Fall 2014

COMPUTATIONAL COMMUNICATION INTELLIGENCE: EXPLORING LINGUISTIC MANIFESTATION AND SOCIAL DYNAMICS IN ONLINE COMMUNICATION

Xiaoxi Xu

COMPUTATIONAL COMMUNICATION INTELLIGENCE: EXPLORING LINGUISTIC MANIFESTATION AND SOCIAL DYNAMICS IN ONLINE COMMUNICATION

A Dissertation Presented

by

XIAOXI XU

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

September, 2014

School of Computer Science

COMPUTATIONAL COMMUNICATION INTELLIGENCE: EXPLORING LINGUISTIC MANIFESTATION AND SOCIAL DYNAMICS IN ONLINE COMMUNICATION

A Dissertation Presented

by

XIAOXI XU

Approved as to style and content by:

Beverly Park Woolf, Chair

David Smith, Member

Victor Lesser, Member

Leah Wing, Member

Lori A. Clarke, Chair
School of Computer Science

To my mother and father,
who empower me with pure love and tranquil harmony.

ACKNOWLEDGMENTS

My Ph.D. journey is an uplifting experience. I began wholeheartedly and finish awakened. I identified the self within me and observed it grow intellectually and emotionally. I attribute my ability to set out and my growth along the way to the people who cultivated my sense of curiosity and hard-working early on, who introduced me to the intriguing world of research, who supported my decision to pursue a Ph.D. in computer science among other academic and career opportunities, and more importantly, who inspired me, guided me, and encouraged me in my Ph.D. career. I also acknowledge the people that surround me, with whom I share my vision, who subscribe to a similar purpose, who demonstrate accountability, who support me in the simplest ways, and who are truly selfless. Below, I would like to introduce you to those individuals and send my loving appreciation to them.

- First, my advisor Beverly Park Woolf – I feel a sea of gratitude to you for your incessant support and encouragements. You always see the positive light in me and tap into my potential by setting no hard limits. You are the most energetic and committed advisor I have ever met. Your pioneer research in intelligent tutoring systems has ignited my ever-expanding interest in artificial intelligence in education. Thank you for being a role model for me. Your swift mind, gracious manner, articulated writing, and eloquent speech are what I hope one day be able to emulate.

- My mentor Victor Lesser – I want to express my lifelong gratitude to you for your infinite help and guidance. Your help and shared wisdoms are no less than the GPS of my Ph.D. journey. Thank you for teaching me to see the connections between two seemingly distant problems. Thank you for constantly caring about my growth in research and life. Thank you for always being available when I need advice. Thank you for broadcasting my research and making connections for my career long before I graduate. Your intelligence operates at a high level beyond reach and yet your kindness touches the heart of everyone that you surround. Your influence will have a lasting impact on my future careers.

- My committee member David Smith – A giant thank you to you, for being on my committee and for your inspiring supervision on my independent study about NLP applied to social meaning and intention. Your vast knowledge about statistical models and your divergent thinking made each our discussion an idea harvest experience for me.

- My committee member Leah Wing – A giant thank you to you for being on my committee. Thank you for the stimulating comments, fresh perspectives, and profound questions on this dissertation. Thank you for inspiring me with each small talk we had about communication studies and online dispute resolution.

- My research collaborator Tom Murray – A giant thank you to you for inviting me on a journey to explore how technologies can support deliberative communication among people. Thank you for the shared interests in the place where education, psychology, cognitive science, social science, computer science, and communication meet. I have learned tremendously from your writing and frequent conversations with you around our research on social deliberation and communication in general. You are the most democratic project manager and the most diligent research scientist I have ever met. I admire you truly for your zen-like life style and productivity in research.

- My mentors in college, at work, in Germany, and at LSU: Shujia Jiu, Zongming Guo, Hong Zhang, Frank Schulz, Robert Lax, Rudy Hirschheim, Jianhua Chen,

and Kevin McCarter – A resounding thank you for believing in me for whatever I engage and for always seeing the positive light in me.

- My mentors of my synthesis project: David Jensen and Edwina Rissland – Thank you for your inspiration, guidance, and valuable time. It was a cheerful experience working with you.

- The Computer Science family: Beverly Park Woolf, Victor Lesser, Sridhar Mahadevan, David Jensen, David Smith, Andrew McGregor, Edwina Rissland, Andrew McCallum, and James Allan – I want to express my sincere gratitude to you for your world-class inspiring courses in artificial intelligence and algorithm.

- The Office family: Leeanne Leclerc, Gwyn Mitchell, Rachel Lavery, Susan Overstreet, Jean Joyce, and CSCF – Thank you, thank you, thank you for your kindest and warmest everyday help in my Ph.D. study, for making sure that my paycheck arrives on time, for providing valuable information of conference and career opportunities, and for all the technical supports.

- The AIED, social computing, machine learning, and computer science communities: Beverly Park Woolf, Gordon Anderson, Mark Floryan, Ivon Arroyo, David Cooper, Qiaozhu Mei, Bing Liu, Brendan O'Connor, Stephanie Kent, Lynn Stephens, Nia Dowell, Art Graesser, Andre Martin, Jiayu Zhou, Ning Zhang, Elif Aktolga, Shiri Dori-Hacohen, Bruno Castro da Silva, Van Dang, Marc Maier, Jacqueline Feild, Marwan Mattar, Lisa Friedland, Huseyin Oktay, Akshat Kumar, Huzaifa Zafar, Mastooreh Salajegheh, Meritxell Vinyals, Juan Liu, and Jingyang Wang – I want to acknowledge you for sharing the same vision and similar research interests with me and for your very kind help.

- My extended family: Yan Dong, Lingge Xu, Jinfeng Wang, Baofang Zhang, and Yizhen Fu – Thank you for your kindest understanding and tremendous support and help, especially at difficult times.

- My beloved parents Jindong Wang and Tonghai Xu – Words are truly inadequate to express my appreciation, but please take my thanks with the understanding that it means so much more. Thank you for nurturing me with the highest selfless love. Thank you for teaching me that life is learning and learning is lifelong. Thank you for allowing me to make my own decisions and supporting me in every possible way. Thank you for taking me to violin lessons for 8 years when I was 4 – it helped greatly with my ability to focus for long stretches of time. Thank you for supporting my dreams at every stage of my life! I am beyond blessed to be your daughter. I am also deeply grateful for everything you have done for my family.

- My eternal cat Feidi – Thank you a million for bringing infinite cheerfulness and warmth to my mom while I study overseas. I wish you peaceful and joyful in the divine world.

- My children Sophia and Collin – Thank you for bringing so much joy and playfulness to my life. You are the song to my soul and smile to my heart. I can not express how you have taught me unconditional love by allowing me the time and space to study and write. You are my angel and miracle.

- Finally, my soul mate and love, Chongjie Zhang – Thank you for inspiring and guiding me in my life and research. Being with you is an everyday learning experience for me. Thank you for being such light. You are intelligence, kind heart, nobility, and humor packed into one great man. You still remain the most amazing person I have ever met.

ABSTRACT

COMPUTATIONAL COMMUNICATION INTELLIGENCE: EXPLORING LINGUISTIC MANIFESTATION AND SOCIAL DYNAMICS IN ONLINE COMMUNICATION

SEPTEMBER, 2014

XIAOXI XU

B.Sc., BEIJING TECHNOLOGY AND BUSINESS UNIVERSITY

M.Sc., LOUISIANA STATE UNIVERSITY

M.Sc., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Beverly Park Woolf

We now live in an age of online communication. As social media becomes an integral part of our life, online communication becomes an essential *life skill*. Online communication has long been considered challenging, largely because participants often have no prior relationship with other participants and therefore lack understanding about their backgrounds, values, and expectations. Missing this prior knowledge often leads to misunderstanding and distrust, which in turn lead to poor group performance in collaboration and unsatisfied decision-making in problem solving (e.g., conflict resolution).

In this dissertation, we aim to understand how people effectively communicate online. We research components of success in online communication and present scientific methods to study the skill of effective communication. This research advances the state of art in machine learning and communication studies.

For communication studies, we pioneer the study of a communication phenomenon we call *Communication Intelligence* in online interactions. We create a theory about communication intelligence that measures participants' ten high-order communication skills, including *restraint*, *self-reflection*, *perspective taking*, and *balance*. We present a multi-perspective analysis for understanding communication intelligence, including its diverse language, shared linguistic characteristics across people, social dynamics, and the effects of communication modality on communication intelligence. We discover that people showing more *perspective taking* behaviors are more popular and influential than others in their communication network. Such people also tend to reach out to people who behave similarly, which implies a like-attracts-like social phenomenon that complies with the Law of Attraction. We furthermore show that participants' communication intelligence is on average scored significantly higher in an asynchronous and facilitated communication mode than is in a synchronous and unfacilitated communication mode.

For the area of machine learning, we contribute new computational models and formulations for addressing multi-label and multi-task machine learning problems. We develop a new hierarchical probabilistic model for addressing the problem of simultaneously identifying multiple intelligence-embodied communication skills from natural language. The model learns the topic assignment for each sentence and provides a practical and simple way to determine document labels without relying on a threshold function. The model performance increases as the number of labels grows, which makes it a promising approach for large-scale data analysis. The model also has high interpretability and its annotated sentences significantly augment the view

of each document with rich contextual information. We also develop a new multi-task formulation for simultaneously identifying multiple intelligence-embodied communication skills from lexical, discourse, and interaction features. The key merit of this model is that it is a general multi-task formulation that unifies many widely used regularization techniques, including Lasso, group Lasso, sparse-group Lasso, and the Dirty model. This model expands the applicability of multi-task learning by allowing analyzing real-world problems where (1) the degree of task relatedness is uncertain and (2) the true structure of the groups in data is not clear ahead of time. Moreover, it can be applied to streaming data to perform large-scale analysis in real time. Beyond the application of studying communication intelligence, the developed models and formulations can also benefit research in other areas where the problems of simultaneously predicting multiple categories are abundant. These areas include, but are not limited to, signal processing, computer vision, computational finance, computational biology, and computational neuroscience.

Keywords: Communication intelligence, multi-task learning, hierarchical probabilistic models, regularized canonical correlation analysis, social network analysis

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

As Web 2.0 gains popularity, social media platforms, including online discussion and support forums, collaboratively edited question and answer sites, chat rooms, and Twitter, have enabled new methods of online interactions through *computer-mediated communication*, which in turn provide researchers new opportunities of analyzing *user-generated content* to study large-scale social phenomena. For example, socialpsychologists nowadays study conversations in online communities to understand opinion formation [83] and analyze Twitter data to understand why people retweet [101]. Sociallinguists study Twitter data to address questions about how language reflects people's social identity [47], communication data in Wikipedia to explore how conversational behavior reveals power relationships [41] and how Wiki mediators reconcile online conflicts and help strengthen community membership [12]. Research [6] has shown that user-generated content provides great opportunities for revealing today's social norms and has profound implications for supporting a literate, respectful, and thriving society.

In this dissertation, we aim to understand how effectively people communicate online. Online communication has long been considered challenging, largely because participants often have no prior relationship with other participants and therefore lack understanding about their backgrounds, values, and expectations. Missing this prior knowledge often leads to misunderstanding and distrust, which in turn lead to poor group performance in collaboration and unsatisfied decision-making in problem solving (e.g., conflict resolution). We believe that effective online communication largely depends upon how *intelligent* a participant is in the area of *communication* in online environments. For example, can participants perceive and respond to the feelings of others, reflect on their own bias, and respect others' perspectives? Previous research [34, 143, 138] has shown that skillful behaviors are useful predictors of intelligences. Drawing on this perspective and based on the theory about multiple intelligences [56], the theory about zone of proximal development [158], and theories about communication as deliberation [57, 150, 116], we develop a new theory about communication intelligence. This new theory initiates a conversation between the disciplines of communication studies and computer science about human communication intelligence and computational methods for measuring it. This dissertation takes the first step to address some of the basic questions:

- What is communication intelligence?
- What are the constructs of communication intelligence, or what are intelligence-embodied communication skills/crafts?
- How can communication intelligence be measured based on the use of these skills?
- How can these skills be identified from online messages computationally?

1.1 A Multi-perspective Approach to Communication Intelligence

Large-scale online communication generally takes place in the form of natural language among multiple parties. *What do people say* and *to whom* provide key data for studying communication behaviors and therefore communication intelligence. In addition, online communication can occur in different communication modalities (e.g., synchronous communication and facilitated communication). Studying the effects of communication modalities on communication intelligence can have important pedagogical implications of how to foster a deliberative and effective communication among people. Wallace Stevens poem "Thirteen Ways of Looking At A Blackbird" shows that the essence of a subject, as simple as a blackbird, can be derived from a number of different perspectives. In this dissertation, we are committed to a thorough study of communication intelligence from the following perspectives.

1.1.1 Understanding the Diverse Language of Communication Intelligence

Language is a phenomenon at the interplay of culture, education, psychology, and communication. The different word choices and diverse ways that people use language to express their thoughts and feelings provide great opportunities for studying communication intelligence. For example, in a negotiation context, where two academic communities negotiate a proper solution to a conference scheduling conflict, some people may show agreement *explicitly* by saying "*I also think that bringing this to the BLUEconf community for discussion vote helps build our community* " and others may use more *implicit* language, such as "*I trust Larry G. in the way he is proceeding to collect data while minimize long iterations and clogging mailboxes.* " In the same context, some people may exhibit the behavior of perspective taking when stating that "*As I understand BLUEorgs work with FocusGroups, they would fully understand our decision, and probably support it*" and others may use the expression "*Perhaps a vote will alter the options, or maybe the BLUEconf community as represented by us will disagree with what I have said.*" Human annotators would annotate messages containing these sentences with multiple labels, including *agreement* and *perspective taking*, and yet it is difficult for a computational model to achieve the same level of competence. This task becomes even challenging when the number of labels associated with each message grow, because the computational model would need to address a more complex associative mapping between labels and sentence statements. In this dissertation, we formulate the computational identification of intelligence-embodied communication skills as a multi-label classification problem, in which words are predictors and labels are skills annotated for each message. We present a new hierarchical probabilistic model, called Constrained Labeled LDA, to address the problem of identifying multiple intelligence-embodied communication skills from natural language. This model reveals the language manifestation of intelligence-embodied communication skills and can support large-scale computational annotations of intelligence-embodied communication skills from text corpora online.

1.1.2 Understanding the Shared Linguistic Characteristics of Communication Intelligence Across People

While it is important to learn the diversity in language among people when a particular intelligence-embodied communication skill is applied, it is equally important to explore the shared linguistic characteristics in skill use across people. High-level features, such as lexical and discourse features, provide a good starting point for this exploration. For example, self-reflection might be characterized as using tentative language (e.g., perhaps, guess) and repetitive grammatical aspect – the use of a verb to express an event related to the flow of time (e.g., "I believed," "now I think"). In this dissertation, we formulate the computational identification of communication skills also as a multi-task learning problem, where tasks are skill labels associated with each message, and predictors are linguistics and interaction features. Interaction features are included in this research to explore language coordination [41], consistently shown in the literature. We present a new multi-task formulation with a novel composite regularizer, called Relaxed Sparse-group Lasso, for identifying multiple intelligence-embodied communication skills from lexical, discourse, and interaction features. The key merit of this model is that it is a general multi-task formulation that unifies many widely used regularization techniques, including Lasso, group Lasso, sparse-group Lasso, and the Dirty model. Moreover, this model can be applied to streaming data to perform *real-time* analysis. This model also reveals psycholinguistic and interaction characteristics of each intelligence-embodied communication skill that, to a great extent, resonate with human understanding.

1.1.3 Understanding the Social Dynamics of Communication Intelligence

Examining the conversational structure of online communication, such as who talks to whom and how such interactions form a social network diagram, provides a tool to understand communication intelligence from the perspective of social interaction patterns. In this dissertation, we study the association between intelligence-embodied communication skills and social dynamics measured by social network metrics using regularized canonical correlation analysis (RCCA). RCCA helps characterize intelligent-embodied skills within a social context. For example, people showing more perspective taking behaviors are found to be more popular and influential than others in their communication network.

1.1.4 Understanding the Effects of Communication Modality on Communication Intelligence

Understanding which intelligence-embodied communication skills are better stimulated in which communication modality has significant pedagogical implications. In this dissertation, we analyze research questions, such as "Within which communication mode is participants' communication intelligence on average higher?" and "Within which communication mode do participants show more self-reflection behaviors?"

1.2 Research Hypotheses

In this dissertation, we evaluate the following four research hypotheses.

- Communication intelligence can be defined with respect to high-order communication skills/crafts and can be measured with respect to the use of those skills.

- Constrained Labeled LDA (CL-LDA) achieves better prediction performance than do state-of the-art text classification methods (i.e., Labeled LDA, SVM) for identifying intelligence-embodied communication skills from online text.

- Multi-task learning with Relaxed Sparse-group Lasso (RSGL), as a general multi-task formulation, unifies other widely used regularization methods (i.e., sparse-group Lasso, the Dirty model), without sacrificing performance in identifying intelligence-embodied communication skills from linguistic and interaction features.

- A statistically significant correlation exists between communication intelligence-embodied skills and social network metrics measured from the same participant.

1.3 Dissertation Contributions

Broadly, in this dissertation, we contribute new theories, methods, models, and formulations for measuring and computing communication intelligence. This dissertation also contributes towards the fields of communication studies, machine learning, and natural language processing. The models that we develop are quite general; it is thus likely that the contributions presented in this dissertation will benefit other areas, such as signal processing, computer vision, computational finance, computational biology, and computational neuroscience. Detailed contributions are summarized below.

- We introduce a new theory of communication intelligence, define its constructs, and propose a method for measuring communication intelligence based on intelligence-embodied communication skills.

- We describe a new hierarchical probabilistic model for addressing the problem of identifying multiple intelligence-embodied communication skills from text. This new model outperforms state-of-the-art multi-class text classification methods by learning topic assignments for each sentence in online communication. The model performance increases as the number of labels grows, which makes it a promising approach for large-scale data analysis. The model is also highly interpretable and its annotated sentences significantly augment the view of each text with rich contextual information.

- We describe a new multi-task formulation with a novel composite regularizer for identifying multiple intelligence-embodied communication skills from lexical, discourse, and interaction features. The key merit of this model is that it is a general multi-task formulation that unifies many widely used regularization techniques, including Lasso, group Lasso, sparse-group Lasso, and the Dirty model. This model expands the applicability of multi-task learning by allowing analyzing real-world problems where (1) the degree of task relatedness is uncertain and (2) the true structure of the groups in data is not clear ahead of time. Moreover, this model can be applied to streaming data to perform *real-time* analysis. It also reveals psycholinguistic and interaction characteristics of each intelligence-embodied communication skill that, to a great extent, resonate with human understanding.

- We report experiments on using regularized canonical correlation analysis to decode the association between intelligence-embodied communication skills and social dynamics, measured by social network metrics. This study complements linguistic discoveries of intelligence-embodied communication skills with social dynamic characteristics.

- We demonstrate that participants' communication intelligence is on average scored significantly higher in an asynchronous and facilitated communication mode than is in a synchronous and unfacilitated mode. We furthermore show that females score consistently higher than do males in communication intelligence regardless of communication modalities.

1.4 A Multifaceted Approach to Big-Data Challenges in Machine Learning

To make this dissertation a coherent document, we omit some of the research work conducted during my doctoral studies, whose theme can be described as *a multifaceted approach to big-data challenges in machine learning*. The different facets of big data challenges explored, or being explored, include (1) large volume, (2) high dimensional features, (3) multiple categories/labels, (4) high dimensional multivariate correlations, (5) multiple latent dependencies between features and labels, and (6) multiple modalities. We remark below the challenges of each type of big-data machine learning challenges and provide proper citations to our work [1] that tackle those challenges.

1.4.1 Data with Large Volumes

- *Examples:* Online communication data from a wide variety of online media, such as discussion forums, negotiation sites, and LinkedIn groups for studying perspective taking behaviors

[1] One work not shown here is an exploration study on identifying discourse predictors for skillful communication in negotiation (AAAI' 12) [166].

- *Challenges*: Data from heterogeneous but related sources (e.g., different online contexts or topical domains) or over time (e.g., data stream). How can a model built from one context generalize well in a new context?

- *Solutions*: Robust machine learning models (EDM' 13 [174])(EEE'13 [170])

1.4.2 Data with High-dimension Features

- *Examples:* Hundreds of lexical and discourse features derived from the textual data in online communication for studying perspective taking behaviors; tens of thousands of words in a thread of online discussion for studying perspective taking behaviors

- *Challenges*: Irrelevant features; unstructured data/text. How can relevant features be selected automatically for the purpose of classification and prediction? How can a projection from high dimensional bag-of-words to low dimensional easy-to-understand themes be learned?

- *Solutions*: regularized machine models (AIED' 13 [117], FLAIRS' 14 [171]); latent variable models (DMIN' 12 [173])

1.4.3 Data with Multiple Categories/Labels

- *Examples:* Multiple skill labels annotated at each online message; multiple diseases associated with each brain scan image; multiple stock options associated with each stock quote

- *Challenges*: Multiple labels/classes potentially follow the power law distribution, and some classification problems are under sampled and yet share a feature space with other related problems. How can classifiers for multiple related problems/tasks be learned jointly?

- *Solutions*: Multi-label learning (CL-LDA); Multi-task learning (RSGL)

1.4.4 Data with High Dimensional Multivariate Correlations

- *Examples:* High dimensional psychological variables (e.g., motivation, self-discipline, and self-esteem) and social network variables (e.g., in-degree, out-degree, and hub) measured on the same participant

- *Challenges*: Correlations among two sets of variables, each with high dimension. How can a compact correlation between two sets of high-dimension intercorrelated variables be identified?

- *Solutions*: (Regularized) canonical correlation analysis (CCA) (ITS' 14 [172])

1.4.5 Data with Multiple Latent Dependencies Between Features and Labels

- *Examples:* Various schools of thought used by physicians to prescribe different treatments for patients based on evaluating their previous medical complications, reported symptoms, and test results; different doctrines followed by Supreme court and federal courts to rule for recovery after evaluating case facts, such as product defect, injuries, and professional duties.

- *Challenges*: Multiple latent dependencies of decision labels on data features. How can these various latent conditional dependency be identified?

- *Solutions*: Discovering latent strategies (clustering conditional dependencies) (AAAI' 11 [168], FLAIRS' 14 [169])

1.4.6 Data of Multiple Modalities

- *Examples:* Neuroimaging features, psychological features, and personality-type features are all available for studying the self-reflection behavior of the same participant

- *Challenges*: Features from different modalities or heterogeneous sources – sometimes two-dimension matrix is not sufficient to represent all the modalities. How can features be represented in a way so that each feature set (i.e., perspective) can contribute to learning a specific task?

- *Solutions*: Tensor decomposition (future work)

Although these omitted papers do not directly relate to communication intelligence being studied in this dissertation, they are the early work on seeking computational predictors of high-order communication skills, exploring computational modeling of *composite* high-order communication skill [2] in various discussion and negotiation contexts, and predicting conflict resolution in an online dispute context. For example, we found evidence about significant statistical correlations between discourse features derived from natural language in the communication text and the higher-order communication skills exhibited in the same text. These communication skills include *reflect back*, *mediate*, *negative emotions about topic*, and *questions about topic*. In another study, we developed robust machine learning models using L_1 regularized logistic regression with lexical, discourse, and gender features to distinguish between the composite high-order communication skill and other speech acts. These models achieve up to 68.5% in-domain accuracy (compared to the 50% baseline), 63.3% in-domain precision, and 90% in-domain recall. In cross-domain identification tasks, the developed models achieve up to 60.9% cross-domain accuracy, 60.9% cross-domain precision, and 89.3% cross-domain recall. In yet another study, we explored the possibility of predicting settlements (i.e., resolved vs. unresolved) in online dispute resolution by performing text-analysis on conflict narratives from disputant parties. The experimental data was from eBay Motor vehicles online disputes, in

[2] Composite high-order communication skill simply treats all studied skills as a whole and therefore is an aggregate of each component skill.

which disputants try to resolve complaints, possibly working with online human mediators. We created an unsupervised disputant negotiation model to represent the negotiation process and analyzed the divergence of topic distributions of each party in the dispute to predict conflict resolution in various negotiation scenarios. The developed model achieves 67% in prediction accuracy (compared to the 50% baseline), 69.43% in precision, and 87.2% in recall, outperforming a state-of-the-art supervised learner (i.e., support vector machine) on both precision and recall.

1.5 Organization of the Dissertation

The remainder of the dissertation is divided into 6 chapters, which we describe below.

- **Chapter 2** elaborates background knowledge and related work on theories about communication as deliberation, text classification and hierarchical probabilistic models, multi-task learning and structured sparsity, and social network analysis.

- **Chapter 3** defines communication intelligence and describe constructs of communication intelligence and measures for computing communication intelligence.

- **Chapter 4** presents a new hierarchical probabilistic model for identifying multiple intelligence-embodied communication skills from text.

- **Chapter 5** presents a new multi-task learning formulation with a novel composite regularizer for identifying multiple intelligence-embodied communication skills from linguistic and interaction features.

- **Chapter 6** describes an advanced correlation analysis between communication intelligence-embodied skills and social network metrics measured from the same participant.

- **Chapter 7** concludes this dissertation and describes future work.

CHAPTER 2

BACKGROUND AND RELATED WORK

2.1 Theories about Communication as Deliberation

One popular school of thought in communication studies is that *communication can be best understood through the lens of deliberative democratic theory* [57]. Deliberative democracy, or deliberation, refers to the concept that democratic practice and policy making should rely on "open and informed" communication on the part of citizenry. "Openness" refers to the ability to allow a voice for multiple perspectives in a discussion, and "informed" refers to the capacity of making rational arguments in the discussion. Deliberation is originally used in the context of politics, where citizens are gathered together in small groups to discuss public or political issues relevant to their communities. Examples of *political deliberation* include National Issue Forums [140], Deliberative Polls [50], and Twenty-first Century Town Meetings [99].

With the rise of socially enabling technologies and the advent of computer-mediated communication, deliberation in online interactions is being explored as a new opportunity to understand participants' communication behaviors online and as a new possibility of promoting the establishment of an increasingly deliberative society. A large body of research in communication studies have provided great insights into the study of the *analytic* aspect (i.e.,"informed") of *online deliberation* [148, 150]. These work include studies on rational argument and consensus [37, 65], problem solving and inquiry [141, 27], critical thinking [87], and metacognition [106, 165]. A burgeoning body of research has examined the *social-relational* aspect (i.e., "openness") of deliberation in group interactions, such as respect [97], conflict management [8, 21, 97], establishing trust [133, 88], and managing group members' expectations [23].

To encourage attention to both analytic and social-relational dimensions of deliberation, Gastil and Black [58, 14] have created a prominent framework that conceptualizes the dual dimensions of deliberation. In their framework, the analytic dimension includes five components: *creating an information base*, *prioritizing key values at stake*, *identifying a wide range of possible solutions*, *weighing the solutions*, and *(in situations that call for decisions) making the best decision possible*. The social-relational dimension involves the following four components: *having equal and adequate speaking opportunities*, *attempting to comprehend anothers views*, *making efforts to fully consider another's input*, and *demonstrating respect for each other*. This framework was created for the purpose of understanding politic conversation and discussion and has been successfully applied to analyzing the communication behavior of team members who collaboratively edit knowledge repositories in a Wikipedia environment. However, the authors note the potential limitation of this framework in trying to apply their framework to other contexts of online interactions.

Murray's theory about social deliberative skills [116] is a conceptual framework that considers both analytic and social-relational dimensions of deliberation, with a focus on the social dimension of deliberation in an online environment. Specifically, this theory focuses on inter-subjectivity (i.e., shared values, meaning, background, relationship), interaction, self-reflection, reciprocal role taking, and cognitive empathy. More importantly, this theory was created based on examining online group conversations in a variety of domains and across both collaborative (e.g., college classroom online discussions, civic deliberation forum discussions) and conflicting contexts (e.g., workplace disputes, e-commerce disputes, and divorce settlements). Murray argues that *social deliberative skills* are the driving force of high-quality communication. In his theory, social deliberative skills include the following three dimensions:

- Social perspective taking, which includes cognitive empathy and reciprocal role taking;
- Social perspective seeking, which includes social inquiry and question asking skills;
- Social perspective monitoring, which includes self-reflection and meta-dialogue.

This theory about social deliberative skills lays the groundwork for studying communication intelligence.

2.2 Text Classification and Hierarchical Probabilistic Models

Machine learning approaches to multi-label text classification have largely relied on discriminative modeling techniques, such as support vector machines (SVM) [39]. In general, discriminative approaches suffer from huge performance loss when the total number of labels and the number of labels per text document grow larger [98]. Their performance degrades even more when the label frequencies follow a highly skewed power-law distribution as often observed in real-world data sets.

Among generative approaches, hierarchical probabilistic models, such as Latent Dirichlet allocation (LDA) [20], have gained widespread popularity in analyzing large-scale text collections in science, humanities, industry, and culture [63, 121, 4, 43, 111]. LDA is a probabilistic model for discovering main themes or topics that pervade a large and otherwise unstructured collection of texts. Therefore, it is also referred to as topic model. The main advantages of topic models are interpretability and extensibility.

- **Interpretability:**
  - Topic models reduce the dimension of text data by projecting high-dimension bag-of-words into low dimensional salient themes, which greatly help interpret, organize, and summarize the text.

  - Topic models describe generative processes about how text arose, which facilitate the "intuitive" understanding of how topic models work.
  - The central computational problem for topic modeling is to "reverse" the generative process – discovering themes through posterior inference. The posterior distributions can be used in creative ways, such as visualization, summarization, and recommendations.

- **Extensibility:**
  - LDA is a simple building block that can be embedded in more complicated models to enable many applications. For example, it can be extended to account for syntax, authorship, interaction, word sense, dynamics, correlation, hierarchies, and other structures [149, 173, 64, 24, 17, 18, 15] and can model a variety of data, including images, social networks, music, software bugs, purchase histories, genetic data, and other types [48, 108, 146, 4, 78, 33].

LDA is an unsupervised model. For data that are paired with response variables, we need a variant of LDA that models both text and responses. In the literature, a number of approaches have been proposed for adapting the unsupervised LDA model to its supervised variants, including supervised topic models [19], DiscLDA [89], and MedLDA [178]. However, these adaptations are designed for single label classification or regression, but not for learning tasks with multiple labels. In [132], Labeled LDA was proposed with the intention to solve multi-label classification problems. In Labeled LDA, supervision is accomplished by constraining the topic model to use only those topics that correspond to labels in the label set. One great advantage of Labeled LDA is that it explicitly assigns individual words to specific labels of a piece of text, rather than assume that all of the words in the text are relevant to each label. (Discriminative models, such as SVM, however, do not model this association explicitly.) This associative mapping between words and labels is critical for accomplishing many real-world tasks, such as tagging documents, tagging webpages, and information retrieval. Despite its capacity of solving multi-label text classification problems, the generative process described in Labeled LDA inherits the naive assumption from LDA that words within a document are assumed to be exchangeable. This bag-of-words assumption is the fundamental limitation to apply topic models to domains, where word order, phrases, or sentences are critical to capturing the meaning of text, such as in this study.

In the rest of this section, we provide an overview of LDA, Labeled LDA, and author-topic model [149]. The author-topic model is reviewed for the reason that the method carries great similarity to Labeled LDA and thus the relationship between these models is important to note.

