Mining Social Deliberation in Online Communication

If You Were Me and I Were You

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

ABSTRACT

Social deliberative skills are collaborative life-skills. These skills are crucial for communicating in any collaborative processes where participants have heterogeneous opinions and perspectives driven by different assumptions, beliefs, and goals. In this paper, we describe models using lexical, discourse, and gender demographic features to identify whether or not participants demonstrate social deliberative skills from various online dialogues. We evaluate our models using three different corpora with participants of different educational and motivational levels. We propose a protocol about how to use these features to build models that achieve the best in-domain performance and identify the most useful features for building robust models in cross-domain applications. We also reveal lexical and discourse characteristics of social deliberative skills.

Keywords

Social deliberative skills, collaborative problem solving, collaborative learning, collaborative knowledge-building, discourse analysis, applied machine learning, feature engineering

1. INTRODUCTION

Learning is often depicted as a social process that includes collaborative knowledge-building and problem solving. In this "situated" perspective on learning, learners often must negotiate differing perspectives or goals to build knowledge or solve problems collaboratively. Previous research [25] has shown that certain communication skills, such as listening with empathy and perceiving and responding to other's emotions, part of the collective intelligence of groups, can improve group performance on a wide variety of tasks, such as brainstorming, making collective moral judgments, and negotiating over limited resources. In this research, we focus explicitly on identifying similar skills that are called for to handle diverse opinions and perspectives. For example, do participants attentively listen to each other's opinions? Do they make a good faith effort to understand perspectives other than their own? These skills, including cognitive empathy, affective empathy, and reciprocal role-taking, are part of what we call *social deliberative skills* [14].

Social deliberative skills are at the overlap of cognitive skills and social/emotional skills. Specifically, a participant should present rational arguments with supporting evidence in order that his view be taken seriously and valued. Similarly, one has to turn down the volume of his own thoughts to attentively listen to other's opinions and has to intentionally switch the channel from "me" to "you" to be able to understand or even appreciate another's perspectives. Indeed, this "cognitive empathy" of "*if you were me and I were you*", the soul of social deliberative skills, is needed in any sphere of human interaction, from collaborative learning, to marriage, to workplace relationship, and to world affairs. The ultimate goal of our research is to support social deliberative skills in online communication. In this research, we explore the possibility of automatically assessing and predicting the occurrence of social deliberative skills.

Creating computational models for assessing social deliberative skills has profound implication on several fronts: it (1) supports more efficient analysis for research purposes into online communication and collaboration in social processes; (2) provides assessment measures for evaluating the quality and properties of group dialogues; and (3) provides tools for informing facilitators to adjust skill support and intervention efforts [13]. Previous research in learning science has extensively focused on creating educational software that supports cognitive skills in collaborative environments, such as inquiry skills, metacognition and self-regulated learning skills, and reflective reasoning skills [24, 3, 5, 19, 11]. Research in these areas has provided a deep theoretical context for studying the cognitive aspect of social deliberative skills. A burgeoning body of research has begun to study the social relational aspect of collaborative processes, such as influence [20] and up-taking [21]. This line of research has mainly used structural features of social interactions, such as reply structure, linking notes in a conceptual framework, as well as spatial and temporal proximity to address the questions of who are the central actors in discussions and whose ideas receive the most development. But, collaboration interactions generally take place in the form of natural language. It is reasonable to suppose that language-level features, including lexical features (i.e., what is said) and discourse features (i.e., how it is said) could provide crucial insights into the characteristics of social deliberative skills that are called for in collaborative problem solving and communication in general.

In this paper, we create computational models for assessing social deliberative skills in online communication. The online dialogues that we study are from participants ranging from undergraduate students of multiple disciplines, to highly-educated academic professionals, to members of the general public. We first analyze these online dialogues through a variety of lexical, discourse, and gender demographic features and then create machine learning classifiers to recognize social deliberative skills. To the best of our knowledge, this research is the first work that implements the state-of-the-art conceptual framework of social deliberation. This paper makes four contributions: (1) development of an automatic system for predicting social deliberation, (2) discovery of which type of features or feature combinations are the best for building social deliberative classifiers and under which conditions, (3) discovery of which type of features are the best for building a *robust* social deliberative classifier across domain changes, and (4) identification of language characteristics of social deliberation. These contributions lie at the intersection of machine learning, education, computational social science, and communication studies.

