January 2019

Machine Learning Methods For The Analysis Of Clinical Conversation

Md Mehedi Hasan
Wayne State University, mehedi2003@gmail.com

MACHINE LEARNING METHODS FOR THE ANALYSIS OF CLINICAL CONVERSATION

by

MD MEHEDI HASAN

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2019

MAJOR: COMPUTER SCIENCE

Approved By:

Advisor	Date

DEDICATION

This dissertation is dedicated to Allah and my parents, for all their love, patience, kindness and support.

I also dedicate this to my daughter and wife who are my everything and always been my greatest inspiration.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my advisor, Dr. Alexander Kotov, who generously offered his always wise guidance. I am very thankful for his support, his patience, and his time. Without his guidance and persistent help, this dissertation is not possible. His deep understanding of his research field, diligent work ethic, and determination to pursue intellectual beauty has set a perfect example for me both in my research and personal life. The most important thing I learned from Dr. Kotov through my Ph.D. journey is how to identify the right problem to solve and how to tackle a problem I have never encountered before. I am also fortunate to have Dr. Dongxiao Zhu, Dr. Ming Dong, and Dr. April Idalski Carcone as my committee members. I also want to thank Dr. Schwiebert, Lanita Stewart, Angelique Meiu, Areej Salaymeh, and other staff in my department for providing all the necessity to make my life easier.

I would like to acknowledge my mother and father, Rijia Parvin and Motiur Rahman. I offer my sincerest gratitude for their unconditional support. I would like to acknowledge my brother, Mahabub Hasan for keeping life interesting. I am also thankful to my mother-in-law, Momotaz Begum who motivated me to accomplish this goal. I would also like to acknowledge my daughter, Mariyah Hasan, and significant other, Ummay Hany Polly, for the emotional support and for keeping me in good spirits.

In addition, I would like to thank my colleagues and friends at Textual Data Analytics (TEANA) Lab: Fedor, Saeid, Indira, and Diana for their help and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

1.1 Problem description

Motivational Interviewing (MI) is an evidence-based communication technique to increase intrinsic motivation and self-efficacy for behavior change[96, 98]. A typical MI interview sessions involve a counselor and a patient. The goal of an MI session is to increase intrinsic motivation for behavior change through the exploration of the patient's own desires, ability, reasons, need for and commitment to the targeted behavior change. These statements, referred to as "change talk" (or CT), consistently predict actual behavior change [10] that can be sustained for as long as 34 months after an interview [132]. However, communication science approaches to understanding the efficacy of MI are inherently limited by traditional qualitative coding methods which is a time-consuming and resource-intensive process. Thus, an efficient method is required to automate the coding process which will accelerate the pace of communication research in behavioral science. The specific provider behaviors responsible for the elicitation of change talk, are also less clear and may vary by treatment context. Therefore, new design objective and perspective are necessary to understand which provider behaviors and in which contexts lead to patient change talk.

1.2 Our contribution

In this section, we summarize our contributions and outline our dissertation. In this dissertation, we deal with two types of clinical conversation, one that involves a face to face dialogue between patient and counselor and another one which involves an email-based conversation between patient and an ecoach. In the following, we summarize our three research projects that we accomplished as part of this dissertation. In the first two projects, we focus on the dialogue-based clinical conversation while the third project mainly focuses on email-based clinical conversation.

- **Automatic annotation of clinical conversation**. As the first step, we ad-

dress the problem of manual behavioral coding of MI. Traditionally, clinical interviews are transcribed and then each utterance is manually annotated with a set of codes from a pre-defined codebook operationalizing specific behavior types. Training human coders to reliably and accurately assign codes to textual fragments requires a large investment of manpower, time and money. For example, in a recent MI study [24], training coders to reliability took about four months and, once trained, coders required five hours to code every recorded hour. A similar study reported requiring 60 hours of training over six weeks to attain coder reliability, and the actual coding involved two coding passes and six coders [101]. This study was using Minority Youth-Sequential Coding of Process Exchanges (MYSCOPE), which is similar to the codebook of the proposed project. In the past decade, machine learning (ML) techniques have begun providing an efficient alternative to intensive cognitive tasks. Therefore, we leveraged eight supervised classification models to automatically code MI counseling sessions with 37 African American adolescents with obesity and their caregivers. This study examined the effectiveness of state-of-the-art supervised machine learning methods in conjunction with different feature types for the task of automatic annotation of fragments of clinical text based on codebooks with a large number of categories. We used a collection of motivational interview transcripts consisting of 11,353 utterances, which were manually annotated by two human coders as the gold standard (a collection of high-quality and accurate labeled data that can be gathered from experts), and experimented with state-of-art classifiers, including Nave Bayes, J48 Decision Tree, Support Vector Machine (SVM), Random Forest (RF), AdaBoost, DiscLDA, Conditional Random Fields (CRF) and Convolutional Neural Network (CNN) in conjunction with lexical, contextual (label of the previous utterance) and semantic (distribution of words in the utterance across the Linguistic Inquiry and Word Count