2.2.1 LDA

LDA [20] and other topic models are part of the larger field of *probabilistic modeling*. In generative probabilistic modeling, data is assumed to arise from a generative process that includes "latent variables." This generative process defines a *joint probability distribution* over both the observed and latent random variables. We uncover the latent variables by computing the *posterior distribution* of the latent variables conditioned on the observed variables. In LDA, the observed variables are words of the documents; the latent variables are the topics. The generative process in LDA can be described as a two-step stochastic process shown below.

1. For every topic β out of K,

   (a) Draw a word distribution $\beta_k \backsim$ Dirichlet (η).

2. For each document d out of D,

   (a) Draw a topic proportion $\theta_d \backsim$ Dirichlet (α).

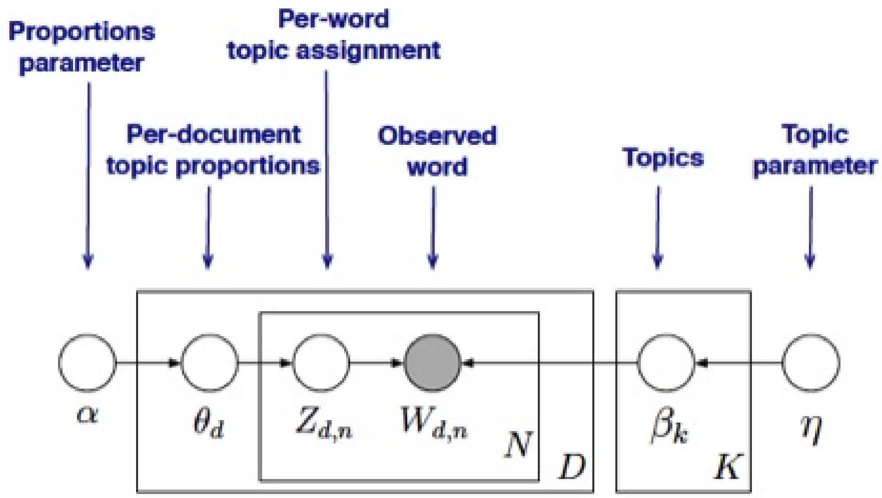

Figure 2.1. The graphical model of LDA (from [16])

(b) For each word n out of N,

(i) Draw a topic assignment $Z_{d,n} \backsim$ Multinomial (θ_d).

(ii) Draw a word $W_{d,n} \backsim$ Multinomial $(\beta_{Z_{d,n}})$.

The graphical model using plate notation shown in Figure 2.1 gives us the following joint probability distribution:

$$\begin{aligned} P(\theta, \beta, z, w) &= P(\theta)P(\beta)P(z|\theta)P(w|z, \beta) \\ &= \prod_d Dir(\theta_d; \alpha) \prod_k Dir(\beta_k; \eta) \prod_n \theta_{z_n|d_n} \prod_n \beta_{w_n|z_n} \end{aligned}$$

To uncover latent topics, we compute the posterior distribution over latent topics given the observed words and model hyperparameters α and η :

$$P(\theta, \beta, z|w, \alpha, \eta) = \frac{P(\theta, \beta, z, w, \alpha, \eta)}{P(w, \alpha, \eta)}$$

In addition to learning high-level topic themes, estimating θ and β provides information about the topics that participate in a corpus and the proportions of those topics in each document respectively. Various learning algorithms have been developed in recent years to estimate latent topics and model parameters, including collapsed Gibbs sampling [63], variational inference [20], and expectation propagation [113]. In [9], empirical evaluations were carried out to compare these algorithms and found that using appropriate hyper parameters causes the performance differences between these algorithm to largely disappear.

As can been seen from the generative process of LDA, each document is assumed to be a finite *mixture* over latent topics. In the next section, we will introduce another topic model that extends LDA to include author information. In that model, each document is assumed to be a mixture over authors who are associated with a mixture over latent topics.

2.2.2 Author-topic model

The author-topic model [149] reduces the process of writing a scientific document to a simple series of probabilistic steps and is aimed at discovering the topical interests of each author. As a result, it not only discovers which topics participate in each document, but also which authors are associated with each topic. In this model, the list of authors is assumed to be observed. When generating a document, an author is chosen uniformly at random. The generative process of the author-topic model, shown in Figure 2.2, can be described as follows. Note that in the author-topic model, we have two sets of latent variables: $\mathbf{z}$ and $\mathbf{x}$ for topics and authors respectively.

1. For every topic ϕ out of T,

   (a) Draw a word distribution $\phi_t \backsim$ Dirichlet (β).

2. For each document d out of D,

   (a) Choose an author x uniformly from an observed list of authors a_d.

   (b) For each author x out of A,

      (i) Draw a topic proportion $\theta_x \backsim$ Dirichlet (α).

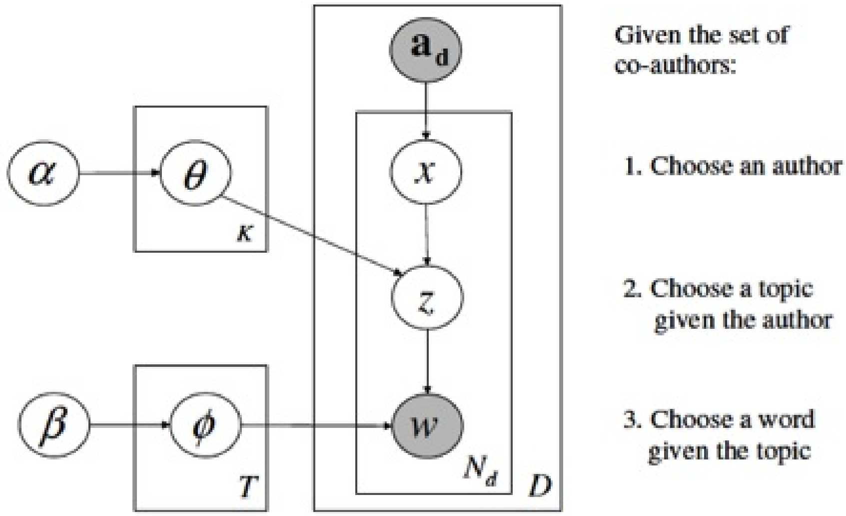

Figure 2.2. The graphical model of author-topic model (from [149])

(ii) For each word n out of N,

(1) Draw a topic assignment $Z_{x,n} \backsim$ Multinomial (θ_x),

(2) Draw a word $W_{x,n} \backsim$ Multinomial $(\phi_{Z_{x,n}})$.

2.2.3 Labeled LDA

Labeled LDA [132] is very similar to the author-topic model from the previous section. The author-topic model is conditioned on the set of *authors* in a document, and therefore, a "topic" is learned for each author in the corpus. Similarly, Labeled LDA is conditioned on the set of *labels* assigned to a document, and a "topic" is learned for each label in the corpus. Labeled LDA describes a process for generating a labeled document collection. Like LDA, Labeled LDA models each document as a mixture of latent topics and generates each word from one topic. Unlike LDA, Labeled LDA incorporates supervision by constraining the model to use topics that are correspondent to a document's observed label set. Therefore, the number of topics K is now the number of labels in the labels set. The graphical model of Labeled LDA

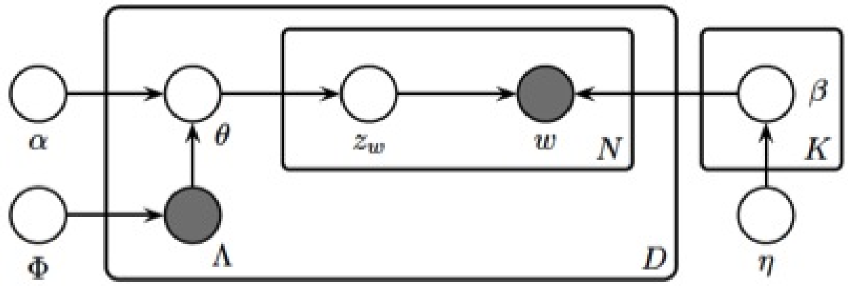

Figure 2.3. The graphical model of Labeled LDA (from [132])

is shown in Figure 2.3. The generative process of Labeled LDA can be described as follows:

1. For every topic β out of K,

    (a) Draw a word distribution $\beta_k \backsim$ Dirichlet (η).

2. For each document d out of D,

    (a) For each topic β out of K,

        (i) Draw topic presence/absence indicator $\Lambda \in \{\ 0,\ 1\ \} \backsim$ Bernoulli (ϕ).

    (b) Project the Dirichlet prior vector into low dimensions $\alpha = L$ x α.

    (c) Draw a topic proportion $\theta \backsim$ Dirichlet (α).

    (d) For each word n out of N,

        (i) Draw a topic assignment $Z_{d,n} \backsim$ Multinomial (θ_d).

        (ii) Draw a word $W_{d,n} \backsim$ Multinomial $(\beta_{Z_{d,n}})$.

As can been seen from the generative process, the constraint that the document's labels are restricted to its own labels is fulfilled by step 2(b), where the Dirichlet prior α is projected from topic dimension K into a low dimension of the size of document labels M. To accomplish this, a projection matrix L of dimension of M by K is created. Each entry of this matrix has value 1 only if the document label m is equal to the topic k, zero otherwise.

2.3 Multi-task Learning and Structured Sparsity

Multi-task learning (MTL) [30] is a learning paradigm where multiple related tasks are jointly learnt. The key idea of MTL is that tasks, if related, learned simultaneously through *parallel inductive transfer* can mutually benefit each other and lead to improved prediction performance. Multi-task learning has been applied to many problems, including those in computer vision, finance, natural language processing, and genomics [155, 130, 59, 3, 124]. Multi-task learning is especially beneficial when the training sample size is relatively small for each task, because learning multiple related tasks simultaneously increases the sample size for each task and consequently improves the performance of the learners. Therefore, it favorably addresses the problem of the skewed class distribution in power-law data, such as unbalanced skill use in life contexts. In the rest of this section, we discuss in detail important notions in MTL, regularization-based approaches to MTL, and its limitations. We hope this discussion can foster a better understanding about the motivations for creating a new model to address the task of predicting multiple high-order communication skills.

2.3.1 Inductive Transfer Through Parallel Learning

In traditional single-task learning, each task is considered to be independent and therefore learned independently. In multi-task learning, multiple tasks are learnt in parallel, by using task relatedness. As shown in Figure 2.4, multi-task learning and single-task learning are different in the training or induction phase. The induction of multiple tasks is performed simultaneously to capture intrinsic relatedness. Multi-task learning emphasizes parallel learning and transfer rather than sequential processes is because if training tasks are executed independently followed by transfer using only

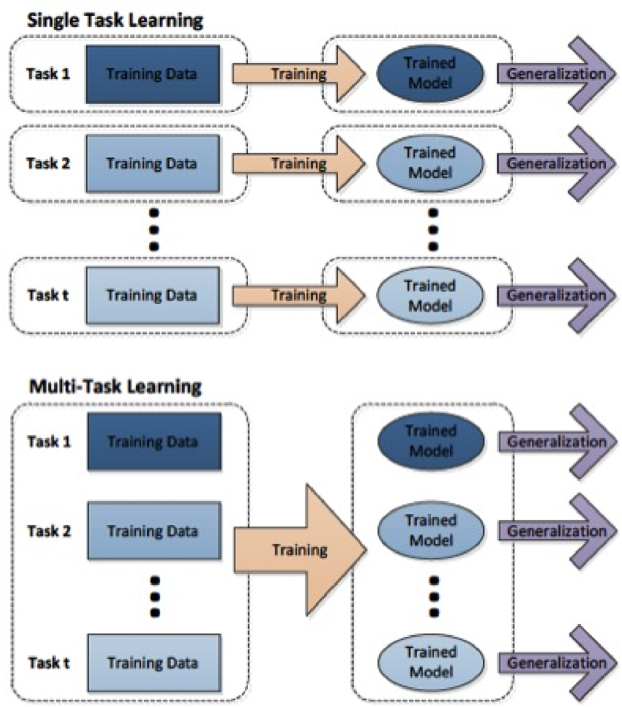

Figure 2.4. A comparison of single-task learning and multi-task learning (from [177])

the models learned for each task, then the computation will result in lost information in the training data that are not captured by those models [30]. In other words, the representations (i.e., models) learned to achieve good performance on tasks trained individually may not be the representation that a learner learning a related task will find most useful.

2.3.2 Task Relatedness vs. Task Dependence

One key assumption of multi-task learning is that tasks relate to each other, so it is critical to understand what we mean by *task relatedness.* Two tasks that are correlated or dependent are related; tasks that have no correlation or dependency can still be related. In the latter case, task relatedness may exist in the feature space they share. It is important to note the difference between *explicit* task dependence or correlation and *implicit* but *intrinsic* task relatedness. In multi-task learning, there

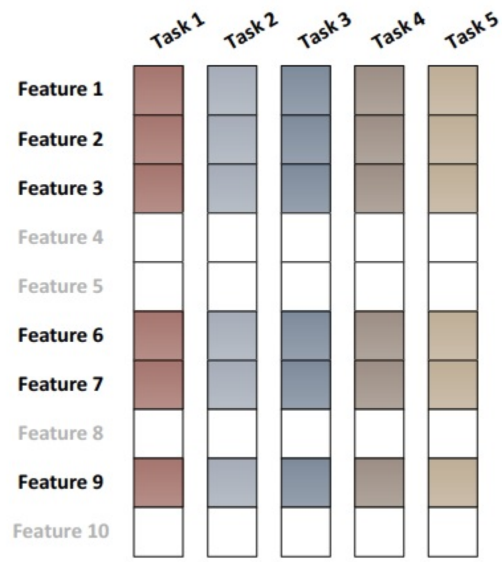

Figure 2.5. Task relatedness through the share of a common set of features (from [177])

are two ways to define task relatedness. The first one assumes that tasks share a common yet latent feature space [5], as shown in Figure 2.5. The second way is to assume that tasks share a low-rank subspace [80]. While the first way learns a shared feature space among tasks, the second learns a model in a black-box way [1]. In this dissertation, we focus on the former paradigm. Consider the task of predicting multiple intelligence-embodied communication skills, commonness among skills may not be shown with respected to language use, yet it does not exclude the possibility that there is one at the level of discourse style, as measured by LIWC and Coh-Metrix systems.

2.3.3 Joint Feature Learning

In the multi-task learning paradigm, joint feature learning, or embedded feature learning, is based on the assumption that *a common yet latent structure in the feature*

[1]For the purpose of feature analysis, we need to perform singular value decomposition on the resulting model to get the basis and their importances (singular values).

space is shared among multiple tasks. A great strength of joint feature learning is to *perform feature selection and learning simultaneously.* Before we introduce how embedded feature selection works, we review existing approaches to automatic feature selection.

- **Filter approach:** The idea of filter approach to feature selection is that for each candidate feature, a heuristic is applied to determine whether or not to include it. Examples of filter-based feature selection include mutual information [126] and correlations [66] between features and labels. This approach to feature selection can be considered as a preprocessing step independent from the learning algorithm used at the learning phase. Although the filter approach tends to be fast, its major drawback is that the optimal subset of features may not be independent of the biases of the learning algorithm. Therefore, the preselected features may not lead to the best learning performance.

- **Wrapper approach:** Wrapper-based feature selection uses a search algorithm to search through the space of possible features and evaluate each subset by running a learning model on the subset. Examples of wrapper-based feature selection include decision tree [131] and random forests [26]. This approach can be computationally expensive and has a risk of overfitting to a particular model.

- **Embedded feature selection:** This approach addresses the limitations in both filter-based and wrapper-based feature selections. Specifically, it selects features while performing learning and it formulates the learning problem as a trade-off between minimizing loss (i.e., achieving good accuracy on training data), and choosing a desirable model (i.e., improving generalization in prediction on unseen data, interpretability, and computational savings.) Formally, embedded feature selection approach to MTL can be described as follows:

$$\min_{\mathbf{w}} L(\mathbf{w}, X, Y) + \Omega(\mathbf{w}) \tag{2.1}$$

where L is an empirical loss function and Ω is a regularizer; X is the inpute feature matrix with dimension N x D (where N is the number of training examples; D is the feature dimension), Y is the response matrix with dimension N x T (where T is the number of tasks), $\mathbf{w}$ is the weight matrix with dimension D x T.

Widely studied loss functions include square error loss, logistic loss, hinge loss, and perceptron loss [13]. Square error loss is typically used in regression analysis, whereas log-liner loss functions as employed in logistic regression [109]; maximum entropy [11] and conditional random fields [90] are often used for classification and structure prediction. It is beyond the scope of this dissertation to review each loss function. For our multi-task text classification task, we will use logistic loss. In the next section, we motivate our proposed work by reviewing the commonly used regularizers in length.

2.3.3.1 Structured Sparsity and Regularization

Sparsity in the context of MTL refers to *model sparsity.* In other words, most dimensions of the feature space are not needed for the learning task and those dimensions can be then set to zero, leading to a sparse model. Model sparsity is desirable, because it leads to a model that is more interpretable and has greater generalizability than a model that is not sparse. Regularization is often achieved by inducing sparsity. Sparsity-induced regularization is often achieved by using norms.

Different choices of norms make a difference at the level of sparsity. For example, classical L_2 norm [72], also referred to as ridge norm, imposes no sparsity.

- **Ridge** (L_2 norm):

$$\Omega(\mathbf{w}) = \frac{\lambda}{2}||\mathbf{w}||_2 \tag{2.2}$$

L_1 norm [154], also called least absolute shrinkage and selection operator (Lasso), yields element-wise sparsity.

- **Lasso** (L_1 norm):

$$\Omega(\mathbf{w}) = \lambda||\mathbf{w}||_1 = \lambda\Sigma_{i=1}^{D}|\mathbf{w}_i| \tag{2.3}$$

where D is the feature dimension. The difference between ridge and Lasso can be better explained from a Bayesian perspective. Specifically, while the ridge penalty produces a Gaussian prior that is near zero, the Lasso penalty produces a laplace (i.e., double exponential) prior that is "pointy" at zero, which allows feature shrinkage and selection. More analysis about when to use which norms can be found in Ng's paper [122].

Another popular norm is group Lasso, which implements the idea of promoting structured patterns by discarding entire group of features.

- **Group Lasso:**

$$\Omega(\mathbf{w}) = \Sigma_{m=1}^{M}||\mathbf{w}_m||_2 \tag{2.4}$$

where M represents groups $G_1, \ldots, G_M$, each $G_m \subseteq \{, 1, \ldots, D \}$, $\mathbf{w}_1, \ldots,$ $\mathbf{w}_M$ are feature sub-matrices.

The group Lasso regularizer is also called composite regularizer, because it is the l_1 norm of the l_q norms, where $q > 1$. Technically, it is still a norm, but a mixed norm, denoted $l_{q,1}$ [2]. In the literature, $l_{\infty,1}$ and $l_{2,1}$ norms are the two commonly used structure sparsity-imposed norms. Although both regularizers induce sparsity on the group level, the $l_{2,1}$ norm penalizes the sum of the group-wise l_2 norms of the regression weights, whereas the $l_{\infty,1}$ norm penalizes the sum of maximum absolute values per group. Group Lasso is widely used in MTL. In [157], it has shown that $l_{2,1}$ consistently outperforms the $l_{\infty,1}$ counterpart

[2]Some researchers use the notation of $l_{1,q}$.

in terms of predicting accuracy for MTL. Using group Lasso for MTL means a feature is either selected as relevant for all tasks simultaneously, or is excluded all-together for all tasks. Thus, it is often referred to as "simultaneous Lasso." It is this all-in-all-out manner that captures the common structure across multiple tasks. Although group Lasso constrains the set of selected features to be identical across tasks, it allows for different amplitudes for the selected regression coefficients.

Despite its structured sparsity, group Lasso finds its limitations in many real world problems where the interest is in identifying *important groups as well as important features within the selected groups.* As an efficient way of addressing this limitation, sparse-group Lasso [54] was proposed to produce the desired effect of group-wise and within group sparsity by combining group Lasso and Lasso.

- **Sparse-group Lasso:**

$$\Omega(\mathbf{w}) = \lambda_1||\mathbf{w}||_{2,1} + \lambda_2||\mathbf{w}||_1 \tag{2.5}$$

In Figure 2.6, we visually illustrate the difference among sparsity-induced norms, namely, Lasso, group Lasso, and sparse-group Lasso.

2.3.4 Sharing vs. Individual Difference

Although "shared common structure" is the key assumption for joint feature learning, research [118] has shown that if the extent of overlap in the feature space is less than a threshold, the group Lasso regularization (and therefore the sparse-group Lasso) could actually perform *worse* than simple separate element-wise l_1 regularizer. Since the choices of regularizer largely depend on the unknown true parameter hidden in the data, we might not know when and which regularizer to apply beforehand.

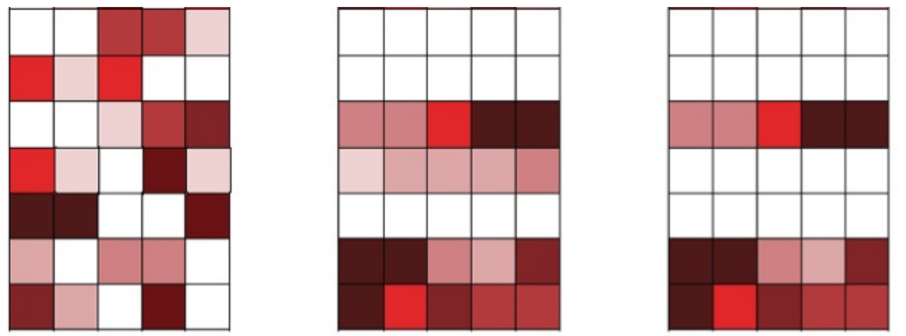

Figure 2.6. A comparison of sparsity-induced norms

Recent research has proposed the idea of decomposing the model (i.e., regression weight matrix) into a task-specific component and a sharing component. As shown in Figure 2.7, an element-wise regularizer (i.e., l_1 norm) is imposed on the task-specific component, and a group Lasso regularizer is imposed on the sharing component. The resulting model is referred to as the " Dirty model [79]." Since the Dirty model uses group Lasso rather than sparse-group Lasso, it does not encourage the sparsity within a group. As a result, important features within the selected groups can not be effectively recognized. In addition, the Dirty model applies the $l_{\infty,1}$ norm, which, as discussed earlier, is not as effective as $l_{2,1}$ norm for MTL.

- **The Dirty Model**:

$$\begin{aligned} &\min_{\mathbf{w}} L(\mathbf{w}, X, Y) + \lambda_1 \parallel \mathbf{u} \parallel_{\infty,1} + \lambda_2 \parallel \mathbf{v} \parallel_1 \\ &\text{subject to: } \mathbf{w} = \mathbf{u} + \mathbf{v} \end{aligned} \tag{2.6}$$

where u denotes common structure; v denotes task-specific structure.

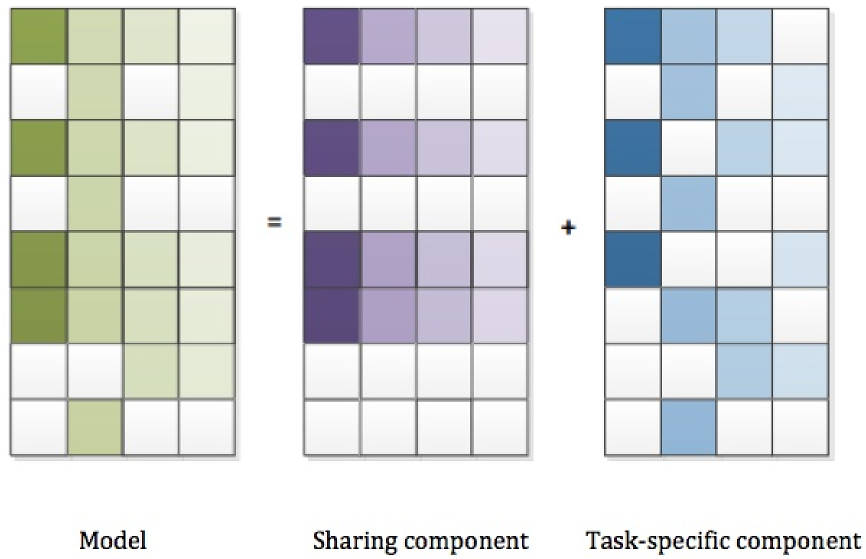

Figure 2.7. An illustration of the Dirty model (adapted from [177])

2.4 Social Network Analysis

Social network analysis (SNA) has its origins in both social science and in the broader fields of network analysis and graph theory. SNA concerns itself with analyzing social relationships by studying individuals who are embedded in a network of relations and by seeking explanations for social behaviors in the structure of these networks [162]. SNA has gained a significant following in anthropology, communication studies, biology, social psychology, sociolinguistics, economics, geography, information science, and organizational studies [142, 162, 22, 49, 110]. Social network analysis is also diverse in perspectives, from ego network to whole network, structure to relation, and behavior to attitude. For example, SNA has been used to examine how organizations interact with each other, characterizing the many informal connections that link executives together, as well as the connections between individual employees in different organizations [162]. In another study, SNA has found that happiness is contagious in social networks – when a person is happy, nearby friends have a 25 percent higher chance of being happy themselves [51]. In this dissertation, we will use SNA to analyze the structure properties of a participant's social interactions (e.g., who talks to whom), which may provide some keen insights into the nature of communication intelligence and its embodied communication skills. In the text below, we will first introduce the definition of social network analysis, and then describe several network properties and structure measures.

Social network analysis views social relationships in terms of network theory, consisting of nodes (representing individuals within the network) and links (representing relationships between individuals [162]). For example, in this dissertation, a social communication network comprises interconnected nodes representing participants in online communication and directed edges representing messages sent from one participant to another. Weights frequency can be added to each edge to represent the number of messages sent in a period of observation. SNA analyzes social connections by studying the following four network properties [164].

- Homophily: The extent to which individuals form connections with similar versus dissimilar others.
- Reciprocity: The extent to which two individuals reciprocate each other's interaction.
- Transitivity: The tendency for a connection to be transitive (A connects to B, and B connects to C, so A connects to C).
- Propinquity: The tendency for individuals to have more connections with other individuals geographically or psychological close to them (i.e., like attracts like).

Note that homophily and propinquity can overlap, when similarity is defined by viewpoints or beliefs. Transitivity is not applicable in this research because communication online is often motivated by social choice and preference and a lack of direct connection implies a social choice of "no."

SNA also uses an array of network metrics to study social interaction characteristics. Different network analysis tools focus on different selections of network metrics. In this research, we choose Gephi [3], an open-source network analysis package widely used in the community of communication studies. In Table 6.1, we list network metrics measured by Gephi and provide their interpretations in the context of this dissertation. Conventional interpretations of social network measures can be found in [163, 125], which have shown that, for example, hub and degree are the constructs of popularity, authority is a measure of influence, and components indicate communities.

Table 2.1: Social network measures and their interpretations in the context of this research

Network structure measures	**Definition**
In_Degree	This metric indicates the number of people, from whom a message is sent to the studied participant.
Out_Degree	This metric indicates the number of people, to whom a message is sent from the studied participant.
Degree	This metric indicates the total number of people that the studied participant has communication with.
Weighted_In_Degree	This metric indicates the number of messages received by the studied participant.

[3]https://gephi.org/

Weighted_Out_Degree	This metric indicates the number of messages sent by the studied participant.
Weighted_Degree	This metric indicates the total number of messages both received and sent by the studied participant.
Eccentricity	This metric indicates the length of the longest directed path (assuming it is the only path) between the studied participant and another participant.
Closeness_Centrality	This metric indicates the average length of the directed path between the studied participant and another participant.
Betweenness_Centrality	This metric indicates on average how possible the studied participant is in the middle of a direct chain between any two other participants.
Authority	This metric indicates how influential the studied participant is.
Hub	This metric indicates how popular the studied participant is.
Modularity_Class	This metric indicates how sophisticated the communication network's internal structure is.

PageRank	This metric indicates on average how influential the participants who send messages to the studied participants are.
Component_ID	This metric describes community.
Strongly_connected_ID	This metric describes how closely members of the community, to which the studied participant is belong, interact.
Clustering_Coefficient	This metric indicates how closely the neighborhoods of the studied participant interact.
Eigenvector_Centrality	This metric also indicates on average how influential are the participants who send messages to the studied participant.

CHAPTER 3

COMMUNICATION INTELLIGENCE

Living, succeeding, and leading in the 21^{st} century hinge closely on consistently acquiring and subsequently enhancing transferable skills – skills that serve purposes in many areas of work and life and can be used from one situation to another. These skills include effective communication, critical thinking, and self-directed learning. Among these skills, the ability to communicate effectively is often treated lightly or even glossed over in our personal agenda for self-improvement and growth. As a result, effective communication endures as one of the central, collective challenges in our society. We think this happens partly because of the lack of a clear, actionable definition of effective communication, and partly because a supportive scaffolding environment for improving communication capacity is not in place.

In this research, we hope to initiate a movement to improve people's communication capacities in online interactions. We research skillful behaviors related to effective communication and develop computational systems for modeling and measuring peoples' communication behaviors online. Although our research focuses on online communication, it is reasonable to assume that the resulting skill improvement is likely to be transferred from online to offline experiences.

Our research starts with a realization that individuals are born with a general intelligence in communication, upon which one can improve with practice, as in other forms of intelligences [45, 138, 143]. Therefore, we take the first steps to define an ability-based model for this intellectual construct of communication that we call *communication intelligence.* In this chapter, we present the definition of communication intelligence and describe a measure to compute it based on participant's use of intelligence-embodied communication skills. In the next two chapters, we will present computational models for identifying intelligence-embodied communication skills from online dialogues. The last chapter of this dissertation will focus on plans of creating a scaffolding environment for improving people's communication intelligence.

3.1 Related Work

Research has shown that people have multiple intelligences, including IQ [100], EQ (i.e., emotional intelligence) [60], and SQ (i.e., social intelligence) [61]. Practical experiences teach us that, instead of one intelligence, it is the combination of both that makes for success in life. The research on growth mindset [45] has suggested that IQ is malleable and improvable. EQ and SQ can also be strengthened *through skill practice*, as shown in [138, 143]. In the theory of multiple intelligences, emotional intelligence is absorbed by *intrapersonal intelligence*, or the ability to have a deep understanding of the self, including one's strengths and weaknesses, knowledge of what makes a person unique, and the ability to self-reflect and to control one's own emotion and reaction. Similarly, social intelligence is subsumed by *interpersonal intelligence*, or the ability to lead and inspire other people through influence, empathy, and care. In this research, we aim to understand the projection of intrapersonal intelligence and interpersonal intelligence onto the space of *communication* [1] in online environments with the ultimate goal of supporting participants to improve their communication skills.

In addition, our research is closely related to existing theories about communication as deliberation [57, 150, 116], which we reviewed in the chapter about Back-

[1] In this research, we focus on online communication and do not explore how these two intelligences affect other important areas of life, such as work and family balance.

ground. With those research, we share our goal of fostering socially literate citizens in a technology-advanced society capable of navigating situations where different perspectives and opinions exist.

