An Exploration of Text Analysis Methods to Identify Social Deliberative Skills

Tom Murray, Xiaoxi Xu, Beverly Park Woolf

Computer Science Dept., Commination Dept.
University of Massachusetts, Amherst, MA

Contact: tmurray@cs.umass.edu

Abstract: We report on a series of evaluations of text processing and machine learning methods with the goal of building classifiers for social deliberative skills, i.e. the capacity to deal productively with heterogeneous goals, values, or perspectives. Our corpus includes online deliberative dialogue from four diverse domain contexts. We use the LIWC and CohMetrix linquistic analysis tools to (1) assess differentiating characterisitcs of these domains, and (2) generte feature sets for machine learning. We report on comparative analysis of various machine learning alrorythms, feature selection methods, and training methods in our attempts to build classifiers for (1) individual skills and (2) a composite total skill measure.

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

1 Introduction

The capacity to flexibly and productively negotiate differences of opinion, belief, values, goals, or world-views, is of critical importance in today's world, and is markedly lacking. In the increasingly global world the economic productivity and security of nations can be linked to citizens' and leaders' capacity to understand and deal productively with diverse perspectives. We use the term "social deliberative skills" (SD-skills) to indicate the capacity to deal productively with heterogeneous goals, values, or perspectives, especially those that differ from ones own.

People are increasingly engaged in dialogue, deliberation, and collaboration online.

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

adfa, p. 1, 2011.

able to measure, identify or assess them. This paper describes our initial 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 (Murray et al. 2013 in submission). 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 use the human-coded ratings of SD-skills as the reference standard and training input for machine learning. Automated assessment would not only facilitate *data analysis* by allowing us to assess more data faster, but, if done in real time, can be used to support deliberative dialogue through tools that visualize dialogue skill and quality metrics for *instructors, facilitators*

In our work we are using the Cohmetrix multiple-level text analysis system (Graesser et al, 2007; 2011), the LIWC "Linguistics Inquiry Word Count" application (Pennebaker et al., 2007), and a variety of machine learning methods. Text analysis in our domain is challenging for several reasons. (1) the behaviors we are investigating (speech acts in categories associated with social deliberative skill and high quality dialogue) have relatively low frequency, both in our data and in communication in general. The resulting skew in the data is a challenge for machine learning methods. (2) We have many categories to differentiate compared with much prior research. (3) The feature space we want to make use of (including LIWC and CohMetrix measures) is quite large. These challenges, as well as the fact that little work in text classification has been done in the area of human-to-human dialogue (and none in deliberative dialogue) indicate that text classification in the domain of supporting SD-skills in deliberative dialogue explores important new ground.

We have experimented with approaches designed to ameliorate each one of these challenges. One first step in simplifying the challenges mentioned above is to create a binary composite code for Total-SD-skill, which is true if a code is one of the SD-skill categories (or, depending on its definition, a high quality dialogue act category). As reported later, we have had encouraging results in predicting Total-SD-skill, and are still improving our multi-class methods to spot individual SD-skills.

2 Background

Text Classification. In the work reported here we focus on text classification methods that analyze individual participant posts or shorter segments of text; and our domains involve online dialogue, usually within discussion forum software.

Text analysis has been used successfully for a wide variety of purposes, including to: grade essays (Shermis & Burstein 2003), analyze content for conceptual understanding (Lintean et al., 2011), score text sophistication, writing quality, and reading grade level (McNamara et al., 2010), and score deliberative, argumentative, and question-answering quality (Rose et al. 2008; Ravi & Kim 2007).

Past research exploring linguistic and discourse features in dialogues has proven moderately successful in predicting complex phenomena such as personality type, status, deception behavior, metacognition, speech acts, intention, and affect states . Therefore, it is plausible to expect that a linguistic and discourse analysis of deliberation dialogues would provide valuable insights into predictors that are diagnostic of deliberation dynamics and skills

Much text classification research involves the comparative exploration a wide variety of machine learning algorithms, algorithm tuning parameters, feature selection methods, and data category weighing methods.

