Text Analysis of Deliberative Skills in Undergraduate Online Dialogue: Using L1 Regularized Logistic Regression with Psycholinguistic Features

Xiaoxi Xu, Tom Murray and Beverly Park Woolf
School of Computer Science
University of Massachusetts, Amherst, MA
xiaoxi@, tmurray@, bev@cs.umass.edu

Abstract: We report on a text analysis and machine learning study of social deliberative skill using online dialogues on controversial topics from a college class. We report on our comparison between using the LIWC and Coh-Metrix text analysis feature sets, as well as demographic feature information in an L1 Regularized Logistic Regression machine learning algorithm.

Keywords: social deliberative skills, online dialogue and collaboration, machine learning, text classification

Introduction

Leaders, organizations, and nations are increasingly faced with complex issues requiring higher order thinking skills, and in particular what we call "social deliberative skills." King & Baxter [1] note that "in times of increased global interdependence, producing interculturally competent citizens who can engage in informed, ethical decision-making when confronted with problems that involve a diversity of perspectives is becoming an urgent educational priority…these skills; however, 'they are what corporations find in shortest supply among entry-level candidates' [2]." Jordan et al. [3] propose two important skill sets for skillfully addressing "complex societal issues, such as gang-related crime, deteriorating residential areas, environmental problems, long-term youth unemployment, [and] racist violence" (p. 34.) Jordan calls these skill sets "complexity awareness" and "perspective awareness," and they have significant overlap with social deliberative skills.[1] Ill-defined (or "wicked") social problems are defined as: have many interacting factors; have multiple stakeholders with heterogeneous viewpoints; are chronic and, while improvable, are not completely 'solvable' in any decisive sense; and require ongoing flexible attention because conditions evolve over time [3] [4]. Though these characteristics are used to describe intransigent social problems, they define many mundane situations as well. Parenting, perusing a career, maintaining intimate relationships, planning and managing a project, and "composing a life" in such a way as to balance one's many needs and constraints—these all present one with mini wicked problems on a regular basis. They require complexity awareness and perspective awareness to address the mental and moral demands of modern life [5]. It is important that we support the development of these skills in the educational systems.

Our overall research goals are to better *understand, assess, and support* SD-skills in online contexts. A prerequisite to researching how to support SD-skills is being able to *measure*, identify or assess them. This paper describes an aspect of our ongoing attempts to assess SD-skills using linguistic models. As part of our work investigating online support of SD-skills we have developed a hand-coding scheme for categorizing segments of online text. It has been used to evaluate software features in college classes, with encouraging results [6]. In parallel we are using text classification tools and machine learning to develop automated methods to categorize text to ascertain SD-skills and related indicators of deliberative dialogue quality, which we report on here. We are using automated assessment of SD-skills for two purposes: (1) to assess skill differences and correlations in our evaluative research [7] [8] [9], and (2) to display facets of

[1] According to Jordan: "Complexity awareness [is] a person's propensity to notice…that phenomena are compounded and variable, depend on varying conditions, are results of causal processes that may be…multivariate and systemic, and are embedded in processes [that involve non-simple information feedback loops]…If a person does not notice the complexity in which an issue is embedded, he or she will fail to consider many conditions, causes and consequences that may be significant for managing the issue (Kuhn, 1991)…Perspective awareness [is] the propensity to notice and operate with properties of one's own and others' perspectives" (Jordan, 2013, p. 41, italics added).

social deliberative skill in a Facilitators Dashboard that gives facilitators and teachers a birds-eye view of important deliberative properties of an online conversation [6].

We focus on the following social deliberative skills or capacities, which are seen repeatedly in the literature (described using a variety of terms):

- Social perspective taking (includes cognitive empathy, reciprocal role taking);
- Social perspective seeking (includes social inquiry, question asking skills);
- Social perspective monitoring (includes self-reflection, meta-dialogue); and
- Social perspective weighing (related to "reflective reasoning" and includes comparing and contrasting the available views, including those of participants and external sources and experts).