This research is also motivated by the well-known theory of the zone of proximal development (ZPD) [158]. ZPD has shown that the performance-competence gap often exists between functional and optimal skill applications, which suggests that people can improve their skills through scaffolding and practice. Improvable skills and abilities lie at the heart of this research on communication intelligence. In addition, our human beings are very rich and flexible by nature, and as a result, we are able to behave differently at different times. Because our behavior can vary in different situations, so can our ability to communicate effectively. Therefore the ability-based model of communication intelligence we develop in this research has the following property: *dynamic and situational (i.e.,context dependent)*. The dynamic of people's communication intelligence in different contexts and from time to time can provide behavior traces about people's improvement of their communication intelligence. In the rest of this dissertation, the terms "communication intelligence" and "contextual communication intelligence" will be used interchangeably. To understand the contexts of communication intelligence, we explain with an simple example. Let us assume that in an online forum setting, Ali contributes 16 posts while interacting with other participants in a thread. The context that influences Ali's application of certain communication skills (and therefore her communication intelligence) in a given post at any micro-moment is called *micro context*. A micro context includes all the posts up to that particular micro-moment. For example, post #1 through post #4 is the micro context associated with Ali's communication intelligence related to her post #5 in a thread. Given a certain context, we can compute Ali's contextual communication intelligence. For example, in the above example, we can compute Ali's communication intelligence associated with her post #5, with respect to a micro context. With respect to an *overall context* (i.e., the whole thread), Ali's contextual communication intelligence can be computed as the average of all the micro contextual communication intelligence. We use this way to compute peoples' communication intelligence in this research. In the future, we will consider a more sophisticated approach to computing participants' communication intelligence by taking account of the following factors: recency effect (i.e., time decay effect), the outcome of the communication if applicable, the level of satisfaction reported by other parties in the communication, and the context effect (i.e., negotiation vs. discussion). For example, we will design a weighted formula of these factors to compute participants' communication intelligence in an overall context.

Lastly, our theory about communication intelligence also benefits from studying various online contexts, including deliberation, negotiation, and inquiry-based learning, and analyzing online interaction experiences of people ranging from undergraduate students of multiple disciplines to highly-educated academic professionals and to members of the general public. These first-hand experiences allow us to see the different orientations in people: while some people are more relationship-oriented or people-oriented, others are more task-oriented or outcome-oriented [32]. Therefore, when we define our ability model of communication intelligence, we also attend to those two orientations.

3.2 Definition

- **Definition**: *Communication intelligence is an intellectual construct that supports intentional dialogue. It is composed of several abilities, including the ability to be mindful about one's own assumptions and emotions and examine them from an objective perspective, to attend to others' views and feelings and respond in a respectful manner, and to present rational ideas and evidence in order to move the conversation toward a meaningful direction.*

3.3 Constructs of Communication Intelligence

Any form of intelligence is challenging to measure and is often measured through quantitative scoring of constructs. Depending on whether assuming intelligence is innate or improvable, research in measuring human intelligence has mostly fallen into three main categories: trait-based model [129], ability-based model [139], or the combination of both [25]. Since a myriad of recent research about neuroplasticity and growth mindsets from neuroscience and psychology (as highlighted in [45]) has suggested that intelligence is improvable, in this research we will focus on creating an ability-based model of communication intelligence and measuring it with skill constructs.

Based on our own research, in particular [116], and many of others introduced in the previous section, we distill an intellectual model of communication intelligence comprising ten interrelated actionable dimensions, illustrated in Figure 3.1. These ten dimensions keep a good balance of acknowledging the different orientations among people. Loosely, the six people-oriented dimensions include connection, restraint, agreement, appreciation, self-reflection, and perspective taking; the four task-oriented dimensions include proof, monitoring, balance, and plan. In appendix, we also provide a chart showing how our communication intelligence model maps to the social deliberative skill framework.

Note that our intellectual model of communication intelligence includes the dimension *agreement* [2]. This is because we believe that agreement is an important intelligence-embodied skill for social communication – it requires attentive listening, analyzing and identifying the shared space in different minds, and acknowledging that shared opinions or feelings to build rapport and harmony in a dialogue. Indeed, the lack of agreement may imply a starting point of divergence in opinions.

[2] The skill of agreement falls outside of the category of social deliberative skills in Murray's framework.

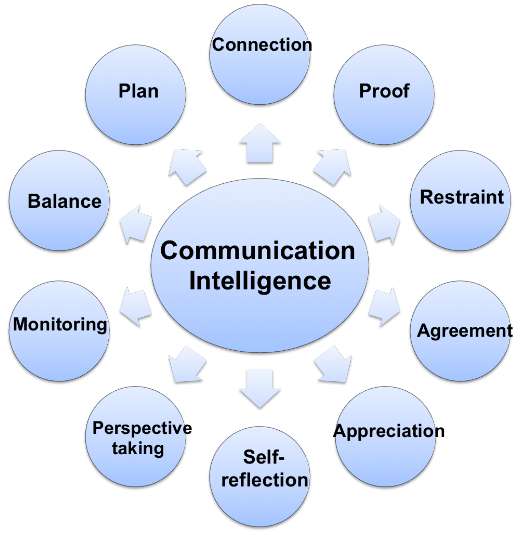

Figure 3.1. An overview of the constructs of communication intelligence

- Connection: connection with ideas of other participants.
- Proof: showing evidence by providing references or noting the source of statements.
- Restraint: controlling negative emotions toward other participants.
- Agreement: expressing agreement to other participants' viewpoints.
- Appreciation: showing appreciation to others' ideas or situations.
- Self-reflection: reflecting on one's own assumptions, values, biases, or emotions.
- Perspective taking: reflecting on the ideas or feelings of others (including participants in or outside of the dialogue).
- Monitoring: reflecting on the quality of discussions and/or suggesting changes.
- Balance: reflecting on the quality of discussed topics and/or weighing alternatives and identifying trade-offs.

- Plan: proposing or suggesting actions for resolving the discussed problem.

3.4 Computing Communication Intelligence

Measuring communication intelligence with respect to a set of skills becomes a manageable task that involves two steps: (1) identifying the set of skills, and (2) computing participants' communication intelligence based on the use of those skills. In Chapter 4 and Chapter 5, we will propose new models for automatically identifying intelligence-embodied communication skills. In those chapters, we will learn binary vectors of skill labels, in which the value of 0 or 1 represents the *absence* or *presence* use of a certain skill. In this section, we focus on computing communication intelligence by assuming that the skills labels are readily available.

Motivated by the theory of zone of proximal development, we first define an optimal state, or ideal state, of communication intelligence. *The ideal state of communication intelligence refers to the use of the whole set of intelligence-embodied skills invariant of contexts.* This definition is reasonable, because ideally, we hope to perform skillful communication regardless of circumstances, whether discussing a simple matter or a complicated issue, with an open-minded listener who respects your perspective or with a stubborn debater who competes with your position. This definition allows us to measure the communication intelligence of an individual in a *micro context* by computing the *set similarity* between the *ideal* set of skills and the *actual* set of skills employed by that participant in that context. As explained early, an individual's communication intelligence in an *overall context* (e.g., a thread) can be simply computed by taking an average of the communication intelligences in micro contexts that are part of the overall context. This definition also permits us to compare the communication intelligence of different individuals based on an overall context, which would be the absolute value of the difference between their communication intelligence with respect to the same overall context. It is worth noting that computing the score of contextual communication intelligence always involves evaluation against the ideal set of skills. Therefore, comparing an arbitrary set of skills with another is meaningless.

3.4.1 Similarity Measures

To compute communication intelligence, we need to measure the similarity between a performance skill vector and the optimal skill vector. Two widely-used similarity metrics for binary vectors are Hamming similarity [147] and Jaccard coefficient [95]. In the rest of this section, we provide definitions for each metric and use examples to illustrate how to use each measure to compute communication intelligence.

3.4.1.1 Hamming Similarity

From information theory, we know that the Hamming distance for 0 and 1 sequences of the same length is defined as the number of positions at which the corresponding symbols/attributes are different. Hamming similarity is then defined as the number of positions where corresponding attributes from two binary vectors agree. Formally, given two binary vectors x and y of the same length, the Hamming similarity between x and y is $f_{0,0} + f_{1,1}$, where $f_{0,1}$ denotes the number of attributes where x is 0 and y is 1; $f_{0,1}$ denotes the number of attributes where x is 0 and y is 1; $f_{0,0}$ denotes the number of attributes where x is 0 and y is 0; and $f_{1,1}$ denotes the number of attributes where x is 1 and y is 1;

We now show an example of computing communication intelligence using Hamming similarity. Suppose $x = 1111111111$, $y = 1110010110$, where x represents the optimal skill vector and y represents the performance skill vector of a single participant. The Hamming similarity in this case is $H = 0 + 6 = 6$. Therefore, the communication intelligence associated with this participant is 6 out of 10. When using Hamming distance, communication intelligence is lower bounded by 0 and upper bounded by 10.

3.4.1.2 Jaccard Coefficient

Jaccard coefficient or Jaccard similarity is an *asymmetric* similarity measure, which is often used in situations that the value of 1 or *the presence* is more important than that of 0 or absence. It computes the number of attributes where two binary vectors agree at 1. Formally, given two binary vectors x and y of same length, the Jaccard coefficient of x and y is $\frac{f_{1,1}}{f_{0,1} + f_{1,0} + f_{1,1}}$[3].

Using the same example from last section, where $x = 1111111111$, $y = 1110100110$, the Jaccard coefficient J is $\frac{6}{0+4+6} = 6/10$. Therefore, the communication intelligence of the participant whose performance skill vector is illustrated by y is 6/10. When using Jaccard coefficient, communication intelligence is lower bounded by 0 and upper bounded by 1.

By comparing Hamming similarity and Jaccard coefficient, we see that both metrics give the same measurement (i.e., 6 out of 10 is equivalent to 6/10). This is because the x vector representing the optimal skill set has a uniform distribution of 1. In this dissertation, we use Jaccard coefficient to measure communication intelligence. This is because if theories about which skills are more valued than others for measuring communication intelligence are developed in the future, we can switch to use *weighted* Jaccard coefficient, where weights are added to take account of the importance of each attribute in the set.

Before closing this chapter, we want to remark that creating a new measure for communication intelligence and then validating it through repeated tests would have profound implications for increasing production in work groups and teams [4] and building a respectful and deliberative society at large.

[4] Recent research from the MIT media lab [167] has verified a new formula for successful working teams: Smart effective teams = people willing to listen and empathize + people able to perceive and respond to others emotions.

CHAPTER 4

MULTI-LABEL LEARNING WITH CONSTRAINT LABELED LDA

In this chapter, we present a new hierarchical probabilistic model for addressing the problem of identifying multiple intelligence-embodied communication skills from natural language. This model reveals the language manifestation of intelligence-embodied communication skills and can support large-scale computational annotation on these high-order skills.

4.1 Motivation and Related Work

Language is to online communication as a seed to a plant – it is the core to understanding communication phenomena in an online environment. Intensive research in anthropology, sociology, linguistic, and communication has studied a wide variety of social phenomena, including leadership [107], power [84], conflicts [152], deception [68], and perspective-taking [74]. This previous research typically looks at language use in situations, where social relationships are known, rather than using language *predictively*.

There is also a body of literature in computational linguistic that uses a two-tier framework for modeling social phenomena. This line of research first creates computational methods for detecting *social language cues*, such as on-topic discussion and involvement, and then uses these language cues to infer high-order social constructs, such as influence and conflicts [151]. Little research has attempted to identify *multiple* social and communicational phenomena, including perspective-taking, monitoring, and balance, both *directly* and *simultaneously* from natural language. In this chapter, we introduce a novel language model for identifying intelligence-embodied communication skills within online communication.

As we learned from the chapter on Background, generative models have great advantages over discriminative models for the task of multi-lable text classification. An extension of LDA, called Labeled LDA, seems to be a plausible choice for identifying high-order communication skills. Labeled LDA is specifically designed for multi-label classification problems, and, as a generative model, it learns an associative mapping between words and document labels that can help understand how language reflects the use of each intelligence-embodied skill. However, Labeled LDA is based on the bag-of-words assumption and is limited in the current task, where *credit attribution* between labels and sentences is desired. For example, human annotators would label the following text from an authentic online interaction as *perspective taking* only after processing the whole sentence.

> *"From both of you I have now a little insight into how you view this problem and what the problem solution could be."*

This observation suggests that we need a model, which, similar to Labeled LDA, but *takes account of word order within a sentence.* As true in many real-world problems, a sentence is the base unit that carries a *meaning*. Such a model, once developed, would find wide applications.

In the literature, two general solutions have been proposed to address the limitation of topic models that the positions of individual words are neglected for inference (i.e., the bag-of-words assumption): (1) modeling the word order (i.e., n-grams) explicitly [159, 160, 64], and (2) imposing constraints to capture the word occurrence in proximity [81]. The former approach mainly relies on extending the LDA model by adding more variables (i.e., indicator variables for each word are used to signify if a bigram should be generated to form an n-gram.) One major drawback of these models is the increased model complexity and therefore decreased computational efficiency in inference. In this chapter, we follow the second approach and present a new model, called constrained Labeled LDA (CL-LDA), which adapts Labeled LDA to impose the constraint that *all words in a sentence are generated from one label*.

Before we formally describe the generative process of CL-LDA, we show the list of research questions we can address with CL-LDA.

- **Research questions:**
  - How can the use of multiple intelligence-embodied skills be automatically and simultanously identified?
  - How does language (at the sentence level) reflect the use of each intelligence-embodied communication skill, respectively?

4.2 Constrained Labeled LDA (CL-LDA)

In this section, we first show the graphical model of CL-LDA (Figure 4.1) and its generative process, and then describe the Gibbs sampling inference for CL-LDA.

CL-LDA assumes the following stochastic process of writing messages on social media sites.

1. For every topic ϕ out of K,

   (a) Draw a word distribution $\phi_k \backsim$ Dirichlet (β).

2. For each message d out of D,

   (a) For each topic ϕ out of K,

      (i) Draw topic presence/absence indicator $\Lambda \in \{ 0, 1 \} \backsim$ Bernoulli (η).

   (b) Project the Dirichlet prior vector into lower dimensions $\alpha = L$ x α.

   (c) Draw a topic proportion $\theta \backsim$ Dirichlet (α).

   (d) For each sentence m out of M,

      (i) Draw a topic assignment $Z_{d,m} \backsim$ Multinomial (θ_d).

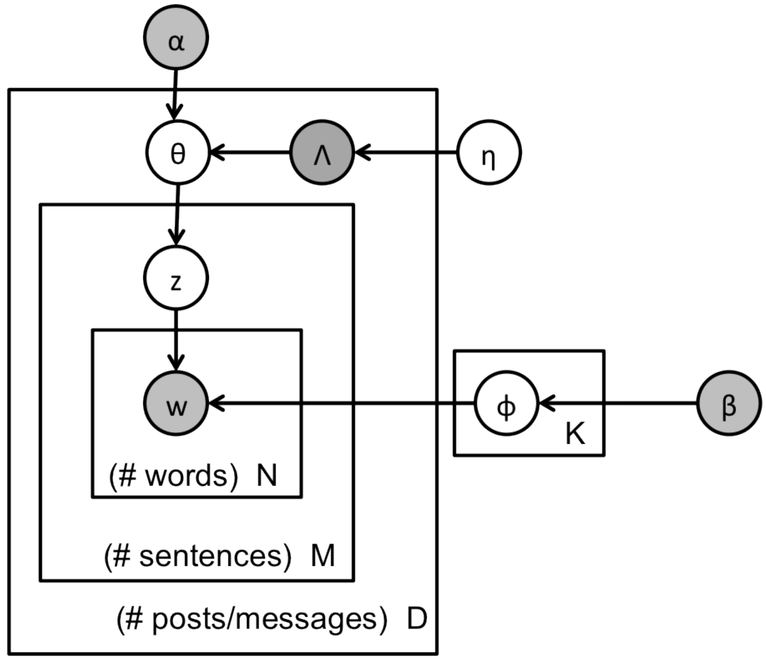

Figure 4.1. The graphical model of constrained Labeled LDA

(ii) For each word n out of N,

(1) Draw a word $W_{d,m,n} \backsim$ Multinomial $(\phi_{Z_{d,m,n}})$.

As can be seen in the graphical model in Figure 4.1, CL-LDA implements the word-order constraint by *adding a sentence-level plate* to the Labeled LDA model. With this modification, the hidden topic variable z now resides in the sentence plate, and thereby all words in a sentence are assumed to be generated from one label. Since the labels are observed, as in LDA, the labeling prior is D-separated from the rest of the model conditioned on the labels, so that we can use collapsed Gibbs sampling to derive the posterior distribution of labels.

During training, CL-LDA adds the restriction that all words in a sentence can only be assigned to the observed labels of the text. At test, CL-LDA infers all the labels associated with each text. Inferring multiple labels of a text is much more challenging than inferring a single label. It involves evaluating all combinations of label assignments, 2^k in total (k is the number of labels). This problem is amplified by the fact that, for each assignment, we need to construct a low-dimension Dirichlet prior vector α (using the projection matrix L) to restrict the model to use only those topics that correspond to document labels. As in Labeled LDA, in this research, we made a simplifying assumption that model inference reduces to standard LDA inference. This is a reasonable assumption, because exploring the whole topic space is in essence similar to exploring all the possible label assignments.

4.2.1 Inference Using Gibbs Sampling

We used collapsed Gibbs sampling [63] to estimate the posterior distribution of hidden variable $\mathbf{z}$ given the input variables $\mathbf{w}$, $\mathbf{\Lambda}$, and hyperparameters, α and β, and then use the results to infer model parameters θ and ϕ. (In the appendix, we detail the derivation of Gibbs sampling for CL-LDA.)

$$P(\theta, \phi, z|w, \Lambda) = \frac{P(\theta, \phi, z, w, \Lambda)}{P(w, \Lambda)}$$

Using Gibbs sampling, we constructed a Markov chain that converges to the posterior distribution on z. The transition between successive states of the Markov chain is achieved by randomly sampling z from its distribution conditioned on all other variables, summing out θ and ϕ. By derivation (shown in the appendix), we get:

$$P(z_i|\mathbf{z_{-i}}, w, \Lambda) \propto \frac{N_{k|d} + \alpha_k}{N_d + \sum_k \alpha_k} \cdot \frac{\Gamma(N_k + \sum_w \beta_w)}{\Gamma(N_k + s_i + \sum_w \beta_w)} \cdot \prod_w \frac{\Gamma(N_{w|k} + s_{wi} + \beta_w)}{\Gamma(N_{w|k} + \beta_w)}$$

where the subscript z_{-i} denotes all topic assignments excluding the ith sentence, $N_{k|d}$ is the number of times that topic k is assigned to message d, excluding the current sentence, $N_{w|k}$ is the number of times that topic k is assigned to word w, excluding the current sentence, s_{wi} is the number of word w in sentence i, and s_i is the number of total words in sentence i.

After the Gibbs sampling process, the model parameters in CL-LDA can be obtained as follows:

$$\phi_{w|k} = \frac{N_{w|k} + \beta_w}{N_k + V\beta_w}$$

$$\theta_{k|d} = \frac{N_{k|d} + \alpha_k}{N_d + K\alpha_k}$$

where $\phi_{w|k}$ is the probability of using word w in topic k, and $\theta_{k|d}$ is the probability of using topic k in message d.

4.2.2 Gibbs Query Sampling for Unseen Data

To estimate the labels (i.e., topics) in unseen data, we need to derive a Gibbs query sampler. In order to find the required counts for a previously unseen data, we follow the approach in [73] to run the inference algorithm on the new data exclusively. Specifically, we first initialize the algorithm by randomly assigning topics to sentences and then perform a small number of iterations through the Gibbs sampling update. For such an inference, the Gibbs query sample takes the following form:

$$P(\tilde{z}_i|\tilde{\mathbf{z_{-i}}}, \tilde{w}, \Lambda) \propto \frac{N_{k|d} + \widetilde{\alpha_k}}{N_d + \sum_k \widetilde{\alpha_k}} \cdot \frac{\Gamma(N_k + \widetilde{N_k} + \sum_w \widetilde{\beta_w})}{\Gamma(N_k + \widetilde{N_k} + \widetilde{s_i} + \sum_w \widetilde{\beta_w})} \cdot \prod_w \frac{\Gamma(N_{w|k} + \widetilde{N_{w|k}} + \widetilde{s_{wi}} + \widetilde{\beta_w})}{\Gamma(N_{w|k} + \widetilde{N_{w|k}} + \widetilde{\beta_w})}$$

where $\widetilde{[*]}$ denote the corresponding quantities in the test corpus and $\sum_w \widetilde{\beta_w}$ counts the words appeared in both training and testing vocabularies.

4.3 Corpora

As a part of a larger research endeavor, we collected online dialogues from a variety of online contexts, including deliberation, negotiation, and inquiry-based learning. In this research, we examined two online corpora: one involving participants in negotiation and another involving participants in an open discussion. Since these two corpora are from two domains, each corresponding to a different online context, we may use the term “domain” to refer “online context” in the text below unless otherwise specified.

In the first domain, professional community negotiation, 72 email exchanges from a faculty listserv with geographically dispersed participants were gathered. Sixteen faculty members from two academic communities negotiated about a proper solution to a conference scheduling conflict. An emerging theme in this dialogue was the tension between democratic decision making versus fiat decision making by those in authoritative roles. Participants were highly educated academic professionals and most of them encouraged democratic decision making about relocating the conference.

In the second domain, civic deliberation, 51 posts were collected from a civic engagement online discussion forum at e-deomocracy.org. Thirty two participants discussed ethnic issues suggesting ways to alleviate tensions about their multi-racial community. This discussion was in response to a post describing negative incidents they believed occurred because of their race, being black in the predominantly white, upper-middle class neighborhood. Several participants attempted to be consoling and supportive, others attempted to frame the social characteristics of the neighborhood in a wider political context, and yet others reacted against this imposed political context.

Two independent trained human judges annotated the two corpora based on Murray’s theory about social deliberative skills [116], which is a hierarchical conceptual coding framework containing over 50 social deliberative skills and other speech acts.

Messages were first segmented manually at speech act boundaries before annotation. The inter-rater reliability of human annotations was 69% (measured in Cohen's Kappa) for the professional communication negotiations and 76% for the civic deliberation discussions. According to [91], the agreement level between human judges was good. For the purpose of this study, we aggregated appropriate social deliberative skills to construct each intelligence-embodied communication skill. We used 0 and 1 to denote the absence and presence of the use of each skill, respectively.

4.4 Experimental Design

4.4.1 Data Preparation

CL-LDA is a supervised machine learning model. For the task of identifying intelligence-embodied communication skills, we need data for training and testing from each domain. In this research, we split data according to the following three principles.

- The ratio of training to testing data is 4:1;
- Each skill label has representations in both training and testing sets;
- For both training and testing set, the number of positive skill labels (i.e., the "1s" in the label matrix) ranges from 2 to 8 (where 2 is the smallest and 8 is the largest number of annotated labels in the actual data).

This last principle was specifically designed with the goal of studying CL-LDA's prediction performance on data whose number of positive labels is across a relatively wide range. In the literature, the majority of studies have focused on corpora with either relatively few labels or many examples of each label [96]. With this study, we wish to study the robustness of CL-LDA in the face of the number of labels in a relatively wide range and with a small number of training data available for each label.

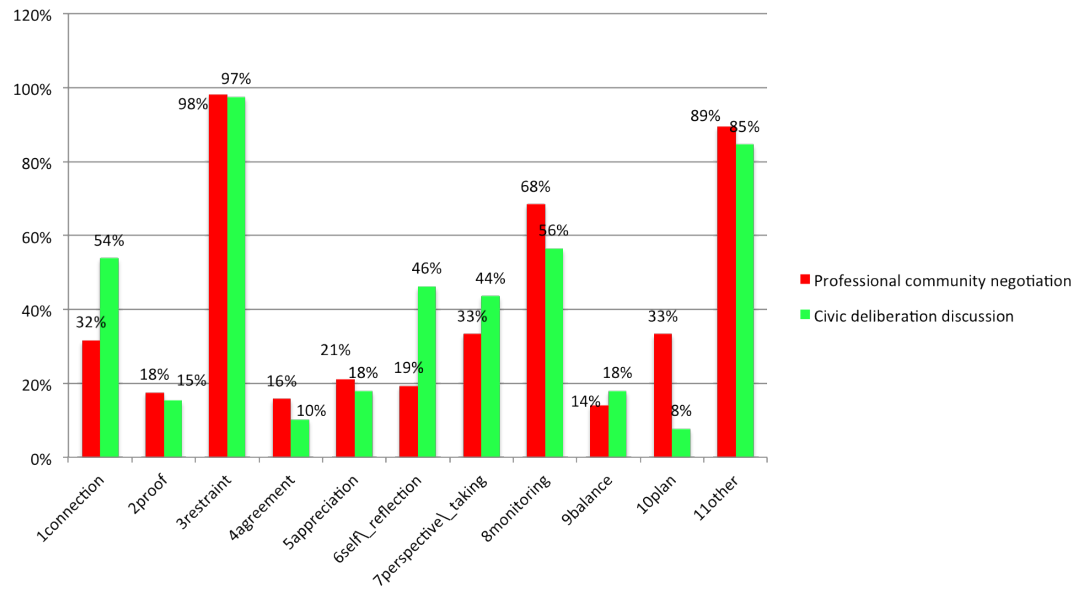

Figure 4.2. An illustration of the training data class distributions in different domains

In the experiments that followed, we have 57 training and 15 testing instances in the professional community negotiation domain. In the civic deliberation discussion domain, we have 39 training and 12 testing instances.

We show in Figure 4.2 the statistics of class distribution in the training data for both domains. Note that in the professional community negotiation domain, the majority classes are *restraint*, *monitoring*, and *other* [1]. In the civic deliberation discussion domain, the majority classes include *connection*, *restraint*, *monitoring*, and *other*. These numbers will be referenced later when we study the prediction performance of some comparison methods whose predictions in unseen data rely heavily on this prior information from the training data.

[1] "Other" includes all other annotated labels that are not related to any of the intelligence-embodied skills.

4.4.2 Data Preprocessing

In order to prepare data for CL-LDA, we first preprocessed the data by splitting data messages into sentences using Stanford Natural Language toolkit (NLTK 2.0) [2]. We realized that function words (e.g., would, but) play an integral role in statements showing the use of intelligence-embodied skills (e.g., self-reflection, perspective taking). Therefore, we did not filter standard English stop words completely. Instead, we set up a threshold to prune words that appear more than 110 times (e.g., a, the) or less than 3 times. We applied the Porter stemmer [3] algorithm on unigram features in this study. In the professional community negotiation domain, we had 13,714 words from 72 messages that contain 741 sentences. In the civic deliberation discussion domain, we had 17,810 words from 51 messages that contain 934 sentences. The training vocabulary in the professional community negotiation domain was 538 and that in the civic deliberation discussion domain was 719.

4.4.3 Parameter Configurations

In CL-LDA, the number of topics (i.e., 11 labels: ten skill labels, one "other") was given. At training, the Dirichlet prior β was set to 0.01. The hyper-parameter η that specifies the total weight contributed by the labels was set to 20. The Dirichlet prior α was projected from the topic dimension (i.e., 11) into a low dimension of the size of the positive labels M (i.e., labels that are present). In this way, the supervision is incorporated so that each message takes on the topics that correspond to the message's positive labels. We run Gibbs sampler with 1000 burn-in iterations and 1000 sampling iterations.

At test, the Dirichlet prior β and α were set to 0.01 and 4, respectively. We used symmetric Dirichlet prior α because we want to challenge CL-LDA to see its ability to

[2]http://nltk.org

[3]http://tartarus.org/martin/PorterStemmer/

identify those communication skills *without being given any prior information about observed frequencies of different skill labels from the training data.* We ran Gibbs sampling with a small number of times (i.e., 100 burn-in iterations and 100 sampling iterations) before we collected the results.

For both domains, we used the same set of hyper-parameters. Note that these hyper-parameters were chosen heuristically and were not optimized. Therefore, we would expect that with hyper parameter optimization, at least a modest improvement in performance over the presented results could be obtained.

4.4.4 Data Postprocessing

Inferring message labels in supervised topic models often involves thresholding probabilities of the label-message distribution. Choosing a threshold-selection method is non-trivial and is a research problem in and of itself [175, 55]. In the literature, there are two main rank-based cutoff approaches to thresholding: *proportional method* and *calibrated method.* The first one sets the cut-off number N_i based on training data frequencies. For example, suppose that 28 training data are assigned label C_i and there are 40 training data and 10 testing data in total, the cut-off number for label C_i will be $28 * \frac{10}{40} = 7$, meaning that we select the top 7 labels with the highest probability over topics from the label-message distribution. The second approach, calibrated method, sets the cut-off number equal to the true number of positive instances in the testing data. In other words, if the testing data has 10 positive labels, then the top 10 labels with the highest probability over topics would be selected as predictions for that data. Note that both methods use prior information – the proportional method uses prior information from the training data; whereas the calibrated method uses an even stronger prior information from the testing data. In this research, we will use the calibrated method to infer labels from a comparison model – labeled-LDA. *We* *equip Labeled-LDA with this strong bias to see what the best this boosted version of labeled-LDA can perform to identify intelligence-embodied communication skills.*

Note that for CL-LDA, we do not need to select a threshold. This is because CL-LDA learns the topic assignment for each sentence. It is practical and easy to obtain the labels for each data by employing a simple aggregation method over sentence labels. For example, a message is determined to have a particular label, only if its containing sentences are assigned that label. In this study, we used this simple method to construct the 0/1 label matrix for each testing message.

4.5 Results and Discussions

In this section, we will evaluate the learning performance of CL-LDA for identifying multiple intelligence-embodied communication skills. Specifically, we will first compare the prediction performance of CL-LDA against state-of-the-art multi-label text classification methods: Labeled LDA and a set of multiple one-vs-rest SVM classifiers. We then evaluate the quality of the most salient words that represent each skill using topic coherence metric. Finally, we illustrate CL-LDA's performance on credit attribution – assigning skills labels to sentences.

