Machine Learning and Student Performance in Teams

Rohan Ahuja, Daniyal Khan, Sara Tahir, Magdalene Wang, Danilo Symonette, Shimei Pan, Simon Stacey(✉), and Don Engel

University of Maryland Baltimore County, Baltimore, MD, USA
{rahuja2,dkhan1,sarata1,wangmag1,danilo2,shimei, spstacey,donengel}@umbc.edu

Abstract. This project applies a variety of machine learning algorithms to the interactions of first year college students using the GroupMe messaging platform to collaborate online on a team project. The project assesses the efficacy of these techniques in predicting existing measures of team member performance, generated by self- and peer assessment through the Comprehensive Assessment of Team Member Effectiveness (CATME) tool. We employed a wide range of machine learning classifiers (SVM, KNN, Random Forests, Logistic Regression, Bernoulli Naive Bayes) and a range of features (generated by a socio-linguistic text analysis program, Doc2Vec, and TF-IDF) to predict individual team member performance. Our results suggest machine learning models hold out the possibility of providing accurate, real-time information about team and team member behaviors that instructors can use to support students engaged in team-based work, though challenges remain.

Keywords: Machine learning · Teamwork · Performance prediction · Text mining

1 Introduction

Teamwork skills are vital for college students, both while they are at university [7] and for their employability and success after graduation [4]. This is true across the board, for students in a wide variety of disciplines [3,5,6,13]. Despite great interest in supporting and developing student teamwork skills, there are relatively few tools available to help instructors do so [2] and the few tools that do exist are often focused on fairly artificial and controlled experimental settings rather than robust teaching environments [10] or suffer from other shortcomings [2,12]. This paper reports on the collection of teamwork data "from the wild" in

This work was supported by the National Science Foundation under Grant No. 1339265. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. Support was also provided through a Hrabowski Innovation Fund Innovation and Research Grant from the University of Maryland, Baltimore County.

I. I. Bittencourt et al. (Eds.): AIED 2020, LNAI 12164, pp. 301–305, 2020.
https://doi.org/10.1007/978-3-030-52240-7_55

a deliberately non-intrusive manner, and the subsequent machine learning driven analysis of these data to identify high performing and non-high performing team members. Although just making this discrimination is not on its own enough to support team members and their teams, this is an important initial step towards developing a more broad-ranging program that can do so. We hope with this effort to begin to remedy the dispiriting conclusion of a recent article that "no study has shown that technological support for group regulation can help teams to improve their course-based, collaborative discourse over time [1]."

2 Data Collection

The data for this project come from two semesters of a mandatory, two credit, Pass or Fail class for freshman students in the Honors College at a midsize American university, enrolling about 100 students each year, divided into 12 teams of 8–9 students, each with a non-freshman team leader. The students in the classes came from a very wide variety of majors, and one of the primary requirements of the class was that each team identify a social issue or problem in the city near the campus, research it, and propose a multidisciplinary approach to addressing it. Teams had the entire fifteen-week semester to work on the project. Team member performance was assessed through the Comprehensive Assessment of Team Member Effectiveness (CATME) tool [9]. Twice during the semester, students completed CATME self and peer-assessments, in which they completed a report on their own and their team-members' contributions to the work of their team. CATME calculates a total for each team member for each dimension on the basis of all the assessments a team member receives (including his or her own), averages those scores and then uses an "adjustment factor" to accommodate the fact that some teams may assess more generously than others. CATME scores form a continuum, so to dichotomously categorize team members for analysis we used CATME's "high performer" definition- team members with an average rating of 3.5 out of the available 5 points, and with an overall rating at least half a point above their teammates' average rating. The Fall 2018 class had 36 high performers, and the Fall 2019 class had 22 high performers. (We used the end rather than middle of semester CATME assessments, when team members had the most information on which to base their evaluations.) The class met for two hours every week, but because little of that class time was available for project work, much of the work on team projects took place online, using the GroupMe messaging platform. Data for the project were collected by adding a dummy member to each team's GroupMe group, after obtaining written informed consent from each student. The 94 students who participated in the Fall 2018 GroupMe chats yielded an approximately 5000 message transcript, and the 100 students who participated in the Fall 2019 GroupMe chats generated an approximately 6000 message transcript.

Table 1. 10-fold cross-validation accuracy and macro-f1 scores for machine learning models that were trained to predict high performing team members. The first column shows the features or combination of features that were used as input for machine learning, and the best models were first found individually for several algorithms such as Logistic Regression, K-Nearest Neighbors, SVM, Naive Bayes and Random Forests, using grid searches for hyper-parameter tuning, followed by selection of the best performing model among these different models.

Method	Accuracy	Macro-F1 score
Dummy classifier with "most frequent" strategy	0.698	0.411
Doc2Vec embedding only	0.762	0.699
LIWC only	0.766	0.714
TF-IDF + Doc2Vec embedding	0.928	0.906
TF-IDF only	**0.959**	**0.947**

3 Methods, Analysis and Results

We explored a range of machine learning models to predict high performing students, including Logistic Regression, K-Nearest Neighbors, SVM, Naive Bayes and Random Forests. Based on ten-fold cross-validated Macro-average F1 scores, SVM with Recursive Feature Elimination proved to be the best-performing model overall, with its tendency to reduce overfitting as an added benefit. With the model selected, we trained it to predict high performers using several features, some in combination with others, with the results reflected in Table 1. Among them, TF-IDF scores are frequently used to represent text in text mining and information retrieval. Linguistic Inquiry and Word Count (LIWC) [11] is an off-the-shelf linguistic analysis tool, which categorizes words into roughly eighty different psychologically meaningful categories, signaling attentional focus, attitudes, perceptions, emotionality, social relationships, thinking styles, and authenticity, etc. Doc2vec is a neural network-based text embedding method that automatically learns a dense vector representation of each document/message [8]. Among all the features, TF-IDF scores proved to be the most effective in predicting high performers (0.959 prediction accuracy and 0.947 F1), followed by LIWC features. Although Doc2vec embedding and LIWC both out-perform the Dummy Classifier with "most frequent" strategy baseline significantly, adding them to TF-IDF does not improve performance (see Table 1).