Capacities implied in the above include: tolerance for uncertainty, ambiguity, disagreement, paradox, and the ability to take first, second, and third-person perspectives on situations or issues (i.e. subjective, intersubjective (you/we/they), and objective).

Here we describe a continuation of prior work using the text analysis systems LIWC and Coh-Metrix to generate features for machine learning methods to assess SD-skills. In this paper we apply our methods to a new domain and add demographic data (gender and grade level) to the machine learning trials to assess the relative effectiveness of various methods in producing the most accurate machine learning model.

Background and Related Work

Automatic text analysis (including a wide variety of computational methods: supervised learning, latent semantic analysis, topic modeling, etc.) has been used successfully for a wide variety of purposes in educational contexts, including to: grade essays [10] [11], analyze content for conceptual understanding [12] [13], discover topics or themes , score text sophistication, writing quality, and reading grade level [14], detect off-topic behavior, assess learning styles [16], and score argumentative and question-answering quality [17] [18] [19] [20], identify dialogic moves and patterns, identify tutorial behaviors [21] [22]. As far as we know, we are the only ones researching text analysis to assess social deliberative skills such as perspective taking and meta-dialogue in educational contexts or in human dialogues of any sort. There has been related work in non-educational and non-dialogical contexts to identify psycho-linguistic and socio-linguistic phenomena such as emotional states and sentiment [21] [22], personality traits; and even to predict health improvement based on essay writing [23]. Text analysis methods have been used to classify speech acts (including dialogue moves, tutorial acts, argument moves, etc.). For example, Rosé et al. [17] achieved 53% accuracy on classifying "epistemic activity," 61% accuracy for "social modes of co-construction."

Our work uses the output of sophisticated text analysis systems (LIWC and Coh-Metrix) as feature inputs for machine learning algorithms. LIWC (Linguistic Inquiry Word Count, [23]) is a well researched but "shallow" dictionary-matching text categorization system yielding about 80 linguistic categories (e.g. positive emotion words, pronouns, and causation words). Coh-Metrix [22] [24] performs a series of deep-processing analysis (including semantic cohesion, latent semantic analysis, and reading complexity level) yielding about 100 metrics. A simplistic view of these systems is the LIWC categorizes speech acts based on *what* participants are saying, and Coh-Metrix produces measurements related to *how* participants are speaking. LIWC features are derived across topic domains and from people from all walks of life; Coh-Metrix features are generated across text genres from a wide spectrum of disciplines. Though LIWC's dictionary-matching method is simple (like keyword-matching), hundreds of studies have been done using it (and contributed to its development) so many of the categories it uses are well researched in terms of how use of these linguistic categories correlate with important psychological or social phenomena. LIWC and Coh-Metrix measurements are ideal for this study, where the discourse data comes from participants across a variety of topic domains and online contexts. Both LIWC and Coh-Metrix features have been shown to be valid and reliable markers of a variety of psycholinguistic phenomena.

In prior studies [9] we used text analysis in conjunction with multi-class machine learning methods to build models for individual deliberative skills. This proved to be challenging for the methods available to us at the time, and we shifted to the more tractable task of building models for a total or composite deliberative skill measure that was the aggregate of the individual sub-skills (later to return to individual skill modeling). A series of experiments, reported in several papers, refined our ability to automatically assess deliberative skill across multiple domains of online engagement. These experiments were conducted with a data corpus consisting of online interactions from three domains. Participant posts were first

partitioned into segments if the type of speech act changed within a post (usually there were 1-4 segments per post). The domains were: an online civic engagement dialog (32 participants with 396 segments of text), two faculty communities engaged in logistical decision making (16 participants and 438 text segments), and, the largest set, college classroom online discussions of controversial topics (90 participants and 1783 text segments). Training was done based on human-rated assessment of deliberative skill, using a coding scheme that had shown inter-rater Cohen's Kappa statistics of 71% on average across the domains (average percent agreement of 76%), which is quite good for a scheme of its complexity [25].[2] Ten-fold cross validation over the data set was used in all cases.