4.5.1 Multi-label Text Classification

4.5.1.1 Category-pivoted Evaluations

In this section, we focus on category-pivoted evaluations – the evaluation of model performance on predicting each skill label. We evaluated model performance quantitatively in terms of sensitivity (the true positive rate), specificity (the true negative rate), and accuracy. Both sensitivity and specificity are valued in this research, because for the purpose of measuring communication intelligence, the *presence* and *absence* use of skills are equally important to identify. As shown in Table 4.1, in the professional community negotiation domain, CL-LDA achieves the best average predic-

Table 4.1: Category-pivoted evaluations in the professional community negotiation domain: A comparison of SVM, Labeled LDA+Calibrated-labels, and CL-LDA

	SVM			**Labeled LDA+Calibrated-labels**			**CL-LDA**		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
1connection	0.00	1.00	0.53	0.14	0.88	0.53	0.71	0.88	0.80
2proof	0.00	1.00	0.80	1.00	0.50	0.47	1.00	0.58	0.67
3restraint	1.00	0.00	0.93	0.71	0.00	0.67	0.57	1.00	0.60
4agreement	0.00	1.00	0.73	0.75	0.73	0.73	0.75	0.45	0.53
5appreciation	0.00	1.00	0.80	0.67	0.50	0.53	0.67	0.67	0.67
6self_reflection	0.00	1.00	0.73	0.50	0.45	0.47	0.50	0.73	0.67
7perspective_taking	0.00	1.00	0.53	0.57	0.75	0.67	0.86	0.88	0.87
8monitoring	1.00	0.00	0.60	0.44	0.67	0.53	0.78	0.67	0.73
9balance	0.00	1.00	0.80	0.33	0.58	0.53	0.67	0.71	0.47
10plan	0.00	1.00	0.67	0.00	0.90	0.60	0.40	0.30	0.33
11other	1.00	0.00	0.87	0.54	1.00	0.60	0.62	0.50	0.60
Min	0.00	0.00	0.53	0.00	0.00	0.47	0.40	0.30	0.33
Max	1.00	1.00	0.93	1.00	1.00	0.73	1.00	1.00	0.87
Avg	0.27	0.73	**0.73**	0.51	0.63	**0.58**	0.68	0.67	**0.63**
Avg (sen+spe)		**0.50**			**0.57**			**0.68**	

tion sensitivity (68%) across all 11 categories, followed by Labeled LDA+calibrated-labels [4](51%.) and SVM (27%). Note that for SVM, different kernels were tested, including linear, polynomial, sigmoid, and radial basis function. Linear kernels yielded the best performance for both domains. As to average prediction specificity, CL-LDA (67%) achieves the second-best followed by Labeled LDA+calibrated-labels (63%). With further examinations, we realized that SVM's good performance on average prediction specificity and accuracy is attributed to the high-skewed class distribution. In other words, SVM predicted all data as coming from the majority class, which confirms the findings from the literature that we suryeved in the chapter on Background. For example, as shown in Figure 4.2, in the professional community negotiation domain, the majority classes are *restraint*, *monitoring*, and *other*. Therefore, only for those 3 classes, SVM achieves 100% sensitivity, and for other classes, it achieves 0% sensitivity. Since SVM is heavily biased by class distribution, we only compare CL-LDA with Labeled LDA+calibrated-labels for average prediction accuracy. CL-LDA achieves 63% (compared to the 9% baseline), whereas Labeled LDA+calibrated-labels achieves 58% in average prediction accuracy. This is an im-

[4]Because the labeled-LDA uses the calibrated method to obtain its labels for the testing data, we refer it as "LLDA+Calibrated-labels."

Table 4.2: Category-pivoted evaluations in the civic deliberation discussion domain: A comparison of SVM, Labeled LDA+Calibrated-labels, and CL-LDA

	SVM			**Labeled LDA+Calibrated-labels**			**CL-LDA**		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
1connection	1.00	0.00	0.33	1.00	0.00	0.33	0.75	0.50	0.58
2proof	0.00	1.00	0.67	1.00	0.50	0.67	0.75	0.63	0.67
3restraint	1.00	0.00	0.92	1.00	0.00	0.92	0.45	1.00	0.50
4agreement	0.00	1.00	0.83	0.50	0.80	0.75	0.50	0.20	0.25
5appreciation	0.00	1.00	0.83	0.50	0.70	0.67	0.50	0.40	0.42
6self_reflection	0.00	1.00	0.58	0.40	1.00	0.75	1.00	0.43	0.67
7perspective_taking	0.00	1.00	0.58	0.80	0.71	0.75	0.80	0.71	0.75
8monitoring	1.00	0.00	0.50	0.17	1.00	0.58	0.67	0.67	0.67
9balance	0.00	1.00	0.83	0.00	0.80	0.67	1.00	0.50	0.58
10plan	0.00	1.00	0.92	0.00	0.91	0.83	0.00	0.64	0.58
11other	1.00	0.00	0.92	0.18	1.00	0.25	0.82	0.00	0.75
Min	0.00	0.00	0.33	0.00	0.00	0.25	0.00	0.00	0.25
Max	1.00	1.00	0.92	1.00	1.00	0.92	1.00	1.00	0.75
Avg	0.36	0.64	**0.72**	0.50	0.67	**0.65**	0.66	0.52	**0.58**
Avg (sen+spe)	**0.50**			**0.59**			**0.59**		

pressive result, given that Labeled LDA+calibrated-labels has supervision from the test phase (through thresholding), while CL-LDA receives no guidance from the testing data. Moreover, CL-LDA does not use observed frequencies of skill labels from training data either.

As can also be seen from Table 4.1, CL-LDA achieves the highest sensitivity (i.e., the true positive rate) (100%) in predicting the skill of *proof* and the lowest sensitivity (40%) in predicting the skill of *plan*. CL-LDA achieves the highest specificity (i.e., the true negative rate) (100%) in predicting the skill of *restraint* and the lowest specificity (30%) in predicting the skill of *plan*. Please be cautious that these observations can not be used as evidence to conclude which intelligence-embodied skills are easy or hard to predict automatically from natural language. This is because these results are tied to a particular online context (e.g., negotiation vs. discussion) that we study in this research. They are also influenced by the number of training and testing data available in each skill/category that is constrained by the three experiment design principles introduced early. This statement also applies to similar observations in the following experiments in this chapter.

Now, let us look at the civic deliberation discussion domain. As can be seen in Table 4.2, in the professional community negotiation, CL-LDA achieves the best average prediction sensitivity (66%) across all 11 categories, followed by Labeled LDA+calibrated-labels (50%) and SVM (36%). We once again observed that SVM's prediction performance highly reflects the bias of class distribution. As a result, we ignore SVM for performance comparisons in the rest of this section. We also observed that Labeled LDA+calibrated-labels outperforms CL-LDA in average prediction specificity and accuracy and we hypothesize that the assumption of CL-LDA is violated in this domain. In other words, sentences in this domain are relatively long, so that all words in a sentence might be generated from more than one label. We tested this hypothesis by computing the average number of words in each domain. We found that the professional community negotiation domain has an average of 18 words per sentence and the civic deliberation discussion domain has an average of 19 words per sentence, which does not support this hypothesis. Furthermore, if this assumption is violated, CL-LDA would not achieve good prediction sensitivity. Another hypothesis is that the better performance of Labeled LDA+calibrated-label in average specificity (i.e., the true negative rate) is largely attributed to the extra supervision from the thresholding method it uses. Specifically, the calibrated labels (i.e., the true number of positive instances for the test data) set the upper-bound for the number of labels to be selected for each test data, which in turn guarantee a certain number of negative predictions. In contrast, CL-LDA – without supervision in the test phase – might assign each sentence in a message with a different label, leading to most positive predictions.

As can also be seen from Table 4.2, in the civic deliberation discussion domain, CL-LDA achieves the highest sensitivity in predicting the skills of *self-reflection* and *balance* and the lowest sensitivity (45%) in predicting the skill of *restraint*. CL-LDA achieves the highest specificity (100%) in predicting the skill of *restraint* and the lowest specificity (0%) in predicting the skill of *others*.

Table 4.3: Message-pivoted evaluations in the professional community negotiation domain: A comparison of SVM, Labeled LDA+Calibrated-labels, and CL-LDA

	SVM			**Labeled LDA+Calibrated-labels**			**CL-LDA**		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Min	0.29	0.75	0.45	0.00	0.00	0.27	0.00	0.00	0.36
Max	0.75	1.00	0.91	1.00	1.00	1.00	1.00	0.89	0.91
Avg	0.52	0.92	**0.73**	0.49	0.57	**0.58**	0.57	0.54	**0.63**
Avg (sen+spe)	**0.72**			**0.53**			**0.56**		

Comparing model performance across domains, we found that all models have lower performance in the civic deliberation discussion domain than in the professional community negotiation domain. The lower model performance might be attributed to fewer training instances (41% less) in the civic deliberation discussion domain.

4.5.1.2 Message-pivoted Evaluations

Category-pivoted evaluations allow us to study a model's predicability for each skill label *separately*. When multi-category classification is concerned, message-pivoted evaluations provide a holistic view on a model's predicability of all skill labels associated with a message.

As shown in Table 4.3, in the professional community negotiation domain, CL-LDA achieves the best average prediction sensitivity (57%) across all the messages in the testing set, followed by SVM (52%) and Labeled LDA+Calibrated-labels (49%). For the same reason that SVM is highly biased by the class distribution, we ignore it in the rest of this section. As to the average prediction specificity, Labeled LDA+Calibrated-labels (57%) outperforms CL-LDA (54%) by 3%. For average prediction accuracy, CL-LDA (63%) has an upper hand and outperforms Labeled LDA+Calibrated-labels by 5%.

In the civic deliberation discussion domain, as shown in Table 4.4, we found the same pattern as in category-pivoted evaluations. CL-LDA outperforms Labeled LDA+Calibrated-labels on prediction specificity, while Labeled LDA+ Calibrated-labels has an upper hand on prediction specificity and accuracy.

Table 4.4: Message-pivoted evaluations in the civic deliberation discussion domain: A comparison of SVM, Labeled LDA+Calibrated-labels, and CL-LDA

	SVM			**Labeled LDA+Calibrated-labels**			**CL-LDA**		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Min	0.33	0.60	0.45	0.25	0.40	0.45	0.00	0.00	0.36
Max	1.00	1.00	0.91	0.88	0.89	0.82	1.00	0.89	0.91
Avg	0.67	0.80	**0.72**	0.53	0.69	**0.65**	0.61	0.49	**0.58**
Avg (sen+spe)	**0.74**			**0.61**			**0.55**		

4.5.1.2.1 The Relationship Between the Number of Positive Labels per Message and Model's Performance When designing experiments, we ensured that, for data in both training and testing sets, the number of positive instance of skill labels spans a spectrum. In doing so, we can study the relationship between a model's prediction performance and the number of positive labels the data has. For example, we ask is CL-LDA more likely to have better prediction when a data has fewer labels or vice versa?

As can be seen in Figure 4.3, in the professional negotiation domain, a statistically significant ($p=0.0005$) positive correlation exists between the number of positive labels per message and CL-LDA's prediction sensitivity. A negative statistically significant ($p = 0.0091$) correlation is also found between the number of positive labels per message and CL-LDA's prediction specificity. No statistically significant relationship exists between the number of positive labels per message and CL-LDA's prediction accuracy.

In the civic deliberation discussion domain, as shown in Figure 4.4, we only observed a positive statistically significant correlation ($p=0.0278$) between the number of positive labels per message and CL-LDA's prediction sensitivity. No statistically significant correlation is found between the number of positive labels per message and other prediction measures.

These observations imply that despite the difficulty of predicting multiple labels simultaneously, CL-LDA's performance increases as the prediction task becomes more challenging (i.e., predicting a large number of labels). It suggests that CL-LDA can

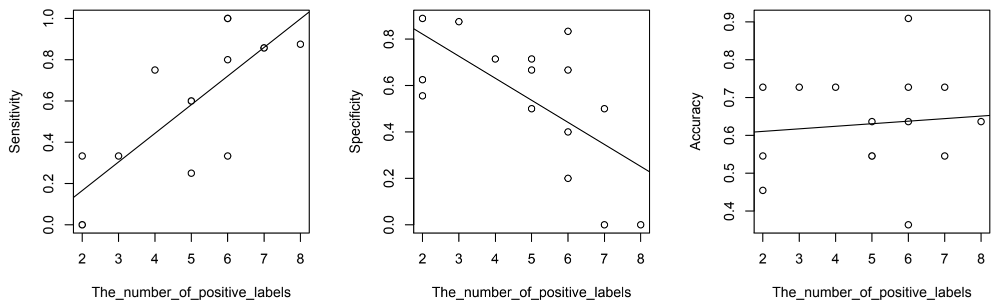

Figure 4.3. The relationship between the prediction performance of CL-LDA and the number of positive labels per message in the professional negotiation domain

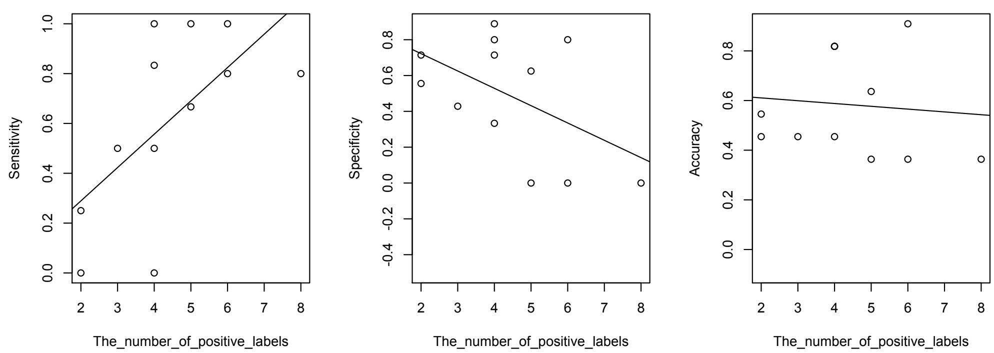

Figure 4.4. The relationship between the prediction performance of CL-LDA and the number of positive labels per message in the civic deliberation discussion domain

be a promising multi-label text classification technique for large-scale applications. This observation also suggests that there might exist an inter-dependency between labels that contribute to the performance gain in predictions.

4.5.2 Word and Sentence Discovery

CL-LDA is a supervised machine learning model. During training, CL-LDA learns label-specific word distributions under the constraint that words in a sentence can only take observed labels of the text. Our experiments showed that function words

Table 4.5: Coherence scores of learned topics using the 5 most salient words in the professional community negotiation domain

Topics	Scores	5 Most Salient Words
1connection	-64.6	confer, would, at, not, or
2proof	-60.8	but, discuss, at, not, are
3restraint	-61.3	on, communiti, discuss, would, not
4agreement	-59.9	would, agre, tri, on, but
5appreciation	-57.3	veri, peopl, would, but, communiti
6self_reflection	-64	not, discuss, on, communiti, confer
7perspective_taking	-37.8	confer, blueorg, not, as, some
8monitoring	-62.9	not, would, are, veri, at
9balance	-65.7	confer, or, will, there, communiti
10plan	-51.5	would, not, you, confer, do
11other	-58	would, as, at, confer, discuss

(e.g., would, but) play an important role in statements/messages showing the use of intelligence-embodied skills (e.g., self-reflection, perspective taking). Because of this reason, it would be hard to evaluate the coherence of each topic (label) qualitatively. For example, a coherent topic "brain" may contain words like neurons, neuronal, brain, axon, neuron, nervous system, which our minds can visualize. However, a coherence topic "self-reflection" containing function words is hard to imagine.

In our previous research [173], we used a quantitative approach – topic coherence metric – to evaluate the quality of learned topics. Topic coherence [112] is based on the assumption that pairs of words belonging to a single topic will co-occur within a single document, whereas word pairs belonging to different topics will not. This assumption is violated in this research, where each text message can have multiple equally appropriate labels. Nevertheless, for lack of a better evaluation method, in Table 4.5 and Table 4.6, we show the 5 most salient words and topic coherence score for each topic. Numbers closer to zero indicate higher coherence. To put the coherence scores in context, we use the following example. In an online dispute domain, we

Table 4.6: Coherence scores of learned topics using the 5 most salient words in the civic deliberation discussion domain

Topic	Scores	5 Most Salient Words
1connection	-58.4	be, as, was, are, or
2proof	-56.1	you, were, can, if, so
3restraint	-64.3	for, my, you, as, be
4agreement	-57.5	was, peopl, not, white, my
5appreciation	-59.5	was, neighborhood, my, on, peopl
6self_reflection	-61	have, be, you, for, with
7perspective_taking	-62.4	are, we, about, all, was
8monitoring	-54.9	you, white, for, we, was
9balance	-54.9	are, from, feel, there, vigil
10plan	-64.2	are, differ, with, have, white
11other	-63.5	as, for, have, about, with

developed a variant of LDA that learned a topic about vehicle transaction on eBay with a coherence score of -58.0, where the top 5 topic words were car, vehicle, seller, buyer, and state.

Because CL-LDA learns a label for each sentence, in the section below, we evaluate the quality of each learned topic (label) with annotated sentences.

4.5.2.1 Credit Attribution – Sentence Discovery

Credit attribution in the context of supervised language models often refers to the ability to associate individual words in a text with their most appropriate labels. To better understand the language characteristics of intelligence-embodied skills, the associations between sentences (vs. words) in a message and their most appropriate labels are desired. CL-LDA's ability to annotate sentences in online text manifests its extraordinary model interpretability.

In Table 4.7 and Table 4.8, we demonstrate that CL-LDA can effectively model the sentence-label associations of online text with multiply labels. We show three examples of learned sentences for each skill label. Those annotated sentences augment the view of each online message with rich contextual information. It is worth noting that, in the supervision phase, CL-LDA is only given label information *at the message* *level* rather than at the sentence level. In other words, at the training stage, CL-LDA still needs to solve the problem of associative mapping between m observed labels and each of the n sentences in the text.

Table 4.7: Examples of learned sentences by CL-LDA for each intelligence-embodied skill in the professional community negotiation domain

Label	**Sentence**
1connection	From some of the emails in this chain from those in the know it sounds like the leadership at REDorg is not interested in co-locating with BLUEconf. However, it sounds like there are some remaining points of confusion. As an aside, the idea of a co-sponsored track at REDorg seems non-controversial.
2proof	This morning I received your email (see below) indicating that BLUEconf is proceeding with BLUEconf as previously planned. 2011 CFP on your institutions internal email lists and other email lists that you know but are not reached by our publicity chairs. I say apparent, as there has been no vote, although one has been suggested.
3restraint	However, it sounds like there are some remaining points of confusion. I repeat my encouragement for a discussion with Silas F. about our reasons for shifting BLUEconf 2011 to early 2012.

	Further, the issue is not that decisions are made behind closed doors.
4agreement	I also think that bringing this to the BLUEconf community for discussion vote helps build our community, lets the leadership respond to the communitys desires, and shows our good will towards REDorg win-win on all counts. I trust Larry G. in the way he is proceeding to collect data while minimize long iterations and clogging mailboxes. The consultation process is still taking place and Larry G. is trying to find a compromise for which most will be satisfied without jeopardizing BLUEconf.
5appreciation	Larry G. has been busy around the clock sending messages to groups and to individuals as well making numerous phone calls. Each meeting will draw on (largely) different groups of people with (largely) different backgrounds. Having been to many REDorg meetings, I think that BLUEconf can will fill a niche and need that does complement REDorg.
6self_reflection	Particularly, if everybody insists on running BLUEconf in the same way as the conference of hisher own area, then there is no way to run BLUEconf. I've been really happy to hear the opinions that have been shared so far.

	That's why each organization has an executive committee to collect information and does internal discussion to figure out what is the best option for that organization by considering all the considerations.
7perspective_taking	Patricia W. is very experienced and will know what can be done (and how), whereas Sally K. will provide the driving force and ideas to SEE what can be done. Perhaps a vote will alter the options, or maybe the BLUEconf community as represented by us will disagree with what I have said. As I understand BLUEorgs work with FocusGroups, they would fully understand our decision, and probably support it.
8monitoring	My impression is that most people had in mind the former. We have also seen some other conferences that are ruined by doing certain things, but it is inappropriate to say the examples in public. Moreover, does the discussion within REDorg happen in this way?
9balance	We will have a REDdomain1 track at our expanded two-day lab symposium on June 4-7. As a result, moving BLUEconf at this point would effectively force us to skip BLUEconf 2011 while disappointing people who have been planning to attend. Moreover, BLUEconf needs to represent interest of many different areas people, rather than one person area comes to insist on doing things in the way of that person area.

10plan	Could someone volunteer to list the URLs of active research groups (university, industry and govt) and courses on RED-domain1 inside and outside of REDdomain. This should be top priority. I look forward to hearing what others have to add.
11other	I have no desire to be ruled. Things take time especially when much is at stake. It is unfortunate that some have hastily decided to resign.

Table 4.8: Examples of learned sentences by CL-LDA for each intelligence-embodied skill in the civic deliberation discussion domain

Label	**Sentence**
1connection	If you have ideas on updating modifying it, I would love to hear from you. Cloe, I think you have a lot to be proud of. Also I think we have to simply admit to ourselves that the artful description and framing of a situation is powerful in creating action.
2proof	They meet the 3rd Tuesday of each month. Finally, I want you all to know that I am still very much in awe that this thread stayed within the boundaries of the forum guidelines for as long as it did.

	There is a huge amount of unexamined white privilege and class privilege in this community, and it was on such display at the vigil that my partner and I left early because we were so uncomfortable.
3restraint	The first step in equality is to realize what were missing up here in the land of class privilege. It is so hard for us white folks to see that were the ones without knowledge, experience or insight into how to integrate–we don't know which of our assumptions are stupid – or what we do to marginalize, stereotype and alienate. The first step in equality is to realize what were missing up here in the land of class privilege.
4agreement	In addition to bringing more attention she also made people feel optimistic, which is a much better way to create action. Maybe theres a next step. Not part of this crowd you mention, Bill, not as far as I can tell, but maybe that cant be seen from the outside.
5appreciation	Love the parade and the community. She told me thank god I'm not white or I'd have to eat with the family like you do - was she ever right. There will always be naysayers and critics for one reason or another, but the bottom line is the event was a great gift for all of the neighborhood.
6self_reflection	I also think it was very important that this event was framed in an optimistic manner early on by the mother.

	Of course this in no way secures my place in my own mind as a good person who tries and anyway that is clearly not enough. I cant help but imagine what that is like, for her and for her family.
7perspective_taking	Poor people are also less likely to have good, warm winter clothes in which they would feel comfortable standing about on a cold winter night. There's only one source of any information and experience and that's from people of color. Puppets are great and fun, but in the face of the real work that needs to happen here, the focus of recent community events really feels to a lot of us (according to the emails Ive received off-list) like a slap in the face.
8monitoring	It is hard work to learn what white privilege is – how it works. We still don't actually know the race of anyone involved in the more recent incident. Typically threads dealing with race or other major societal issues are dominated by a few authors, so for so many people add their thoughts is pretty amazing.
9balance	Encouraging our neighbors of color to join committees is only one possible option, and it's insufficient to tell people they are welcome. I'm afraid this one topic could go on into eternity. Additionally, I would like to note that saying that we have now reclaimed the park is a really problematic statement in and of itself and claiming territory is not for any of us to do.

10plan	There are plenty of people who are interested in what goes on here, for a variety of reasons, and if they want to add their thoughts, as long as they follow the same guidelines as everyone else that is fine. The only catch is that whomever is doing the posting is responsible for the content. The neighborhood needs people like you, Michelle, and others who will come forward and provide leadership in order to make our neighborhood the sort of special place that so many of us choose to live in.
11other	Glad you were able to bring out your family. I'm willing to meet up with folks for something like that. Be Civil - No insults, name calling or inflamed speech.

4.6 Conclusion and Future Work

In this chapter, we presented a new hierarchical probabilistic model for identifying multiple intelligence-embodied communication skills simultanously from natural language. This model, called Constrained Labeled LDA (CL-LDA), learns the topic assignment of each sentence so it provides a practical and simple way to determine document labels without relying on a threshold function. CL-LDA has high interpretability and its annotated sentences significantly augment the view of each document with rich contextual information. CL-LDA outperforms state-of-the-art multi-label text classification methods on prediction sensitivity, specificity, and accuracy in an online negotiation context. Experimental results also show that CL-LDA's performance increases as the number of labels grows, which makes CL-LDA a promising approach for large-scale data analysis. The results of comparing LDA-based models with SVM models indicate that a set of binary SVM models performed poorly in the face of many labels and only a small number of training instances.

We note that multi-label classification problems are by no means exclusive of natural language processing. Therefore, CL-LDA is general enough to be applied to other domains where the research interest lies in predicting multiple labels simultaneously, such as signal processing, computer vision, and computational neuroscience.

In future work, we will apply CL-LDA to more online contexts and data sets where people are from diverse culture backgrounds so that we can explore the effect of culture differences on peoples' communication intelligence. In addition, we will extend CL-LDA with the ability to model label associations. The statistically significant positive correlation between the number of positive labels per message and the prediction sensitivity of CL-LDA provides evidence that an inter-dependency may exist between multiple labels and can help the model learn better. Moreover, we are interested in augmenting CL-LDA with temporal information. The resulting dynamic CL-LDA can address questions, such as whether the use of intelligence-embodied communication skills follow certain pattern? For example, is a *perspective taking* statement/message always followed by an *appreciation* message?

CHAPTER 5

MULTI-TASK LEARNING WITH RELAXED STRUCTURED SPARSITY REGULARIZATION

The learning approach, constraint labeled LDA (CL-LDA), introduced in the previous chapter can support large-scale computational annotations of intelligence-embodied communication skills from text corpora in online communication. CL-LDA is a powerful language model in that, it, for the first time, illustrates the diverse language (i.e., sentences) that people use when applying high-order communication skills in online interactions. The extensive model evaluations provide evidence about the inter-dependency between labels and show that this dependency can help the model learn better. In this chapter, we introduce a learning approach that exploits label dependency to improve prediction performance of multi-class learning and can be used for *real-time* analysis.

5.1 Motivation and Related Work

Language is a phenomenon at the interplay of culture, education, psychology, and communication. The different word choices and diverse ways that people use language to express their thoughts and feelings provide windows into people's cognitive and emotional worlds. While it is important to learn the diversity in language among people when a particular high-order communication skill is applied, it is equally important to explore the *shared linguistic characteristics in skill use across people*. High-level features, such as lexical and discourse features, provide a good starting point for this exploration. For example, self-reflection might be characterized as using tentative language (e.g., perhaps, guess) and using repetitive grammatical aspect – the use of a verb to express an event related to the flow of time (e.g., "I believed," "now I think"). In this research, we also explore the language coordination phenomenon [41] and priming effects (e.g., semantic priming) by using interaction features (i.e., the skill labels of prior message).

In this chapter, we introduce a novel machine learning approach that simultaneously identifies multiple intelligence-embodied communication skills from online messages. This approach has two improvements over CL-LDA introduced in the previous chapter. First, it does not assume label independence. Indeed, it exploits the relationship among multiple tasks/labels to learn them simultaneously, so that tasks can mutually benefit from each other leading to improved prediction performance. Second, it can be used in *real-time* to identify those intelligence-embodied communication skills from online communication. In other words, it can be applied to domains where messages become available one at a time, such as online social media.

State-of-the-art approaches that explore label relationship to simultaneously learn multiple labels mostly fall under the paradigm of multi-task learning (MTL) [30]. In the chapter on Background, we surveyed a number of multi-task learning approaches that employ advanced regularization techniques to induce structured sparsity. Applying sparse-group Lasso to identifying multiple skills using lexical, discourse, and interaction features seems ideal at the first thought. This is because with sparse-group Lasso, we can identify both important feature groups and important features within the selected groups while learning multi-task problems. However, although we expect that a certain level of sharing exists among the studied skills in terms of discourse styles, it is always sensible to take account of the individual differences by allowing tasks to learn independently when appropriate. The Dirty model allows us to explicitly model the sharing among tasks as well as task specificity. However, it does not induce sparsity within a group, as a result, important features within the selected groups can not be effectively recognized. To get the best of both worlds, we develop a new multi-task learning formulation based on a novel composite regularization technique, called relaxed sparse-group Lasso (RSGL) (because the between-group sparsity is relaxed to model task specificity). RSGL combines the advantages of SGL and the Dirty model. As a result, it encourages between-group sparsity, within-group sparsity, and also takes account of both task sharing and task specificity. More importantly, the key merit of RSGL is that it is a general multi-task formulation that is able to unify many widely used regularization techniques, including Lasso, group Lasso, sparse-group Lasso, and the Dirty model, as we show in the next section.

Before we formally describe its model formulation, we show the list of research questions we can address with RSGL.

- **Research questions:**
  - How can the use of multiple intelligence-embodied skills be automatically and simultaneously identified?
  - What are the shared linguistic characteristics (e.g., lexical and discourse) across people with respect to each intelligence-embodied skill?

5.2 Features

Computational understanding of intelligence-embodied communication skills is an unexplored research area. We turned to the literature of social, psycholinguistic, and communication studies to explore possible feature sets. In this research, we take the first step by using LIWC, Coh-Metrix, and interaction features to identify linguistic characteristics of each intelligence-embodied communication skill.

5.2.1 Lexical Features – LIWC

The ways that individuals communicate provide windows into their cognitive and emotional worlds. Methods for studying the various emotional, cognitive, social, and psychological process components in individuals' language allow researchers to understand mechanisms and strategies of effective communication. LIWC, Linguistic Inquiry Word Count [127], is a discourse analytic system generated based on analyzing the utterances of over 24,000 writers or speakers totaling over 168 million words. LIWC produces groups of words from 82 language dimensions through a word counting approach. As shown in Figure 5.1, these 82 groups fall into 5 general categories: *linguistic processes* (e.g., total word count, words per sentence, percentage of words in the text that are pronouns), *psychological processes* (e.g., affect, cognition, biological processes), *personal subjects*, *paralinguistic dimensions*(e.g., assents, fillers, nonfluencies), and *punctuations* (e.g., quotation marks, exclamation marks). LIWC categories have been shown to be valid and reliable markers of a variety of psycholinguistic phenomena. For example, when investigating gender differences in linguistic styles using LIWC features, researchers 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 [10]. In the work of [161], LIWC features helped find the roles played by emotional and informational support in participants' commitment in online health support groups. In another study [67], 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 from the language they use, we believe that LIWC features can contribute to demystifying the link between language and effective communication.

5.2.2 Discourse Features – Coh-Metrix

Successful discourse communication often occurs when discourse participants understand one another, agree on the subject matter, or even agree to "disagree [120]." Discourse communication may easily break down when participants encounter substantial differences in language, common ground, prior knowledge, or discourse skills,

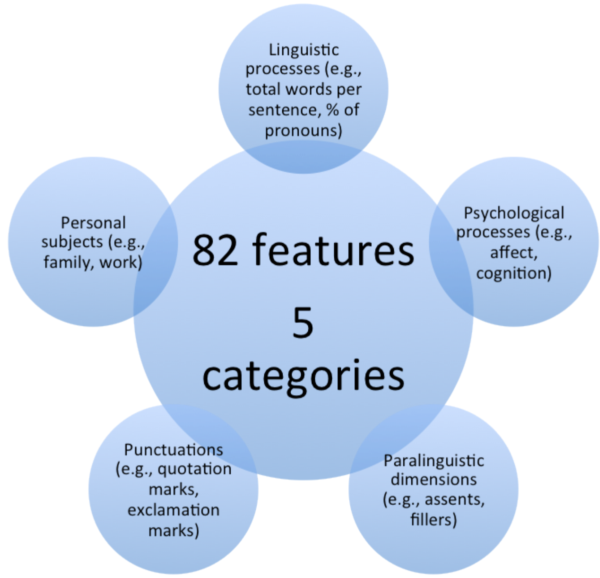

Figure 5.1. An overview of LIWC features

which is often true in computer-mediated communication [62]. Coh-Metrix is a discourse model aimed at better understanding of discourse comprehension, communication breakdowns and misalignments. It operationalizes a multilevel discourse framework using a number of advanced language models (e.g., latent semantic analysis) and text processing algorithms (e.g., syntactic parser). Coh-Metrix focuses on *text cohesion*, or "characteristics of the text that play some role in helping the reader mentally connect ideas in the text." Coh-Metrix was developed by analyzing the TASA (Touchstone Applied Science Associates, Inc.) corpus of 37,520 texts. Coh-Metrix outputs over 80 measurements [1] about text cohesion that fall under 8 categories, shown in Figure 5.2: *narrativity*, *referential cohesion*, *syntactic simplicity*, *word concreteness*, *causal cohesion*, *logical cohesion*, *verb cohesion*, and *temporal cohesion*.

Despite its academic roots, Coh-Metrix has been widely validated as a computational psycholinguistic tool for predicting complex phenomena, such as personality,

[1] The second-order features derived from principal component analysis were excluded in this research to remove redundancy.