Social deliberative skills. We frame SD-skills in terms of these capacities (see Murray et al., 2013 submitted): perspective taking (includes cognitive empathy, reciprocal role taking); perspective seeking (includes social inquiry, question asking skills); perspective monitoring (includes self-reflection, meta-dialogue); and perspective weighing (related to "reflective reasoning" and includes comparing and contrasting the available views, including those of participants and external sources and experts) . SD-skills overlaps with but is distinct from other cognitive constructs that have been studied in depth, including collaboration skills, metacognition, reflective reasoning, social intelligence, argumentation skills, and critical thinking (Lin & Sullivan, 2008; King & Kitchener, 1994; Kuhn, 2000; Graesser et al., 2008). We differentiate our research from others that focus on *argumentation*, which aims to help learners generate logical, well-formed, well-supported explanations and justifications (Andriessen et al., 2003;), usually framed in objective rather than intersubjective terms. That is, they are about finding the right answer or the most efficient and effective solution to a technical or scientific question—but don't address, as we do, the specific moments of deliberation or collaboration where opportunities for mutual understanding and mutual recognition arise.

LIWC and CohMetrix linguistic analysis tools. Our primary goal is to build domain-independent classifier models that will predict SD-skill components or Total-SD-skill. Perhaps the most prominent machine learning method used in natural language processing, information retrieval, and document/text classification is the 'bag of words" unigram method, in which the feature set for the learning algorithm consist of an unordered set of all the words in a document (preprocessed with stemming etc. as necessary). However, we have much more information available with which to build our predictive models. In particular, CohMetrix and LIWC are text analysis systems that output a number of linguistic metrics. We hypothesize that using these as feature inputs to machine learning models would increase their accuracy and efficiency vs bag-of-words methods. Thus we can do a two-step analysis, in which we extract the CohMetrix and/or LIWC features, and then use these features as inputs to machine learning methods.

We are also interested in CohMetrix and LIWC measurements in their own right.

LIWC (Linguistic Inquiry Word Count; Pennebaker et al., 2007) 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). CohMetrix (Graesser et al, 2007; 2011) performs a series of deep-processing analysis (including semantic cohesion, latent semantic analysis, and reading complexity level) yielding about 100 categories. LIWC features are derived across topic domains and from people from all walks of life; CohMetrix features are generated across text genres from a wide spectrum of disciplines. LIWC and CohMetrix 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 CohMetrix features have been shown to be valid and reliable markers of a variety of psycholinguistic phenomena.

3 Methods: Coding and Corpora

Coding scheme. We are using both manual coding and automated text analysis. We have developed and refined a 30-category hierarchical coding scheme for human raters to code segments of the text. This scheme was derived from several prominent schemes found in the literature (Black et al., 2011;Stromer-Galley, 2007; Stolcke et al. 2000) and adds codes specific to social deliberative skills (Murray et al. 2012). With three trained coders, the scheme is showing inter-rater Cohen's Kappa statistics of 71% on average in these domains (average percent agreement of 76%), which is quite good for a scheme of this complexity (Altman, 1991).

Our coding scheme includes 10 codes for social deliberation behaviors (including perspective taking, asking clarifying questions, mediation actions, and meaning generation and repair actions), 7 codes for additional deliberation quality indicators (including weighing alternatives, citing sources, changing ones mind, and apologizing), and other categories for action negotiation and argumentation. For some of our analysis we constructed a Total-SDSkill metric, which aggregated the SD core set and the additional deliberation quality indicators. For analysis focusing on the segment level Total-SDSkill is a Boolean indicating whether or not the code falls into one of these two sets, and for analysis at the post or participant level Total-SDSkill is a sum of how many of these were contained in that unit (e.g. the value would be 3 if a post had 5 segments, 3 of which were in the Core or Additional code sets).