dictionaries) features. We found out that, when the number of classes is large, the performance of CNN and CRF is inferior to SVM. When only lexical features were used, interview transcripts were automatically annotated by SVM with the highest classification accuracy among all classifiers of 70.8%, 61% and 53.7% based on the codebooks consisting of 17, 20 and 41 codes, respectively. Using contextual and semantic features, as well as their combination, in addition to lexical ones, improved the accuracy of SVM for annotation of utterances in motivational interview transcripts with a codebook consisting of 17 classes to 71.5%, 74.2%, and 75.1%, respectively. With no modification, the SVM model also tested with other studies, in which SVM model correctly classified 72.0% and 79.8% of patient-provider utterances in HIV clinical encounters and eCoaching sessions, respectively. These results demonstrate the potential of using machine learning methods in conjunction with lexical, semantic and contextual features for automatic annotation of clinical interview transcripts with near-human accuracy.

- **Sequential analysis of clinical conversation**. In our previous project, we automatically annotate MI transcripts which are used by this project for the sequential analysis of MI. In this project, we focus on predicting the outcome of patient-provider communication sequences in the context of the clinical dialog, which is the first part of the sequential analysis process, establishing the sequencing of behaviors to generate evidence for the causal sequencing of communication behaviors. Specifically, we consider the prediction of the motivational interview success (i.e. eliciting a particular type of patient behavioral response) based on an observed sequence of coded patient-provider communication exchanges as a sequence classification problem. We proposed two solutions to this problem, one that is based on Recurrent Neural Networks (RNNs) and another that is based on Markov Chain (MC), a probabilistic model that con-

ditions each observation in a sequence only on preceding observation and not on any other past observation and Hidden Markov Model (HMM), a probabilistic generative model for sequence data, for modeling sequences of behavior codes. We compared the accuracy of these solutions using communication sequences annotated with behavior codes from the motivational interviews. Our experiments indicate that the deep learning-based approach is significantly more accurate than the approach based on probabilistic models in predicting the success of motivational interviews (0.8677 versus 0.7038 and 0.6067 F1-score by RNN, MC and HMM, respectively, when using under-sampling to correct for class imbalance, and 0.8381 versus 0.7775 and 0.7520 F1-score by RNN, MC and HMM, respectively, when using over-sampling). These results indicate that the proposed method can be used for real-time monitoring of progression of clinical interviews and more efficient identification of effective provider communication strategies, which in turn can significantly decrease the effort required to develop behavioral interventions and increase their effectiveness. Although there is strong empirical evidence linking "MI-consistent" counselor behaviors and patient motivational statements (i.e., "change talk"), the specific counselor communication behaviors effective for eliciting patient change talk vary by treatment context and, thus, are a subject of ongoing research. An integral part of this research is the sequential analysis of pre-coded MI transcripts. In the second part of our sequential analysis process, we evaluated the empirical effectiveness of the Hidden Markov Model and closed frequent pattern mining, a method to identify frequently occurring sequential patterns of behavior codes in MI communication sequences to inform MI practice. We conducted experiments with 1,360 communication sequences from 37 transcribed audio recordings of counseling sessions with African-American adolescents with obesity and their caregivers. Transcripts had been previously annotated with patient-counselor

behavior codes using an MYSCOPE codebook. Empirical results indicate that the Hidden Markov Model and closed frequent pattern mining techniques can identify counselor communication strategies that are effective at eliciting patients' motivational statements to guide clinical practice.

