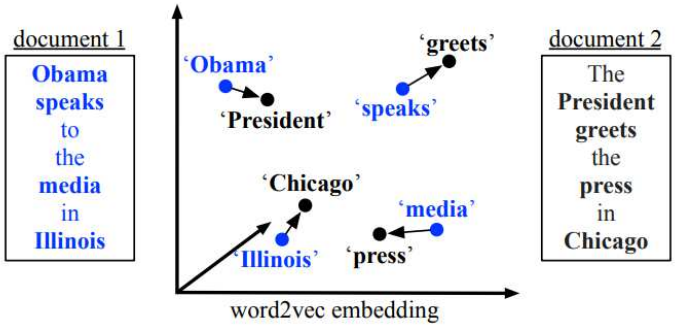
We were unable to generate successful predictions of lab report scores with c-rater (Jiang et al., 2020). Moreover, we wanted to develop an approach that would provide interpretability, a feature that is not provided in c-rater (Leacock & Chodorow, 2003). We developed two computational approaches, both of which use supervised contrastive learning to train an encoder to convert lab report text into a cluster of points in a multi-dimensional representation space. The workflow for these methods is shown schematically in Figure 1. First, text from lab reports was processed by removing punctuation and stop words, converting to lowercase, and applying stemming, and lemmatization using standard tools from the Natural Language ToolKit (NLTK) package. Then we used word embeddings (Pennington et al., 2014), pre-trained on a generic corpus, to map the lab reports to point clouds in a Euclidean space.

Figure 1
Schematic of the two supervised contrastive learning methods



We then used supervised, contrastive learning (Saunshi et al., 2019) to train an encoder. Contrastive learning trains an encoder that learns to push the documents (point-clouds) with different scores apart in a second Euclidean representation space while pulling the documents with similar scores together. In our context, the contrast is measured as the Word Mover's Distance (WMD) (Kusner et al., 2015) between the representations of the documents. Thinking of a document as point cloud generated from its words, WMD measures the distance between documents as the Wasserstein or Optimal Transport (OT) distance (Peyré & Cuturi, 2019) between the two corresponding clouds. Intuitively, WMD is the total cost of minimal matching between the words in the two documents, measured as the sum of the squared lengths of the vectors that match the words (see Figure 2).

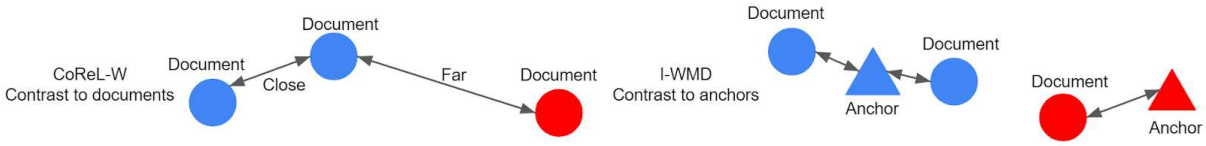
Figure 2
Example of the WMD as a minimal matching between two documents. Arrows represent minimal matching between words.



We next applied two methods that use two different contrastive loss functions for training the encoder and predicting the score (Figure 3). The first method (Figure 3a) used standard contrastive loss (Saunshi et al., 2019) with a bidirectional long-short term memory (bi-LSTM) as an encoder. Prediction was performed using k-Nearest Neighbor (kNN) with WMD as the measure of nearness. Further details can be found in (Jiang et al., 2020).

Figure 3
Two contrastive loss functions. Color denotes the categorical score given by human coders. (a) The encoder minimizes the WMD between documents with the same scores and maximizes the WMD between documents with

different scores. (b) The encoder learns "anchors" (triangles) such that the WMD between reports from the same category and the anchor is minimized.

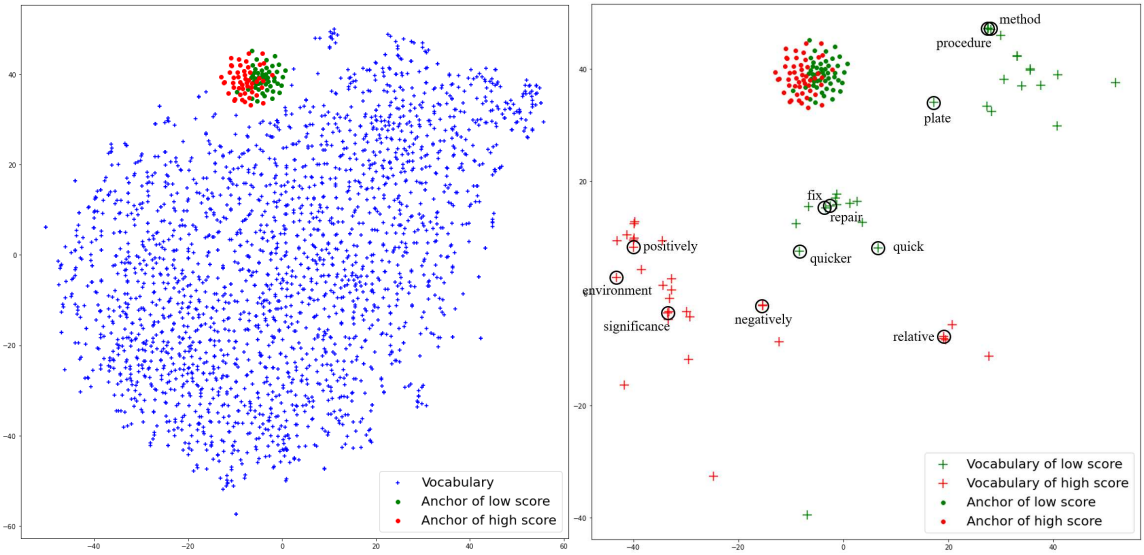


For the second method, we modified the contrastive loss to incorporate a clustering mechanism in the representation space that we used to identify the most informative words for discriminating between documents with different scores (see also Jiang et al., 2020). This is done by learning an "anchor," a cluster of points that minimizes the distance (in WMD) to all reports in the same category, as shown in Figure 3b. Because the dataset was limited in size, we combined levels to create two categories: high-scoring (assigned scores of 3 or 4 by coders) and low-scoring (score of 1 or 2) reports. This clustering also reflected a meaningful distinction in the original coding scheme between reports that presented single claims and those that considered and to some extent integrated two claims.

Given the learned anchors for the two categories corresponding to the low (C_L) and high (C_H) scores, we then computed the distance of each word (w) to each cluster and denoted these distances as $d(w, C_H)$ and $d(w, C_L)$ respectively. Then, the importance of each word was computed as $I(w) = d(w, C_L) - d(w, C_H)$. A low value for $I(w)$ means the word w is important or discriminatory for predicting a low score and vice versa for a high score. Prediction is again performed using kNN. We repeated the training four times identifying the top 30 words in each run. Figure 4 visually represents the output of a single run as t-Stochastic Neighborhood Embedding (t-SNE) visualization plots (Van der Maaten & Hilton, 2008).

Figure 4

Visualization plots of anchors and important words for a single training run. Left plot is a visualization of the complete vocabulary of the lab report text (blue) and the learned anchors for high- (red) and low- (green) scoring reports. Right plot depicts the top 30 words identified as important for each category.



Findings

Baseline score agreement and recapitulating trends

Our first step was to evaluate the ability of the statistical machine learning approach to predict argumentation scores that approached agreement with human coders. Given the relatively small size of the dataset (146 scored reports), ML predictions were calculated using a 10-fold cross-validation process in which 70% of the data was used as the training set and the other 30% as the test set. Each predicted score was calculated as an average across the runs in the test set. We then calculated the quadratic weighted kappa (QWK), as a measure of the degree of agreement between the average ML predicted scores and the consensus score assigned by human coders.[1] QWK between the ML prediction and human consensus scores was estimated at 0.645, which is considered to be a

"substantial" agreement and approaches the cut-off recommended for high-stakes assessment (QWK = 0.70) (Williamson et al., 2012). Because our aim was to examine average shifts in scores rather than predict individual scores for assessment, we considered the estimated QWK to be an acceptable level of agreement for our purposes.

In addition, we examined whether the ML-predicted scores would recapitulate the differences in scores across implementation years of the laboratory curriculum redesign. Mirroring the qualitative analysis, we found that the mean ML-predicted scores increased with implementation year (Y1 = 2.18; Y2 = 3.12; Y3 = 3.95) and that these differences were statistically significant ($F = 474$; $df = 2$, $p < .0001$; post hoc Tukey test $p < .0001$ for all comparisons). These results suggest that the contrastive learning method can assign scores at an acceptable level of reliability with human analysts, which is notable given the complex nature of the original coding scheme.

Identifying mismatches to improve scoring accuracy

We identified nine reports for which the ML prediction consistently differed from the consensus score across multiple validation runs. For five of these, the ML prediction was consistently higher than the consensus score and for four the ML prediction was consistently lower. We presented these nine reports along with a random sample of 21 additional reports to three of the four original qualitative analysts for re-scoring. This re-scoring took place over a year after the original scoring, and coders were blind to both the original score and the ML prediction. Each report was re-scored by at least two coders independently and was then discussed by all three coders to reach a consensus.