Domains. We examined four corpora of online conversations that our team has procured:

1) **Civic Deliberation**: postings from a neighborhood civic engagement online discussion forum at e-democracy.org. The participants were discussing racial issues and tensions about their multi-racial community.
2) **Faculty listserve**: email exchanges from a faculty listserv with geographically dispersed participants. Two research communities were engaged in a discussion about how to organize a conference addressing overlapping interests. The discussion became contentious.
3) **Classroom discussions**: postings from 7 online discussions on controversial issues from three college classrooms, assigned as homework. The topics include "should the legal drinking age be lowered in Massachusetts?" and "pros and cons of using FaceBook or other social networking software as part of high school curriculum".
4) **Workplace dispute mediation**: exchanges from an online dispute mediation session about a workplace dispute, involving one employee, one supervisor, and

one mediator. This data was not used for some of our analysis because it has unique characteristics compared to the others.

Two independent trained human judges annotated the 4 corpora based on the coding scheme described above. The Cohens' Kappa inter-rater reliability statistics for the four domains are: 73.49%, 68.36%, 68.7%, 46%, respectively. The low inter-rater score for the Workplace domain is another reason we excluded it from some of the analysis reported.

Domain	Posts	Seg-ment	Partic-ipants	SD-Skill seg	% SD-Skills	Words/ Post	Posts / Partic	Seg. / post
Civic deliberation	51	396	31	225	57%	352	1.6	7.8
Faculty negotiation	72	438	16	231	53%	195	4.5	6.1
College discussions	768	1783	90	565	32%	88	8.5	2.3
Workplace	56	637	3	251	39%	200	18.7	11.4
All domains	947	3254	140	1272	39%	117	6.8	3.4

Table 1: Descriptive Statistics for Four Domains

4 Results

We report on several streams of analysis: (1) a comparison of characteristics of the domains, to (a) help interpret later results that are domain-specific, and (b) take initial steps toward a large goal of designing a domain-characteristics scheme for use in deliberative skill research; (2) an analysis of code frequencies in each domain to investigate the relationship between domain characteristics and deliberative skillfulness; (3) investigations of the correlations between automated linguistic measurements and hand-coded skill scoring; (4) efforts to build a classifier that will predict overall SD-skill, (5) efforts to build classifiers to predict individual SD-skills; (5) efforts to improve total-SD-skill classification by finding domains that serve as the best training sets.

Code frequencies. Our analysis of the frequencies of codes in each of the four domains (and total for all domains) shows several things. The Arg_Gen, or generic argumentation codes predominate, as might be expected. The more reflective speech acts we are interested occur much less frequently than acts such as making, explaining, or defending a fact or opinion (the bar chart fo ARG_GEN is not shown, to make it easier to see the relative frequencies of other codes). Intersubjective speech acts were the second higher frequency overall, and were also second in the classroom (14%) and workplace (20%) domains. This is a good indication that the conversations were at least moderately reflective.

For the Faculty domain Meta_dialogue was second (19%) and INTERSUB was third (16%). It makes sense that the highly educated Faculty group would have a high level of meta-dialogue. The Civic domain was especially high on SELF_REFLection, which tends to include a move away from certainty into more nuance and contingency of opinions. In this domain participants were more part of an established community with longer term relationships. As mentioned, data skew or imbalance is one of the challenges of working in this domain, and Table 2 shows a comparison of domains that speaks to this issue.

	Civic	Class	Faculty	Work	Total
Total_SDskill %	55%	29%	53%	39%	38%
ARG_GEN%	27%	48%	24%	22%	37%
OTHERS	18%	23%	24%	39%	26%

Table 2: Analyses of Data Frequency Skew by Domain

SD-skill correlations with LIWC/CohMetrix measures. Above we mentioned that we expected that using LIWC and CohMetrix measures as machine learning features would allow for more accuracy and/or fast models than the more traditional bag-of words features. We also mentioned that we were interested in these linguistic measurements in their own right.