- **Segmentation of clinical conversation**. The annotation model and sequential analysis models represent two critical processes necessary to automate behavioral coding. However, a segmentation model is needed to process the email conversation for developing autocoding and sequence analysis models to fully automate behavioral counseling. In this project, we propose various segmentation models to facilitate behavioral coding of e-Coaching sessions, behavior interventions delivered via email and grounded in the principles of MI. Segmentation process partitions emails into fragments that correspond to MI behaviors which is more challenging in eCoaching sessions because eCoaching data differs from traditional face to face counseling. Unlike transcribed in-person exchanges, email correspondence is not clearly segmented into codable speech acts (i.e., utterances). Thus, the unstructured nature of e-Coaching exchanges poses a unique set of analytic challenges. Traditionally, trained coders manually segmented emails before applying the annotation model to predict behavioral code. Therefore, there is a need for segmentation model to fully automate the behavioral coding. This project frames email segmentation task as a classification problem, in which each word or punctuation mark is annotated with one of the two classes: "new segment" and "same segment". Our proposed method utilizes word and punctuation mark embeddings in conjunction with part-of-speech features to address the segmentation problem. We evaluate the performance of conditional random fields (CRF) as well as multi-layer perceptron (MLP), bi-directional recurrent neural network (BRNN) and convolutional recurrent neural network (CRNN) for the task of email segmentation. Results

show that CRNN outperformed CRF, MLP and BRNN achieving 98.9% overall and 86.4% and 99.3% accuracy for detecting "new segment" and "same segment", respectively. Segmentation was also a concern in our dialogue-based clinical conversation although this project focuses on eCoaching session. Actually, we also need segmentation for face to face sessions because we allow counselors' speech to be segmented, a fact that was ignored in earlier two studies where the data had previously been parsed or segmented by human coders.

1.3 Organization

The rest of this dissertation is organized as follows. We present our motivational interview-based clinical conversation works in Chapter 2 and 3. In Chapter 2, we present eight state-of-the-art machine learning methods and their experimental results for the automatic annotation of patient-provider clinical conversation. We perform two sequential analysis on MIs which is described in Chapter 3. We present deep learning and probabilistic models to analyze the sequencing of patient-provider communication. We further investigate sequential patterns from the identified sequence of patient-provider communication. In Chapter 4, we propose traditional machine learning based approach as well as deep learning approaches for the segmentation of email-based patient-provider clinical conversation. We conclude and discuss some possible future research directions in Chapter 5.

CHAPTER 2 AUTOMATIC ANNOTATION OF CLINICAL CONVERSATION

Annotation of clinical interview transcripts to distinguish different behavior types is an important and integral part of clinical research aimed at designing effective interventions for many conditions and disorders. This chapter describes our research work to automate the process of behavioral coding, which has been traditionally done by a trained coder. We examine the effectiveness of eight state-of-the-art supervised machine learning methods in conjunction with different feature types for the task of automatic annotation of fragments of clinical text based on codebooks with a large number of categories. We believe that automatic annotation of clinical conversation can significantly accelerate the pace of research in behavioral science.

2.1 Introduction

Annotation (or labeling) of fragments of clinical text with the categories (or labels, codes) from a predefined codebook is an integral part of qualitative research. It can also be viewed as classification of textual fragments into a predefined number of categories (classes). Textual annotation has been traditionally performed manually by trained coders, which is a tedious, costly and time-consuming process. Furthermore, manual annotation increases the likelihood of errors due to coder fatigue and bias associated with human subjectivity. To automate tedious cognitive tasks such as classification, supervised machine learning methods have been recently proposed. These methods have been shown to be successful at binary (two-class) classification [109, 104] (e.g. classifying textual fragments as neutral or opinionated) but failure for textual classification tasks involving large number of classes. Such tasks, however, are fairly common in clinical setting (e.g. annotation of clinical interviews, assignment of ICD-9/10 codes to patient records). Our recent work address this limitation by utilizing contextual and semantic features and present the results of an extensive experimental evaluation of state-of-the-art supervised machine learning

methods in conjunction with lexical and the proposed features for the task of automatic annotation of utterances in clinical interview transcripts with the codebooks consisting of large number of classes. This chapter provides a guideline for clinical informatics researchers and practitioners, who consider an option of using machine learning methods for automatic annotation of clinical text in their projects.

In this chapter, we focused on the transcripts of motivational interviews with obese adolescents (teens) and their caregivers. Automatic annotation of patient utterances in clinical communication is a challenging task, since patients usually come from a variety of cultural and educational backgrounds and their language use can be quite different [127]. This problem is exacerbated when the interviews are conducted with children and adolescents due to their tendency to use incomplete sentences and frequently change subjects.