The rest of the paper is organized as follows. In Section 2, we introduce the concept of social deliberative skills. In Section 3, we describe three experimental domains. Section 4 introduces experimental design and methodology. We discuss experimental results in Section 5 and conclude with future work in Section 6.

2. SOCIAL DELIBERATIVE SKILLS

Social deliberative skills involve the application of cognitively-oriented higher-order skills to thinking about the perspectives of others and, consequently, of the self as well. In other words, social deliberative skills require that a speaker reflect not only upon a purely *objective* idea (e.g., a topic) but also upon *my* ideas, *your* ideas, *our* ideas, and *their* ideas. Tracing the origins of this phrase also describes its meaning: to live with others (social) and to balance (deliberative) differences (skills). Our prior research [14] has defined a theoretical framework for social deliberative skills, which includes a group of high-order communication skills that are essential for different tasks and stages of communication that involves a disequilibrium of diverse perspectives. These component skills include social perspective seeking (i.e., social inquiry), social perspective monitoring (i.e., self-reflection, weighting opinions, meta-dialogue, meta-topic, and referencing sources for supporting claims), as well as cognitive empathy and reciprocal role-taking (i.e., appreciation, apology, inter-subjectivity, and perspective taking). Here is an example of "perspective taking" from authentic dialogue in our corpora: *I can't help but imagine what that is like, for her and for her family.* As an another example, the following statement is about "self reflection": *I am probably extremely bias because I am under 21 years old and in college. I wonder if as a 45 year old I will feel differently.*

Social deliberative skills can also be seen as a composite skill [15], which, though less precise can serve as a general marker of social deliberation, for use in evaluation and real-time feedback in intervention. In this study, we focus on creating computational models to assess whether participants of online dialogues demonstrate the use of *composite social deliberative skill* (or social deliberative behavior, SBD).

3. CORPORA

Problem solving and negotiating with others at some level are a regular part of our lives. These actions represent typical everyday communication situations where social deliberative behavior is needed. In this study, we examined three online corpora, two of which involve participants in discussions addressing separate ill-defined problems and one of which involves participants in a negotiation.

In the first domain, **civic deliberation**, posts were collected from a civic engagement online discussion forum at *e-democracy.org*. Thirty two participants discussed ethnic issues and suggesting ways to alleviate tensions about their multi-racial community. These participants were self-selected with an implicit goal to improve community relations. Participants were mostly level-headed and demonstrate social deliberative behavior (SDB) repeatedly. In this domain, SDB occupied 57% of the total 396 annotated segments [1], see Table 1.

In the second domain, **college dialogues**, posts were collected from college students participating in computer-mediated discussions. Ninety undergraduate students from a variety of disciplines discussed about controversial topics. The topics included "should the legal drinking age be lowered in Massachusetts?" and "what are the pros and cons of using FaceBook or other social networking software as part of high school curriculum?" These discussions were part of experimental trials with the goal of assessing online educational software tools that support SDB. In contrast to participants from the other two domains that were self-motivated to be deliberative, participants in this group received class credit and were encouraged to participate. In this domain, SDB occupied only 32% of the total 1783 annotated segments.

In the third domain, **professional community negotiation**, email exchanges were collected from sixteen geographically dispersed faculty participants who did not know each other and who were from two academic communities. These faculty members negotiated about a proper solution to a conference scheduling conflict. An emerging theme in this dialogue was the tension between democratic decision-making versus top down fiat decision-making by those in authority. Participants were highly educated academic professionals, most of whom encouraged democratic decision making about relocating the conference, which partly led to a more deliberative dialogue. In this domain, SDB occupied 53% of the total 438 annotated segments.

In order to provide training data for machine learning models to automatically assess whether or not participants perform social deliberation, two independent trained human judges had annotated the three corpora based on the social deliberative skill scheme [2]. We achieved good inter-rater re-

[1]Posts were segmented manually at speech act boundaries, and there are typically 3-5 segments per post.

[2]We developed a hand coding scheme containing over 50

Table 1: Data statistics with various domains

Domain	Social deliberative behavior	Other speech acts	Total segment count	Participant count
Civic deliberation	225 (57%)	171 (43%)	396	32
Professional community negotiation	231 (53%)	207 (47%)	438	16
College dialogues	565(32%)	1218(68%)	1783	90
All	1021(39%)	1596(61%)	2617	138

liability scores for both composite and component social deliberative skills as measured using Cohen's Kappa statistics across domains. Note that the social deliberative behavior (or the composite social deliberative skill) is an aggregate over component social deliberative skills. The inter-rater reliability scores of social deliberative behavior for the civic deliberation domain, college dialogues domain, and professional community negotiation domain were 73.5%, 64.3%, and 68.4%, respectively.