As part of the L1 machine learning analysis described later, we produced values indicative of the correlations between the *Total*-SD-skill composite metric and the LIWC and CohMetrix measures. The results are summarized below in terms of trends:

Total-SD-Skill Positive Correlations	Total-SD-Skill Negative Correlations
More negative addictive connectives	Less narrativity
More negations	Less lexical diversity
More pronouns	Lower reading ease
More second person pronouns	Less connectives
More meaningful words	Less concrete words
More punctuation	Lower number of words
More verbs	Less spatial location & motion words (i.e., upon, into)
Older age of acquisition for words	Less causal verb and particles
	Less nouns

Table 3. Correlations between Total-SD-skill and LIWC/CohMetrix measures

We would expect that SD-Skill, by the definition of the component scores, would correlate with pronoun use (i.e. intersubjectivity). Several of the correlations align with intuitions that those with higher SD-skills would in general be more sophisticated or mature, including: word acquisition age (i.e. grade level), more meaningful words (i.e. more concrete and less vague words), lower reading ease (i.e. more complex sentence structure), and more punctuation. Some correlations are counter to our current intuitions, including: less lexical diversity, less connectives, and lower number of words. **Predicting Total-SD-skill**. One of our goals is to build (multi-class) models that will classify SD-skills (or a subset of them) with performance comparable to human raters. As mentioned above this is a very challenging problem. A more tractable first step is to build a *binary* model for recognizing deliberative skillfulness. Such a model would be useful for assessing the general deliberative skill of participants, groups, phases of a dialogue. This information could be used both to evaluate experimental interventions and as information given to facilitators in real time.

We compared a number of machine learning methods and feature sets in an attempt to build a binary classifier for generally skillful dialogue (Total-SD-skill). (Note that in this document we used 10-fold cross validation where applicable on all machine-learning methods, unless otherwise stated.) We started by trying the standard SVM (support vector machine) methods. We chose SVM because it has outperformed other methods in related research, but later we will compare it with other methods.) We tried various features combinations with bag-of-words unigram, LIWC measurements, and CohMetrix measurements. Including many spurious features in a model can result in overfitting so we also tried to narrow down the LIWC/CohMetrix feature set by using only those features that were highly correlated with the total-skill measure (we tried this using a $p < .01$ threshold, and again using a $p < .05$ threshold). Interestingly, none of the models using LIWC and CohMetrix measurements did as well as the unigram bag-of-words features. The best binary classification model was SVM with unigram features (using TF-IDF) yielding these performance metrics: Accuracy 74%, Precision 73%, Recall 81%, and F-measure 77%. These results were quite encouraging given the exploratory nature of our work. However, during this early stage of the work their were more codes included in the Total-Skill composite than we eventually settled on, and we consider these results incomparable to succeeding results, and also not quite representative of total deliberative skillfulness as we now define it.

Predicting individual SD-skills. Next we tried to move from binary classification to multi-class (and hierarchical) classification methods, to build classifiers for individual skills. We will say up front that these attempts were not particularly successful, but we feel that it is useful to report because we tried many methods and the negative results might inform future research by others. (We continue to experiment and develop new methods for this task.) Preliminary work showed that building classifiers to differentiate the entire set of codes in our scheme was implausible, especially given the modest size of our data set, so we did two things to reduce the number of categories. First, we employed a hierarchical modeling strategy to recognize all codes marked as indicative of high quality deliberation (a binary classification). We used only this reduced data set to train the model at the next step. Second, we reduced the code set further to six classes: intersubjective speech acts, meta-dialogue, self reflection, meta-topic, topic questions, and Others, which, because of the hierarchical method means other *quality deliberation* speech acts, rather than other speech acts.

The skewed nature of the data presented a significant challenge, which we approached from three directions: algorithm types, hyper-parameter level, and sampling methods. We tried a variety of machine learning algorithms at this phase, including SVM, Naïve Bayes, and boosting (a decision tree method using the Random Forest model to iteratively adjust feature weights). We used the Linguisitc feature sets, not bag-of words, for the multi-class analysis. At the hyper-parameter level we tried adjusting class salience. Only the SVM algorithm (polynomial kernel with degree 4) achieved good training accuracy, recall, and precision of over 70%, but the testing accuracy was only at about about 30% (i.e. it had a large variance). For the rest of the classifiers, even the training accuracy for each class was only about 30%.

Cross-domain Training. Next we turn to the question of whether some deliberative domains make better training sets for a domain-independent model. In this section we return to binary classification of Total-SD-skill (leaving further experiments with multi-classification of individual skills for future work). Other researchers have found that certain domains have better characteristics for use as training sets. We hypothesize that domains that have either the most representative distribution of codes, or those that have the least skew (imbalanced frequency distributions) might serve as better training sets.

In this phase of our research we also addressed another concern. Earlier results indicated that using LIWC and CohMetrix measures as features did not do quite as well as the unigram bag-of-words method. But we became increasingly concerned about runtime performance in addition to accuracy performance. Performance time degrades non-linearly with the number of features, and the bag-of-words unigram method uses a very large feature set and is not very time efficient (bigram and N-gram methods are even more time-intensive). Therefore we return to looking at LIWC and CohMetrix features as potential feature sets.

As we prepared for this series of experiments we discovered the L1 regularized logistic regression class of machine learning algorithms which showed promise for addressing several of the challenges we have faced (Tibshirani, 1996).L1 RLR tends to have superior generalizability, interpretability, and scalability.

For these tests the Naïve Bayes and L1 LRL models used the LIWC+CohMetrix feature set, while the SVM model used the bag-of-words unigram feature set. (Using the best tuning parameters as described above; we did not try SVM with LIWC+CohMetrix because prior trials showed that bag-of-words did better. L1 is not designed to do bag-of-words type analysis so we did not try L1 with bag of words). In addition to comparing these three learning algorithms, we wanted to test *cross-domain training* in which data from one domain is used to train the model, and testing occurs on other domains or on all domains. We used the three domain corpuses shown in Figure 1, Civic deliberation, Faculty dialogue, and Classroom discussion (we did not use the Workplace dispute domain because of its outlier nature as described earlier). Using a factorial experimental design, we tried all combinations of the three learning algorithms with all combinations of each of the three domains serving as a training domain, with all domains individually plus all domains together serving as the test domain. Thus we constructed a 3x3x4=36 cell matrix, and determined the Accuracy, Precision, Recall, and F2 measure in each cell. The results are reported in detail in Xu et al. (2013 submitted), and are summarized below. The primary result we are interested in is the average performance over all three domains used as test domains.

1) Overall using the **Civic domain** as the training set did much better than using the Faculty domain, the Classroom domain, or all of the data as the training set. This was true for all three learning algorithms and all four performance measures. Our hypothesis that the domain with the least skew would serve as the best cross-domain training set was confirmed.

2) Overall the **L1 RLR algorithm** significantly outperformed Naïve Bayes and SVM (this was true when the Civic or Faculty domains were used to train). This confirms our expectation that L1 RLR has performance characteristics addressing for the modeling challenges we face.

3) From #1 and #2 above we see that the best model for domain-independent prediction, i.e. prediction that worked best averaged over all three domains, was L1 RLR using the Civic domain for training: accuracy 51%, precision 49%, recall 82%, and F2 71%.

4) *Cross-training* proved to have advantages. For precision, recall, and F2-measure (but not accuracy) using the Civic domain as a training set outperformed using the *same* domain to train as was tested on. I.E. for performance on the Faculty domain by itself, training with Civic was better than training with Faculty. Similarly with the Classroom domain.

These overall results for binary classification of Total-SD-Skill, accuracy 51%, precision 49%, recall 82%, and F2 71%, are encouraging for our exploratory study, but not particularly impressive for a binary classifier (accuracy and precision are at chance levels).

5 Conclusions

A key human capacity is the ability to negotiate situations involving differing opinions where a resolution of ideas is sought, e.g., in dispute resolution, collaborative problem solving, bargaining, and civic deliberation processes. The need for this deliberative capacity, which we call social deliberative skill, is seen in all realms of human activity from international politics, to collaborative work, to mundane familial squabbles. As communication, collaboration, and deliberation occur increasingly on the internet we believe that there is great potential to design software that supports skillful deliberation through gentle prompts and scaffolds, especially for groups of interlocutors who, acknowledging that deliberation in complex and stressful situations can be challenging, are interested in putting some attention and effort on the quality of their communication. One component of this vision is being able to assess deliberative skill. Such models can help evaluate deliberation quality for engagement evaluation, experimental trials of software or other supportive interventions. In addition, we have prototyped a Facilitators Dashboard that gives facilitators or teachers (and eventually participants and peers) a birds eye view of a conversation. We have begun to feed the results of text analysis into the Dashboard, and will continue to work toward real-time versions of our models that enable such visualization tools.