4 Conclusion and Discussion

This project investigates whether machine learning analysis of the text messages of online team member exchanges can discriminate high performing from non-high performing team members. The work demonstrates the potential of such automatic assessments of online student teamwork, and provides some initial

pointers about which machine learning approaches are most effective. Near term future work will involve refining these most promising approaches.

One major potential benefit of automatically assessing online teamwork is that it can provide instructors with this information on a real-time or near real-time basis (e.g., in a team performance dashboard), which is important to their making timely decisions about what corrective or supportive actions to take. Furthermore, this benefit is available without the significant outlay of time or energy by instructors it would take for instructors to attempt to assess the quality and trajectory of a team's work themselves. That time and energy can then be devoted to instruction and to the more challenging tasks of determining whether, when and how to intervene.

But several challenges remain. First, we have so far explored only data generated by team members using text-based platforms. This simplified the data collection process, but limited the range of data we had to analyze. In particular, we have so far collected team member interactions neither from online verbal conversations between team members (on Zoom, WebEx, Blackboard Collaborate, etc), nor from in-person conversations between team members. Such conversations are likely to be richer in data, but are technically more challenging to capture and process. In addition, the capture of conversations of this type also raises more serious questions about student expectations of and rights to privacy. Still, as the COVID-19 crisis forces universities to move classes online in the Northern hemisphere's 2020 summer (and perhaps fall), an important if regrettable opportunity to collect data from classes with a teamwork component is presenting itself.

The second challenge is of a different sort- how to represent the findings of these models to instructors in ways which are intelligible and actionable. Using SVM with recursive feature elimination and focusing on TF-IDF features produced the best predictions of high performing team members, but it would be difficult for an instructor to know what to do to support student team members identified as non-high performing, because the features used to make the predictions are so low-level. No matter how predictively potent it is, it is likely that instructors, especially in non-STEM fields, will resist adopting a pedagogical tool if its workings are opaque to them. The challenge, then, is to retain the accuracy of a model like the one that performed best, while making its findings intelligible and usable. For example, somehow grouping the features the model relies on in understandable categories (perhaps, even, categories of the kind employed by LIWC) would allow instructors to identify the kinds of missteps in communicative behavior occurring in student teams. An important focus of future work, then, will be to try to retain the predictive power of low level feature-based models but to add to those models a measure of interpretability and intelligibility that makes them useful instructional tools.

References

1. Borge, M., Ong, Y.S., Rosé, C.P.: Learning to monitor and regulate collective thinking processes. Int. J. Comput.-Support. Collab. Learn. **13**(1), 61–92 (2018). https://doi.org/10.1007/s11412-018-9270-5
2. Britton, E., Simper, N., Leger, A., Stephenson, J.: Assessing teamwork in undergraduate education: a measurement tool to evaluate individual teamwork skills. Assess. Evaluation High. Educ. **42**(3), 378–397 (2017)
3. Earnest, M.A., Williams, J., Aagaard, E.M.: Toward an optimal pedagogy for teamwork. Acad. Med. **92**(10), 1378–1381 (2017)
4. Hart Research Associates: Raising the Bar: Employers' Views on College Learning in the Wake of the Economic Downturn. Hart Research Associates (2009)
5. Hastie, C., Fahy, K., Parratt, J.: The development of a rubric for peer assessment of individual teamwork skills in undergraduate midwifery students. Women Birth **27**(3), 220–226 (2014)
6. Ibrahim, B., DeMiranda, M.A., Lashari, T.A., Siller, Y.J.: Teamwork and engineering design outcomes: examining the relationship among engineering undergraduate students. In: 2017 7th World Engineering Education Forum (WEEF), pp. 628–635. WEEF, Kuala Lumpur (2017)
7. Kuh, G.: High-Impact Educational Practices: What They Are, Who Has Access to Them, and Why They Matter. Association of American Colleges and Universities, Washington, DC (2008)
8. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on International Conference on Machine Learning, ICML 2014, vol. 32, p. II-1188–II-1196. JMLR.org (2014)
9. Ohland, M.W., et al.: The comprehensive assessment of team member effectiveness: development of a behaviorally anchored rating scale for self- and peer evaluation. Acad. Manag. Learn. Educ. **11**(4), 609–630 (2013)
10. Stewart, A.E.B., et al.: I say, you say, we say: Using spoken language to model sociocognitive processes during computer-supported collaborative problem solving. In: Proceedings of the ACM on Human-Computer Interaction, pp. 1–19. CSCW (2019)
11. Tausczik, Y., Pennebaker, J.: The psychological meaning of words: LIWC and computerized text analysis methods. J. Lang. Soc. Psychol. **29**(1), 24–54 (2010)
12. Vivian, R., Falkner, K., Falkner, N.: Analysing computer science students' teamwork role adoption in an online self-organized teamwork activity. In: Proceedings of the 13th Koli Calling International Conference on Computing Education Research, Koli Calling 2013, pp. 105–114, Koli, Finland (2013)
13. Weinstein, J., Morton, L., Taras, H., Reznik, V.: Teaching teamwork to law students. J. Leg. Educ. **63**(1), 36–64 (2013)