4. EXPERIMENTAL DESIGN AND METHODOLOGY

The goal of experiments in this section is to address the following two research questions. First, which type of features (i.e., lexical, discourse, and gender demographic features) or feature combinations are the best for building social deliberative classifiers for each domain? Second, which type of features are the best for building a *robust* social deliberative classifier across domain changes? To this end, we designed two experimental scenarios.

- **Scenario 1: In-domain analysis for each domain**
- **Scenario 2: Cross-domain analysis for each pair of domains**

In both scenarios, we study *feature effects* on prediction performance of machine learning models. Specifically, we build machine learning models using different feature sets and their possible combinations to see which leads to the best prediction performance. We have three types of features (i.e., lexical, discourse, and gender demographic), so we evaluate a group of 6 possible feature configurations.

These two scenarios differ in the following way. The first scenario allows features comparison for each domain and provides the basis for evaluating cross-domain performance. The second scenario offers a systematic view of how machine learning models built with different feature sets perform across domains. We have three domains, so we will evaluate all 6 possible combinations of domain pairs.

With respect to performance measures, we use accuracy (% of correct identification of social deliberative behavior (SDB)), precision (% correct of identified as SDB), recall (% labeled as "SDB" that were predicted to be "SDB"), and F_2 measure (the harmonic mean of precision and recall that weights recall twice as high as precision). Recall is more valued than precision in this study for two reasons. First, the social deliberative skill scheme is expanding, so the SDB considered in this research is by no means complete. The second reason is relevant to our planned applications. Our first planned application to real-time deliberation is through a Facilitators Dashboard. The Dashboard will alert facilitators to potentially important patterns and metrics in the dialogues they are monitoring, in order to help them decide when and how to perform interventions. Because facilitators can intelligently filter out dubious analysis, our algorithms should err on the side of identifying all important patterns, at the risk of including some false positives.

annotations on social deliberative skills and other speech acts.

4.1 Features

Computational understanding of social deliberation is an unexplored research territory. In choosing features for this study, we recognize that we lack sufficient knowledge of what features might be predictive of social deliberative behavior. Therefore, to explore possible features, we turned to the literature of social, psychology, and psycholinguistic studies. This research is the first to use lexical, discourse, and gender demographic features to characterize linguistic patterns of social deliberative behavior.

4.1.1 Lexical features – LIWC

LIWC, Linguistic Inquiry Word Count [18], is a lexicon based linguistic system. It was created by analyzing the utterances of over 24,000 participants totaling over 168 million words. LIWC produces groups of words from 82 language dimensions through a word counting approach. These 82 groups fall into 10 general categories: *linguistic processes*, *social processes*, *affective processes*, *cognitive processes*, *perceptual processes*, *biological processes*, *relativity*, *personal concerns*, *spoken categories*, and *punctuation*.

LIWC has gained a trusted reputation for tracking linguistic features that are indicative of social and psychological phenomena. For example, when investigating gender differences in linguistic styles using LIWC features, researchers in [1] found significant differences between genders for the use of self references, but not for the use of social words and positive and negative emotion words. In [23], LIWC features helped find the roles that emotional and informational supports play in participants' commitment in online health support groups. In another study [8], LIWC helped identify the communication characteristics of terrorists and authoritarian regimes. Given a wealth of evidence of the effectiveness of LIWC features in decoding people's communication and interaction styles from the language they use, we expect that LIWC features can contribute to demystifying the link between language and social deliberation.

4.1.2 Discourse features – Coh-Metrix

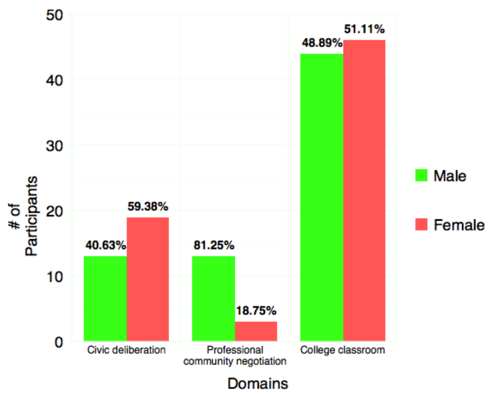

Figure 1: Gender distributions in various domains

Coh-Metrix [7] is a discourse model aimed at better understanding of discourse comprehension, communication breakdowns and misalignments. It was initially developed to explore cognitive constructs of cohesion in written text. Cohesion here refers to the linguistic features that explicitly link words, propositions, and events in a text, which in turn facilitate a reader's coherent mental representation of a text [7]. Coh-Metrix tracks word-level features that are similar to LIWC, but also incorporates modules and algorithms that assess collocations of words. Specifically, Coh-Metrix produces approximately 100 measurements that fall under 8 categories: *narrativity*, *referential cohesion*, *syntactic simplicity*, *word concreteness*, *causal cohesion*, *logical cohesion*, *verb cohesion*, and *temporal cohesion*.