We have described on our attempts to build robust classifiers for individual skills and total composite deliberative skill, reporting what we have learned from a number of comparative experiments with different modeling methods. Our results are encouraging but exploratory and incomplete, and we continue to try new methods. Finally, we have also begun experimenting with machine learning and text classification methods to recognize factors in deliberative success, participant roles, and important transition points or intervention opportunities in a dialogue.

References

1. Altman DG (1991) Practical statistics for medical research. London: Chapman and Hall.
2. Andriessen, J., Baker, M., & Suthers, D. D. (Eds.). (2003). Arguing to Learn: Confronting Cognitions in Computer-Supported Collaborative Learning Environments. Dordrecht: Kluwer.

3. Black, L., Welser, H., Cosley, D., and DeGroot, J., Self (2011). Governance Through Group Discussion in Wikipedia Measuring Deliberation in Online Groups. Small Group Research 42(5) pp. 595-634.
11. Graesser, A. C., Jeon, M., Yang, Y., & Cai, Z. (2007). Discourse cohesion in text and tutorial dialogue. Information Design Journal, 15, 199–213.
12. Graesser, A.C., Dowell, N., & Moldovan, C. (2011). A computer's understanding of literature. Scientific Study of Literature 1:2 (2011).
16. Ravi, S., Kim, J. (2007). *Profiling student interactions in threaded discussions with speech act classifiers*. IOS Press.
17. King, P.M. and Kitchener, K.S. (1994). Developing reflective judgment: Understanding and promoting intellectual growth and critical thinking in adolescents and adults. San Francisco: Jossey-Bass.
20. Kuhn, D. (2000). Metacognitive Development. Current Directions in Psychological Science, 9(5), pp. 178-181.
21. Lin, X. & Sullivan, F. (2008). Computer contexts for supporting metacognitive learning. In J. Voogt, G. Knezek (eds.) International Handbook of Information Technology in Primary and Secondary Education, 281–298. Springer Science+Business Media, LLC.
22. Lintean, M., Rus, V., & Azevedo, R. (2011). 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.
25. McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. Written Communication, 27(1), 57-86.
28. Murray, T., Woolf, B., Xu, X., Shipe, S., Howard, S., Wing, L. (2012) Supporting social deliberative skills in online classroom dialogues: preliminary results using automated text analysis. In: Intelligent Tutoring Systems. pp. 666–668. Springer.
29. Murray, T., Woolf, B., Wing, L., Shipe, S., Shrikant, N. (in preparation). A manual coding scheme for social deliberative skills and other attributes related to dialogue quality.
30. Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A. L., & Booth, R. J. (2007). The development and psychometric properties of LIWC2007. Austin, TX: www.LIWC.net.
32. Rosé, C., Wang, Y.-C., Cui, Y., Arguello, J., Stegmann, K., Weinberger, A., et al. (2008). Analyzing collaborative learning processes automatically: Exploiting the advances of computational linguistics in computer-supported collaborative learning. International Journal of Computer-Supported Collaborative Learning. Volume 3, Issue 3, September 2008.
35. Shermis, M. & J. Burstein (Eds.) (2003). *Automated essay scoring: A cross-disciplinary perspective*. Hillsdale, NJ: Erlbaum.
36. Stephens. L., Murray, T., Shrikant, N., Wing, L., Xu, X. & Woolf, B.P (2013 in submission). An Evaluation of Online Forum Features Supporting Social Deliberative Skills: Preliminary Results.
37. Stolcke, A., Ries, K., Coccaro, N., Shriberg, J., Bates, R., Jurafsku, D., et al. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics, 26(3), 39– 373.
38. Stromer-Galley, J. (2007). Measuring Deliberation's Content: A Coding Scheme. Journal of Public Deliberation, 3(1).
39. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological) pp. 267–288 (1996)
40. Xu, X., Murray, T., & Woolf, B. (in submission 2013). Adaptive Models for Social Deliberative Behaviors in Online Communication. Submitted to 2013 International Conference on Education Data Mining.