Early work compared various machine learning methods including Naïve Bayes, Support Vector Machine, Topic Modeling, and Regularize Logistic Regression methods (experimenting with a number of parameters within each). We found L1 Regularized Logistic Regression to be the preferred model (though we continued to include comparison with other models though subsequent experiments to validate this finding). Next we compared the success of various feature sets including bag-of-words, LIWC, Coh-Metrix, and combinations of these. We found that using text analysis (LIWC or Coh-Metrix) outperformed bag-of-words methods, that LIWC features usually outperformed the Coh-Metrix features, and that combining these feature sets lead to worse performance that using them individually. Finally, we did cross-domain studies showing that superior models resulted from using certain domains as the training set [26]. Specifically, the model developed using the faculty community showed better performance on all three domains than either drawing training data from the entire corpus or drawing the training sample from the domain to be tested. It appears that this is because the faculty domain had the most balanced (least skewed) data, i.e. there was a sufficiently large percentage of text segments tagged as deliberative skills vs. others (about half).

We continue our research in the study reported here by: (1) applying methods developed previously to a new set of classroom online dialogue data and (2) adding demographic information, gender and grade level, to the models feature set. In this study we extend out prior research on building machine learning models to predict an aggregate (total) social deliberative skill measure.

Method

Data set. Twenty six students in a college Alternative Dispute Mediation class discussed two topics (the Trayvon Martin killing in Florida and Gun Control, one each week over two weeks) in using the Mediem deep dialogue discussion software. Students were randomly broken into three discussion groups of 8-9 members each, with all groups discussing these topics. There were 8 males and 14 females ranging in undergraduate grade level from sophomores to seniors, with one non-degree student. Each of the three groups used a different set of software features based on our protocol for an experimental study of the effects of tools to support social deliberative skills. In Murray et. al. [6] we discuss our findings that "reflective tools" showed a significant effect size in deliberative skills as measured by human coding, but for this paper we ignore the grouping of students as we are only interested in trying to model the human rating of total deliberative skill using computational methods. The data set consisting of 829 text segments from 369 posts. 43% of the segments were coded under the "deliberate skill" meta-category (vs. 57% "other").

Machine learning method. In this study, we used our highest performing machine learning method, L_1 regularized logistic regression (L_1RLR) [27] to model social deliberative behavior and predict its occurrences. L_1RLR is also preferred in this research because it not only works well with high dimension feature space and small data sets, but also is able to automatically select features and learn an easy-to-interpret (transparent) model. Being able to automatically select features mitigates the problem that little precedent research exists in this new area that is suggestive of features predictive of social deliberative behavior. Being able to yield an interpretable model presents fewer challenges for researchers in social science and communication science to understand the efficacy of a computational model for social deliberative behavior.

Before we describe L_1RLR, let us recall that the logistic loss function is defined as:

$$p(y|x; \mathrm{W}) = \frac{1}{1 + \exp(-\mathbf{W}^T x)}$$

[2] Our coding scheme has 42 categories, 17 of which indicate deliberative skills.

where x is the training data, y is the response variable, and W is the model we learn.

In regularized logistic regression, we solve the following optimization problem:

$$\underset{\mathbf{W}}{\operatorname{argmax}} \sum_{i} \log(p(y_i|x_i;\mathbf{W}) - \lambda * \Omega(\mathbf{W})$$

where Ω(W) is a regularization term used to penalize large weights.

In the case of L_1 regularized logistic regression, L_1 norm [27], or least absolute shrinkage and selection operator (Lasso) is used to induce the penalty. Previous research [28] has shown that L_1 regularization logistic regression requires the number of training examples that grows logarithmically with the number of features to learn well, which favors this study.

In our experiments, we used the l_1 regularized dual averaging algorithm [29] for solving l_1RLR. We trained l_1RLR (i.e., λ=1, γ=2) with various feature sets and carried out 10-fold stratified cross-validation.

Results and Discussion

We performed a set of experiments by exploring the effectiveness of different types of features on predictive accuracy, precision, recall, and F_2 measure (the harmonic mean of precision and recall that weights recall twice as high as precision). In Table 1, we report the average performance across cross-validation runs.