Of the nine reports for which the ML prediction consistently differed from the original human-coded consensus score, four were revised in re-coding in the direction of the ML prediction (see Table 2). In three examples the re-coding increased the score and in one example the score was lowered. The remaining five examples were not changed from their original scores.

Table 2
*Original and re-coded scores for which ML prediction consistently differed from original score (*indicates change in score in the direction of the ML prediction).*

	ML prediction higher than original score						**ML prediction lower**		
Original score	1	2	2	2	2.5	3.5	3.5	3.5	4
ML prediction	3.0	2.9	3.0	3.0	4.0	4.0	2.0	2.5	3.2
New score	2.5*	2	2	3*	4*	3.5	2.5*	3.5	3.5*

That the ML prediction matched revised scores from the re-coding process suggests the possibility that ML may be useful for identifying examples that human coders may want to revisit, either because they were coded inconsistently or because they represent difficult to code "borderline" examples. For example, an excerpt from a report that was originally scored as level 1, begins with a tautological claim about how bacteria with a higher-mutation rate mutated more often. However, at the end of the report the student included the following:

> Yet, looking at lab group six (my group) in particular, given that it is the only group with accessible photo documentation, not all dishes adhered to the expected trend of increased growth in the mutagenic [fast mutating] strain. [The fast-mutating strain] fared worse in the plain nutrient agar medium than [the slow-mutating strain], which may have been a result of a harmful mutation limiting growth in otherwise ideal conditions.

Here, the student includes an unexpected data pattern and provides a potential explanation that raises the possibility of harm due to a higher rate of mutation. Attention to this part of the text prompted coders to revise the score from 1 to 2.5, which matches the direction of the ML score of 3.

Identifying discriminatory words

The anchor-based WMD method generated a set of 72 unique important words for high-scoring reports and 64 unique words for low-scoring reports. To make meaning of these words, we first located all instances of each word in context to review its use in students' writing. For example, we found that all instances of the word "positive" referred to positive or adaptive effects of mutation. We then clustered synonymous words together. For example, the words "negative", "detrimental", and "harmful" were each used in high-scoring reports to refer to the negative effects of mutation. Finally, we counted frequency with which the discriminatory words appeared in high and low scoring reports. Table 3 shows the top five most frequent sets of discriminatory words for both high- and low-scoring reports.

Table 3
Most common clusters and counts of discriminatory words in each set of reports

High-scoring reports		Low scoring reports	
"environment"	444	"plate", "dish", or "tube"	583
"positive" and synonyms	239	"procedure" "method" and synonyms	550
"negative" and synonyms	114	"repair" "fix" and synonyms	274
"comparative" "relative" and related words	91	[repair] "mechanism" or "system"	139
"importance" "significance" and related words	32	"quick" or "quicker" [adapting]	46

For both high- and low-scoring reports, the word clusters reflect features of writing that align with categories from the original coding scheme. A key feature of higher scoring arguments was the integration of two possible claims about mutation rate. One type of claim focused on the positive effects of mutation such as allowing organisms to generate novel mutations and adapt to new environments. Another type of claim attended to how mutations can disrupt function and lead to a decrease in fitness. The word clusters for "positive" and "negative" respectively, align with these two sides of the argument represented in the "Argument Structure" dimension of the coding scheme. Moreover, a hallmark of high-scoring arguments was an attempt to relate these two claims together to discuss their relative effects. Words such as "comparative" and "relative" reflect the relational structure of these for high-scoring arguments. Similarly, words such as "importance" and "significance," while not inherently comparative, also appear to reflect students' attempts to emphasize certain claims or evidence over others. Finally, the word "environment" is associated with conditions on claims. Students referred to specific types of environments (e.g., antibiotic) in specifying conditions in which a higher or lower rate of mutation would have an advantage. The following excerpt from a high-scoring report, illustrates how words that differentiate high and low scoring reports appear in the context of a Level 4 report (discriminatory words underlined).

> Moreover, using computer simulation of bacterial growth, which was another part of the lab, supported the idea that mutation rate is advantageous on an environmental level, and the fitness of an organism is comparative to the environmental conditions as there are situations where the lower mutator strain has higher fitness or the higher mutator strain has higher fitness, and where both strains will co-exist. Overall, the high mutator strain seemed to outcompete the low mutator strain in situations where the mutations gave the organism an advantageous mutation, such as metabolic benefit.