Much like LIWC, Coh-Metrix has been widely used as a computational psycholinguistic tool for predicting complex phenomena, such as affect states, personality, deception, and even physical and mental health outcomes [9, 12, 2, 4]. Given that Coh-Metirx provides a platform for a systematic and deeper analysis of discourse contents, we believe that it can uncover subtle linguistic characteristics relevant to social deliberative behavior.

4.1.3 Demographic feature – gender

Previous research [25] has revealed a formula for successful teams in group environments (e.g., business, classroom, or at home). The formula indicates: *People willing to listen and empathize + people with social sensitivity (i.e., perceive and respond to other's emotions) = smart effective teams able to achieve in any environment.* That research concludes by noting that adding women to a team helps improve group performance. This is because women were found to score higher on average on social sensitivity. Motivated by theses research findings, we decided to incorporate gender as a factor in our analysis. In Figure 1, we show gender distributions in different domains. In future research we will collect other demographic data, such as education level, age, and political orientation, and test their predictive power of social deliberative behavior.

4.2 Machine Learning Models

In this study, we face the problems of small training data and high dimension feature space. In choosing machine learning models to identify social delineative behavior, we prefer a model that meets the following requirements. First, the model is able to select important features automatically during learning. Second, the model performs well with a low ratio of training data size to the number of feature variables. Third, the learnt model is transparent and easy to interpret (i.e., "glass box" model).

As we show below, L_1 Regularized Logistic Regression (L_1RLR) is a model that satisfies our needs. L_1RLR performs feature selection and learning simultaneously. It formulates the learning problem as a trade-off between minimizing loss (i.e., achieving good accuracy on training data) and choosing a sparse model (i.e., improving generalization in prediction on unseen data, higher interpretability, and computational savings).

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

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

where x is the training data, y is the response variable, and $\mathbf{W}$ is the model we learn.

In L_1 Regularized Logistic Regression, we solve the following optimization problem:

$$\arg\max_{\mathbf{W}} \sum_i \log(p(y_i|x_i; \mathbf{W}) - \lambda * \Omega(\mathbf{W})$$

where $\Omega(\mathbf{W})$ is a regularization term used to penalize large weights. In L_1RLR, $\Omega(\mathbf{W})$ is the L_1 norm [22], which is also called least absolute shrinkage and selection operator (Lasso), described below:

$$\Omega(\mathbf{W}) = ||\mathbf{W}||_1 = \Sigma_i |\mathbf{W}_i|$$

Lasso produces a laplace (i.e., double exponential) prior that is "pointy" at zero, which allows feature shrinkage and selection. It is different from the classical L_2 norm [10], also referred to as ridge norm, because L_2 norm produces a Gaussian prior that is near zero and therefore imposes no sparsity. Previous research [16] has shown that L_1 regularization requires the number of training examples that grows logarithmically with the number of features to learn well.

In this study, we used the l_1 regularized dual averaging algorithm [26] for solving L_1 Regularized Logistic Regression. For results reported in Figure 2, we trained l_1RLR (i.e., λ=1, γ=2) [3] with various feature sets. For all in-domain experiments, we report average performance over 10-fold stratified cross-validation within the same domain. For cross-domain experiments, we report results following the training-validating-testing protocol. We trained and validated on the training corpus and tested on the testing corpus.

5. RESULTS AND DISCUSSIONS

Experimental results (Figure 2) reveal a number of interesting patterns. One of the most salient patterns is that imbalanced class/label distribution hurts predictive performance

[3]We also experimented with other values (0.01, 0.1, 10) of λ and found slightly worse performance than the results reported here.

(more on recall than precision), regardless of feature configurations. This can be seen in the third sub-column (i.e., the college dialogues domain) of in-domain results. This observation suggests that before creating a model, it is important to strategically solve the imbalanced data problem, either from the algorithm level (e.g., adjusting class weights or priors) or from the data level (e.g., up-sampling or down-sampling).

Other important yet subtle patterns are explained below. Note that, we use "performance" and "recall" interchangeably in this discussion because recall is the most valued among all the performance measures in this application, as explained earlier.

Gender's effect on predicting social deliberative behavior. The first row in Figure 2 shows that gender *alone* has no predictive power of social deliberative behavior. Specifically, the classification results in each domain reflect the bias of class distribution on training machine learning models toward predicting all data as coming from the majority class. For example, in the college dialogue domain, as shown in Table 1, the majority class is "other speech acts." Classifiers built with various feature configurations unanimously used this bias without any corrections from the gender feature to predict every instance as "other speech acts." This means that every cell in the confusion matrix [4] is zero except the false negative, and therefore recall and precision are zero. This pattern also applies to other domains. We speculate that because social deliberative behavior (as a composite skill) contains skills that greatly overlap cognitive and social/emotional skills, features correlated with only *emotional* related skills, such as social sensitivity, are not effective in predicting social deliberative behavior.

Different capacities of lexical and discourse features in different domains. First, we examine the performance of each feature alone, ignoring feature combinations. As can be seen from in-domain results, compared to LIWC features (70.7% at recall), Coh-Metrix features (83.6% at recall) have the best predictive power on the *civic deliberative domain*. The performance of the model built with LIWC features added on top of Coh-Metrix features has a slight ($< 1\%$) increase in this domain. In the *professional community negotiation domain*, compared to Coh-Metrix features (74.0 % at recall), LIWC features (90.0% at recall) have the upper hand. The performance of the model built with Coh-Metrix features added on top of LIWC features has a drastic ($> 15\%$) decrease in this domain. The *college dialogues domain* has similar patterns as the professional community negotiation domain. In other words, LIWC features are the most predictive for the college dialogues domain. These patterns suggest that lexical and discourse features have different capacities in different domains for the task of predicting social deliberative behavior.

Next, we look at feature combinations. For the *civic deliberation domain*, Coh-Metrix and LIWC features combined, among all 6 feature configurations, led to the best model in that domain. For the *professional community negotiation domain*, LIWC features alone, among all 6 feature configurations, led to the best model in that domain. For the *college dialogues domain*, LIWC and gender feature combined, led to the best model in that domain. This implies that determining which features or feature combinations to use and in which order has an impact on whether and when we will attain the best model. We will explore this point in the text below.

Features for building robust models. Now, we look at the feature effects on predictive performance for cross-domain analysis. The model built with LIWC features using the data from the *professional community negotiation domain* achieved the best cross-domain performance [5]. For example, this model, when applied to the *civic deliberation domain*, achieves 89.3% on cross-domain recall, which is even better than the best in-domain recall (83.6%) achieved by using Coh-Metrix features in this domain. In addition, this model, when applied to the *college dialogues domain*, achieves 86.9% on cross-domain recall, which is much better than the best in-domain recall (9.9%) achieved by using LIWC features in this domain. This observation concludes that LIWC features seem to be the most useful features for building robust models in cross-domain applications. Moreover, when averaging in-domain and cross-domain performance for each feature and feature combinations for each domain, we observe that LIWC features achieved the highest recall (88.8%), followed by Coh-Metrix features (84.5%).

Protocols for using linguistic features to predict social deliberative behavior. The results in Figure 2 imply a protocol about how to use lexical and discourse features to build a model (i.e., l_1RLR) in order to achieve the best in-domain performance. This protocol can be described as follows:

1. Use LIWC features to build a model, whose performance (i.e., recall) is denoted by $p(l)$.
2. Use Coh-Metrix features to build a model, whose performance is denoted by $p(c)$.
3. If $p(l) > p(c)$, the best performance is $p(l)$; otherwise combine LIWC and Coh-Metrix to build a model, whose performance, denoted as $p(lc)$, is the best performance.

This protocol is the most efficient way to find the *right* feature sets for building a model with the best predictive performance. This protocol also suggests that for certain domains LIWC features – features related to "what is said" – are sufficient to predict social deliberative behavior. In these domains, Coh-Metrix features – features related to "how it is said" – might be too overwhelming for the model to achieve good performance. For other domains, the LIWC features are not close enough for identifying the sophistication of social deliberative behavior, and combining with Coh-Metrix features can greatly help increase model performance. In a

[4]In a confusion matrix, each column represents the instances in a predicted class, while each row represents the instances in an actual class.