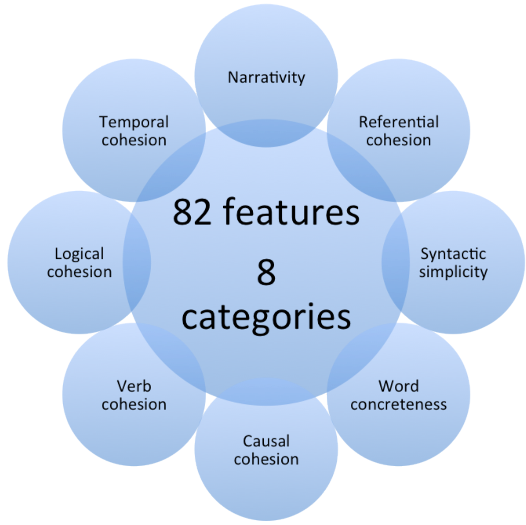

Figure 5.2. An overview of Coh-Metrix features

deception, affect states, and even physical and mental health outcomes [68, 102, 28, 46]. Given that Coh-Metirx provides a platform for a systematic and deeper analysis of discourse contents, we believe that it can help uncover subtle linguistic characteristics related to intelligence-embodied communication skills.

5.2.3 Interaction Features

Communication is an interactive event – it occurs in a dialogue between interlocutors. Numerous researchers argued that mutual influence between conversational parties created an interdependent relationship in language use [36], known as linguistic style matching [123], or language coordination [41]. In other words, people tend to unconsciously take on the linguistic features of the person immediately preceding them. Thinking in this line, we created a feature, called labels of prior message, to investigate whether the labels of prior message would help to predict the labels of the current message. Because we have 11 skill labels, 11 corresponding interaction features were used in this study.

5.3 Relaxed Sparse-group Lasso (RSGL)

Research in regularization-based approaches to multi-task learning (MTL) has attempted to relax two underlying assumptions in MTL to allow its wider applicability. These two assumptions are (1) task relatedness and (2) well-defined group structure [2]. The Dirty model relaxes the first assumption by modeling task-specificity in addition to task sharing, which allows it to solve problems where tasks only relate to each other "to some extent." The sparse group Lasso (SGL) relaxes the second assumption by allowing some features to be "added" back to groups entirely shrunk early and some other features to be "shrunk" from groups entirely selected early. The relaxed sparse group lasso (RSGL) developed in this research relaxes these two assumptions all together and therefore can find applications in a wide range of real-world problems where (1) the degree of task relatedness is uncertain and (2) the true structure of the groups in data is not clear ahead of time. Therefore, RSGL contributes to the research in multi-task learning by expanding its applicability to data and situations that were previously not applicable.

Another way to understand the differences between RSGL and the other two methods is by looking at the technical challenges they tackle. The Dirty model solves a constrained optimization problem where the penalties are separated over two feature spaces (i.e., task-specific and task-sharing). SGL solves a problem of structured sparsity involving overlapping group structures – a key challenge in many multi-task learning formulations [104]. The compound challenges from penalties separated over two feature spaces and groups overlapping one another are what RSGL is set to solve.

[2] Choices of groups are problem dependent. Prior knowledge can be used to favor certain structure patterns but may not be always available. In this research, the group structures of features were obtained from the predefined categories in each feature set. For example, Coh-Metrix has 12 feature groups, because Coh-Metrix further divides its 8 overarching categories into 12 subgroups that include lexical diversity, latent semantic analysis, and connectives. Similarly, the LIWC system outputs 10 subgroups (e.g., affect process, cognitive process) out of its 5 main categories. As to the 11 interaction features, we had them in one group.

Table 5.1: A comparison of SGL, the Dirty model, and RSGL

	SGL	**The Dirty Model**	**RSGL**
Assuming task relatedness	Yes	No	*No*
Assuming well-defined groups	No	Yes	*No*
Sovling a constrained optimization problem (with penalties separated over two feature spaces)	No	Yes	*Yes*
Solving a structured sparsity problem (with overlapping group structures)	Yes	No	*Yes*

Adapting the Dirty Model solution to RSGL is not feasible. This is because their solution approach is tied to the $l_{\infty,1}$ norm used in the Dirty model formulation, and the $l_{\infty,1}$ norm is shown to be less effective than the $l_{2,1}$ norm for multi-task learning, as we learned from the chapter about Background. A comparison chart of these models is shown in Table 5.1. In this section, we present a simple solution for solving RSGL.

In the text below, we first define relaxed sparse-group Lasso (RSGL) formally, then illustrate its solution with a working example, and finally describe an online learning algorithm for RSGL.

Formally, the multi-task learning model with relaxed sparse-group Lasso penalties can be described as follows:

$$\begin{aligned} \min_{\mathbf{w}} L(\mathbf{w}, X, Y) + \lambda_1 ||\mathbf{u}||_1 + \lambda_2 ||\mathbf{u}||_{2,1} + \lambda_3 ||\mathbf{v}||_1 \\ \text{subject to: } \mathbf{w} = \mathbf{u} + \mathbf{v} \end{aligned} \tag{5.1}$$

where u and v denote the common structure and the task-specific structure, respectively.

Relaxed sparse-group Lasso has a key property: *it subsumes the most widely-used norms and the Dirty model* [3]*, as special cases.*

[3] Here, we consider substituting $l_{2,1}$ norm for $l_{\infty,1}$ norm.

- If $\lambda_1 = \lambda_2 = u = 0$, it becomes Lasso;
- If $\lambda_1 = \lambda_3 = v = 0$, it becomes group Lasso;
- If $\lambda_3 = v = 0$, it becomes sparse-group Lasso;
- If $\lambda_1 = 0$, it becomes the Dirty model.

In this research, we use the logistic loss function for the task of multi-skill classification, where the logistic loss function is as follows.

$$-y * log(\frac{1}{1 + exp(-w^T x)}) - (1 - y) * log(1 - \frac{1}{1 + exp(-w^T x)}) \tag{5.2}$$

5.3.1 Solving RSGL – Reducing a Constrained Optimization Problem to an Unconstrained One

The new formulation of multi-task learning with relaxed sparse-group Lasso is a constrained optimization problem where the penalties are separated over u and v. Different from the solution to the Dirty model, we propose a simple method to *reduce this constraint optimization problem to an unconstrained problem* by exploiting the good property of $\mathbf{w} = \mathbf{u} + \mathbf{v}$. Specifically, defining $\mathbf{w} = \mathbf{u} + \mathbf{v}$ is equivalent to *duplicating the feature space*, since $w'^* f(x) = u'^* f(x) + v'^* f(x) = [u, v]'^*[f(x), f(x)]$. Therefore, we can reason about a feature space that is twice the original dimension, a double-size weight matrix that concatenates u and v.

5.3.2 A Working Example

In this section, we use an example to illustrate the feature-duplicating approach to feature selection with RSGL. We also motivate the use of algorithms that can handle a special property of RSGL: *overlapping groups*.

Suppose that we have 4 features, f_1, f_2, f_3, f_4, and 3 tasks. The feature space then has 12 features to begin with – let them be f_{11}, f_{12}, f_{13}, f_{21}, $\dots$, f_{43}, where the second index denotes the task.

Then we double the feature space and get 24 features in total. Let us use letter g for the copied features: g_{11}, g_{12}, g_{13}, g_{21}, $\ldots$, g_{43}.

Suppose the groups in this example are:

- 4 groups that group the f-features over the tasks, $G1 = \{f_{11}, f_{12}, f_{13}\}$, $G2 = \{f_{21}, f_{22}, f_{23}\}$, $G3 = \{f_{31}, f_{32}, f_{33}\}$, $G4 = \{f_{41}, f_{42}, f_{43}\}$,

- 12 groups for the individual f-features, $G5 = \{f_{11}\}$, $G6 = \{f_{12}\}$, $\ldots$, $G16 = \{f_{43}\}$,

- 12 groups for the individual g-features , $G17 = \{g_{11}\}$, $G18 = \{g_{12}\}$, $\ldots$, $G28 = \{g_{43}\}$.

Let us assume that the f-features are from the sharing component u, and the g-features are from the task-specific component v. We further assume that the groups that selected groups are $G4 = \{f_{41}, f_{42}, f_{43}\}$, $G14 = \{f_{41}\}$, and $G28 = \{g_{43}\}$. Then the selected feature in u will be f_{41}, and the selected feature in v will be g_{43}. When summing u and v the selected features will be f_{41} and g_{43}.

Observations:

- f_{42} and f_{43} are shrunk. RSGL inherits the good property from SGL, so it allows for the shrinkage of features within the selected group. Technically, this is because f-features have an *overlapping-group* (i.e., each feature may appear in more than one group) structure that makes the within-group sparsity possible.

- g_{43} is put back. RSGL inherits the desirable property from the Dirty model, so it allows individual features to be selected even though its belonging group is shrunk. Technically, we can attribute this property to the inclusion of a task-specific feature space.

  non-overlapping group structure of g.

5.3.3 Online Learning for RSGL

Different from the Dirty model that has only disjoint groups, RSGL has both disjoint groups and *overlapping groups*. In the literature, a number of algorithms have proposed to solve problems of structured sparsity with overlapping groups. In this research, we use an algorithm, called online proximal-gradient algorithm (OPG) [104], which unifies many well-know learning algorithms for multi-task learning with structure sparsity, including online projected sub gradient algorithm [179], $PEGASOS$ [144], truncated gradient descent [92], and $FOBO$ [44]. We chose OPG because of two reasons. First, the OPG algorithm uses an easy way to handle overlapping groups with the application of $\Phi-$proximity operators [114], so that mixed-norm proximity operators can be applied *sequentially*. Second, as an online algorithm, OPG allows RSGL to be applied in a real-time manner, so that data can be analyzed as it comes.

In Algorithm 1, we adapt the OPG algorithm to solve RSGL with logistic loss.

5.4 Experiments and Results

In this section, we evaluate the learning performance of RSGL for simultaneously identifying multiple intelligence-embodied communication skills. Specifically, we first compare the prediction performance of RSGL against state-of-the-art multi-task learning formulations: sparse-group Lasso (SGL) and the Dirty model. We then examine the feature space *shared* by all the skill labels and those *specific* to each skill label in order to study the attributions of RSGL's performance gain over the Dirty model and SGL. Lastly, we show the learned features with respect to each communication skill and evaluate features' quality qualitatively.

Our experimental data are from the two online corpora described in the previous chapter. They are the professional community negotiation and the civic deliberation discussion domains. The training and testing sets remain the same as before for each

Algorithm 1:

Input: data matrix X with dimension N * $2D$ (doubling feature space);
response matrix Y with dimension N x M;
regularization constraint constants:
λ_1 (controlling within-group sparsity in u);
λ_2 (controlling between-group sparsity in u);
λ_3 (controlling element-wise sparsity in v);
T : the number of epochs * training examples (N);
α_t : the learning rate at time t.
Output: weight matrix: sum (u_{T+1}, v_{T+1})
Initialization: Initialize u_1 = 0, v_1 =0
for $t \leftarrow 1$ **to** T **do**
  //computing the gradient of logistic loss
  g= $\nabla(u$+v; x_t, $y_t)$; //concatenate u and v
  $\tilde{u_t} = u_t - \alpha_t * g$
  $\tilde{v_t} = v_t - \alpha_t * g$
  //proximal mapping for u, first at the group level, then within-group
  $\tilde{u_t}$ = soft-thresholing($\tilde{u_t}$,λ_2)
  $\tilde{u_t}$ = soft-thresholing($\tilde{u_t}$,λ_1)
  //proximal mapping for v
  $\tilde{v_t}$ = soft-thresholding($\tilde{v_t}$,λ_3)
  //projection (optional for learning speedup)
  $u_{t+1} = \tilde{u_t} * \min(1, (\lambda_1+\lambda_2)/ \parallel \tilde{u_t} \parallel)$
  $v_{t+1} = \tilde{v_t} * \min(1, \lambda_3/ \parallel \tilde{v_t} \parallel)$
end
return u_{T+1}, v_{T+1}

domain. For both domains, we acquired lexical and discourse features from the LIWC and Coh-Metrix systems. The interaction features were readily available, as they are skill labels of the message that immediately precedes the message being studied. We had 175 features – 82 LIWC features, 82 Coh-Metrix features, and 11 interaction features in total. We preprocessed the data by standardizing all feature variables to have zero mean and unit variance. In doing so, we avoided imposing priors on any features based on their numerical values.

To select the best model parameters of RSGL, we performed a grid search for the optimal learning rate α from values {0.001, 0.01, 0.1, 1, 10} and searched for the optical values of each regularization constant λ (λ_1, λ_2, and λ_3) from {0.0001, 0.001,

Table 5.2: Category-pivoted evaluations in the professional community negotiation domain: A comparison of SGL, Dirty+, and RSGL

	SGL			**Dirty+**			**RSGL**		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
1connection	1.00	0.50	0.73	1.00	0.50	0.73	1.00	0.50	0.73
2proof	0.67	0.50	0.53	0.67	0.50	0.53	0.67	0.50	0.53
3restraint	0.64	0.00	0.60	0.57	1.00	0.60	0.57	1.00	0.60
4agreement	0.50	0.55	0.53	0.50	0.55	0.53	0.50	0.64	0.60
5appreciation	0.33	0.50	0.47	0.33	0.50	0.47	0.33	0.50	0.47
6self_reflection	1.00	0.45	0.60	1.00	0.36	0.53	1.00	0.45	0.60
7perspective_taking	1.00	0.63	0.80	1.00	0.63	0.80	1.00	0.63	0.80
8monitoring	0.78	0.83	0.80	0.89	0.67	0.80	0.89	0.67	0.80
9balance	1.00	0.33	0.47	0.67	0.42	0.47	1.00	0.33	0.47
10plan	0.60	0.50	0.53	0.60	0.50	0.53	0.60	0.50	0.53
11other	0.85	1.00	0.87	0.92	0.50	0.87	0.92	0.50	0.87
Min	0.33	0.00	0.47	0.33	0.36	0.47	0.33	0.33	0.47
Max	1.00	1.00	0.87	1.00	1.00	0.87	1.00	1.00	0.87
Avg	0.76	0.53	**0.63**	0.74	0.56	**0.62**	0.77	0.57	**0.64**
Avg (sen+spe)	**0.64**			**0.65**			**0.67**		

0.01, 0.1, 1}. Specifically, in our experiments, the learning rate was set to decrease as $\alpha_t = \alpha_0/\sqrt{t}$, while all combinations of possible λ values were tested. We looked at the training objective and picked α_0 that yields the best objective value after 2 epochs. When choosing each regularization parameter λ, we used the chosen learning rate and performed leave-one-out cross validation on the training set for 2 epochs and selected the one that has the smallest misclassification error on the validation data. We applied the best model parameters sets (α= 10, $\lambda_1 = 0.01$, $\lambda_2 = 0.01$, and $\lambda_3 = 1$) to the testing data for 1 epoch and omitted the optional projection step. We followed the same procedure for both domains and arrived at the same best parameter setting. For other comparison approaches SGL and the Dirty model, similar experimental figurations were applied.

5.4.1 Evaluating Classification Performance

5.4.1.1 Category-pivoted Evaluations

In this section, we focus on category-pivoted evaluations – the evaluation of model performance on predicting each skill label. We evaluated model performance quantitatively in terms of sensitivity (the true positive rate), specificity (the true negative rate), and accuracy. Both sensitivity and specificity are valued in this research because, for the purpose of measuring communication intelligence, the presence and absence of skills are equally important to identify. Note that the original Dirty model minimizes the least squared objective and uses $l_{\infty,1}$ norm. As discussed in the chapter on Background, $l_{\infty,1}$ norm is not as effective as $l_{2,1}$ norm for multi-task learning. We implemented a variation of Dirty model with logistic loss function and $l_{2,1}$ norm so that we can compare methods on the same footing. In the text below, for simplicity, we refer to this improved variation as Dirty+. As shown in Table 5.2, in the professional community negotiation domain, RSGL achieves the best average prediction sensitivity (77%) across all 11 categories, followed by SGL (76%) and Dirty+ (74%). The similar performance of RSGL and other methods in predicting some of the intelligence-embodied skills is because the feature spaces learned by the relevant methods are similar, as we will show in the next section. RSGL also achieves the best average prediction specificity (57%) and the best average prediction accuracy (64%) (compared to the 9% baseline).

As can also be seen in Table 5.2, in the professional community negotiation domain, RSGL achieves the highest sensitivity (100%) in predicting the skills of *connection*, *self-reflection*, and *perspective taking*, and achieves the lowest sensitivity (33%) in predicting the skill of *appreciation*. RSGL achieves the highest specificity (100%) in predicting the skill of *restraint* and the lowest specificity (33%) in predicting the skill of *balance*. Please be cautious that these observations can not be used as evidence to conclude which intelligence-embodied skills are easy or hard to predict automatically from linguistic and interaction features. This is because these results are tied to a particular online context (e.g., discussion vs. negotiation) that we study in this research. They are also influenced by the number of training and testing data available in each skill/category which is constrained by the three experiment design principles we introduced early. This statement also applies to similar observations in the following experiments in this chapter.

Table 5.3: Category-pivoted evaluations in the civic deliberation discussion domain: A comparison of SGL, Dirty+, and RSGL

	SGL			**Dirty+**			**RSGL**		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
1connection	0.75	0.25	0.42	0.50	0.13	0.25	0.50	0.13	0.25
2proof	0.50	0.50	0.50	0.25	0.50	0.42	0.50	0.38	0.42
3restraint	0.55	0.00	0.50	0.91	1.00	0.92	0.91	1.00	0.92
4agreement	0.50	0.60	0.58	0.50	0.60	0.58	0.50	0.60	0.58
5appreciation	0.50	0.40	0.42	0.50	0.40	0.42	0.50	0.40	0.42
6self_reflection	0.60	0.29	0.42	0.60	0.43	0.50	0.60	0.29	0.42
7perspective_taking	0.80	0.57	0.67	1.00	0.57	0.75	1.00	0.57	0.75
8monitoring	0.33	0.17	0.25	0.50	0.17	0.33	0.50	0.17	0.33
9balance	0.00	0.30	0.25	0.00	0.20	0.17	0.00	0.20	0.17
10plan	1.00	0.36	0.42	1.00	0.46	0.50	1.00	0.36	0.42
11other	0.55	0.00	0.50	0.55	0.00	0.50	0.64	0.00	0.58
Min	0.00	0.00	0.25	0.00	0.00	0.17	0.00	0.00	0.17
Max	1.00	0.60	0.67	1.00	1.00	0.92	1.00	1.00	0.92
Avg	0.55	0.31	**0.45**	0.57	0.40	**0.49**	0.60	0.37	**0.48**
Avg (sen+spe)	**0.43**			**0.49**			**0.49**		

Now, let us look at the civic deliberation discussion domain. As can be seen in Table 5.3, RSGL achieves the best average prediction sensitivity (60%) across all 11 categories, followed by Dirty+ (57%) and SGL(55%). Dirty+ achieves the best average prediction specificity (40%), followed by RSGL (37%) and SGL (31%). RSGL achieves the second-best (48%) in average prediction accuracy (compared to the 9% baseline), which is 1% lower than Dirty+.

Given that Dirty+ outperforms RSGL slightly on the average specific and accuracy in the civic deliberation discussion domain, we performed a *multivariate permutation test* [128] to see whether a statistically significant difference exists between the predictions from RSGL and those from Dirty+. Permutation test is a Monte Carlo procedure that shuffles the data to test the equality of two sample distributions. Multivariate permutation test was used here because the prediction includes a group of 11 skill labels. The result of the multivariate permutation test (based on 10,000 permutations) shows that there is no statistically significant difference (P=0.52) between the performance of RSGL and that of Dirty+.

As can also be seen in 5.3, in the civic deliberation discussion domain, RSGL achieves the highest sensitivity in predicting the skills of *perspective taking* and *plan* and about 50%-60% in predicting other skills. RSGL achieves the highest specificity

Table 5.4: Message-pivoted evaluations in the professional community negotiation domain: A comparison of SGL, Dirty+, and RSGL

	SGL			**Dirty+**			**RSGL**		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Min	0.40	0.17	0.36	0.25	0.00	0.45	0.50	0.00	0.45
Max	1.00	0.83	0.82	1.00	0.83	0.82	1.00	1.00	0.91
Avg	0.79	0.52	**0.63**	0.79	0.50	**0.62**	0.81	0.51	**0.64**
Avg (sen+spe)	**0.65**			**0.65**			**0.66**		

(100%) in predicting the skill of *restraint*, and achieves the lowest specificity (13%) in predicting the skill of *connection*.

Comparing model performance in two domains, we once again found again that all models have lower performance in the civic deliberation discussion domain. In addition to the fact that fewer training instances are available in the civic deliberation domain, we also speculate that, in the civic deliberation domain, the linguistic characteristics of intelligence-embodied skills might be more distinct from one another in a negotiation context than might in a discussion context. In other words, each high-order communication skill appears to have more unique linguistic characteristics in a communication environment that is more controversial or with greater tension.

5.4.1.2 Message-pivoted Evaluations

Category-pivoted evaluations allow us to study a model's predicability for each skill label separately. When multi-category classification is concerned, message-pivoted evaluations provide a holistic view on a model's predicability of all skill labels associated with a message. Moreover, message-pivoted evaluations are suitable when data becomes available one at a time (e.g., in online messages), and therefore great for *real-time* analysis.

As shown in Table 5.4, in the professional community negotiation domain, RSGL achieves the best average prediction sensitivity (81%) across all the messages in the testing set, followed by SGL and Dirty+ (79%). RSGL achieves the second-best (51%) in average prediction specificity, which is 1% lower than SGL. RSGL also achieves the best average prediction accuracy (64%).

Table 5.5: Message-pivoted evaluations in the civic deliberation discussion domain: A comparison of SGL, Dirty+, and RSGL

	SGL			**Dirty+**			**RSGL**		
	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy	Sensitivity	Specificity	Accuracy
Min	0.00	0.00	0.18	0.33	0.00	0.27	0.33	0.00	0.27
Max	1.00	0.67	0.64	1.00	0.71	0.82	1.00	0.60	0.73
Avg	0.57	0.37	**0.45**	0.66	0.38	**0.49**	0.69	0.34	**0.48**
Avg (sen+spe)	**0.47**			**0.52**			**0.52**		

In the civic deliberation discussion domain, as shown in 5.5, RSGL achieves the best average prediction sensitivity (69%) across all the messages in the testing set, followed by Dirty+ (69%) and SGL (57%). Dirty+ achieves the best average prediction specificity (38%), followed by SGL (37%) and RSGL (34%). RSGL achieves the second-best (48%) in average prediction specificity, which is 1% lower than Dirty+.

We found consistently in both domains that RSGL outperforms other comparison methods on predicting sensitivity – the ability to correctly predict the use of skills, with little sacrifice on prediction specificity.

5.4.1.2.1 The Relationship Between the Number of Positive Labels per Message and Model Performance In the previous chapter, we found statistically significant correlations between a model's prediction performance and the number of positive labels each message has, which implies that an inter-dependency might exist between labels and an help the model learn better. Multi-task learning naturally exploits such interrelationship to perform parallel learning across tasks. Not surprisingly, we found no significant correlation ($\alpha = 0.05$) for any measures in either the professional community negotiation domain or the civic deliberation discussion domain.

5.4.2 Evaluating Feature Compression Capacity

To study the performance differences between RSGL and Dirty+, we focus on the technical differences between the two models. Specifically, Dirty+ uses group Lasso to impose between-group sparsity in the feature space shared by tasks (e.g.,

Table 5.6: Feature compression evaluations in the professional community negotiation domain (percentage shrinkage of feature space shared by skill labels): A comparison of Dirty+ and RSGL

	SGL	**Dirty+**	**RSGL**	**Shrinkage (%)**
1connection	171	172	172	0.00
2proof	172	172	172	0.00
3restraint	172	172	172	0.00
4agreement	170	172	170	1.14
5appreciation	172	172	171	0.57
6self_reflection	172	172	172	0.00
7perspective_taking	172	172	172	0.00
8monitoring	172	172	172	0.00
9balance	172	172	172	0.00
10plan	171	172	171	0.57
11other	172	172	172	0

skill labels), whereas RSGL imposes both between-group sparsity and within-group sparsity in that feature space. As we learn from the chapter on Background, the natural consequence of group Lasso is that a feature is either selected as relevant for all tasks simultaneously, or is excluded all-together for all tasks, implying that all tasks share the same number of active features. As shown in Table 5.6, in the professional community negotiation domain, compared to Dirty+, RSGL achieves up to 1.14% compression of the task-sharing feature space, while improving prediction sensitivity by 3%, predicting specificity by 1%, and predicting accuracy by 2%. In the civic deliberation discussion domain, shown in Table 5.7, compared to Dirty+, RSGL achieves up to 14.86% compression of the task-sharing feature space, while improving prediction sensitivity by 3%, with a decrease of predicting specificity by 3% and predicting accuracy by 1%.

We observe that the level of feature shrinkage is agreed by SGL, Dirty+, and RSGL. This observation implies that simultaneously identifying multiple intelligence-embodied communication skills from online communication is a challenging task, and therefore most of the linguistic and interaction features are needed to achieve the competence we reported in the last section. The discrepancy in feature compression

Table 5.7: Feature compression evaluations in the civic deliberation discussion domain (percentage shrinkage of feature space shared by skill labels): A comparison of Dirty+ and RSGL

	SGL	**Dirty+**	**RSGL**	**Shrinkage (%)**
1connection	170	170	170	0.00
2proof	165	170	165	2.86
3restraint	169	170	170	0.00
4agreement	170	170	170	0.00
5appreciation	170	170	169	0.57
6self_reflection	161	170	161	5.14
7perspective_taking	163	170	163	4.00
8monitoring	170	170	170	0.00
9balance	144	170	144	14.86
10plan	168	170	165	2.86
11other	165	170	170	0

in different online contexts suggests that skill predictions in a negotiation context need more features than do predictions in a discussion context, which implies that skill predictions in a negotiation context are more complex than those in a discussion context.

5.4.3 Evaluating the Importance of Task-specific Feature Space

As shown early, RSGL consistently outperforms SGL on both domains under category-pivoted evaluations. In this section, we study the attribution of RSGL's performance gain over SGL. Recall that the difference between these two models is that RSGL models task-sharing and task specificity, whereas SGL only models task-sharing. As shown in Table 5.8, in the professional community negotiation domain, for some tasks, RSGL adds to the task-specific feature space as many as 84 features. A further examination shows that these added features are all kept (not shrunk) in the task-sharing feature space, so *what is being added is feature weights rather than the number of features*. It is important to note that, in multi-task learning, multiple tasks (e.g., skill labels) are learnt in parallel in order to capture the interrelation between tasks, meaning that *task-specific features are not only useful to learn its own task*

Table 5.8: An illustration of the number of features in the task-specific feature space (the professional community negotiation domain): A comparison of SGL and RSGL

	SGL	**Dirty+**	**RSGL**
1connection	0	1	1
2proof	0	0	0
3restraint	0	84	84
4agreement	0	0	0
5appreciation	0	0	0
6self_reflection	0	6	6
7perspective_taking	0	0	0
8monitoring	0	83	84
9balance	0	0	0
10plan	0	0	0
11other	0	75	75

Table 5.9: An illustration of the number of features in the task-specific feature space (the civic deliberation discussion domain): A comparison of SGL and RSGL

	SGL	**Dirty+**	**RSGL**
1connection	0	2	2
2proof	0	0	0
3restraint	0	127	127
4agreement	0	128	128
5appreciation	0	32	32
6self_reflection	0	0	0
7perspective_taking	0	0	0
8monitoring	0	9	9
9balance	0	0	0
10plan	0	2	2
11other	0	1	1

(i.e., skill labels), but also contribute to the learning of other labels. As can be seen in Table 5.8, the task-specific feature weights learned by RSGL lead to an improvement of 1% in prediction specificity, 3% in predicting specificity, and 1% predicting accuracy over SGL. Similarly, in the civic deliberation discussion domain, shown in Table 5.9, RSGL adds feature weights for up to 128 features from the task-specific feature space. These feature weights lead to an improvement of 5% in prediction specificity, 6% in predicting specificity, and 3% predicting accuracy over SGL.

5.4.4 Evaluating Learned Features

In Table 5.10, we present top 10 learned features by RSGL for each communication skill in the professional community negotiation domain. In the text below, we explain some highlights of our findings.

- *Connection*: We observed that the use of this skill can be attracted by the use of skills of *proof* and *appreciation*. In other words, when a participant demonstrates the use of either *proof* or *appreciation*, others in the group may be motivated to do *connection*. This is reasonable, because when one participant provides factual information or references, others follow up with inquiries, or, when one participant shows appreciations in his message, another connects back. This skill is also found to involve cognitive processes and the use of insightful and inclusive words.

- *Proof*: We observed that this skill is positively correlated to the use of colon. This is reasonable, because colon is often used in the situation of citing references and other resources. We also found high semantic similarity of messages showing the use of this skill, which implies the presented idea is coherent.

- *Restraint*: We observed that this skill – emotional control – is negatively correlated with the use of discrepancy words (e.g., should) and anger words.

- *Agreement*: We observed that this skill involves the use of assent words and words related to positive emotions. This skill has negative correlations with the use of adversative and contrastive connectives (e.g., although). We also found that the use of this skill can be attracted by the use of skills of *monitoring* and *self-reflection*. In other words, when one participant demonstrates the use of either *monitoring* or *self-reflection*, others in the group may be motivated to show *agreement*. This is reasonable, because when a participant presents his thoughts on how the dialogue goes or when a participant shows some reflective thoughts, others may follow with consensus statements.

- *Appreciation*: We observed that the use of this skill can be attracted by the use of skill of *self-reflection*. In other words, when a participant demonstrates the use of *self-reflection*, others in the group may be motivated to show *appreciation*. This is reasonable, because when a participant shows some reflective thoughts on his previous ideas, others may follow with an appreciative note. This skill is also shown to be negatively correlated with the use of negation words.

- *Self-reflection*: We observed that the use of this skill can be attracted by the use of a number of other communication skills, including *appreciation*, *agreement*, *proof*, *connection*, *monitoring*, and *plan*. In other words, when a participant demonstrates the use of one of the listed skills, others in the group may be motivated to do *self-reflection*. This skill is also shown to involve the use of tentative words (e.g., perhaps) and adversative and contrastive connectives (e.g., although). This is reasonable, because self-reflection can sometimes involve uncertainty and adaption.

- *Perspective taking*: We observed that the use of this skill can be attracted by the use of skills of *self-reflection* and *perspective taking*. In other words, when a participant demonstrates the use of either *self-reflection* or *perspective taking*, others in the group may be motivated to do *perspective taking*. This finding, to some extent, resonates with the Law of Attraction: you attract who you are. This skill is also shown to involve the use of parenthesis – a punctuation often used in referencing other's thoughts and ideas.

- *Monitoring*: We observed that this skill is positively correlated with the use of third person pronoun singular (e.g., she, he). This is reasonable, because monitoring how the dialogue has gone may involve referencing other's opinions. This skill is also shown to be positively correlated with sentence length. This is reasonable, because summarizing or synthesizing ideas can lead to long statements.