In this excerpt, the student is presenting an argument for context-dependent advantage, specifically arguing that, while there are many possible outcomes, a high mutator would be expected to have an advantage in environments in which there is an opportunity to benefit from a novel mutation (in this case the ability to metabolize a new nutrient). Note also that this excerpt contains other markers of a high scoring reports including the integration of evidence beyond the experimental data (computer simulation) and hedging language that keeps the overall conclusion open (e.g., "seemed"), but that these words were not identified as discriminatory.

Low scoring reports were characterized by arguments that restated claims given by the lab manual. Most often this meant stating that the experiment showed that bacteria with dysfunctional DNA repair systems would mutate more often (a fact given in the lab manual). Words such as "repair" or "fix" and "system" or "mechanism," indicate that this phrase was important for identifying low-scoring reports in the anchor-based ML model. Level 2 reports additionally included claims that mutations can be adaptive (without considering harmful consequences of mutation). Use of the words "quick" or "quicker" (which were always associated with the word "adapt" in student text) reflect claims that bacteria with higher rates of mutation are also able to more quickly adapt to new situations. The following examples illustrate discriminatory words in the context of lower-scoring reports. The first illustrates the pattern of aligning the experiment with the given fact about DNA repair:

> The results above conform to the hypothesis that as the efficiency of DNA repair mechanisms increases the mutation rate decreases. Part one of the experiment showed that under the environment of antibiotics, [the fast mutating] strain with suboptimal DNA repair mechanisms survived more than [the slow mutating] strain with functional DNA repair mechanisms.

Note also the use of words such as "conform" and "showed" that signal a high level of certainty in this conclusion.

A second excerpt illustrates how words related to experimental materials (plate, dish, tube) and words like "procedure" and "method" were used.

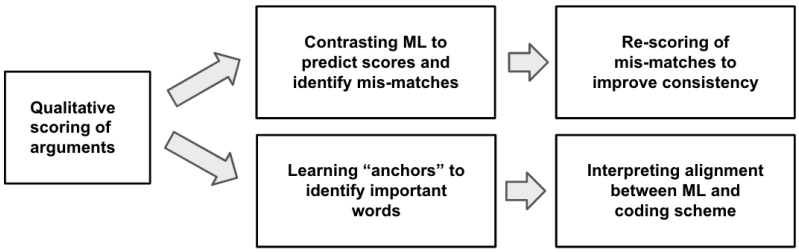
> Were there to be improper plating or diluting procedures, this would also affect the data because there may be fewer or more bacteria in the plate than the experiment calls for, which would lead to underestimating or overestimating growth rates and thereby mutation rates.

In this excerpt, the student references procedures to discuss possible sources of error involving experimental materials (diluting and then pipetting bacteria onto the plate). This is an example of listing possible errors that do not alter the conclusions of the report. The error in this case could have resulted in under or over-estimation of bacteria. While this is a reasonable attribution of error, it amounts to a general statement about improving data quality that does not impact the main claims of the argument.

Discussion

In this work, we examined students' writing as a means of gaining insight into their engagement with scientific argumentation. These texts were lengthy (500 words on average) and contained many different features. Our aim was to examine the utility of combining computational and qualitative analyses of these texts (see Figure 5). The purpose of the qualitative analysis was to learn about how students in different design-iterations of a laboratory course were engaging in argumentation through an examination of three inter-related features of arguments; claim structure, scope of evidence used, and consistency and closure of conclusions. These dimensions were collapsed into a "best fit" score to compare reports across different years.

Figure 5
Summary of our collaborative workflow



The comparative ML approach was relatively successful in predicting scores that agreed with human coders, where other established ML methods, such as c-rater, were not. We believe that the reason for this success is that our approach used point cloud representations of documents rather than collapsing them into a single vector. While this method did a reasonable job of predicting scores, we are not suggesting that ML methods replace human analysts. Rather, in our workflow, the ML algorithm functioned as an additional coder whose conjectures about how to categorize the data could inform discussions that increased our overall confidence in our ability to reliably categorize texts. This pathway suggests a possible role for ML methods complementing the skills of human analysts by identifying empirical outliers that may require additional discussion.

Second, the anchor-based WMD approach provided us with a way to interpret ML predictions by associating them with clusters of words that we could compare to our original coding scheme. This analysis reflected important themes from the qualitative coding scheme. For example, high-scoring words reflected the relational quality of high-scoring arguments, while low scoring words reflected the tendency to repeat given claims. Interestingly, words that conveyed certainty (i.e., "showed", "proved) or uncertainty ("seemed", "may have") did not emerge as important for discriminating high and low scoring reports, as we might have expected given their importance in the original coding scheme. This may be due to the small size of the original dataset, or it may be that these words are used in both high- and low-scoring reports, but potentially in different ways. Investigating how uncertainty is expressed by students and detected by ML approaches is an area for future work.