[5]We witnessed an unstable performance of combining gender with other features. For example, the cross-domain performance of the model built with LIWC and gender combined, compared to that of the model built with LIWC alone, decreases in the civic deliberation domain but increases in the college dialogues domain. Due to the unstable performance of the gender feature, we ignore it for the rest of this study.

broader sense, this protocol evaluated on different corpora provides evidence that determining the right feature set with the best model performance can be streamlined to improve work efficiency. The streamlined processes need to be designed by taking advantage of feature capacities.

Now, we examine the linguistic characteristics of social deliberative behavior. We learnt earlier that, when considering each feature alone, LIWC and Coh-Metrix features have different capacities in different domains for the task of predicting social deliberative behavior. Specifically, Coh-Metrix features are the most predictive in identifying social deliberative behavior in the civic deliberation domain; LIWC features are the most predictive in identifying social deliberative behavior in the professional community negotiation domain and produce the best model for cross-domain prediction tasks. In Table 2, we show the top 10 Coh-Metrix features learnt by L_1 regularized logistic regression built from the civic deliberation domain. Similarly, in Table 3, we show the top 10 LIWC features learnt by L_1 regularized logistic regression built from the professional negotiation domain. Below, we summarize the lexical characteristics of social deliberative behavior and the discourse characteristics of social deliberative behavior.

Lexical characteristics of social deliberative behavior. The *lexical* characteristics of social deliberative behavior, compared to that of "other speech acts," are as follows: shorter message, more dictionary words, fewer big words, fewer words per sentence, more adverbs, fewer pronouns, fewer punctuations, fewer cognitive processes words, fewer space words, fewer auxiliary verbs.

Discourse characteristics of social deliberative behavior. The *discourse* characteristics of social deliberative behavior, compared to that of "other speech acts," include more negative additive connectives, higher negation density, less lexical diversity, less narrativity, shorter message, more pronouns (especially more second person pronouns), fewer spatial motion words, lower word concreteness, fewer connectives.

Examining the linguistic patterns of social deliberative behavior, we found that the LIWC system and the Coh-Metrix system agreed on some features (e.g., a few spatial motion words) and produced incongruent results for others. For example, a few pronouns found by LIWC, whereas many pronouns found by Coh-Metrix. This incompatible finding suggests that social deliberative behavior may have different appearances in different domains. Therefore, in this study we found no conclusive linguistic characteristics of social deliberative behavior.

6. CONCLUSION AND FUTURE WORK

In this paper, we built machine learning models to identify social deliberative behavior from various online dialogues using lexical, discourse, and gender demographic features. We recognized the different capacities of lexical and discourse features in different domains and proposed a protocol about how to use them to build models that achieve the best in-domain performance. We also found that lexical features (i.e., LIWC) were the most useful features for building robust models in cross-domain applications.

Table 2: Top 10 Coh-Metrix features learnt by L_1 regularized logistic regression built from the civic deliberation domain

Coh-Metrix feature	*Interpretation*	*Weight*
CONLOGi	negative additive connectives	9.044
DENNEGi	negation density	8.582
LEXDIVVD	lexical diversity	- 8.122
PNar	narrativity	-7.774
READNW	total number of words	-7.317
PRO2i	second person pronouns	6.03
DENPRPi	pronouns	5.503
SPATlpi	spatial motion words	-4.889
WRDCacwm	word concreteness	-4.69
CONi	all connectives	-4.24

Table 3: Top 10 LIWC features learnt by L_1 regularized logistic regression built from the professional community negotiation domain

LIWC feature	*Interpretation*	*Weight*
WC	word counts	-0.043
Dic	dictionary words	0.037
Six|tr	big words	-0.011
WPS	words/sentence	-0.01
adverb	adverbs	0.009
pronoun	pronouns	-0.009
AllPct	total punctuations	-0.009
cogmech	cognitive processes	-0.007
space	space	-0.004
auxverb	auxiliary verbs	-0.004

In future work, we will include semantic features (e.g., name entity relations) in our models to predict social deliberative behavior. In addition, we will build models using interaction features and structure features to study whether mutual influences, such as linguistic style matching [17], and group dynamics are predictive of social deliberative behavior. We will also investigate whether combining language features and structure features for building models can lead to performance gains. Moreover, we will evaluate the proposed protocol on more data sets to test its external validity and to identify the characteristics of domains with which each feature type (i.e., lexical vs. discourse) works the best. Furthermore, we will create multi-task machine learning models with advanced regularizers (e.g., sparse group Lasso [6]) to simultaneously identify each component social deliberative skills from online dialogues. We hope these endeavors can increase our understanding of the nature of social deliberative behavior and thereby inform the design and development of educational tools to support social deliberative behavior in collaborative processes, from knowledge building, to problem solving, and to communication in general.