- *Balance*: We observed that this skill is positively correlated with the use of tentative words and causal connectives. This is reasonable, because weighing different viewpoints may involve causal reasoning and a sense of uncertainty.
- *Plan*: We observed that the use of this skill can be attracted by the use of skills of *monitoring*, *balance*, and *agreement*. In other words, when a participant demonstrates the use of any one of the listed skills, others in the group may be motivated to do planning – proposing actions for the next steps. This skill is also shown to be positively correlated with the use of tentative words and words related to positive emotions. This is reasonable, because plans can be tentative and are often filled with positive expectations.

Table 5.10: Learned features by RSGL for each intelligence-embodied skill in the professional community negotiation domain

	Top 10 features	Interpretations	Feature source	Weight
1connection	2proof	Interaction effects	Interaction	1.19
	cogmech	Cognitive process	LIWC	1.06
	insight	Insight words	LIWC	1.05
	5appreciation	Interaction effects	Interaction	0.99

	percept	Perceptual process	LIWC	-0.98
	incl	Inclusive	LIWC	0.94
	LEXDIVVD	Lexical diversity	Coh-Metrix	0.93
	work	Work related words	LIWC	0.93
	achieve	Achievement	LIWC	0.86
	swear	Swear words	LIWC	0.84
2proof	ADJi	Adjectives	Coh-Metrix	0.58
	OtherP	Other punctuations	LIWC	0.52
	Colon	Colon	LIWC	0.48
	anx	Anxiety words	LIWC	0.45
	READASW	Word length	Coh-Metrix	0.42
	QMark	Question mark	LIWC	0.41
	LexDensity	Lexical density	Coh-Metrix	0.36
	FRCLaewm	CELEX Log frequency for all words	Coh-Metrix	-0.36
	LSApssd	Semantic similarity (all sentences in a paragraph)	Coh-Metrix	0.34

	funct	Function words	LIWC	-0.34
3restraint	discrep	Discrepancy	LIWC	-2.23
	they	They	LIWC	-2.00
	CAUSWN	Wordnet verb overlap	Coh-Metrix	-1.98
	Quote	Quotation	LIWC	1.97
	CONCAUSi	Causal connectives	Coh-Metrix	1.88
	LSAassa	Semantic similarity (all sentences in a paragraph)	Coh-Metrix	-1.82
	anger	Anger words	LIWC	-1.81
	SPATmpi	Motional preposition	Coh-Metrix	-1.80
	PRO1i	First person pronoun	Coh-Metrix	-1.78
	LEXDIVVD	Lexical diversity	Coh-Metrix	-1.77
4agreement	assent	Assent words	LIWC	0.82
	hear	Perceptual process	LIWC	0.68
	8monitoring	Interaction effects	Interaction	0.64
	we	We	LIWC	0.57

	WRDFacwm	Familiarity of content words	Coh-Metrix	0.55
	6self_reflection	Interaction effects	Interaction	0.49
	percept	Perceptual process	LIWC	0.49
	posemo	Positive emotion	LIWC	0.44
	article	Article	LIWC	-0.43
	CONADVCONi	Adversative and contrastive connective (Although, Whereas)	Coh-Metrix	-0.39
5appreciation	CONCAUSi	Causal connectives	Coh-Metrix	1.33
	conj	Conjunctions	LIWC	-1.22
	6self_reflection	Interaction effects	Interaction	1.16
	number	Numbers	LIWC	1.14
	leisure	Personal words	LIWC	1.13
	cogmech	Cognitive process	LIWC	-1.12
	negate	Negations	LIWC	-1.05

	CONTEMPEXi	Temporal connectives	Coh-Metrix	1.03
	we	We	LIWC	-1.01
	bio	Biological process	LIWC	-1.01
6self_reflection	5appreciation	Interaction effects	Interaction	1.04
	4agreement	Interaction effects	Interaction	0.99
	excl	Exclusive	LIWC	0.98
	Exclam	Exclamation	LIWC	0.92
	tentat	Tentative	LIWC	0.91
	2proof	Interaction effects	Interaction	0.90
	1connection	Interaction effects	Interaction	0.88
	8monitoring	Interaction effects	Interaction	0.86
	CONADVCONi	Adversative and contrastive connective (Although, Whereas)	Coh-Metrix	0.81
	10plan	Interaction effects	Interaction	0.77

7perspective_taking	6self_reflection	Interaction effects	Interaction	1.58
	leisure	Personal words	LIWC	1.24
	7perspective_taking	Interaction effects	Interaction	1.15
	Numerals	Numbers	LIWC	1.13
	they	They	LIWC	1.11
	CONLOGi	Logical connectives	Coh-Metrix	1.08
	CONADVCONi	Adversative and contrastive connective (Although, Whereas)	Coh-Metrix	1.07
	relig	Religion words	LIWC	1.02
	CRFBN1um	Noun overlap, adjacent sentences	Coh-Metrix	0.97
	Parenth	Parenthesis	LIWC	0.92
8monitoring	filler	Fillers	LIWC	1.34
	shehe	She/he	LIWC	1.31
	see	Perceptual process	LIWC	-1.23

	INTEi	Intentional actions	LIWC	-1.21
	WPS	Sentence length	LIWC	1.22
	3restraint	Interaction effects	Interaction	-1.15
	Dic	Dictionary words	LIWC	-1.11
	Comma	Comma	LIWC	0.98
	WRDMacwm	Meaningfulness	Coh-Metrix	-0.97
	future	Future tense	LIWC	0.95
9balance	death	Personal words	LIWC	0.45
	future	Future tense	LIWC	0.36
	DATTIMi	Date time	Coh-Metrix	0.36
	HYVERBaw	Mean hypernym values of verbs	Coh-Metrix	-0.30
	verb	Verb	LIWC	0.28
	tentat	Tentative words	LIWC	0.28
	sexual	Biological process	LIWC	0.28
	auxverb	Auxiliary verbs	LIWC	0.25
	cause	Causal connectives	LIWC	0.24

	excl	Exclusive	LIWC	0.22
10plan	9balance	Interaction effects	Interaction	1.10
	8monitoring	Interaction effects	Interaction	0.79
	Numerals	Numbers	LIWC	0.69
	INFi	Infinitives	Coh-Metrix	0.56
	4agreement	Interaction effects	Interaction	0.56
	tentat	Tentative	LIWC	0.52
	CRFBAaum	Argument overlap	Coh-Metrix	-0.51
	ingest	Biological process	LIWC	0.50
	posemo	Positive emotion	LIWC	0.50
	MEDawm	Minimum edit distance of all words	Coh-Metrix	-0.50
11other	adverb	Adverbs	LIWC	2.41
	Period	Period	LIWC	2.10
	CAUSC	Ratio of causal particles to causal verbs	Coh-Metrix	2.08
	GERUNDi	Gerund density	Coh-Metrix	1.96

CAUSWN	Wordnet verb overlap	Coh-Metrix	-1.94
CONADVCONi	Adversative and contrastive connective (Although, Whereas)	Coh-Metrix	1.90
CONCAUSi	Causal connectives	Coh-Metrix	1.90
HYNOUNaw	Mean hypernym values of nouns	Coh-Metrix	-1.81
SPATmpi	motional preposition	Coh-Metrix	-1.77
LSAassa	LSA overlap (adjacent sentences)	Coh-Metrix	-1.76

Now, we look at the civic deliberation discussion domain. Some result highlights are as follows.

- *Connection*: We observed that the use of this skill can be attracted by the use of skills of *proof* and *restraint*. In other words, when a participant demonstrates the use of either *proof* or *restraint*, others in the group may be motivated to do *connection*. This is reasonable, because when one participant provides factual information or references, others follow up with inquiries – people are willing to connect with others who are able to control their emotions and not judge.

- *Proof*: We observed that this skill is positively correlated with the process of causal inference. This is reasonable, because during causal inference, citing facts and providing references are often used.

- *Restraint*: We observed that this skill – emotional control – is positively correlated with affective process. This is reasonable, because the presence of emotion precedes the control of emotion.

- *Agreement*: We observed that this skill involves the use of words related to positive emotions and personal pronouns.

- *Appreciation*: We observed that this skill is positively correlated with the process of causal inferences and the use of positive emotional words.

- *Self-reflection*: We observed that the use of this skill can be attracted by the use of *agreement*. In other words, when a participant demonstrates the use of the skill *agreement*, others in the group may be motivated to do self-reflection. This skill is also shown to involve the use of verbs and inclusive words. More interestingly, this skill is found to be positively correlated with repetitive grammatical aspect – the use of a verb to express an event related to the flow of time (e.g., "I believed", "now I think"). This linguistic phenomenon reveals precisely how one's thinking evolves during self-reflection.

- *Perspective taking*: We observed that this skill is positively correlated with the use of hypernym (i.e., generic words), the word "feel," and positive emotional words. This is reasonable, because perspective taking involves the ability to intuit another person's thoughts and feelings and see them from a positive light.

Perspective taking may also involve reframing, which is reflected by the use of hypernym for nouns and verbs.

- *Monitoring*: We observed that this skill is positively correlated with stem overlap in adjacent sentences and with paragraph length. This is reasonable, because monitoring how the dialogue has gone involves referencing and commenting, which, in tandem with summarizing or synthesizing ideas, can lead to long statements.

- *Balance*: We observed that the use of this skill can be attracted by the use of *self-reflection*. In other words, when a participant demonstrates the use of the skill *self-reflection*, others in the group may be motivated to do balance – weighing different opinions about the topic being discussed. This skill is also found to be positively correlated with the imaginability of content words (e.g., a vivid description).

- *Plan*: We observed that this skill is positively correlated with the use of discrepancy (e.g., should) words, which are often used in planning statements. We also found that the semantic similarity of messages showing the use of this skill is high, which means the presented idea is coherent.

As we can see from Table 5.10 and Table 5.11, for some skills, the learned features of some skills learned from the professional community negotiation domain more closely conform to human understanding than those from the civic deliberation discussion domain. This comes as no surprise, given that the model performance in the professional community negotiation domain is about 20% better than that in the civic deliberation discussion domain. Nevertheless, we found some evidence consistent with both domains. For example, the use of *proof* attracts the use of *connection*, and the use of *agreement* attracts the use of *self-reflection*.This observation also implies that interaction features are more robust than psycholinguistic features for predicting intelligence-embodied communication skills at the change of domain or online context. In addition, as feature rankings interaction features are not as high as that of pyscholinguistic features, future studies using interaction features alone might provide some insights into the predictive power of these features.

Table 5.11: Learned features by RSGL for each intelligence-embodied skill in the civic deliberation discussion domain

	Top 10 features	Interpretations	Feature source	Weight
1connection	2proof	Interaction effects	Interaction	13.60
	READFKGL	Reading easiness	Coh-Metrix	12.71
	Numerals	Numbers	LIWC	11.24
	CRFBSaum	Argument Overlap, all distances, unweighted	Coh-Metrix	11.14
	inhib	restraint	LIWC	10.96
	LSAppa	Semantic similarity (at the paragraph level)	Coh-Metrix	10.24
	future	Future tense	LIWC	10.00
	INFi	The use of infinitives	Coh-Metrix	-9.86

	3restraint	Interaction effects	Interaction	-9.85
	Period	Period	LIWC	-8.50
2proof	friend	Words about friend and neighborhood	LIWC	8.55
	health	Words about health	LIWC	-5.02
	swear	Swear words	LIWC	4.84
	LEXDIVVD	Lexical diversity	Coh-Metrix	4.21
	CAUSV	Causal inference	Coh-Metrix	4.16
	past	Past tense	LIWC	4.08
	percept	Perceptual words	LIWC	3.94
	see	Perceptual words	LIWC	3.61
	sad	Sadness	LIWC	-3.75
	space	Spacial words	LIWC	3.23
3restraint	CONCAUSi	Causal connectives	Coh-Metrix	-18.78
	future	Future tense	LIWC	-18.51
	TEMPtaa	Temporal cohesion	Coh-Metrix	-18.42
	LOCi	Location words	Coh-Metrix	-18.13

	POLm	Polysemy for content words, mean	Coh-Metrix	-16.12
	FRCLmcsm	CELEX Log minimum frequency for content words, mean	Coh-Metrix	15.43
	TEMPtrs	Tense repetition	Coh-Metrix	14.49
	affect	Affective process	LIWC	14.27
	CRFBA1um	Argument Overlap	Coh-Metrix	-12.97
	money	Money words	LIWC	-12.42
4agreement	WRDAacwm	Age of acquisition for content words, mean	Coh-Metrix	10.09
	death	Death words	LIWC	8.32
	sexual	Love words	LIWC	7.76
	posemo	Positive emotion	LIWC	6.80
	Numerals	Number	LIWC	6.15
	excl	Exclusive	LIWC	6.05
	READASL	Sentence length	Coh-Metrix	5.98

	ppron	Personal pronouns	LIWC	5.93
	anger	Anger words	LIWC	5.74
	AllPct	All punctuations	LIWC	5.65
5appreciation	cause	Causal inference	LIWC	12.59
	CONADDi	Additive connectives	Coh-Metrix	11.67
	posemo	Positive emotion	LIWC	11.48
	nonfl	Nonfluencies	LIWC	-11.15
	HYPm	Hypernymy for nouns and verbs, mean	Coh-Metrix	10.51
	GERUNDi	Gerund density, incidence	Coh-Metrix	-10.22
	hear	Perceptual words	LIWC	10.05
	READFKGL	Reading easiness	Coh-Metrix	-10.00
	WRDAacwm	Age of acquisition for content words, mean	Coh-Metrix	9.77
	pronoun	Pronoun	LIWC	-9.66
6self_reflection	verb	Verb	LIWC	13.80

	SYNLE	Number of words before the main verb	Coh-Metrix	12.67
	READL2	Second language comprehension	Coh-Metrix	11.62
	humans	Social processes	LIWC	10.58
	READFRE	Reading easiness	Coh-Metrix	10.33
	incl	Inclusive	LIWC	9.82
	4agreement	Interaction effects	LIWC	9.54
	TEMPars	Aspect repetition	Coh-Metrix	8.61
	READNP	Number of paragraphs	Coh-Metrix	8.54
	friend	Friend words	LIWC	8.48
7perspective_taking	HYPm	Hypernymy for nouns and verbs, mean	Coh-Metrix	16.19
	feel	Perceptual words	LIWC	15.56
	posemo	Positive emotion	LIWC	15.18
	SPATmpi	Motional preposition	Coh-Metrix	11.17

	leisure	Personal words	LIWC	10.46
	sexual	Love words	LIWC	10.07
	TYPTOKc	Vocabulary variation	Coh-Metrix	-9.95
	DENPRPi	Personal pronoun	Coh-Metrix	-9.89
	DENSPR2	Ratio of pronouns to noun phrases	Coh-Metrix	-9.41
	verb	Verb	LIWC	8.82
8monitoring	TEMPta	Tense and aspect repetition	Coh-Metrix	-13.45
	PRO1i	First person pronoun	Coh-Metrix	-11.32
	CRFBS1um	Stem overlap, adjacent sentences	Coh-Metrix	11.29
	READAPL	Paragraph length	Coh-Metrix	10.56
	verb	Verb	LIWC	10.52
	LSAppd	Semantic similarity (adjacent paragraphs)	Coh-Metrix	-10.13

	LSAassd	Semantic similarity (adjacent sentence)	Coh-Metrix	9.72
	AllPct	All punctuations	LIWC	9.35
	LSAassa	Semantic similarity (all sentences in a paragraph)	Coh-Metrix	-8.82
	funct	Functional words	LIWC	8.78
9balance	CRFBAaum	Argument overlap	Coh-Metrix	6.57
	Exclam	Exclamation marks	LIWC	5.14
	6self_reflection	Interaction effects	Interaction	5.11
	WRDIacwm	Imaginability for content words, mean	Coh-Metrix	4.02
	SPATlpi	motional preposition	Coh-Metrix	3.71
	DENSPR2	Ratio of pronouns to noun phrases	Coh-Metrix	3.70

	cogmech	Cognitive processes	LIWC	3.36
	they	They	LIWC	3.23
	body	Biological process	LIWC	3.21
	LSAppa	Semantic similarity (at the paragraph level)	Coh-Metrix	2.97
10plan	LSAassd	Semantic similarity (adjacent sentences)	Coh-Metrix	11.94
	WRDAacwm	Age of acquisition for content words, mean	Coh-Metrix	8.61
	sexual	Love words	LIWC	7.42
	Numerals	Numbers	LIWC	5.96
	discrep	Discrepancy	LIWC	5.54
	AllPct	All punctuations	LIWC	5.41
	ingest	Biological process	LIWC	5.04
	CAUSWN	Wordnet word overlap	Coh-Metrix	-4.99

	ppron	Personal pronouns	LIWC	4.92
	READASW	Average Syllables per word	Coh-Metrix	-4.72
11other	i	First person pronoun singular	LIWC	21.84
	TEMPtrs	Tense repetition	Coh-Metrix	21.28
	tentat	Tentative	LIWC	19.22
	MEDwtm	Minimum edit distance of all words	Coh-Metrix	-19.22
	CONCAUSi	Causal connectives	Coh-Metrix	-19.08
	CONi	All connectives	Coh-Metrix	-18.63
	cause	Causal inference	LIWC	17.88
	8monitoring	Interaction effects	Interaction	17.75
	CONLOGi	Logical connectives	Coh-Metrix	-17.58
	humans	Social processes	LIWC	17.18

5.5 Conclusion and Future Work

In this chapter, we present a new multi-task formulation with a novel composite regularizer, called relaxed Sparse-group Lasso (RSGL), for simultaneously identifying multiple intelligence-embodied communication skills using lexical, discourse, and interaction features. The key merit of RSGL is that it is a general multi-task formulation that unifies many widely used regularization techniques, including Lasso, group Lasso, sparse-group Lasso, and the Dirty model. Moreover, RSGL can be applied to streaming data to perform large scale analysis in real time. Empirical results show that, as a more general framework in multi-task learning, RSGL does not sacrifice performance. In fact, RSGL outperforms state-of-the-art multi-tasking learning formulations on prediction sensitivity, specificity, accuracy, and feature compression capacity in an online negotiation context. Future studies involving cross-validations and complying with our three experiment design principles may be used to study the performance variance of each model and provide evidence for the significance of performance improvement of RSGL. Finally, RSGL also revealed psycholinguistic and interaction characteristics of each of the intelligence-embodied communication skill that, to a great extent, resonate with human understanding.

We note that multi-task classification problems are by no means exclusive for understanding intelligence-embodied skills. The developed model is general enough to be applied to any other domains where the research interest includes predicting multiple interrelated labels simultaneously, including signal processing, computer vision, and computational neuroscience.

In future work, we will apply RSGL to more online contexts and data sets where people are from diverse culture backgrounds with a hope to explore the effect of culture differences on peoples' communication intelligence. In addition, we will extend RSGL with the ability to handle tasks embedded in a graph structure. Our experimental results have shown that certain skills attract the use of other skills.

This inter-dependency between skills can be well captured in a graphical framework. Therefore, multi-task learning on identifying multiple intelligence-embodied communication skills with graph structures is a promising direction to explore. Moreover, lexical and discourse features provide a good starting point for the initial exploration of linguistic manifestation of intelligence-embodied skills. The next steps will include incorporating to the model other types of features, such as semantic features to see whether prediction performance can be improved. Although less common, another possibility is to use word features in RSGL and compare it with the results from the previous chapter. The benefits of using word features is to save feature generation time by using LIWC and Coh-Metrix systems for real-time applications. It is also possible to use CL-LDA (the model from the previous chapter) with linguistic and interaction features when a mapping can be appropriately designed between those features and sentences/words in the context of text analysis.

CHAPTER 6

UNDERSTANDING COMMUNICATION INTELLIGENCE AND ITS EMBODIED SKILLS THROUGH SOCIAL NETWORK ANALYSIS

In previous chapters, we presented novel contributions of multi-task learning and multi-label learning, with applications to simultaneously identifying multiple intelligence-embodied communication skills from online dialogues. Those novel models provide keen insights into the language characteristics (Chapter 4) as well as lexical, discourse, and interaction characteristics (Chapter 5) of each intelligence-embodied communication skill. In this chapter, we aim to deepen our understanding of the nature of intelligence-embodied communication skills by analyzing the structure properties (e.g., degree, clustering coefficient) of participants' interactions in a social communication network context.

6.1 Motivation and Related Work

Understanding communication intelligence and its embodied communication skills requires multiple-perspective analysis of participants' online behaviors. Because online interactions generally take place in the form of natural language, analyzing human languages naturally becomes the first step in evaluating the language and linguistics landscapes of intelligence-embodied communication skills. From a different perspective, examining the conversational structure of online communication, such as who talks to whom and how such interactions form a social network diagram, provides an opportunity to understand communication intelligence from the lens of social interaction patterns.

Previous research in conversational analysis [137] has explored the structure of interactions in a communication phenomenon, called turn-taking. That line of research was concerned with the systematic analysis of turn-turning, such as when interruptions occurred and how repairs were signaled. Another line of research in online deliberation has studied up-taking [153], whose main goal was to reveal participants' roles and contributions (e.g., who are the central actors in the discussions and what ideas (from whom) receive the most development). Little research has attempted to characterize a group of high-order social constructs, such as intelligence-embodied communication skills, from the perspective of the structure of interactions. The present research as described in this chapter focuses on addressing the following intriguing question: *what are the network signatures of intelligence-embodied communication skills?*

The remaining of this chapter is organized as follows. In Section 6.2, we describe two experimental domains and provide analysis on comparing communication intelligence scores across communication modalities and across genders. In Section 6.3, we use canonical correlation analysis to test the hypothesis that *statistically significant correlations exist between participants' communication skills and network metrics.* We summarize our findings in Section 6.4.

6.2 Data and Experiments

We collected activity logs of participants' communication from two different communication modalities (i.e., synchronous communication without facilitation vs. asynchronous communication with facilitation) within classroom experimental trials, whose goal was to assess the effectiveness of educational software tools for supporting high-order communication skills. Subjects in the experimental trials were college students of two different majors. The students were engaged in computer mediated communication discussing about ill-defined topics. In the synchronous communication mode, students took part in the discussion at the same time from difference places; whereas in the asynchronous communication mode, students joined he discussion at different times and places. In the synchronous communication mode, the total number of contributing messages was 489, the average number of posts per participant was 19, and the number of words per post was 26. The 25 participating students were from a mixed majors of communication and pre-law. There were 52% females and 48% males and 92% of the students were juniors/seniors. The topic of their discussion was "right to die." In the asynchronous communication mode, the total number of contributing messages was 93, the average number of posts per participant is 5, and the number of words per post is 80. The 19 participating students were from the communication major. There were 58% were females and 42% were males and 63% of the students were juniors/seniors. The topic of their discussion was "internet free speech." Professional mediators were only present remotely and facilitated the discussion in the asynchronous communication mode. Facilitators' messages were excluded for this study.

One trained human judge annotated this data based on Murray's theory about social deliberative skills [116]. We aggregated appropriate social deliberative skills to construct each intelligence-embodied communication skill at the message level. We computed the overall communication intelligence for each participant across their messages in the whole discussion based on the measuring protocol introduced in Chapter 3.

6.2.1 Understanding the Effects of Communication Modalities on Communication Intelligence

We computed descriptive statistics of the average level of participants' communication intelligence and the average use of each skill across communication modalities. As shown in Figure 6.1, participant's communication intelligence is scored 47.1%

higher in an asynchronous and facilitated communication mode (0.25) than is in a synchronous and unfacilitated communication mode (0.17). Similar patterns were also found in each of the ten intelligence-embodied communication skills, with only one exception – the skill of *appreciation* is shown to be used slightly more often in a synchronous and unfacilitated mode (0.04) than is in an asynchronous and facilitated mode (0.03). We found that the asynchronous and facilitated communication mode is related to the creation of longer posts, which may partly contribute to the use of more intelligence-embodied skills. In addition, although in both communication modalities most of the participants were females and most participants were juniors/seniors, participants' majors are not same, which may, to some degree, contribute to the differences in participants' communication intelligence scores in different modalities. We think this effect is likely mitigated and absorbed by the facilitation element of the mode of the communication. In other words, participants' higher communication intelligence is more likely attributed to the fact that they were situated in a "facilitated" discussion of an asynchronous mode than to the fact that they were majoring in communication studies (vs. pre-law). This is because the communication skills of students within the same major may also vary to some degree. To test this hypothesis, we run a simple analysis and the results have shown that the communication intelligence scores of students from the same major is scored consistently higher in an asynchronous and facilitated communication mode than is in a synchronous and unfacilitated communication mode. (Similar patterns were also found for scores of all ten intelligence-embodied skills.) These results provide evidence that the asynchronous and facilitated communication mode may be a valuable means for promoting a deliberative communication as measured by communication intelligence. In addition, asynchronous interactions may benefit facilitators in a way that facilitators are less likely to be caught up in the immediacy of response and therefore having more time to reframe participant's sentiments in a more productive way. In future work

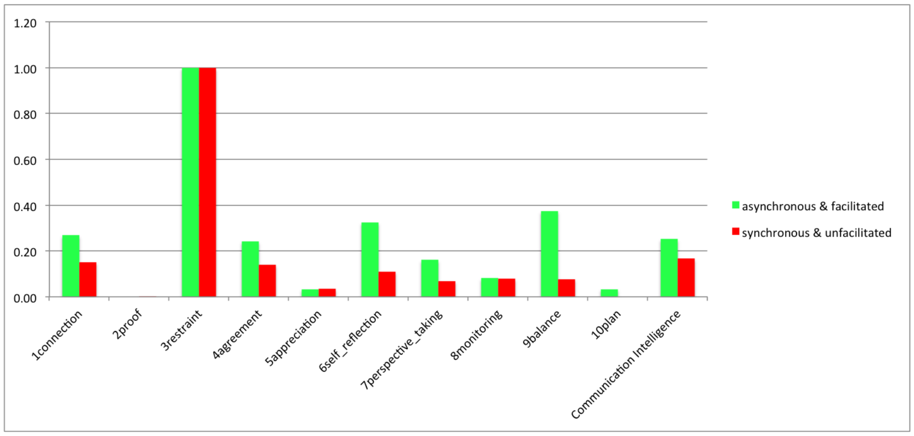

Figure 6.1. A comparison of the scores of communication intelligence and its embodied skills across communication modalities (In the asynchronous & facilitated communication mode the discussion topic was "internet free speech;" in the synchronous & unfacilitated communication mode the discussion topic was "right to die.")

we will study the effect of communication modalities on communication intelligence by controlling the topic variable. In addition, rather than use asynchronous and facilitated communication modality as a joint experimental condition, we will improve our experimental design to allow for studying the separate effects of communication modality (i.e., synchronous, asynchronous) and intervention (i.e., facilitation) on communication intelligence.

It can also be seen from Figure 6.1 that, in both communication modalities, participants are shown to use the skill of *proof* the least often. In the asynchronous and facilitated communication mode, participants are shown to use the skill *balance* the most, when discussing "internet free speech." In the synchronous and unfacilitated communication mode, participants are shown to use the skill *connection* the most, when discussing "right to die."

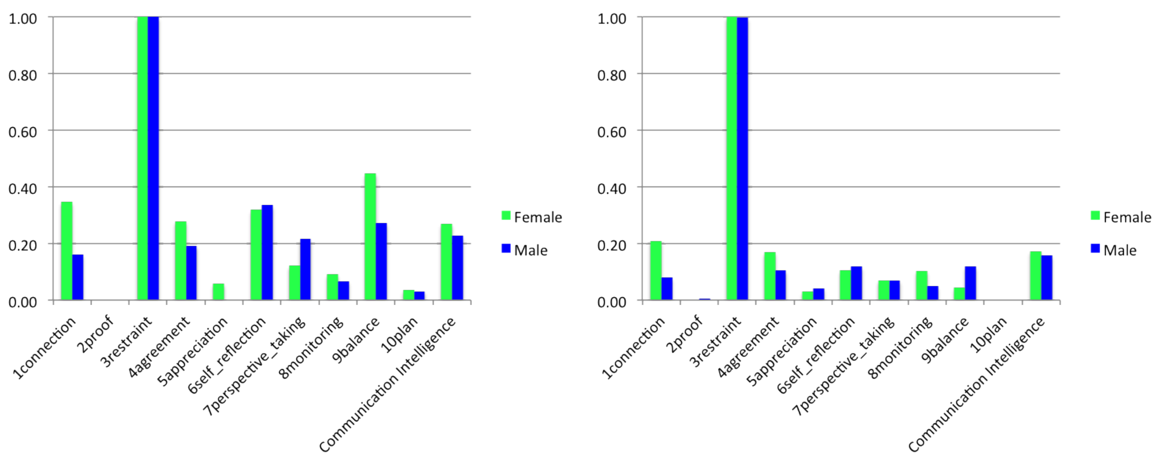

Figure 6.2. A Comparison of the scores of communication intelligence and its embodied skills across gender: the asynchronous and facilitated communication mode with topic "internet free speech" (left panel), the synchronous and unfacilitated mode with topic "right to die" (right panel)

6.2.2 Understanding the Gender Difference of Communication Intelligence

In this section, we examine the gender difference in the use of high-order communication skills and associated communication intelligence. As shown in Figure 6.2, females' communication intelligence scores are consistently higher than those of males across communication modalities. This finding is consistent with the recent research from MIT which concludes that adding women to a team helps improve group performance [167]. Moreover, the gender gap in communication intelligence appears to be widened in the case of asynchronous and facilitated mode. One possible explanation for this observation is that females are more sensitive to the external environment and more adaptive in response to external stimuli in communication, such as distance, the passage of time, and facilitated interventions. When comparing the use of each communication skill across genders, we found that males appear to use *self-reflection* slightly more often than females, regardless of communication modalities. In the asynchronous and facilitated mode, males are also shown to use *perspective* *taking* more often than females. In the synchronous and unfacilitated mode, males are shown to use *appreciation* and *balance* more often than females. These interesting observations shed light on which communication modality works the best for each gender to stimulate the use of which communication skills.

6.2.3 Studying the Relationship Between Intelligence-embodied Communication Skills and Social Interaction Patterns

To decode the link between intelligence-embodied skills and social interaction patterns, we extracted the structural information of "reply-to" from participants' activity logs. We used the "reply-to" feature because it provides important evidence about participants' social choices in online communication, which, to some degree, can speak about their communication intelligence. We measure participant's communication network quantitatively and generate a social network profile for each participant. The profile includes a group of network metrics, including degree, weighted degree, and eccentricity, as shown in Table **??** (also shown in the chapter about Background).

Table 6.1: Social network measures and their interpretations in the context of this research

Network structure measures	**Definition**
In_Degree	This metric indicates the number of people, from whom a message is sent to the studied participant.
Out_Degree	This metric indicates the number of people, to whom a message is sent from the studied participant.

Degree	This metric indicates the total number of people that the studied participant has communication with.
Weighted_In_Degree	This metric indicates the number of messages received by the studied participant.
Weighted_Out_Degree	This metric indicates the number of messages sent by the studied participant.
Weighted_Degree	This metric indicates the total number of messages both received and sent by the studied participant.
Eccentricity	This metric indicates the length of the longest directed path (assuming it is the only path) between the studied participant and another participant.
Closeness_Centrality	This metric indicates the average length of the directed path between the studied participant and another participant.
Betweenness_Centrality	This metric indicates on average how possible the studied participant is in the middle of a direct chain between any two other participants.
Authority	This metric indicates how influential the studied participant is.