	LIWC features	Coh-Metrix features	LIWC+gender+gradLevel features
Accuracy	61.41	60.68	60.81
Precision	54.30	54.31	53.78
Recall	68.52	57.94	67.41
F_2 measure	65.11	57.18	64.16

Table 1: Predictive performance (in %) of
L1 regularized logistic regression built using different type of features

Predictive performance and feature comparisons. As can be seen in Table 1, with computational models, we are able to predict social deliberative behavior with up to 61% accuracy, 54% precision, 68% recall, and 65% F_2 measure. LIWC features outperformed Coh-Metrix features by a slight margin overall, which confirms earlier findings (we did not model using combined LIWC and Coh-Metrix features as prior work suggested this would not help [26]). Surprisingly, adding the demographic information of gender and grade level as machine learning inputs did not improve performance (it degraded it slightly).[3] This suggests that variations due to grade and gender are already encoded in the text analysis features (of both LIWC and Coh-Metrix)—a hypothesis we will pursue in further research.

The performance of the L1-RLR on this data set outperformed the models reported en earlier studies of classroom data. In general, prior studies of multi-domains showed that prediction in the classroom domain was worse than in the other domains (civic engagement and faculty logistical decision-making). More specifically, the results reported here improved over previous results of classroom domains by 8% on precision and 64% on recall. We believe that this is mostly due to the newer data set having less data skew (43% deliberative skill on this set vs. 32% on the prior classroom data set). We are looking into methods to compensate for data skew, including training our models on the most robust data sets as opposed to the testing data sets [26].

In a larger sense, the results suggest that it may be feasible to train machine learning models to automatically analyze conversations in online communication to identify high-order communication skills such as social deliberative behavior.

Parameters in the learned model. As mentioned, one of the benefits of using L1-RLR is that the relative importance or weights of each feature can be inspected (they are related to the coefficients of the regression

[3] Indeed, when examining the learned feature space, we found that both gender and grade level features were shrunk by the L_1RLR model. In other words, both features were assigned zero weights in the final model.

model). The L_1 regularized logistic regression learned a model with 27 features in this domain. In other words, 55 out of the 82 LIWC features were shrunk by L_1RLR (which automatically prunes features, another advantage vs. other machine learning methods). In Table 2, we show the top 10 most salient features of the learned model. The rest of the 17 features have absolute feature weights less than 0.01 and are commented below.[4]

LIWC feature	Interpretation	Weight
assent	assent	0.335
WC	word counts	0.223
social	social processes	-0.051
Dic	dictionary words	-0.045
i	1st pers singular	0.028
funt	total function words	-0.024
posemo	positive emotion	0.023
AllPct	total punctuations	0.023
affect	affective processes	0.023
period	punctuation	0.022

Table 2: Top 10 LIWC features learnt by L1 regularized logistic regression

Next we summarize the characteristics of social deliberative behavior in the language of LIWC features. LIWC was not designed to measure deliberative skill or any sort of dialogue-quality related speech act categories, and predictive relationships between its categories and deliberative skill are expected to be secondary (i.e. resulting from more clearly relevant intermediate factors). Compared to "other speech acts", social deliberative behavior has: more assent words, longer messages, more 1st person pronouns, more positive emotions, more total punctuations, more affective processes, more certain words, more pronouns (i.e., personal pronouns and impersonal pronouns), more cognitive process, more auxiliary verbs, fewer social processes, fewer dictionary words, fewer functional words, fewer relative words, fewer words per sentence, fewer prepositions, fewer big words, fewer dashes, fewer words about time, fewer commas, fewer space words, fewer present tense, and fewer articles.