7. ACKNOWLEDGEMENT

We thank all anonymous reviewers for their thorough evaluation and constructive recommendations for improving this paper. This research is supported by the US National Science Foundation under grant #0968536 from the division of Social and Economic Sciences. Any opinions or conclusions expressed are those of the authors, not necessarily of the funders.

		In-domain			Cross-domain					
	Training corpus	*Civic deliberation*	*Professional community negotiation*	*College dialogues*	*Civic deliberation*	*Civic deliberation*	*Professional community negotiation*	*Professional community negotiation*	*College dialogues*	*College dialogues*
	Testing corpus	*Civic deliberation*	*Professional community negotiation*	*College dialogues*	*Professional community negotiation*	*College dialogues*	*Civic deliberation*	*College dialogues*	*Civic deliberation*	*Professional community negotiation*
Gender	*Accuracy*	56.8	52.7	68.3	52.7	31.7	56.8	31.7	43.2	47.3
	Precision	56.8	52.7	0.0	52.7	31.7	56.8	31.7	0.0	0.0
	Recall	100.0	100.0	0.0	100.0	100.0	100.0	100.0	0.0	0.0
	F2	86.8	84.8	0.0	84.8	69.9	86.8	69.9	0.0	0.0
LIWC	*Accuracy*	55.3	55.3	67.9	56.2	47.1	59.6	39.3	42.4	47.3
	Precision	58.9	54.6	46.1	56.9	33.4	59.6	32.7	46.8	50.0
	Recall	70.7	**90.0**	9.9	69.7	67.6	**89.3**	**86.9**	9.8	2.6
	F2	68.0	79.7	11.8	66.7	56.1	81.2	65.2	11.6	3.2
Cohmetrix	*Accuracy*	62.1	54.8	68.0	53.4	36.4	53.3	39.0	41.9	47.0
	Precision	61.4	55.3	44.7	53.5	30.7	58.3	30.2	36.8	42.9
	Recall	**83.6**	74.0	3.7	90.0	80.0	62.2	70.3	3.1	1.3
	F2	77.9	69.3	4.6	79.2	60.5	61.4	55.5	3.8	1.6
LIWC+Gender	*Accuracy*	52.3	54.6	68.0	56.2	47.1	60.9	38.3	42.4	48.0
	Precision	56.8	54.2	49.2	56.8	33.2	60.6	32.6	46.8	60.0
	Recall	67.1	89.6	10.6	70.6	66.6	88.9	88.9	9.8	1.3
	F2	64.8	79.3	12.6	67.3	55.4	81.3	66.1	11.6	1.6
Cohmetrix +Gender	*Accuracy*	62.4	55.5	68.2	53.9	35.8	53.8	38.8	40.9	46.2
	Precision	62.7	55.6	46.5	53.9	30.8	58.5	30.1	32.3	33.5
	Recall	83.6	77.5	3.5	87.0	77.9	64.0	70.3	4.0	1.3
	F2	78.3	71.8	4.3	77.5	59.6	62.8	55.5	4.8	1.6
LIWC +Cohmetrix	*Accuracy*	62.4	54.5	68.3	54.1	38.8	55.6	39.4	41.2	48.0
	Precision	63.3	55.1	48.9	54.0	30.9	60.0	30.3	35.7	66.1
	Recall	84.4	74.5	4.1	87.0	75.2	65.3	70.4	4.4	1.3
	F2	79.2	69.6	5.0	77.5	58.4	64.2	55.7	5.4	1.6
All	*Accuracy*	62.1	54.3	68.5	52.7	36.9	54.8	39.8	41.4	47.7
	Precision	62.3	55.0	48.2	53.1	30.6	59.7	31.6	38.7	42.9
	Recall	84.4	74.0	4.6	88.7	78.6	63.1	71.0	5.3	1.3
	F2	78.8	69.2	5.6	78.2	59.8	62.4	56.8	6.4	1.6

Figure 2: Predictve performance (in %) of L_1 regularzied logistic regression built using different feature configurations in different scenarios

8. REFERENCES

[1] C. Bell, P. McCarthy, and D. McNamara. Using liwc and coh-metrix to investigate gender differences in linguistic styles. In *Applied natural language processing and content analysis: Identification, investigation, and resolution*, pages 545–556. Hershey, PA: IGI Global, 2012.