Hub	This metric indicates how popular the studied participant is.
Modularity_Class	This metric indicates how sophisticated the communication network's internal structure is.
PageRank	This metric indicates on average how influential the participants who send messages to the studied participants are.
Component_ID	This metric describes community.
Strongly_connected_ID	This metric describes how closely members of the community, to which the studied participant is belong, interact.
Clustering_Coefficient	This metric indicates how closely the neighborhoods of the studied participant interact.
Eigenvector_Centrality	This metric also indicates on average how influential are the participants who send messages to the studied participant.

In the following sections, we will present analyses for understanding the associations between intelligence-embodied communication skills and participants' network metrics. In order to find patterns that hold true across communication modalities, we aggregate the data for the analysis below.

6.3 Research Method

6.3.1 Regularized Canonical Correlation Analysis

There are several measures of correlation to express the relationship between two or more variables. For example, Pearson correlation [93] measures the extent to which two variables are related; multiple regression [38] measure the relationship between a dependent variable and a set of independent variables; multivariate regression [103] computes how two sets of variables are associated. Canonical correlation analysis (CCA) [76] is a method for exploring the relationships between two sets of variables, all measured on the same experimental unit. CCA is both a regression technique and a dimension reduction technique – *it determines the relationship between two sets of variables and computes how many dimensions are necessary to understand the association between them.* CCA is different from other dimension reduction techniques, such as principal component analysis [82] and factor analysis [70], because those two techniques examine relationships within a single set of variables, whereas CCA looks at the relationship between two sets of variables.

CCA finds its limitations in applications where multicolliearity [1] is present within either or both sets of variables, or the number of experimental units is less than the number of measuring variables. To efficiently address these limitations, regularized canonical correlation analysis (RCCA) [69] is developed by imposing a ridge penalty [72] (as discussed in the chapter about Background) to CCA. In this research, we use RCCA to identify associative patterns between participants' use of intelligence-embodied skills and their network metrics.

Below we first introduce some key concepts in RCCA.

- **Canonical variate (dimension)**: Canonical variate is the weighted sum of a set of variables, a form of a latent variable. For each canonical correlation be-

[1] Multicollinearity refers to the situation where one or more variables are linearly dependent on other variables.

tween two sets of variables, there are two canonical variates, each corresponding to one set of variables. For example, suppose we have two sets of measuring variables X and Y.

Set 1 has p variables: $\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{pmatrix}$, and

set 2 has q variables: $\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_q \end{pmatrix}$

We look at linear combinations of the data, as in principal components analysis. Canonical variate U corresponds to the linear combinations of the first set of variables X, and canonical variate V corresponds to the linear combinations of the second set of variables Y. For example, U_1 below is a linear combination of the p variables in X and V_1 is the corresponding linear combination of the q variables in Y. Similarly, U_2 is a different linear combination of the p variables in X, and V_2 is the corresponding linear combination of the q variables in Y.

$$\begin{aligned} U_1 &= a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\ U_2 &= a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\ &\vdots \\ U_p &= a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p \end{aligned}$$

$$\begin{aligned} V_1 &= b_{11}Y_1 + b_{12}Y_2 + \cdots + b_{1q}Y_q \\ V_2 &= b_{21}Y_1 + b_{22}Y_2 + \cdots + b_{2q}Y_q \\ &\vdots \\ V_q &= b_{q1}Y_1 + b_{q2}Y_2 + \cdots + b_{qq}Y_q \end{aligned}$$

Each member of U is paired with a member of V. The goal of RCCA (and CCA) is to find weights in the linear equations so as to maximize the correlation between canonical variate U and V.

- **Wilks's lambda significant test**: The number of canonical correlations between two sets of variables is determined by the number of variables in the smaller set. In other words, the maximum number of canonical variate pairs is the same as the number of variables in the smaller set. The number of canonical variate pairs is often referred to as dimension. Wilks's lambda is commonly used to test the significance of each dimension. Specifically, it determines how many dimensions are needed to account for the relationship between canonical variates. For example, in the case of only one significant dimension, the relationship between two set of variables can be easily examined in a one dimensional space. When more than one significant dimension is present, the first dimension is always the one which explains the most of the relationship, followed by the second dimension, and so on. The canonical correlation is interpreted in the same way as in Pearson's correlation – the square of the correlation is the percent of shared variance along that dimension. For example, a canonical correlation of 0.9 in the first dimension represents 81% of the shared variance between the two sets of variables. In other words, the effect size of the canonical correlation is 0.81.

- **Canonical coefficients**: Canonical coefficients are used to assess the relative importance of individual variables' contributions to a given canonical correlation. Canonical coefficients are the weights in the linear equation of variables that creates the canonical variates. In general, the larger the absolute value of the weight, the greater is the respective variable's unique positive or negative contribution to the weighted sum (canonical variate). To facilitate comparisons

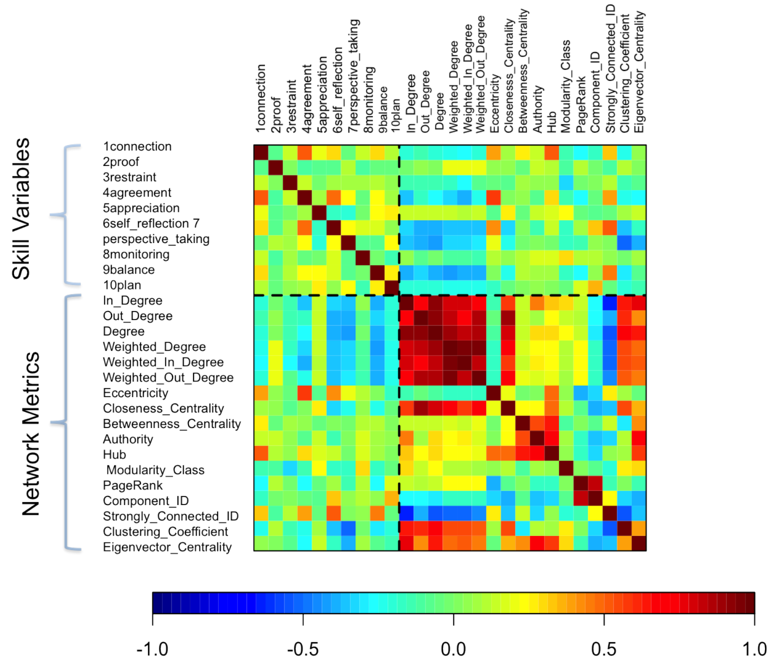

Figure 6.3. Pearson correlations of two variables (1) within the set of skill variables (upper left corner), (2) within the set of network metrics (lower right corner), and (3) between the two sets (lower left and upper right corner)

between weights, the standardized canonical weights are usually used for interpretation (i.e., using the z transformed variables with a zero mean and unit standard deviation).

6.4 Experiments and Results

6.4.1 Understanding the Association Between Intelligence-Embodied Communication Skills and Network Metrics

Before we perform canonical correlation analysis, we first want to determine if any relationship exists between communication skill variables and network metric variables. To test for independence, we performed Pearson correlation analysis for these two sets of variables. In Figure 6.3, we reveal the correlation within communication skill variables, within network metric variables, and between these two sets of variables. For example, we found that *agreement* has a high positive correlation

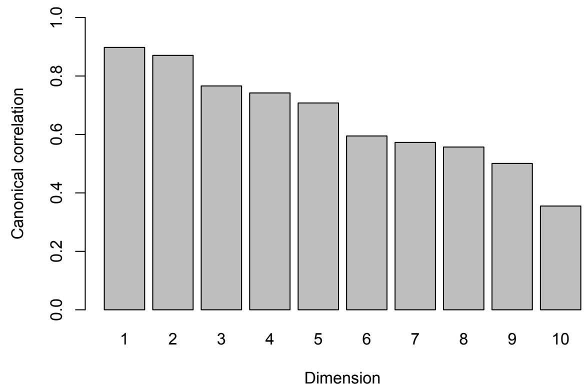

Figure 6.4. Canonical coefficients of each dimension for the correlation between skill variables and network metrics

with *connection* and *self-reflection*, which implies that, for example, people who use the skill of *agreement* also often use the skill of *self-reflection*. Within network metric variables, we found that, for example, eccentricity has a positive correlation with hub and a negative correlation with page rank. More importantly, we are reassured that correlations appear to exist between communication skill variables and network metric variables (i.e., cells in the lower left corner in Figure **??** are not uniformly green). Given that multicollinearity is present in both sets of variables, we move forward with regularized canonical correlation analysis.

With regularized canonical correlation analysis, ($\lambda_1 = 0.0001$ for communication skill variables and $\lambda_2 = 0.00001$ for network metric variables), we found one statistically significant (α=0.1) canonical dimension. This canonical dimension has a canonical correlation of 0.90 (p=0.08) and large effect size of 0.81. Figure 6.4 shows all 10 canonical dimensions.

Table 6.2 presents the standardized canonical coefficients for the first significant dimension across both sets of variables. For the communication skill variables, the first canonical dimension is most strongly influenced by *perspective taking* (0.63). Similarly, for the network metrics, the first dimension is mainly influenced by page

Table 6.2: Standardized canonical coefficients for the first dimension across skill variables and network metrics

Communication skill variables	Canonical coefficient
Average_of_1connection	0.24
Average_of_2proof	-0.04
Average_of_3restraint	0.01
Average_of_4agreement	0.35
Average_of_5appreciation	-0.04
Average_of_6self_reflection	0.13
Average_of_7perspective_taking	0.63
Average_of_8monitoring	0.11
Average_of_9balance	0.21
Average_of_10plan	0.01
Social network metrics	
In_Degree	0.12
Out_Degree	-0.02
Degree	0.20
Weighted_Degree	-0.06
Weighted_In_Degree	-0.34
Weighted_Out_Degree	0.31
Eccentricity	0.50
Closeness_Centrality	-0.12
Betweenness_Centrality	-0.04
Authority	0.22
Hub	0.38
Modularity_Class	-0.07
PageRank	-0.87
Component_ID	0.86
Strongly_Connected_ID	-0.36
Clustering_Coefficient	-0.52
Eigenvector_Centrality	-0.52

rank (-0.87), component ID (0.86). This result provides an exciting way to study *perspective taking* through the lens of social network metrics. Below, we present some highlights of the social dynamics of people showing more *perspective taking* behaviors.

- Popular – people showing more *perspective taking* behaviors are more popular (i.e., positive correlations with hub, degree) than others in their communication network.

- Influential – people showing more *perspective taking* behaviors are more influential (i.e., a positive correlation with authority). Their neighborhoods interact not much themselves (i.e., a negative correlation with clustering coefficient). They contribute to a large local community (i.e., a positive correlation with eccentricity) that has more communication (i.e., a correlation with strongly connected [2]).

- Like-attracts-like – people showing more *perspective taking* behaviors are more likely to communicate with people who behave similarly. Their communication network demonstrates propinquity and homophily. In other words, people tend to communicate with others who demonstrate similar level of perspective taking (i.e., a positive correlation between perspective taking and network component). Furthermore, they only interact with a subgroup of people rather than all people in the network (i.e., a positive correlation with weighted out degree and a negative correlation with closeness centrality). Most of the people who send messages to them are not influential (i.e., negative correlations with page rank and eigenvector eccentricity). This finding is consistent with the findings from

[2]In Gephi, component ID and strongly connected ID are used to name communities. There is no specific meaning about the number itself. The high *absolute* value of the correlation between perspective taking and component ID indicates people doing perspective taking tend to form a community. The high *absolute* value of the correlation between perspective taking and strongly connected ID indicates people doing perspective taking tend to contribute to a community that has more communication.

Chapter 4, where people showing perspective taking behaviors (measured from their discourse) in an online message were observed to attract other people to match the same behavior in the message immediately following it.

6.5 Summary

In this chapter, we first studied communication intelligence across communication modalities and across genders. We showed that participants' communication intelligence is scored 47.1% higher in an asynchronous and facilitated communication mode (0.25) than is in a synchronous and unfacilitated one (0.17). This observation provides evidence that the asynchronous and facilitated communication mode may be a valuable means for promoting a deliberative communication as measured by communication intelligence. We further showed that females' communication intelligence scores are consistently higher than those of males across communication modalities. Moreover, the gender gap in communication intelligence appears to be widened in the case of asynchronous and facilitated mode. One possible explanation for this observation is that females are more sensitive to the external environment and more adaptive in response to external stimuli in communication, such as distance, the passage of time, and facilitated interventions. When comparing the use of each communication skill across genders, we found that males are shown to use *self-reflection* slightly more often than females, regardless of communication modalities.

We then studied the association between intelligence-embodied communication skills and participants' social network metrics using regularized canonical correlation analysis. We showed that a statistically significant correlation exists between intelligence-embodied communication skills and social network metrics with a large effect size of 0.8, which provides an opportunity to characterize the skill of *perspective taking* from the perspective of social interaction patterns. Specifically, people showing more *perspective taking* behaviors are found to be more popular and influential than others in their communication network. They also tend to reach out to people who behave similarly, which implies a like-attracts-like social phenomenon that complies with the Law of Attraction. These observations complement discoveries from previous chapters about the linguistic manifestation of intelligence-embodied communication skills with social dynamic characteristics.

Future research will expand this study along several dimensions. To further validate of our results, we will replicate the above experiments with a larger sample size. In addition, in the present research, the experimental subjects are college students mostly from the United States. We need to study a larger sample of populations and possibly from diverse cultures to reach solid conclusions. We would also like to improve our experimental design by controlling the topic effect on communication intelligence and by allowing for studying the separate effects of communication modality (synchronous vs. asynchronous) and intervention (i.e., facilitation) on communication intelligence.

CHAPTER 7

CONCLUSION AND FUTURE WORK

In this chapter, we summarize our contributions and suggest possible directions for future research.

7.1 Summary of Contributions

In this dissertation, we have advanced the state of the art in communication studies and machine learning. Generally speaking, our contributions fall into two realms:

- We pioneered the study of a communication phenomenon: *communication intelligence*, in the world of online interactions. We created the first theory about communication intelligence that measures the use of ten high-order communication skills, including *connection*, *proof*, *restraint*, *agreement*, *appreciation*, *self-reflection*, *perspective taking*, *monitoring*, *balance*, and *plan*. We presented a multi-perspective analysis for understanding communication intelligence, including its diverse language, shared linguistic characteristics across people, social dynamics, and the effects of communication modality on communication intelligence.

- We contributed new machine models and formulations for addressing multi-label and multi-task machine learning problems. Those models and formulations can simultaneously identify multiple intelligence-embodied communication skills from natural language, linguistic features, and interaction features. Beyond these applications, they can also benefit research in other areas where the problems of simultaneously predicting multiple categories are abundant. These areas include, but are not limited to, signal processing, computer vision, computational finance, computational biology, and computational neuroscience.

Regarding the new theory about communication intelligence, in Chapter 3, we presented an intellectual model of communication intelligence comprising ten interrelated actionable dimensions/skill constructs. These ten dimensions keep a good balance of acknowledging the different orientations (relationship vs. tasks) among people. We also described a key property of communication intelligence (i.e., dynamic and contextual) and introduced a method for measuring communication intelligence.

Regarding understanding the diverse language of communication intelligence, in Chapter 4, we presented a new hierarchical probabilistic model for addressing the problem of simultaneously identifying multiple intelligence-embodied communication skills from natural language. That model, called Constrained Labeled LDA (CL-LDA), learns the topic assignment for each sentence and provides a practical and simple way to determine document labels without relying on a threshold function. CL-LDA's performance increases as the number of labels grows, which makes CL-LDA a promising approach for large-scale data analysis. CL-LDA also has high interpretability and its annotated sentences significantly augment the view of each document with rich contextual information.

Regarding understanding the shared linguistic characteristics of communication intelligence across people, in Chapter 5, we presented a new multi-task formulation for simultaneously identifying multiple intelligence-embodied communication skills, from lexical, discourse, and interaction features. The key merit of this model is that it is a general multi-task formulation that unifies many widely used regularization techniques, including Lasso, group Lasso, sparse-group Lasso, and the Dirty model. This model expands the applicability of multi-task learning by allowing analyzing real-world problems where (1) the degree of task relatedness is uncertain and (2) the true structure of the groups in data is not clear ahead of time. Moreover, this model can be applied to streaming data to perform *real-time* analysis. Moreover, it can be applied to streaming data to perform large scale analysis in real time. It also reveals psycholinguistic and interaction characteristics of each intelligence-embodied communication skill that, to a great extent, resonate with human understanding.

Regarding understanding the social dynamics of communication intelligence and the effects of communication modalities on communication intelligence, in Chapter 6, we presented an advanced correlation analysis, called regularized canonical correlation analysis (RCCA), for studying the association between intelligence-embodied communication skills and social network metrics, measured on the same participant. RCCA finds a statistically significant correlation between intelligence-embodied communication skills and social network metrics (effect size=0.81), which provides an opportunity to characterize the skill of *perspective taking* from the perspective of social interaction patterns. Specifically, people showing more *perspective taking* behaviors are more popular and influential than others in their communication network. They also tend to reach out to people who behave similarly, which implies a like-attracts-like social phenomenon that complies with the Law of Attraction. We furthermore showed that participants' communication intelligence is on average scored significantly higher in an asynchronous and facilitated communication mode than is in a synchronous and unfacilitated mode. Females' communication intelligence scores are shown to be consistently higher than those of males across communication modalities.

7.2 Future Work

7.2.1 Modeling Multi-modal Data With Tensor Decomposition

A tensor [86] is a multidimensional array. More formally, an N-way or Nth-order tensor is an element of the tensor product of N vector spaces, each of which has its own coordinate system. Tensor decomposition is a technique that applies decomposition to data arrays for extracting and explaining their properties. There has been active research on developing tensor decomposition algorithms and models, including CP [29], Tucker decomposition [156], and PARAFAC2 [71]. Tensor decomposition has shown to be an effective technique for feature extraction, classification, dimensionality reduction, and multiway clustering [86]. Decomposition of higher-order tensors (i.e., N-way arrays with $N \geq 3$) has found wide applications in psychometrics, signal processing, computer vision, data mining, and neuroscience [86, 35]. In these applications, data are often in three or more modes, and therefore, a two diminutional matrix is not sufficient for the purpose of data representation. For example, in data mining, web graph data have three modalities: source, destination, and text. Similarly, in neuroscience, brain imaging data have three modalities: channel, frequency, and time frame. In computer vision, face recognition data have four modalities: people, expression, head pose, and illumination. Current substitutes for tensor techniques include separating the data through one dimension (so that the rest two dimensions can be represented with matrix) and studying them separately, or aggregating the data all together along one dimension and studying them as a whole. These work-around approaches miss the opportunity of studying the split (or aggregate) dimension as part of the data and therefore exclude the possibility to discover the interactions between the split (or aggregate) dimension and other dimensions.

Tensor decomposition is a promising method to further our understanding of communication intelligence, as we collect data from more perspectives in the future. For example, we may consider semantic features as a feature dimension in addition to lexical, discourse, and interaction features for studying the linguistic manifestation of communication intelligence. We may also consider collecting data about participants' personality types [40] and reflective judgment stages [85] in addition to social network metrics for studying the psychosocial characteristics of communication intelligence. Moreover, we will also collaborate with researchers from other disciplines to deepen our understanding of communication intelligence from other perspectives. For example, research in psychology [176] has started to study visual social cognition. They monitored and analyzed how people move their eyes as they perform perspective taking and measured how such intimate behavior changes in space and time. Research in social neuroscience [42] has also studied the link between mental states and social behaviors. They performed functional-imaging studies to understand how people regulate emotional process and show empathy. They created a functional model that highlighted the role of specific brain regions responsible for empathetic behaviors. Similar visual and neuroimaging data are useful to broaden our perspectives and deepen our understanding of intelligence-embodied communication skills. Eventually, we will combine tensor decomposition with multi-task learning to create multi-modal multi-task learning models to jointly predict multiple intelligence-embodied communication skills from multi-modal data. We are also interested in developing multi-modal canonical correlation analysis model to explore multi-way multivariate correlations among different perspectives (i.e., modalities).

7.2.2 Building an Intelligent Tutoring System for Deliberative Communication

Research in the field of intelligent tutoring system (ITS) has been developing interactive education systems to provide personalized scaffolding on knowledge learning [7] and skill improvement [2]. The research on ITS has largely focused on improving students cognitive abilities [7], metacognitive skills [2], and inducing the optimal emotional state for learning [31]. Little research has attempted to develop an intelligent tutoring system for supporting effective communication. In the field of human-computer interaction, some recent research has designed automated personal agents [75] to help people conquer fear and increase self-confidence in pressured social interaction situations, such as public speaking and job interviews. In future work, we will work toward creating an ITS for deliberative communication with the goal of supporting large-scale, multiple-party online communication.

A key component of an ITS is its student model, which is often used to guide pedagogical decision makings, such as providing feedback or hints. In order to inform a pedagogical model, the student model needs to recognize incorrect student answers by checking against a domain model. Similarly, an ITS for deliberative communication needs to have an *intelligent interlocutor model* and a *facilitator model*. The main purpose of the interlocutor model is to inform decision-makings of the facilitator model. The model presented in Chapter 4, regularized sparse-group Lasso (RSGL), is a good candidate for an intelligent interlocutor model in an ITS for deliberative communication. This is because it can, in real time, identify high-order communication skills being used or not used by each participant and provide evidence about the linguistic characteristics associated with each skill use, which guide the facilitator model. For example, within a time window, if the majority of participants did not use certain skills, or, if one participant did not use certain skills over an extended period of time, the facilitator model may be called to intervene. Ultimately, we will create an intelligent interlocutor model that can recognize interlocutors' absence of certain communication skills when called for, so it can provide even valuable information to the facilitator model suggesting when and how to intervene in real time.

7.2.3 Improving Communication Intelligence through Brain-based Education

The ultimate goal of this research is to support people in improving their communication intelligence. The need for being an effective communicator resides at a higher tier in Maslows hierarchy of needs [105], which is often portrayed as a pyramid with the most fundamental level of needs (e.g., breathing, food) at the bottom, and the need for esteem (e.g., respect others) and self-actualization (e.g., acceptance of facts) at the top. Maslow has shown that a person's motivation generally moves through this pyramid, implying that the motivation to achieve goals at the higher tiers of the pyramid is often lacking or not sufficient.

A vast literature in human psychology and motivation [136] has shown that external rewards are not as effective as are *intrinsic needs or desires.* The Google way is an effective example. At Google, people are encouraged to take 20 percent of their time (the Google Way) to work on company-related projects that speak to their own desires. And some of those personal projects have now become Google's most popular products, including Gmail and Google news. In addition, as Maslow's hierarchy of needs suggest, motivation works in a highly individualized fashion– what motivates one individual may not motivate others.

In future research, we plan to work toward creating an *effective motivation strategy* to improve people's communication intelligence, and propose to use brain-based education to support people in improving their communication intelligence. Specifically, we will study the *behavior-brain-benefit connection* by illustrating the *benefits* of using certain intelligence-embodied skills in relation to state-of-the-art research in brain studies. For example, we will study *the neural signature of perspective taking* and research into the *benefits of using the the same part of the brain (e.g., happiness, longevity).* If these results are found, people would more likely be motivated to behave intelligently in communication, because living longer and being happy (as basic needs in the pyramid) are more likely to be their intrinsic desires. Therefore, this motivation strategy would work *universally* well across people.

The past two decades have witnessed a remarkable rise in the number of research published under the rubric of social neuroscience. Fruitful discoveries have been found about the neural basis of love, gratitude, trust, altruism, and empathy [52, 1, 134, 145]. For example, prefrontal cortex, the "executive" part of the brain, has been consistently shown to govern planning, impulse control, and empathy, closely related to some of the intelligent-embodied skills that we study, including *plan*, *restraint*, and *perspective taking*. Research in health psychology has also recently revealed a surprising truth about longevity. This eight-decade study [53] has discovered that health and long life are significantly correlated with living conscientiously, the use of forethought, planning, and perseverance in all aspects of one's life. This behavior-brain-benefit (i.e., planning – prefrontal cortex – longevity) connection shows a good promise for motivating people to improve their communication intelligence.

To motivate the use of some other intelligence-embodied communication skills, we still need to first understand which brain regions govern which behaviors. Measuring how the brain functions during an activity involves the use of brain-scanning instruments, such as functional magnetic resonance imaging (fMRI), magnetoencephalography (MEG), electroencephalography (EEG), and single-photon emission computed tomography (SPECT). These brain-scanning instruments have long been used to understand the relationship between brain function and behaviors [77, 135, 115, 94]. The recent availability of simple, low-cost, portable EEG monitoring devices (audio headsets) [119] makes it feasible to take EEG from the lab into schools, offices, and home for more widespread brain research. These portable headsets require no expertise to wear, and, although they record from only a single sensor with untrained users, they can distinguish two fairly similar mental states (neutral and attentive) with 86% accuracy [119]. These portable EEG monitoring devises are promising for studying brain activities of intelligence-embodied communication behaviors *in vivo*.

Moreover, recently the BRAIN Initiative (Brain Research through Advancing Innovative Neurotechnologies) [1] is announced by the Obama administration. This initiative has sets top priorities to study the *link between neuronal activity and behaviors*. This agenda creates a supportive environment for collaborating with other re-

[1]http://www.whitehouse.gov/share/brain-initiative

search institutes, funding agencies, and individuals to study brain activities related to intelligence-embodied communication skills. For example, in response to that initiative, the US Defense Advanced Research Projects Agency (DARPA) has developed another prototype of low-cost EEG devices aiming to spark a do-it-yourself (DIY) revolution in neuroscience in the society.

It is our hope that this dissertation can help launch a movement to improve communication intelligence with the goal of co-creating a respectful, deliberative, and thriving society. We believe motivating social behaviors through brain-based education can accelerate the progress of achieving this grand goal.

Now we have come to the end of this dissertation, and yet we just begin a journey to awaken the communication intelligence within us.

APPENDIX A

GIBBS SAMPLING DERIVATION OF CONSTRAINT LABELED LDA

Goal: Find posterior distribution over latent variables given the observed variables (omitting hyperparameters).

$$P(\theta, \phi, z | w, \Lambda) = \frac{P(\theta, \phi, z, w, \Lambda)}{P(w, \Lambda)}$$

Graphical model gives us:

$$\begin{aligned} P(\theta, \phi, z, w, \Lambda) &= P(\theta)P(\phi)P(\Lambda)P(z|\theta)P(w|z, \phi) \\ &= \prod_d Dir(\theta_d; \alpha) \prod_d P(\Lambda_d) \prod_k Dir(\phi_k; \beta) \prod_m \theta_{z_m|d_m} \\ &\quad \prod_n \phi_{w_n|z_n} \end{aligned}$$

We use collapsed Gibbs sampling to integrate out model parameters ϕ and θ, and just sample z.

Sample z for $P(z|w, \Lambda)$.

$$P(z|w, \Lambda) = \frac{P(z, w, \Lambda)}{P(w, \Lambda)}$$

Numerator:

$$
\begin{aligned}
P(z, w, \Lambda) &= \int \mathrm{d}\theta \int \mathrm{d}\phi P(\theta, \phi, z, w, \Lambda) \\
&= \int \mathrm{d}\theta \int \mathrm{d}\phi \prod_d Dir(\theta_d; \alpha) \prod_k Dir(\phi_k; \beta) \\
&\quad \prod_m \theta_{z_m|d_m} \prod_n \phi_{w_n|z_n} \prod_d P(\Lambda_d) \\
&= \int \mathrm{d}\theta \int \mathrm{d}\phi \prod_d Dir(\theta_d; \alpha) \prod_k Dir(\phi_k; \beta) \\
&\quad \prod_d \prod_k \theta_{k|d}^{N_{k|d}} \prod_k \prod_w \phi_{w|k}^{N_{w|k}+s_{wi}} \prod_d P(\Lambda_d) \\
&= \int \mathrm{d}\theta \prod_d \left[Dir(\theta_d; \alpha) \prod_k \theta_{k|d}^{N_{k|d}} \right] \\
&\quad \int \mathrm{d}\phi \prod_k \left[Dir(\phi_k; \beta) \prod_w \phi_{w|k}^{N_{w|k}+s_{wi}} \right] \prod_d P(\Lambda_d) \\
&= A \times B \times \prod_d P(\Lambda_d)
\end{aligned}
$$

where $A = \int \mathrm{d}\theta \prod_d \left[Dir(\theta_d; \alpha) \prod_k \theta_{k|d}^{N_{k|d}} \right]$, $B = \int \mathrm{d}\phi \prod_k \left[Dir(\phi_k; \beta) \prod_w \phi_{w|k}^{N_{w|k}+s_{wi}} \right]$.

Now we expand term A and B. Since we use symmetric Dirichlet priors, α and β are scalars.