Assent-words (31 word stems including absolutely, agree, alright*, haha*, ok, yes, yup…) and the segment word count (WC) were by far the largest factors in this model. Pennebaker & King [30] say the following about assent and word count. Higher word count is related to better group performance. Lots of assents and questions stimulate better team performance. "Later in a group task, assents may signal consensus, early assents may indicate blind agreement by unmotivated group members" [31, p 33]; and "in a cooperative coordination context, higher total word count may signal better communication and agreement, whereas in a negotiation context it may signal a breakdown in agreement." (p. 35). Our related analysis of the faculty dialog also showed that word count was highly related to human assessment of deliberative skill, but, curiously assent was not so related [26]. Further work is ongoing to determining the domain-dependent aspects of deliberative behaviors.

Discussion and Conclusions

We have seen encouraging results in our attempts to model an aggregate classification for total social deliberative skill in a number of online deliberation domains, including in college classroom discussions. We believe that we will be able to improve the accuracy and recall values of the model substantially with additional research. We will continue to do research on modeling individual deliberative sub-skills and dialogue quality indicators, though it is not clear yet whether we will be successful with many of these (some, such as "appreciation", are not as difficult).

In future studies, we will perform similar tests on more domains and in various online contexts (e.g., collaborative problem-solving, negotiation, and disputation) to study the role that demographic features (e.g., gender, age, race) play in predicting social deliberative behavior.

[4] Note, the absolute value of the weights is meaningless and dependent on tuning parameters of the algorithm, and in general are not comparable from one model to the next. Only the relative sizes of the weights within a model are meaningful.

One of the most exciting applications of this work has been in the design and evaluation of a Facilitators Dashboard that shows a birds-eye view of certain dialogue parameters [6]. See Figure 1. We have begun to visualize some of the text analysis in this tool. Early comments from instructors and professional facilitators and mediators indicate that such analysis will be very useful for their work.

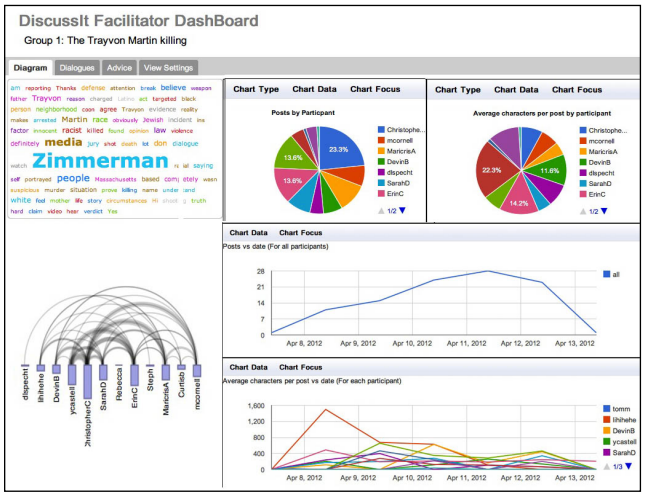

Figure 1. Facilitators Dashboard

References

[1] King, P. M. & Baxter Magolda, M. A developmental model of intercultural maturity. Journal of College Student Development, 46 (6), 571-592, 2005

[2] Bikson, T. K., & Law, S. A. Global preparedness and human resources. Rand Corporation, 1994.

[3] Jordan, T., Andersson, P. & Ringnér, H. The Spectrum of Responses to Complex Societal Issues: Reflections on Seven Years of Empirical Inquiry. Integral Review, February 2013, Vol. 9, No. 1, 2013

[4] Conklin, J. Wicked Problems & Social Complexity. Chapter 1 of Dialogue Mapping: Building Shared Understanding of Wicked Problems, Wiley, 2005

[5] Kegan, R. In over our heads: The mental demands of modern life. Cambridge, MA: Harvard University Press, 1994

[6] Murray, T., Wing, L., Woolf, B., Wise, A., Wu, S., Clark, L., Osterweil, L., Xu, X. A Prototype Facilitators Dashboard: Assessing and visualizing dialogue quality in online deliberations for education and work. International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government, EEE-2013.

[7] Murray, T., Woolf, B., Xu, X., Shipe, S., Howard, S. & Wing, L. "Towards Supporting Social Deliberative Skills in Online Group Dialogues." Presented at The 7th Annual Interdisciplinary Network for Group Research Conference (InGroup). Chicago July 12-14, 2012.