[2] R. Campbell and J. Pennebaker. The secret life of pronouns flexibility in writing style and physical health. *Psychological Science*, 14(1):60–65, 2003.

[3] C. Conati, K. Vanlehn, et al. Toward computer-based support of meta-cognitive skills: A computational framework to coach self-explanation. *International Journal of Artificial Intelligence in Education (IJAIED)*, 11:389–415, 2000.

[4] S. D'Mello, N. Dowell, and A. Graesser. Cohesion relationships in tutorial dialogue as predictors of affective states. In *Proc. 2009 Conf. Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling, V. Dimitrova, R. Mizoguchi, B. du Boulay, and AC Graesser, eds*, pages 9–16, 2009.

[5] D. C. Edelson, D. N. Gordin, and R. D. Pea. Addressing the challenges of inquiry-based learning through technology and curriculum design. *Journal of the Learning Sciences*, 8(3-4):391–450, 1999.

[6] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010.

[7] A. Graesser and D. McNamara. Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, 3(2):371–398, 2010.

[8] J. Hancock, D. Beaver, C. Chung, J. Frazee, J. Pennebaker, A. Graesser, and Z. Cai. Social language processing: A framework for analyzing the communication of terrorists and authoritarian regimes. *Behavioral Sciences of Terrorism and Political Aggression*, 2(2):108–132, 2010.

[9] J. Hancock, L. Curry, S. Goorha, and M. Woodworth. On lying and being lied to: A linguistic analysis of deception in computer-mediated communication. *Discourse Processes*, 45(1):1–23, 2007.

[10] A. Hoerl and R. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

[11] M. C. Linn. Designing the knowledge integration environment. *International Journal of Science Education*, 22(8):781–796, 2000.

[12] F. Mairesse, M. Walker, M. Mehl, and R. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30(1):457–500, 2007.

[13] T. Murray, L. Wing, B. Woolf, A. Wise, S. Wu, L. Clarke, L. Osterweil, and X. Xu. A prototype facilitators dashboard: Assessing and visualizing dialogue quality in online deliberations for education and work. In *Proceedings of the 12th International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government*, 2013.

[14] T. Murray, B. Woolf, X. Xu, S. Shipe, S. Howard, and L. Wing. Supporting social deliberative skills in online classroom dialogues: preliminary results using automated text analysis. In *Intelligent Tutoring Systems*, pages 666–668. Springer, 2012.

[15] T. Murray, X. Xu, and B. Woolf. An exploration of text analysis methods to identify social deliberative skills. In *Proceedings of the 16th International Conference on Artificial Intelligence in Education*, 2013.

[16] A. Ng. Feature selection, l_1 vs. l_2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78. ACM, 2004.

[17] K. G. Niederhoffer and J. W. Pennebaker. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360, 2002.

[18] J. Pennebaker, M. Francis, and R. Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 2001.

[19] B. J. Reiser, I. Tabak, W. A. Sandoval, B. K. Smith, F. Steinmuller, and A. J. Leone. Bguile: Strategic and conceptual scaffolds for scientific inquiry in biology classrooms. *Cognition and instruction: Twenty-five years of progress*, pages 263–305, 2001.

[20] T. Strzalkowski, G. A. Broadwell, J. Stromer-Galley, S. Shaikh, S. Taylor, and N. Webb. Modeling socio-cultural phenomena in discourse. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1038–1046. Association for Computational Linguistics, 2010.

[21] D. D. Suthers, N. Dwyer, R. Medina, and R. Vatrapu. A framework for conceptualizing, representing, and analyzing distributed interaction. *International Journal of Computer-Supported Collaborative Learning*, 5(1):5–42, 2010.

[22] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[23] Y. Wang, R. Kraut, and J. Levine. To stay or leave? the relationship of emotional and informational support to commitment in online health support groups. In *Proceedings of the ACM conference on computer-supported cooperative work*, 2011.

[24] B. Y. White, T. A. Shimoda, J. R. Frederiksen, et al. Enabling students to construct theories of collaborative inquiry and reflective learning: Computer support for metacognitive development. *International Journal of Artificial Intelligence in Education (IJAIED)*, 10:151–182, 1999.

[25] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone. Evidence for a collective intelligence factor in the performance of human groups. *science*, 330(6004):686–688, 2010.

[26] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *The Journal of Machine Learning Research*, 11:2543–2596, 2010.