Towards the Automatic Assessment of Student Teamwork

Rohan Ahuja
Daniyal Khan
Danilo Symonette
University of Maryland, Baltimore County
Baltimore, MD 21250, USA
rahuja2@umbc.edu
dkhan1@umbc.edu
danilo1@umbc.edu

Shimei Pan
Simon Stacey
Don Engel
University of Maryland, Baltimore County
Baltimore, MD 21250, USA
shimei@umbc.edu
spstacey@umbc.edu
donengel@umbc.edu

GROUP'20 Companion, January 6–8, 2020, Sanibel Island, FL, USA.

ACM ISBN 978-1-4503-6767-7/20/01.
http://dx.doi.org/10.1145/3323994.3369894

Abstract

Teamwork skills are crucial for college students, both at university and afterwards. At many universities, teams are increasingly using discussion platforms such as GroupMe and Slack to work virtually. However, little has been done so far to understand how to use the data these platforms generate to analyze student teamwork behaviors, and so to support or improve those behaviors. Furthermore, these data have not been exploited to determine whether effective student team members share any other traits. This project therefore attempts to determine (a) whether there are any characteristics common to the online discussion behaviors displayed by high-performing vs non high-performing student team members and (b) whether high-performing vs non high-performing student team members share any apparently teamwork-exogenous attributes. We find that the features of team member communication that best predict team member performance are sentence length and the number of words contributed to the team's discussion, with a range of other features playing a smaller role. We also find that teamwork-exogenous factors (such as pre-college ACT score, and number of credits attempted during the semester) were only moderately predictive of team member performance.

CCS Concepts
•Human-centered computing → Computer supported cooperative work; *Collaborative interaction;* **•Applied computing → Collaborative learning;**

Author Keywords
Teamwork; teamwork assessment; machine learning; online collaboration

Introduction
Teamwork skills are crucial for college students, both for their learning while at university and for their employability and career success after graduation. [2, 1] Students often choose or are required to use online discussion platforms such as Slack and GroupMe to work on team projects. The record of their communications on these platforms could provide valuable insights about effective team membership and behaviors. This paper collects the exchanges of twelve teams of undergraduate first year students using GroupMe as they completed a semester-long project in an introductory class in Fall 2018, and analyzes student contributions to answer two questions. First, is there a relationship between the characteristics of the messages sent by team members and how they were rated by their peers on the Comprehensive Assessment of Team Member Effectiveness (CATME)? Specifically, do the messages sent by high-performing team members have any characteristics in common? Second, is there any correlation between a team member's CATME rating and that team member's broader academic characteristics and trajectory? That is, did team members who were judged by their peers to have performed well exhibit any commonalities beyond the project and the class in which it took place? This is useful information to have for a variety of reasons, but at least because it might help to make team composition/team member selection more scientific.[3] More generally, building a model to automatically predict a student's performance as a team member based on his/her exchanges with teammates is itself a significant contribution.

Methods and Analysis
The data for the project come from a mandatory class for freshman students in the Honors College at a midsize American university. This two-credit pass/fail class enrolled 96 students in Fall 2018, divided into 12 teams of 8-9 students, each with an upper-class team leader. A requirement of the class was that each team complete a semester-long project, culminating in a final presentation. The class met only once a week for just two hours, which left little time for students to work on their projects in class. As a result, much of each team's project work happened online. Data were collected by adding a dummy member to each team's GroupMe group, after obtaining written informed consent from each student.

Students completed two self and peer-assessments using CATME, one in the middle and one at the end of the semester. We used the final assessment, which allowed students to make more informed and more accurate judgments. Grades were not awarded in this pass-fail class, so we used only the CATME score to assess the performance of team members. CATME calculates a total score for each team member on the basis of all the assessments a team member receives (including his or her own), and then uses an "adjustment factor" ("the average rating of the student divided by the overall average rating for all members of the team") to accommodate the fact that some teams may assess more generously than others. CATME scores form a continuum, so determining how to categorize team members is somewhat complicated. We addressed this problem using CATME's high-performer definition- team members with an average rating of 3.5 out of 5, and with an overall rating more than half a point above their teammates' aver-

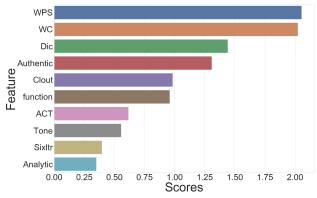

Figure 1: Importance of LIWC features to CATME score prediction, ranked by chi-squared scores.

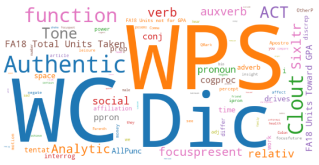

Figure 2: A word cloud representing relative importance of LIWC features to team member performance.

age rating. We compared the contributions of the 36 high performers in the class to the contributions of all other students.

We analyzed the content and form of team member contributions using Linguistic Inquiry and Word Count (LIWC), an off-the-shelf computational linguistics program [4], and a range of machine learning techniques. To answer our first question, we included data about formal characteristics like frequency/number/length of team member contributions; total words contributed per team member; average word length; etc. We also included data about the content of user messages using LIWC's categorization of user messages according to about eighty different psychologically meaningful categories, signaling attentional focus, attitudes, perceptions, emotionality, social relationships, thinking styles, authenticity, etc. To answer our second question, we considered two aspects of team members' academic histories: ACT scores (or concordance table SAT equivalents), and the number of credits they attempted in Fall 2018.

We used a feature selection algorithm which ranks features based on chi-squared scores. This produced a list of the LIWC features most important to accurately predicting high-performing team members. However, this algorithm provides no indication of whether these important features are positively or negatively correlated with high performance, so we performed a grid search to identify the logistic regression model with the best cross-validation accuracy score, and then determined the polarity of the coefficients. This model has a ten-fold cross-validation accuracy score of 78.13%, a score markedly better than the 63.54% that would have been obtained simply by labelling all participants as non high-performers.

Results

The nine features most strongly correlated with CATME high performance are listed below in decreasing order of importance, along with a description of a message exhibiting that feature (where necessary):

- Words per sentence (WPS)
- Total word count (WC)
- Dictionary words (Dic): Message contains a high proportion of words in LIWC's Internal Dictionary, which contains a large range of standard English words (almost 6400).
- Authentic: Message is honest, personal, and unguarded.
- Clout: Message indicates expertise and confidence on the part of its author.
- Function: Message contains a large proportion of words playing a syntactical, but not content-conveying, role in communication.
- Tone: Message is positive and upbeat.
- Six letter (sixltr): Message contains a large proportion of words of six letters or greater.
- Analytic: Message expresses formal, logical, and hierarchical thinking.

Logistic regression showed positive coefficients for word count, dictionary words, function, tone, and analytic; and negative coefficients for words per sentence, authentic, clout, and six letter words. Figure 2 represents the relative importance of these features to team member performance in a word cloud.

We also explored whether two teamwork-exogenous attributes were related to team performance. We found ACT score and the number of credits attempted by the team member in the semester of the project to be correlated with teamwork performance. ACT score was the seventh most important feature when incorporated into the model, and

Feature	Coefficient
Words per sentence	-9.78
Word Count	10.11
Dictionary Words	13.13
Authentic	8.55
Clout	-12.35
Function	1.14
Tone	14.45
Six letter	-12.22
Analytic	11.95
ACT	-28.66
Credits Attempted	13.41

Table 1: Logistic Regression coefficients for important features.

	Accuracy	F1-score
Baseline	63.54%	0%
SVM	82.29%	76.44%
LR	78.13%	70.15%
Naive Bayes	76.04%	64.38%
KNN	73.96%	63.62%
RF	71.88%	56.95%

Table 2: A table showing 10-fold cross validation accuracy and F1 scores of machine learning algorithms such as Support Vector Machines(SVM), Logistic Regression (LR), Naive Bayes, K-Nearest Neighbors (KNN) and Random Forests(RF).

total credits attempted was a fairly distant 16th. Logistic regression showed - surprisingly - a negative coefficient for ACT score and a positive coefficient for credits attempted.

We also performed a grid search on several machine learning models beyond the one reported on here (SVM, random forests, naive Bayes, K-nearest neighbors), and SVM had the highest ten-fold cross-validation accuracy at 82.29%, with precision of 77.05% and recall of 82.29%.

Discussion, Limitations and Future Work

A crude way to represent some of our findings is that students with lower ACT scores and (maybe) higher course loads who communicate a lot, but in short sentences that are upbeat and analytical, are better team members than students with higher ACT scores and (perhaps) lower course loads who communicate less, but in longer, authoritative, authentic sentences with longer words. There may be several explanations for these findings. First, the class was pass-fail, which meant that several good potential measures of team-member performance were not available, such as individual student grades and instructor assessments of final projects. This left us with CATME scores, and these may simply be a poor measure of team member ability. Other factors may have played a role. In a pass-fail class, some students may not have taken the teamwork project very seriously, which may in turn mean their communications were not representative of those of truly participatory team members. Also, the fact that these students were all Honors College students may also have been a confounding factor. For instance, the ACT scores of these highly capable students are clustered toward the upper end of the range, which may mean the discriminating effect one would have expected from them was diminished. Some elements of this picture make sense, but there is much that is puzzling about it, and this suggests that this early, proof-of-concept level work has some way to go.

A first step is to gather data from a more diverse, and graded, range of classes, and data are currently being collected from two engineering and two computer science classes. More data may confirm some of our so far fairly counter-intuitive findings, but we could at least then defend them more robustly. Second, LIWC is a powerful general purpose linguistic analysis tool, but a tool specifically designed for analyzing teamwork interactions, or customizable to be so, may generate more accurate predictions about team member performance. Third, while insights of this type may be useful to instructors, they may benefit team members too, and innovative, non-threatening ways to convey insights tailored to specific team members could be developed.

REFERENCES

[1] Hart Research Associates. 2009. *Raising the Bar: Employers' Views on College Learning in the Wake of the Economic Downturn*. Hart Research Associates, Washington, DC.

[2] The Conference Board. 2008. *New Graduates' Workforce Readiness: The Mid-Market Perspective*. The Conference Board, New York, NY.

[3] Emily Britton, Natalie Simper, Andrew Leger, and Jenn Stephenson. 2017. Assessing Teamwork in Undergraduate Education: A Measurement Tool to Evaluate Individual Teamwork Skills. *Assessment and Evaluation in Higher Education* 42, 3 (2017), 378–397.

[4] Yla Tausczik and James Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29, 1 (2010), 24–54.