[8] Murray, T., Stephens, A.L., Woolf, B.P., Wing, L., Xu, X., & Shrikant, N. Supporting Social Deliberative Skills Online: the Effects of Reflective Scaffolding Tools. Proceedings of HCI International 2013, July, 2013, Las Vegas, 2013

[9] Murray, T., Xu, X. & Woolf, P.B. An Exploration of Text Analysis Methods to Identify Social Deliberative Skills, In Proceedings of the 16th International Conference on Artificial Intelligence in Education (AIED), 2013

[10] Shermis, M. & J. Burstein (Eds.) Automated essay scoring: A cross-disciplinary perspective. Hillsdale, NJ: Erlbaum, 2003

[11] Dikli S. An overview of automated scoring of essays. The Journal of Technology, Learning and Assessment, 5(1), 2006

[12] Lintean, M., Rus, V., & Azevedo, R. Automatic Detection of Student Mental Models Based on Natural Language Student Input During Metacognitive Skill Training. International Journal of Artificial Intelligence in Education, 21(3), 169-190, 2011

[13] Azevedo, R., Guthrie, J.T., & Seibert, DThe role of self-regulated learning in fostering students' conceptual understanding of complex systems with hypermedia. Journal of Educational Computing Research, 30 (1), 87-111, 2004

[14] McNamara, D. S., Crossley, S. A., & McCarthy, P. M. Linguistic features of writing quality. Written Communication, 27(1), 57-86, 2010

[16] Ozpolat, E., & Akar, G.B. Automatic detection of learning styles for an e-learning system. Computers & Education, 53(2), 355-367, 2009

[17] Rosé, C., Wang, Y. C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., & Fischer, F. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. International journal of computer-supported collaborative learning, 3(3), 237-271.

[18] Kim, J., and Shaw, E. Pedagogical Discourse: Connecting students to past discussions and peer mentors within an online

discussion board, Innovative Applications of Artificial Intelligence, IAAAI 09, 2009
[19] Ravi, S., Kim, J. Profiling student interactions in threaded discussions with speech act classifiers. IOS Press, 2007
[20] Rus, V., Cai, Z., & Graesser, A. C. Evaluation in natural language generation: The question generation task. In Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation (pp. 20-21), 2007
[21] D'Mello, Sidney, Olney, Andrew, & Person, Natalie. Mining collaborative patterns in tutorial dialogues. Journal of Educational Data Mining, 2(1), 1-37, 2010
[22] Graesser, A. C., Jeon, M., Yang, Y., & Cai, Z. Discourse cohesion in text and tutorial dialogue. Information Design Journal, 15, 199–213, 2007
[23] Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A. L., & Booth, R. J. The development and psychometric properties of LIWC2007. Austin, TX: www.LIWC.net, 2007
[24] Graesser, A., & McNamara, D. Computational analyses of multilevel discourse comprehension. Topics in Cognitive Science 3(2), 371–398. 2010.
[25] Altman, DG, Practical statistics for medical research. London: Chapman and Hall, 1991
[26] Xu, X., Murray, T. Woolf, B. & Smith, D. Mining Social Deliberation in Online Communication -- If You Were Me and I Were You , In Proceedings of the 6th International Conference on Educational Data Mining (EDM), 2013
[27] Tibshirani, R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) pp. 267–288, 1996
[28] Ng, A. Feature selection, l1 vs. l2 regularization, and rotational invariance. In Proceedings of the twenty-first international conference on Machine learning, page 78. ACM, 2004
[29] Xiao, L. Dual averaging methods for regularized stochastic learning and online optimization. The Journal of Machine Learning Research, 11:2543–2596, 2010
[30] Pennebaker, J. W., & King, L. A. Linguistic styles: Language use as an individual difference. Journal of personality and social psychology, 77(6), 1296-1312, 1999.
[31] Tausczik, Y. R. & Pennebaker, J. W. The psychological meaning of words: LIWC and computerized text analysis methods. Journal of Language and Social Psychology, 2010.