As others have argued, combining qualitative and computational analyses does not necessarily make the process of data analysis faster or more efficient (Sherin, 2013; Sherin et al., 2018). Indeed, a limitation of this work is that supervised ML requires large, coded datasets in order to learn patterns. Thus, there is a relatively significant activation energy to starting this work. Additional work is then required to check the meaning of patterns, as we did in this case by re-coding mismatches and by examining how the discriminatory words function in student text. Despite this effort, this work contributes to a growing understanding of how to combine approaches to manage the challenges of extracting meaningful patterns from complex student-generated texts.

Endnotes

(1) The QWK among human coders was estimated as 0.839 (compared with LWK = 0.74). Note that quadratic-weighted kappa scores are often higher than linear-weighted scores.

References

Berland, L. K., & Hammer, D. (2012). Framing for scientific argumentation. *Journal of Research in Science Teaching*, *49*(1), 68–94. https://doi.org/10.1002/tea.20446

Berland, L. K., & Reiser, B. J. (2011). Classroom communities' adaptations of the practice of scientific argumentation. *Science Education*, *95*(2), 191–216. https://doi.org/10.1002/sce.20420

Biggs, J., & Collis, K. F. (1982). *Evaluating the Quality of Learning: The SOLO Taxonomy (Structure of the Observed Learning Outcome)*. Academic Press.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213–220.

Gouvea, J., Appleby, L., Fu, L., & Wagh, A. (under revision). Motivating and shaping argumentation in lab reports.

Jiang, R., Gouvea, J., Hammer, D., Miller, E., & Aeron, S. (2020, November 26). Automatic coding of students' writing via Contrastive Representation Learning in the Wasserstein space. arXiv preprint arXiv:2011.13384.

Kusner M, Sun Y, Kolkin N, & Weinberger K. (2015, June). From word embeddings to document distances. *International conference on machine learning* (pp. 957-966). PMLR.

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, *37*(4), 389-405.

Lee, H.-S., McNamara, D., Bracey, Z. B., Wilson, C., Osborne, J., Haudek, K. C., Liu, O. L., Pallant, A., Gerard, L., Linn, M. C., & Sherin, B. L. (2019). Computerized text analysis: Assessment and research potentials for promoting learning. *Computer-Supported Collaborative Learning Conference, CSCL*, *2*, 743–750.

Manz, E. (2015). Representing Student Argumentation as Functionally Emergent From Scientific Activity. *Review of Educational Research*, *85*(4), 553–590. https://doi.org/10.3102/0034654314558490

Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. *In Proceedings of the 2014 conference on empirical methods in natural language processing* (EMNLP) (pp. 1532-1543).

Peyré, G. & Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning, 11*(5-6), 355-607.

Rosenberg, J. M., & Krist, C. (2021). Combining Machine Learning and Qualitative Methods to Elaborate Students' Ideas About the Generality of their Model-Based Explanations. *Journal of Science Education and Technology*, *30*(2), 255–267. https://doi.org/10.1007/s10956-020-09862-4

Saunshi N, Plevrakis O, Arora S, Khodak M, & Khandeparkar H. (2019, May). A theoretical analysis of contrastive unsupervised representation learning. *International Conference on Machine Learning* (pp. 5628-5637). PMLR.

Sherin, B. L. (2013). A Computational Study of Commonsense Science: An Exploration in the Automated Analysis of Clinical Interview Data. *Journal of the Learning Sciences*, *22*(4), 600–638. https://doi.org/10.1080/10508406.2013.836654

Sherin, B. L., Kersting, N. B., & Berland, M. (2018). Learning Analytics in Support of Qualitative Analysis. *Rethinking Learning in the Digital Age: Making the Learning Sciences Count, 13th International Conference of the Learning Sciences (ICLS) 2018*, *1*, 464–471.

Van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, *9*(11).

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Acknowledgments

We acknowledge funding support for J. Gouvea by the Davis Educational Foundation and seed funding from Tufts University Innovates, to R. Jiang by NSF DRL 1931978 and NSF EEC 1937057, and to S. Aeron by NSF CAREER CCF 1553075, NSF DRL 1931978, and NSF EEC 1937057. We wish to thank Eric Miller and David Hammer for discussions that contributed to this manuscript.