Note that

$$
\begin{aligned}
\int \mathrm{d}\theta Dir(\theta_d; \{N_{k|d} + \alpha\}) &= 1 \\
\Longrightarrow \int \mathrm{d}\theta \frac{\Gamma(N_d + \sum_k \alpha_k)}{\prod_k \Gamma(N_{k|d} + \alpha_k)} \prod_k \theta_{k|d}^{N_{k|d}+\alpha_k-1} &= 1 \\
\Longrightarrow \frac{\Gamma(N_d + \sum_k \alpha_k)}{\prod_k \Gamma(N_{k|d} + \alpha_k)} \prod_k \int \mathrm{d}\theta_{k|d}^{N_{k|d}+\alpha_k-1} &= 1 \\
\Longrightarrow \prod_k \int \mathrm{d}\theta_{k|d}^{N_{k|d}+\alpha_k-1} &= \frac{\prod_k \Gamma(N_{k|d} + \alpha_k)}{\Gamma(N_d + \sum_k \alpha_k)}
\end{aligned}
$$

First, we start with A

$$
\begin{aligned}
A &= \int \mathrm{d}\theta \prod_d \left[Dir(\theta_d;\alpha) \prod_k \theta_{k|d}^{N_{k|d}} \right] \\
&= \prod_d \int \mathrm{d}\theta Dir(\theta_d;\alpha) \prod_k \theta_{k|d}^{N_{k|d}} \\
&= \prod_d \int \mathrm{d}\theta \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \theta_{k|d}^{\alpha_k - 1} \prod_k \theta_{k|d}^{N_{k|d}} \\
&= \prod_d \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \prod_k \int \mathrm{d}\theta_{k|d}^{N_{k|d}+\alpha_k - 1} \\
&= \prod_d \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{\prod_k \Gamma(N_{k|d} + \alpha_k)}{\Gamma(N_d + \sum_k \alpha_k)}
\end{aligned}
$$

Next, we look at B

$$
\begin{aligned}
A &= \int \mathrm{d}\phi \prod_k \left[Dir(\phi_k;\beta) \prod_w \phi_{w|k}^{N_{w|k}+s_{wi}} \right] \\
&= \prod_k \int \mathrm{d}\phi Dir(\phi_k;\beta) \prod_w \phi_{w|k}^{N_{w|k}+s_{wi}} \\
&= \prod_k \int \mathrm{d}\phi \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \prod_w \phi_{w|k}^{\beta_w - 1} \prod_w \phi_{w|k}^{N_{w|k}+s_{wi}} \\
&= \prod_k \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \prod_w \int \mathrm{d}\phi_{w|k}^{N_{w|k}+s_{wi}+\beta_w - 1} \\
&= \prod_k \frac{\Gamma(\sum_w \beta_w)}{\prod_w \Gamma(\beta_w)} \frac{\prod_w \Gamma(N_{w|k} + s_{wi} + \beta_w)}{\Gamma(N_k + s_i + \sum_w \beta_w)}
\end{aligned}
$$

Denominator: $P(w,\Lambda) = \sum_z P(z,w,\Lambda)$ requires Gibbs sampling. We use the full conditional $P(z_i|z_{-i},w,\Lambda)$ to simulate $P(z|w,\Lambda)$.

$$
\begin{aligned}
P(z_i|z_{-i},w,\Lambda) &= \frac{P(w,z,\Lambda)}{P(w,z_{-i},\Lambda)} \\
&\propto \frac{P(w|z,\Lambda)P(z,\Lambda)}{P(w_{-i}|z_{-i},\Lambda)P(z_{-i},\Lambda)} \\
&= \frac{P(w,z,\Lambda)}{P(w_{-i},z_{-i},\Lambda)}
\end{aligned}
$$

We know that $P(w, z, \Lambda) = A \times B \times \prod_d P(\Lambda_d)$. $P(w_{-i}, z_{-i}, \Lambda)$ is the same except with $N_{k|d} - 1, N_d - 1, N_{w|k} - s_{wi}, N_k - s_i$. Because $x\Gamma(x) = \Gamma(x+1)$, $\frac{\Gamma(x+1)}{\Gamma(x)} = x$. After canceling terms, we have

$$\frac{\frac{\prod_k \Gamma(N_{k|d}+\alpha_k)}{\Gamma(N_d+\sum_k \alpha_k)} \cdot \frac{\prod_w \Gamma(N_{w|k}+s_{wi}+\beta_w)}{\Gamma(N_k+s_i+\sum_w \beta_w)}}{\frac{\prod_k \Gamma(N_{k|d}+\alpha_k-1)}{\Gamma(N_d+\sum_k \alpha_k-1)} \cdot \frac{\prod_w \Gamma(N_{w|k}+\beta_w)}{\Gamma(N_k+\sum_w \beta_w)}} = \frac{N_{k|d}+\alpha_k}{N_d+\sum_k \alpha_k} \cdot \frac{\Gamma(N_k+\sum_w \beta_w)}{\Gamma(N_k+s_i+\sum_w \beta_w)} \cdot \prod_w \frac{\Gamma(N_{w|k}+s_{wi}+\beta_w)}{\Gamma(N_{w|k}+\beta_w)}$$

The posterior on θ and ϕ using the fact that the Dirichlet is conjugate to the multinomial.

$$\phi|z, w, \beta \sim Dir(N_k + \beta)$$

$$\theta|z, w, \alpha \sim Dir(N_d + \alpha)$$

Evaluating the posterior mean of θ and ϕ:

$$\mathrm{E}[\phi_{w|k}|z, w, \beta] = \frac{N_{w|k} + \beta_w}{N_k + V\beta_w}$$

$$\mathrm{E}[\theta_{k|d}|z, w, \alpha] = \frac{N_{k|d} + \alpha_k}{N_d + K\alpha_k}$$

APPENDIX B

UNDERSTANDING THE RELATIONSHIP BETWEEN THE CONSTRUCTS OF COMMUNICATION INTELLIGENCE AND SKILLS IN THE CONCEPTUAL SOCIAL DELIBERATIVE SKILL FRAMEWORK

Table B.1: The correspondence between intelligence-embodied communication skills and skills contained in the social deliberative skill framework

Communication Intelligence Skill Constructs	**Social Deliberative Skills (SDS) and Other Speech Acts**	**Definitions provided by SDS**
1. Connect	Q_INTERL	Asking questions to discover more about a single interlocutor's thoughts or feelings (curiosity)(see Stromer-Galley Directive Questions.
	REF_INTERL	Referencing what others said, including quoting, summarizing, and mentioning it.

2. Proof	FACT_SRC	Stating a fact and noting the source in the same post.
	SOURCE_REF	Mentioning a source, with a pointer or quote.
3. Restraint [1]	NEGEMO_INT	Negative emotion about interlocutors or dialog process.
4. Agreement	AGREE	Expressing agreement with interlocutors.
5. Appreciation	APPREC	Showing appreciation, gratitude, affirmation of another's idea or situation.
6. Self-reflection	SELF_REFL	Reflecting on (or commenting on) one's own assumptions, values, biases, or emotions.
7. Perspective taking	PERSPECTIVE	Putting oneself in another's shoes (of an interlocutor OR a group you are not a member of).
	OTHERS_THNK	Assessing or reflecting on the ideas, assumptions, values, biases of others (individuals and groups – generally outside of the dialogue).

[1] In this research, "restraint" is encoded as the opposite of NEGEMO_INT.

8. Monitoring	MEDIATE	Making suggestions about how interlocutors should communicate or how the conversation should proceed; redirecting conversation toward higher quality.
	META_CONS	Highlighting agreement or consensus on some point, for entire group or part of group (not just self and self and one other).
	META_CONFL	Highlighting disagreement on some point, for entire group or part of group (not just self and self and one other)
	META_SUM	Summarizing the conversation – may include both consensus and conflict.
	META_CHECK	Asking "how are we doing" questions or reflecting about the whole group or context.
9. Balance	WEIGH	Weighing alternatives; identifying trade-offs.

	SYSTEM	System-thinking about the topic (not the dialogue). Introducing (for the first time in a dialog) a larger set of concerns in: time; geography; causality; level; part-to-whole systems. Moving the conversation from individual examples and factors to more inclusive or big picture systems of things or factors.
10. Plan	ActPropose	Proposing or suggesting action or solution planning.

BIBLIOGRAPHY

[1] Adolphs, Ralph. Trust in the brain. *Nature Neuroscience 5*, 3 (2002), 192–193.

[2] Aleven, Vincent A.W.M.M, and Koedinger, Kenneth R. An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive science 26*, 2 (2002), 147–179.

[3] Ando, R.K., and Zhang, T. A framework for learning predictive structures from multiple tasks and unlabeled data. *The Journal of Machine Learning Research 6* (2005), 1817–1853.

[4] Andrzejewski, D., Mulhern, A., Liblit, B., and Zhu, X. Statistical debugging using latent topic models. *Machine Learning: ECML 2007* (2007), 6–17.

[5] Argyriou, A., Evgeniou, T., and Pontil, M. Convex multi-task feature learning. *Machine Learning 73*, 3 (2008), 243–272.

[6] Arriaga, M. *Social dynamics in user-generated content platforms*. PhD thesis, New York University, Graduate School of Business Administration, 2010.

[7] Arroyo, Ivon, Beal, Carole, Murray, Tom, Walles, Rena, and Woolf, Beverly P. Web-based intelligent multimedia tutoring for high stakes achievement tests. In *Intelligent Tutoring Systems* (2004), Springer, pp. 468–477.

[8] Asen, R. Toward a normative conception of difference in public deliberation. *Argumentation and Advocacy 35*, 3 (1999), 115–29.

[9] Asuncion, A., Welling, M., Smyth, P., and Teh, Y.W. On smoothing and inference for topic models. *arXiv preprint arXiv:1205.2662* (2012).

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

[11] Berger, A.L., Pietra, V.J.D., and Pietra, S.A.D. A maximum entropy approach to natural language processing. *Computational linguistics 22*, 1 (1996), 39–71.

[12] Billings, M., and Watts, L.A. Understanding dispute resolution online: Using text to reflect personal and substantive issues in conflict. In *Proceedings of the 28th international conference on Human factors in computing systems* (2010), ACM, pp. 1447–1456.

[13] Bishop, C.M. *Pattern recognition and machine learning*, vol. 4. springer New York, 2006.

[14] Black, L.W., Welser, H.T., Cosley, D., and DeGroot, J.M. Self-governance through group discussion in wikipedia measuring deliberation in online groups. *Small Group Research 42*, 5 (2011), 595–634.

[15] Blei, D.M. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference* (2004), vol. 16, MIT Press, p. 17.

[16] Blei, D.M. Probabilistic topic models. Slides for machine learning summer school in Kyoto, 2012.

[17] Blei, D.M., and Lafferty, J.D. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning* (2006), ACM, pp. 113–120.

[18] Blei, D.M., and Lafferty, J.D. A correlated topic model of science. *The Annals of Applied Statistics* (2007), 17–35.

[19] Blei, D.M., and McAuliffe, J.D. Supervised topic models. *arXiv preprint arXiv:1003.0783* (2010).

[20] Blei, D.M., Ng, A.Y., and Jordan, M.I. Latent dirichlet allocation. *the Journal of machine Learning research 3* (2003), 993–1022.

[21] Bohman, J. *Public deliberation: Pluralism, complexity, and democracy*. MIT press, 2000.

[22] Borgatti, Stephen P, Mehra, Ajay, Brass, Daniel J, and Labianca, Giuseppe. Network analysis in the social sciences. *science 323*, 5916 (2009), 892–895.

[23] Bosch-Sijtsema, P. The impact of individual expectations and expectation conflicts on virtual teams. *Group & organization management 32*, 3 (2007), 358–388.

[24] Boyd-Graber, J., Blei, D., and Zhu, X. A topic model for word sense disambiguation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (2007), pp. 1024–1033.

[25] Bradberry, Travis, and Greaves, Jean. *Emotional Intelligence 2.0*. TalentSmart, 2009.

[26] Breiman, L. Random forests. *Machine learning 45*, 1 (2001), 5–32.

[27] Bromme, R., and Stahl, E. The impact of epistemological beliefs on e-learning: The case of help-seeking. In *The joint workshop of cognition and learning through media-communication for advanced e-learning* (2003), pp. 29–35.

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

[29] Carroll, J Douglas, and Chang, Jih-Jie. Analysis of individual differences in multidimensional scaling via an n-way generalization of eckart-young decomposition. *Psychometrika 35*, 3 (1970), 283–319.

[30] Caruana, R. Multitask learning. *Machine learning 28*, 1 (1997), 41–75.

[31] Chaffar, Soumaya, and Frasson, Claude. Inducing optimal emotional state for learning in intelligent tutoring systems. In *Intelligent Tutoring Systems* (2004), Springer, pp. 45–54.

[32] Charvet, Shelle. *Words that change minds*. Kendall/Hunt Publishing Company, 1997.

[33] Chen, X., Hu, X., Shen, X., and Rosen, G. Probabilistic topic modeling for genomic data interpretation. In *Bioinformatics and Biomedicine (BIBM), 2010 IEEE International Conference on* (2010), IEEE, pp. 149–152.

[34] Ciarrochi, J., Chan, A.Y.C., and Bajgar, J. Measuring emotional intelligence in adolescents. *Personality and individual differences 31*, 7 (2001), 1105–1119.

[35] Cichocki, Andrzej. Tensor decompositions: A new concept in brain data analysis? *arXiv preprint arXiv:1305.0395* (2013).

[36] Clark, Herbert H. *Using language*, vol. 4. Cambridge University Press Cambridge, 1996.

[37] Cohen, J. Procedure and substance in deliberative democracy. *Deliberative Democracy: Essays on Reason and Politics, MIT Press, Cambridge* (1997), 407–437.

[38] Cohen, Jacob, and Cohen, Patricia. *Applied multiple regression/correlation analysis for the behavioral sciences.* Lawrence Erlbaum, 1975.

[39] Cortes, C., and Vapnik, V. Support-vector networks. *Machine learning 20*, 3 (1995), 273–297.

[40] Costa, Paul T, and MacCrae, Robert R. *Revised NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory (NEO FFI): Professional Manual.* Psychological Assessment Resources, 1992.

[41] Danescu-Niculescu-Mizil, C., Lee, L., Pang, B., and Kleinberg, J. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of the 21st international conference on World Wide Web* (2012), ACM, pp. 699–708.

[42] Decety, Jean, and Jackson, Philip L. A social-neuroscience perspective on empathy. *Current directions in psychological science 15*, 2 (2006), 54–58.

[43] Doyle, G., and Elkan, C. Financial topic models. In *Working Notes of the NIPS-2009 Workshop on Applications for Topic Models: Text and Beyond Workshop* (2009).

[44] Duchi, J., and Singer, Y. Efficient online and batch learning using forward backward splitting. *The Journal of Machine Learning Research 10* (2009), 2899–2934.

[45] Dweck, Carol. *Mindset: The new psychology of success*. Random House Digital, Inc., 2006.

[46] DMello, S., Dowell, N., and Graesser, A. 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* (2009), pp. 9–16.

[47] Eisenstein, J., Smith, N.A., and Xing, E.P. Discovering sociolinguistic associations with structured sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (2011), Association for Computational Linguistics, pp. 1365–1374.

[48] Feng, Y., and Lapata, M. Topic models for image annotation and text illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (2010), Association for Computational Linguistics, pp. 831–839.

[49] Fisher, Danyel. Using egocentric networks to understand communication. *Internet Computing, IEEE 9*, 5 (2005), 20–28.

[50] Fishkin, J.S. *Democracy and deliberation: New directions for democratic reform*, vol. 217. Cambridge Univ Press, 1991.

[51] Fowler, James H, and Christakis, Nicholas A. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *BMJ: British medical journal 337* (2008).

[52] Fredrickson, Barbara L. *Love 2.0: How Our Supreme Emotion Affects Everything We Feel, Think, Do, and Become*. Penguin. com, 2013.

[53] Friedman, Howard S, and Martin, Leslie R. *The longevity project: Surprising discoveries for health and long life from the landmark eight-decade study*. Penguin, 2011.

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

[55] Fürnkranz, Johannes, Hüllermeier, Eyke, Mencía, Eneldo Loza, and Brinker, Klaus. Multilabel classification via calibrated label ranking. *Machine Learning 73*, 2 (2008), 133–153.

[56] Gardner, H.E. *Intelligence reframed: Multiple intelligences for the 21st century.* Basic Books, 2000.

[57] Gastil, J. Communication as deliberation. *Communication as...: Perspectives on Theory* (2005), 164.

[58] Gastil, J., and Black, L.W. Public deliberation as the organizing principle of political communication research. *Journal of Public Deliberation 4*, 1 (2007), 3.

[59] Ghosn, J., and Bengio, Y. Multi-task learning for stock selection. *Advances in Neural Information Processing Systems* (1997), 946–952.

[60] Goleman, D. *Emotional intelligence: ; why it can matter more than IQ.* Bantam, 2006.

[61] Goleman, D. *Social intelligence.* Arrow Books, 2007.

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

[63] Griffiths, T.L., and Steyvers, M. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America 101*, Suppl 1 (2004), 5228–5235.

[64] Griffiths, T.L., Steyvers, M., and Tenenbaum, J.B. Topics in semantic representation. *Psychological review 114*, 2 (2007), 211.

[65] Habermas, J. *The structural transformation of the public sphere: An inquiry into a category of bourgeois society.* mit Press, 1991.

[66] Hall, M.A. *Correlation-based feature selection for machine learning.* PhD thesis, The University of Waikato, 1999.

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

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

[69] Hardoon, David R, Szedmak, Sandor, and Shawe-Taylor, John. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation 16*, 12 (2004), 2639–2664.

[70] Harman, Harry H. Modern factor analysis.

[71] Harshman, Richard A. Parafac2: Mathematical and technical notes. *UCLA working papers in phonetics 22* (1972), 30–44.

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

[73] Hofmann, Thomas. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (1999), ACM, pp. 50–57.

[74] Holtgraves, Thomas M. *Language as social action: Social psychology and language use*. Psychology Press, 2001.

[75] Hoque, Mohammed Ehsan, Courgeon, Matthieu, Martin, J, Mutlu, Bilge, and Picard, Rosalind W. Mach: My automated conversation coach. In *International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2013)* (2013).

[76] Hotelling, Harold. Relations between two sets of variates. *Biometrika 28*, 3/4 (1936), 321–377.

[77] Hutchinson, Rebecca A, Niculescu, Radu Stefan, Keller, Timothy A, Rustandi, Indrayana, and Mitchell, Tom M. Modeling fmri data generated by overlapping cognitive processes with unknown onsets using hidden process models. *NeuroImage 46*, 1 (2009), 87–104.

[78] Iwata, T., and Sawada, H. Topic model for analyzing purchase data with price information. *Data Mining and Knowledge Discovery* (2012), 1–15.

[79] Jalali, A., Ravikumar, P., Sanghavi, S., and Ruan, C. A dirty model for multitask learning. *Advances in Neural Information Processing Systems 23* (2010), 964–972.

[80] Ji, S., and Ye, J. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning* (2009), ACM, pp. 457–464.

[81] Jo, Y., and Oh, A.H. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM international conference on Web search and data mining* (2011), ACM, pp. 815–824.

[82] Jolliffe, Ian. *Principal component analysis*. Wiley Online Library, 2005.

[83] Kaiser, C., Piazza, A., Krockel, J., and Bodendorf, F. Are humans like ants?–analyzing collective opinion formation in online discussions. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)* (2011), IEEE, pp. 266–273.

[84] Kamişi, Sibel, and Doǒançay-Aktuna, Seran. Effects of social power on language use across speech communities. *International Journal of Applied Linguistics 6*, 2 (1996), 199–222.

[85] Kitchener, Karen Strohm, and King, Patricia M. Reflective judgment: Concepts of justification and their relationship to age and education. *Journal of applied developmental psychology 2*, 2 (1981), 89–116.

[86] Kolda, Tamara G, and Bader, Brett W. Tensor decompositions and applications. *SIAM review 51*, 3 (2009), 455–500.

[87] Kuhn, D. A developmental model of critical thinking. *Educational researcher 28*, 2 (1999), 16–46.

[88] Kuo, F., and Yu, C. An exploratory study of trust dynamics in work-oriented virtual teams. *Journal of Computer-Mediated Communication 14*, 4 (2009), 823–854.

[89] Lacoste-Julien, S., Sha, F., and Jordan, M.I. Disclda: Discriminative learning for dimensionality reduction and classification. *Advances in Neural Information Processing Systems (NIPS) 21* (2008).

[90] Lafferty, J., McCallum, A., and Pereira, F.C.N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

[91] Landis, J Richard, and Koch, Gary G. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[92] Langford, J., Li, L., and Zhang, T. Sparse online learning via truncated gradient. *The Journal of Machine Learning Research 10* (2009), 777–801.

[93] Lee Rodgers, Joseph, and Nicewander, W Alan. Thirteen ways to look at the correlation coefficient. *The American Statistician 42*, 1 (1988), 59–66.

[94] Leibovitch, FS, Black, SE, Caldwell, CB, Ebert, PL, Ehrlich, LE, and Szalai, JP. Brain-behavior correlations in hemispatial neglect using ct and spect the sunnybrook stroke study. *Neurology 50*, 4 (1998), 901–908.

[95] Levandowsky, M., and Winter, D. Distance between sets. *Nature 234*, 5323 (1971), 34–35.

[96] Lewis, David D, Yang, Yiming, Rose, Tony G, and Li, Fan. Rcv1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research 5* (2004), 361–397.

[97] Littlejohn, S.W., and Pearce, W.B. *Moral conflict: When social worlds collide.* Sage Publications, Incorporated, 1997.

[98] Liu, T.Y., Yang, Y., Wan, H., Zeng, H.J., Chen, Z., and Ma, W.Y. Support vector machines classification with a very large-scale taxonomy. *ACM SIGKDD Explorations Newsletter 7*, 1 (2005), 36–43.

[99] Lukensmeyer, C.J., Goldman, J., and Brigham, S. A town meeting for the twenty-first century. *The Deliberative Democracy Handbook. Strategies for Effective Civic Engagement in the Twenty-First Century. San Francisco: Jossey-Bass, S* (2005), 154–163.

[100] Mackintosh, N.J. *IQ and human intelligence.* Oxford University Press, 1998.

[101] Macskassy, S.A., and Michelson, M. Why do people retweet? anti-homophily wins the day! In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (2011), pp. 209–216.

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

[103] Mardia, Kantilal Varichand, Kent, John T, and Bibby, John M. Multivariate analysis.

[104] Martins, A.F.T., Figueiredo, MAT, Aguiar, PMQ, Smith, NA, and Xing, EP. Online learning of structured predictors with multiple kernels. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics* (2011).

[105] Maslow, Abraham Harold. A theory of human motivation. *Psychological review 50*, 4 (1943), 370.

[106] Mayer, R.E. Cognitive, metacognitive, and motivational aspects of problem solving. *Instructional science 26*, 1 (1998), 49–63.

[107] Mayfield, Jacqueline Rowley, Mayfield, Milton Ray, and Kopf, Jerry. The effects of leader motivating language on subordinate performance and satisfaction. *Human Resource Management 37*, 3 (1998), 235–248.

[108] McCallum, A., Wang, X., and Corrada-Emmanuel, A. Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research 30*, 1 (2007), 249–272.

[109] Menard, S. *Applied logistic regression analysis*, vol. 106. Sage Publications, Incorporated, 2001.

[110] Milroy, Lesley, and Milroy, James. Social network and social class: Toward an integrated sociolinguistic model. *Language in society 21*, 1 (1992), 1–26.

[111] Mimno, D. Computational historiography: Data mining in a century of classics journals. *Journal on Computing and Cultural Heritage (JOCCH) 5*, 1 (2012), 3.

[112] Mimno, D., Wallach, H.M., Talley, E., Leenders, M., and McCallum, A. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2011), Association for Computational Linguistics, pp. 262–272.

[113] Minka, T., and Lafferty, J. Expectation-propagation for the generative aspect model. In *Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence* (2002), Morgan Kaufmann Publishers Inc., pp. 352–359.

[114] Moreau, J. Fonctiones convexes duales et points proximaux dans un espace hilbertien. *Comptes Rendus de lAcadmie des Sciences de Paris 255* (1962), 2897–2899.

[115] Mostow, Jack, Chang, Kai-min, and Nelson, Jessica. Toward exploiting eeg input in a reading tutor. In *Artificial Intelligence in Education* (2011), Springer, pp. 230–237.

[116] Murray, Tom, Woolf, Beverly.P, Xu, Xiaoxi, Shipe, Stefanie, Howard, Scott, and Wing, Leah. Supporting social deliberative skills in online classroom dialogues: preliminary results using automated text analysis. In *Intelligent Tutoring Systems* (2012), Springer, pp. 666–668.

[117] Murray, Tom, Xu, Xiaoxi, and Woolf, Beverly Park. An exploration of text analysis methods to identify social deliberative skills. In *Proceedings of the 16th International Conference on Artificial Intelligence in Education* (2013).

[118] Negahban, S., and Wainwright, M.J. Joint support recovery under high-dimensional scaling: Benefits and perils of $l_{1,\infty}$-regularization. *Advances in Neural Information Processing Systems 21* (2008), 1161–1168.

[119] NeuroSky. Neuroskys esensetm meters and detection of mental state. neurosky, inc., 2009.

[120] Newcomb, Theodore M. An approach to the study of communicative acts. *Psychological review 60*, 6 (1953), 393.

[121] Newman, D.J., and Block, S. Probabilistic topic decomposition of an eighteenth-century american newspaper. *Journal of the American Society for Information Science and Technology 57*, 6 (2006), 753–767.

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

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

[124] Obozinski, G., Taskar, B., and Jordan, M.I. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing 20*, 2 (2010), 231–252.

[125] Ozyer, Tansel, Rokne, Jon G, Wagner, Gerhard, and Reuser, Arno HP. *The influence of technology on social network analysis and mining.* Springer, 2013.

[126] Peng, H., Long, F., and Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 27*, 8 (2005), 1226–1238.

[127] Pennebaker, James W, Francis, Martha E, and Booth, Roger J. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* (2001), 71.

[128] Pesarin, Fortunato. *Multivariate permutation tests: with applications in biostatistics*, vol. 240. Wiley Chichester, 2001.

[129] Petrides, Konstantin Vasily, Pita, Ria, and Kokkinaki, Flora. The location of trait emotional intelligence in personality factor space. *British Journal of Psychology 98*, 2 (2007), 273–289.

[130] Quattoni, A., Collins, M., and Darrell, T. Transfer learning for image classification with sparse prototype representations. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on* (2008), IEEE, pp. 1–8.

[131] Quinlan, J.R. Induction of decision trees. *Machine learning 1*, 1 (1986), 81–106.

[132] Ramage, D., Hall, D., Nallapati, R., and Manning, C.D. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1* (2009), Association for Computational Linguistics, pp. 248–256.

[133] Rico, R., Alcover, C.M., Sánchez-Manzanares, M., and Gil, F. The joint relationships of communication behaviors and task interdependence on trust building and change in virtual project teams. *Social science information 48*, 2 (2009), 229–255.

[134] Rilling, James K, and Sanfey, Alan G. The neuroscience of social decision-making. *Annual review of psychology 62* (2011), 23–48.

[135] Rosenberg-Lee, Miriam, Lovett, Marsha C, and Anderson, John R. Neural correlates of arithmetic calculation strategies. *Cognitive, Affective, & Behavioral Neuroscience 9*, 3 (2009), 270–285.

[136] Ryan, Richard M, and Deci, Edward L. Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology 25*, 1 (2000), 54–67.

[137] Sacks, Harvey, Schegloff, Emanuel A, and Jefferson, Gail. A simplest systematics for the organization of turn-taking for conversation. *Language* (1974), 696–735.

[138] Salovey, P., and Mayer, J.D. Emotional intelligence. *Imagination, cognition and personality 9*, 3 (1990), 185–211.

[139] Salovey, Peter, and Grewal, Daisy. The science of emotional intelligence. *Current directions in psychological science 14*, 6 (2005), 281–285.

[140] Schenck-Hamlin, W. Deliberation quality in national issues forums and midwest democracy project blogs. 2010.

[141] Schommer-Aikins, M., and Hutter, R. Epistemological beliefs and thinking about everyday controversial issues. *The Journal of Psychology: Interdisciplinary and Applied 136*, 1 (2002), 5–20.

[142] Scott, John, and Carrington, Peter J. *The SAGE handbook of social network analysis*. SAGE publications, 2011.

[143] Seal, C.R., Boyatzis, R.E., and Bailey, J.R. Fostering emotional and social intelligence in organizations. *Organization Management Journal 3*, 3 (2006), 190–209.

[144] Shalev-Shwartz, S., Singer, Y., and Srebro, N. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th international conference on Machine learning* (2007), ACM, pp. 807–814.

[145] Singer, Tania, and Lamm, Claus. The social neuroscience of empathy. *Annals of the New York Academy of Sciences 1156*, 1 (2009), 81–96.

[146] Spiliopoulou, A., and Storkey, A. A topic model for melodic sequences. *arXiv preprint arXiv:1206.6441* (2012).

[147] Steane, A.M. Error correcting codes in quantum theory. *Physical Review Letters 77*, 5 (1996), 793–797.

[148] Steenbergen, M.R., Bächtiger, A., Spörndli, M., and Steiner, J. Measuring political deliberation: a discourse quality index. *Comparative European Politics 1*, 1 (2003), 21–48.

[149] Steyvers, M., Smyth, P., Rosen-Zvi, M., and Griffiths, T. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining* (2004), ACM, pp. 306–315.

[150] Stromer-Galley, J. Measuring deliberation's content: A coding scheme. *Journal of Public Deliberation 3*, 1 (2007), 12.

[151] Strzalkowski, Tomek, Broadwell, George Aaron, Stromer-Galley, Jennifer, Shaikh, Samira, Taylor, Sarah, and Webb, Nick. Modeling socio-cultural phenomena in discourse. In *Proceedings of the 23rd International Conference on Computational Linguistics* (2010), Association for Computational Linguistics, pp. 1038–1046.

[152] Suleiman, Yasir. *A war of words: Language and conflict in the Middle East*, vol. 19. Cambridge University Press, 2004.

[153] Suthers, Daniel D, and Desiato, Caterina. Exposing chat features through analysis of uptake between contributions. In *System Science (HICSS), 2012 45th Hawaii International Conference on* (2012), IEEE, pp. 3368–3377.

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

[155] Torralba, A., Murphy, K.P., and Freeman, W.T. Sharing visual features for multiclass and multiview object detection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 29*, 5 (2007), 854–869.

[156] Tucker, Ledyard R. Some mathematical notes on three-mode factor analysis. *Psychometrika 31*, 3 (1966), 279–311.

[157] Vogt, J. E., and Roth, V. The group-lasso: $l_{1,\infty}$ regularization versus $l_{1,2}$ regularization. In *Pattern Recognition*. Springer Berlin Heidelberg, 2010, pp. 252–261.

[158] Vygotsky, L.S., and Cole, M. *Mind in society: The development of higher psychological processes*. Harvard university press, 1978.

[159] Wallach, H.M. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning* (2006), ACM, pp. 977–984.

[160] Wang, X., McCallum, A., and Wei, X. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on* (2007), IEEE, pp. 697–702.

[161] Wang, Y., Kraut, R., and Levine, J. 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).

[162] Wasserman, Stanley. *Social network analysis: Methods and applications*, vol. 8. Cambridge university press, 1994.

[163] Wikipedia. Centrality — Wikipedia, the free encyclopedia, 2013. [Online; accessed 22-September-2013].

[164] Wikipedia. Social network analysis — Wikipedia, the free encyclopedia, 2013. [Online; accessed 22-September-2013].

[165] Winne, P.H. Self-regulated learning viewed from models. *Self-regulated learning and academic achievement: Theoretical perspectives* (2001), 153.

[166] Woolf, B., Murray, T., Xu, X., Osterweil, L., Clarke, L., Wing, L., and Katsh, E. Computational predictors in online social deliberations. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media* (2012).

[167] Woolley, Anita Williams, Chabris, Christopher F, Pentland, Alex, Hashmi, Nada, and Malone, Thomas W. Evidence for a collective intelligence factor in the performance of human groups. *science 330*, 6004 (2010), 686–688.

[168] Xu, Xiaoxi. Discovering latent strategies. In *AAAI* (2011).

[169] Xu, Xiaoxi, Jensen, David, and Rissland, Edwina L. Strategy mining. In *The Twenty-Seventh International Flairs Conference* (2014).

[170] Xu, Xiaoxi, Murray, Tom, and Woolf, Beverly. P. Text analysis of deliberative skills in undergraduate online dialogue: Using l1 regularized logistic regression to model psycholinguistic features. *Proceedings of The 2013 International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government (EEE-2013)* (2013).

[171] Xu, Xiaoxi, Murray, Tom, Woolf, Beverly Park, and Smith, David A. Identifying social deliberative behavior from online communicationa cross-domain study. In *The Twenty-Seventh International Flairs Conference* (2014).

[172] Xu, Xiaoxi, Murray, Tom, Woolf, Beverly Park, and Smith, David A. Social network signatures of effective online communication. In *Intelligent Tutoring Systems* (2014), Springer, pp. 621–622.

[173] Xu, Xiaoxi, Smith, David. A, Murray, Tom, and Woolf, Beverly. P. Analyzing conflict narratives to predict settlements in ebay feedback dispute resolution. In *Proceedings of the Eighth International Conference on Data Mining* (2012).

[174] Xu, Xiaoxi, Smith, David. A, Murray, Tom, and Woolf, Beverly. P. If you were me and i were you mining social deliberation in online communication. *Proceedings of EDM-13, Educational Data Mining* (2013).

[175] Yang, Yiming. A study of thresholding strategies for text categorization. In *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (2001), ACM, pp. 137–145.

[176] Zelinsky, Gregory. J. Perspective taking about interacting characters: Detecting theory of mind inferences from eye movements, 2013.

[177] Zhou, J., and Ye, J. Multi-task learning: Theory, algorithms, and applications. Tutorials for SIAM international conference on data mining, 2012.

[178] Zhu, J., Ahmed, A., and Xing, E.P. Medlda: maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th Annual International Conference on Machine Learning* (2009), ACM, pp. 1257–1264.

[179] Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent.