Automated Predictive Assessment from Unstructured Student Writing

Norma C. Ming[1, 2]
[1] Nexus Research & Policy Center
San Francisco CA, USA

[2] Graduate School of Education
UC Berkeley, Berkeley CA, USA
Norma@NexusResearchCenter.org

Vivienne L. Ming[3, 4]
[3] Socos LLC
Berkeley CA, USA

[4] Redwood Center for Theoretical Neuroscience
UC Berkeley, Berkeley CA, USA
neuraltheory@socos.me

***Abstract*—We investigated the validity of applying topic modeling to unstructured student writing from online class discussion forums to predict students' final grades. Using only student discussion data from introductory courses in biology and economics, both probabilistic latent semantic analysis (pLSA) and hierarchical latent Dirichlet allocation (hLDA) produced significantly better than chance predictions which improved with additional data collected over the duration of the course. Hierarchical latent Dirichlet allocation yielded superior predictions, suggesting the feasibility of mining student data to derive conceptual hierarchies. Results indicate that topic modeling of student-generated text may offer useful formative assessment information about students' conceptual knowledge.**

Keywords-Predictive assessment; learning analytics; text mining; topic modeling; online discussion.

I. INTRODUCTION

Effective instruction depends on formative assessment to discover and monitor student understanding [1]. By revealing what students already know and what they need to learn, it enables teachers to build on existing knowledge and provide appropriate scaffolding [2]. If such information is both timely and specific, it can serve as valuable feedback to teachers and students and improve achievement [3][4].

Yet incorporating and interpreting ongoing, meaningful assessment into the learning environment remains a challenge for many reasons. Most teachers lack training in assessing understanding beyond the established testing culture [5]. Designed as summative assessments for an outside audience, externally designed tests offer limited information to teachers and students, with reduced opportunities for more varied and frequent assessment, long gaps between taking a test and receiving feedback from it, and often only coarse-grained feedback on the performance of groups of students on broad areas. Even for highly skilled teachers who can infer their students' knowledge from informal assessment activities, aggregating and examining data in detail is both time-consuming and difficult.

Further, testing is often intrusive, demanding that teachers interrupt their regular instruction to administer the test. Assessments that have been developed in conjunction with prepackaged curricula typically require adapting one's own instruction to incorporate at least some of their learning activities. Whether due to differences in state standards, particular student needs, or unique local contexts, teachers may not always be free to adopt externally developed teaching and testing materials.

Our proposed solution to these problems is to build a system which relies on the wealth of unstructured data that students generate from the learning activities their teachers already use. Using automated machine intelligence to analyze large quantities of passively collected data can free up instructors' time to focus on improving their instruction, informed by their own data as well as those of other teachers and students. Building an assessment tool which they can invisibly layer atop their chosen instructional methods affords them both autonomy and information.

This paper describes the design of the system, the data source, and the techniques used. A discussion of the results and their implications for future work follow.

II. DESIGN OF THE SYSTEM

Validating any assessment requires aligning it with outcomes of value. What those outcomes are or should be can vary; our intent here is simply to demonstrate that unstructured student data have predictive value, not to make any claims about what those desirable learning outcomes should be. As a proof of concept, we are predicting end-of-course grades, although the same approach may be applied to many other assessments.

As inputs, our system relies on what students actually do, rather than information associated with their identities and backgrounds. Other predictive analytics systems include data on demographics, schooling history, and measures of motivation [6][7], variables which have been shown to predict student retention and performance but which also may reflect prior social, economic, or cultural inequities. Since our goal is to provide predictive information to the teachers and students, we hope to avoid exacerbating these inequities by minimizing the visibility (but not denying the reality) of these influences [8].

Numerous other academic analytics systems incorporate measures of student activity and course performance [9][10][11][12]. We seek to go beyond simple quantity-based metrics of effort, participation, and engagement, by analyzing the semantic content of student-generated products in order to assess what students know, not how active they

ISBN: 978-1-61208-242-4

are. This also enables deeper insights into the ideas which students are addressing, rather than vocabulary, punctuation, sentence complexity, or other linguistic features that signal writing quality [13]. While many computer-aided or intelligent tutoring systems incorporate sophisticated analyses of students' performance on prespecified problems [14], our goal is to explore nuances in student knowledge from a wider diversity of learning experiences.

To elucidate the semantic content of unstructured text data, we employ probabilistic latent semantic analysis (pLSA) and hierarchical latent Dirichlet allocation (hLDA) [15][16]. These techniques yield topic models of the student-generated texts by analyzing word co-occurrence within documents, specifically discussion forum posts in this case. Both pLSA and hLDA are generative, probabilistic models which provide low-dimensional descriptions of text by inferring small sets of latent factors, or topics, which explain the distribution of words in the analyzed documents. Both employ the simplifying "bag-of-words" assumption that a document can be represented as an unordered count of words. Each document is "generated" by mixing topics and then selecting words from those topic mixtures. We expanded on this by using collocation information to automatically select a set of domain-specific phrases (*n*-grams) of arbitrary word length [17]. Topics then describe the distribution of these *n*-grams, including single words and phrases.

For pLSA, the distributions describing topic and *n*-gram likelihoods are assumed to be Gaussian, providing a simple model of document composition. Although pLSA has its limitations, it and its predecessor LSA are widely used to model the semantic content of text. We will use the topics it infers from student posts, specifically the inferred coefficients of the latent factors, as the predictor variables for students' final grades. In other words, over the duration of a multi-week class, do the concepts discussed by students as inferred by pLSA predict their course outcomes? If so, how does the accuracy of these predictions change over time as more student work is analyzed?

The hLDA model provides a much more complex model of the text, combining a more reasonable multinomial model of word occurrence with the ability to infer an arbitrary hierarchy to the topics. Documents are not just mixtures of topics but mixtures of topic branches in a semantic tree structure. A draw from the distribution over branches defines a general topic, while a draw from the distribution of branch depth defines the specificity of the *n*-gram within a branch, such as science → biology → neuroscience → sensory neuroscience → cortical vision → etc. This hierarchical organization, as learned from the data, provides an additional piece of information to test for our outcome predictions. That is, does the specificity of a topic in student posts, as represented by depth in the hLDA tree, aid in predicting course outcomes?

TABLE I. SUMMARY STATISTICS ON THE TWO DATASETS ANALYZED

	Biology	**Economics**
Course length (in weeks)	5	6
# of discussion question threads per class	10	12
# of classes	17	45
# of students (after filtering)	230	970
# of posts by students	9118	44345

III. DATASET AND METHODS

We independently applied pLSA and hLDA to archived data from the online discussion forums of two introductory courses at a large for-profit university, an undergraduate biology course and an MBA-level economics course (**Table 1**). The biology course sample was limited to focus on instructors analyzed in previous research [18], while the economics course sample represented all such classes available in a particular archived database. Throughout both courses, students were required to respond to two discussion questions per week in these forums. Although individual instances (classes) of a course could vary slightly in the specific questions and assignments posed to students, all adhered to a standard course outline and schedule with regard to learning goals, topics covered, and texts used. We removed data from students who dropped out before earning a final grade and normalized final grades to be between [0,1].

We trained a logistic regression model to predict the final grades based on the accumulated weekly topic coefficients. While logistic regression certainly will not yield best-in-class performance, its simplicity and transparency allow for a cleaner analysis of the topic models. Both models were trained in batches with the student posts using the same *n*-gram dictionaries. No normative material was used for training, only student posts. We used five-fold cross-validation and trained pLSA and hLDA on individual posts independently for each course. The data were partitioned into five sets, and on five separate rounds of training and test, a different set was held out for testing while the remainder was used for both unsupervised and supervised training. The results reported below were averaged across the five training rounds.

The results from pLSA had 30 factors, while hLDA produced a tree composed of 50 nodes with a maximum depth of four layers. Once the pLSA topics and hLDA trees were learned, we took weekly posts by the students and projected them into topic space and the semantic tree, respectively. We then used the inferred coefficients for the topic factors from the current week's posts, concatenated with any from previous weeks, as predictor variables. For hLDA, each topic produced pairs of inferred coefficients representing both the loading of the topic and the topic's depth in the tree hierarchy.

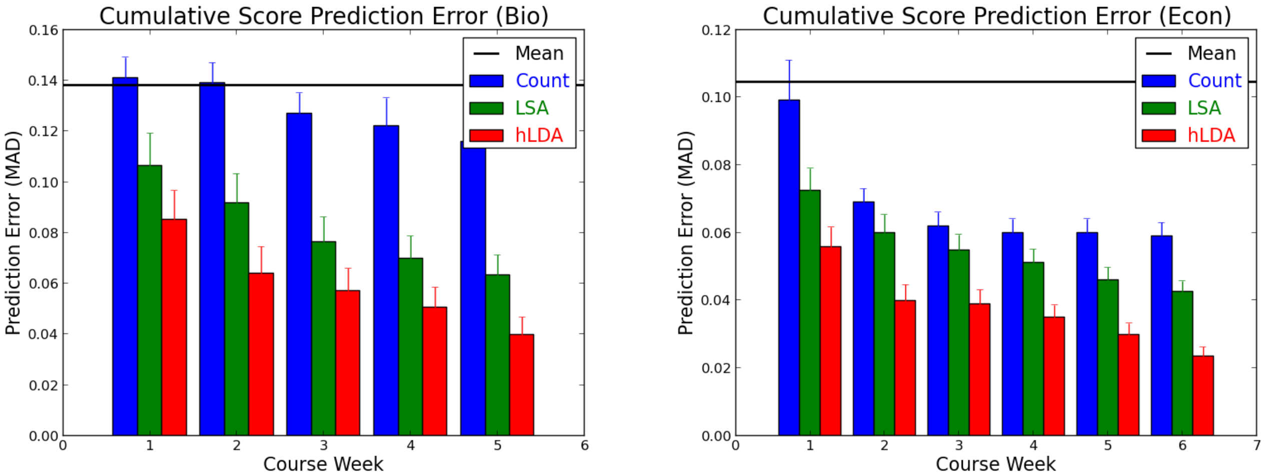

Figure 1. Accuracy of pLSA and hLDA in predicting students' final grades from the topics in their discussion posts (MAD = mean absolute deviation).

IV. Results

The results shown in Figure 1 illuminate all three of our research questions regarding the efficacy of topic modeling in predicting students' grades. The graphs depict the mean absolute deviation (MAD) between students' predicted and actual final grades for each model over the duration of the course. For reference, the black line shows the error that would result from predicting the course mean, and the blue bars show the prediction based on word count per post.

The first finding is that topic modeling using pLSA produces significantly better than chance predictions of students' course grades, even from the first week of the course. Topic modeling also produces consistently better predictions than post length for the biology data, with a smaller advantage for the economics data. Second, accumulating data over additional weeks of the course yields significant modest gains. For example, by the end of the economics course, the pLSA model's prediction is approximately within ±0.05 of the actual grade (or within one letter grade). Third, the hierarchical modeling of hLDA gives better predictions than pLSA.

Additional examination of the data also reveals that higher course grades are correlated with a slightly higher mean of the depth parameter in hLDA. Topics in the hLDA model are structured in a hierarchy learned from the data, with more specialized topics being represented deeper in the hierarchy than more general topics. The central topic is the most general language that appears in all documents, and thus appears at the topmost level (lowest level of depth) in the hierarchy. Figure 2 depicts the percentage of *n*-grams used by students receiving letter grades of A, B, and C at each of the four depth levels specified in the hierarchy. As shown, most of the language used by students who receive C's resides at the topmost level, while relatively greater percentages of the language used by students receiving A's and B's reside at deeper levels in the hierarchy.

A preliminary reading of selected discussion posts indicated that higher grades correlated with more technically proficient language use. Posts containing more general language tended to include more anecdotal comments, whereas posts with more technically specific language addressed course concepts in greater depth. Deeper analysis of potential relationships between these metrics and post quality will be valuable for elucidating how hierarchy depth may correspond with discussion and course characteristics of interest.

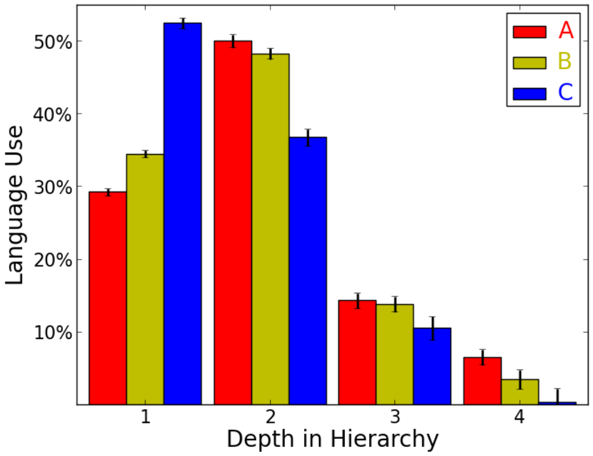

Figure 2. Correlation between language depth and course grade.

V. Conclusion and Future Work

This work demonstrated that unstructured student data in the form of discussion forum posts can be used to predict assessment outcomes of interest (final course grades). It extends previous research investigating LSA and related computational approaches to predicting student outcomes from text data [19][20][21], illustrating the predictive value of adding more data over time as well as the utility of hierarchical topic modeling. The improvements over the duration of the course may reflect the benefits of including a broader range of course concepts and using more recent student performance data. The advantages of hLDA reveal the possibility and benefit of algorithmically discovering domain-specific conceptual hierarchies in student-generated text. The correspondence between hierarchical depth and course grade suggests just one dimension of knowledge which hierarchical modeling reveals from students' writing; further analysis may enable identifying and interpreting other dimensions.

As such these methods show potential for application as a type of formative assessment, to provide more content-relevant feedback to students and teachers about students' thinking in order to better guide learning and instruction. Continued research will be worthwhile for exploring the impact of changes to the algorithms, as well as the inputs (*e.g.*, essays, responses to short-answer questions) and outcomes (*e.g.*, course retention, scores on exams or other assignments). Additional steps include exploring how to present this feedback usefully and possible interventions for students and teachers to follow. While we opted to focus on semantic content alone for this project, future work may investigate the relative value of combining semantic data with other features of performance and additional student data.

REFERENCES

[1] J. W. Pellegrino, N. Chudowsky, and R. Glaser, Knowing What Students Know: The Science and Design of Educational Assessment. Washington, DC: National Academy Press, 2001.

[2] L. A. Shepard, The role of classroom assessment in teaching and learning. (CSE Technical Report 517). Los Angeles CA: Center for the Study of Evaluation, 2000.

[3] A. N. Kluger and A. DeNisi, "Effects of feedback intervention on performance," Psychological Bulletin, vol. 119(2), 1996, pp. 254-284.

[4] P. Black and D. Wiliam, Assessment and classroom learning, in Assessment in Education: Principles, Policy, and Practice, vol. 5(1), 1998, pp. 7-74.

[5] M. C. Ellwein and M. E. Graue, "Assessment as a way of knowing children," in Making schooling multicultural: Campus and classroom, C. A. Grant and M. L. Gomez, Eds. Englewood Cliffs, NJ: Merrill, 1996.

[6] J. P. Campbell, P. B. DeBlois, and D. G. Oblinger, "Academic analytics: A new tool for a new era," EDUCAUSE Review, vol. 42(4), 2007, pp. 41-57.

[7] L. V. Morris, S. S. Wu, and C. Finnegan, "Predicting retention in online general education courses," American Journal of Distance Education, vol. 19(1), 2005, pp. 23-36.

[8] C. M. Steele and J. Aronson, "Stereotype threat and the intellectual test-performance of African-Americans," Journal of Personality and Social Psychology, vol. 69(5), 1995, pp. 797-811.

[9] K. E. Arnold, "*Signals*: Applying academic analytics," EDUCAUSE Quarterly, vol. 33(1), 2010.

[10] E. J. M. Lauría and J. Baron, Mining Sakai to measure student performance: Opportunities and challenges in academic analytics, 2011. Retrieved from http://ecc.marist.edu/conf2011/materials/LauriaECC2011-%20Mining%20Sakai%20to%20Measure%20Student%20Performance%20-%20final.pdf

[11] L. P. Macfadyen and S. Dawson, "Mining LMS data to develop an ‘‘early warning system" for educators: A proof of concept," Computers and Education, vol. 54, 2010, pp. 588-599.

[12] H. Zhang, K. Almeroth, A. Knight, M. Bulger, and R. Mayer, "Moodog: Tracking Students' Online Learning Activities," in Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, Vancouver, Canada, 2007.

[13] S. A. Crossley, R. Roscoe, A. Graesser, and D. S. McNamara, "Predicting human scores of essay quality using computational indices of linguistic and textual features," in G. Biswas, S. Bull, J. Kay, and A. Mitrovic, Eds. Proceedings of the 15th International Conference on Artificial Intelligence in Education, Auckland, New Zealand: AIED, 2011, pp. 438-440.

[14] K. VanLehn, "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems," Educational Psychologist, vol. 46(4), 2011, pp. 197-221.

[15] T. Hofmann, "Probabilistic Latent Semantic Indexing," Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, 1999, pp. 50-57.

[16] D. Blei, T. Griffiths, and M. Jordan, "The nested Chinese restaurant process and Bayesian nonparametric inference of topic hierarchies," Journal of the Association for Computing Machinery, vol. 57(2), 2010, pp. 1-30.

[17] D. Blei and J. Lafferty, Visualizing topics with multiword expressions, 2009. arXiv:0907.1013

[18] N. C. Ming and E. P. S. Baumer, "Using text mining to characterize online discussion facilitation," Journal of Asynchronous Learning Networks, vol. 15(2), 2011.

[19] B. Rehder, M. E. Schreiner, M. B. W. Wolfe, D. Laham, T. K. Landauer, and W. Kintsch, "Using latent semantic analysis to assess knowledge: Some technical considerations," Discourse Processes, vol. 25(2-3), 1998, pp. 337-354.

[20] P. W. Foltz, D. Laham, and T. K. Landauer, "Automated essay scoring: Applications to educational technology," in Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications, B. Collis and R. Oliver, Eds. Chesapeake VA: AACE, 1999, pp. 939-944.

[21] M. J. Ventura, D. R. Franchescetti, P. Pennumatsa, A. C. Graesser, G. T. Jackson, X. Hu, Z. Cai, and the Tutoring Research Group. "Combining computational models of short essay grading for conceptual physics problems," in Intelligent Tutoring Systems, J. C. Lester, R. M. Vicari, and F. Paraguacu, Eds. Berlin, Germany: Springer, 2004, pp. 423-431.