We reported the results of comprehensive evaluation of 8 state-of-the-art classifiers (Naïve Bayes [115, 92, 70], Support Vector Machine [34, 42], Conditional Random Fields [79, 123], J48 [119], AdaBoost [47], Random Forest [21], DiscLDA [77] and Convolutional Neural Network [71]) for the task of annotating clinical interviews with a codebook, consisting of a large number of classes. We also offer and experimentally evaluate two novel features for this task: contextual features based on the label of the preceding textual fragment and semantic features based on the distribution of words in the annotated fragment over a linguistic lexicon.

2.2 Related work

Several prior works have reported the results of adopting machine learning methods, such as topic models [75, 51, 66, 74, 12], classification methods [62, 22, 113, 112] and neural networks [62, 124, 125] to the tasks of annotating MI transcripts for the assessment of intervention fidelity. Perez-Rosas et al. [113] developed a natural language processing system to evaluate counselor fidelity to the MI framework. Their system employed a Support Vector Machines (SVM) classifier based on n-grams

(contiguous sequences of words of a specified length), syntactic (structure of the clinician statements) and semantic (cognitive state) features. In our own recent work, we evaluated the accuracy of state-of-the-art classification methods and deep neural networks in conjunction with the lexical (words expressed), contextual (prior code) and semantic (inferred cognitive state based on Linguistic Inquiry and Word Count dictionaries [126]) features for the task of automated annotation of MI transcripts using codebooks with varying numbers of behavior codes [62]. An SVM model with the aforementioned features achieved 75% accuracy for automated annotation of MI transcripts with 17 behavior codes, accuracy comparable to human coders.

Previous quantitative studies of clinical conversation have resulted in creation of Generalized Medical Interaction Analysis System (GMIAS) [80], which uses a codebook with generic hierarchical categories. The small-size codebook in Comprehensive Analysis of the Structure of Encounters System (CASES) [81] was designed to annotate several meta-discursive aspects of medical interviews, such as assigning "ownership" of topics and partitioning them into distinct segments (speech acts). It was also shown that the fragments of transcripts of routine outpatient visits consisting of several speech acts coded using GMIAS and CASES can be annotated as "information giving" and "requesting information" [91]. Other related previous studies focused on categorizing assertions of medical problems in clinical narrative into 5 classes (present, absent, possible, hypothetical, conditional and associated with someone else) using SVM [116] and annotating the utterances in hemodialysis phone dialogue with 3 categories using AdaBoost classifier [78].

2.3 Methods

2.3.1 Data collection and preprocessing

The golden standard for evaluation of machine learning methods was created based on the transcripts of motivational interviews conducted by the clinicians at the Pediatric Prevention Research Center (PPRC) of Wayne State University. Each interview

is comprised of two parts: conversation of a clinician with an adolescent followed by a conversation of a clinician with the adolescent's caregiver. All adolescents in this project were between the ages of 12 and 17 (M = 14.7, SD = 1.63) and most were female (n = 27). Most caregivers were biological mothers (n = 36), who were married or living with a partner (n = 25). The median family income was $16,000–$21,999 ranging from less than $1,000 to $50,000–$74,999. The audio recordings of the interviews were first transcribed and segmented into utterances belonging to adolescents, caregivers, and counselors, preserving the sequence of utterances. Transcripts were then manually annotated by trained human coders according to MYSCOPE [24], a specialized codebook including a large number of behavior codes, which was developed by an interdisciplinary team including a clinical psychologist, a nutrition scientist, a communication scientist, a linguist and a community health worker specifically for annotating motivational interviews with obese adolescents. The MYSCOPE is an adaptation of the original MI-SCOPE [90], a qualitative code scheme to characterize patient-counselor communication during MI treatment sessions. The MYSCOPE was informed by MI fidelity code schemes including the MI Treatment Integrity Scale (MITI) [102], the MI Skill Code (MISC) [9] and Amrhein's conceptualization of change talk and commitment [7]. A primary coder independently coded interview sessions and a secondary coder co-coded a randomly selected 20% of the transcripts to monitor reliability ($\kappa = 0.696$) [24]. The MYSCOPE codebook contains a total of 115 different codes that are grouped into the youth, caregiver, and counselor code groups. The experimental datasets for this work were constructed based on the transcripts of 37 motivational interview sessions, which include a total of 11,353 segmented and annotated utterances. These utterances have been further partitioned into two subsets based on the structure of motivational interview sessions: one dataset that includes all utterances from the adolescent sessions (6,579 samples) and the other dataset that includes all utterances from the caregiver sessions (4,774 samples). A fragment of an

adolescent session transcript is presented in Table 2.1.

Table 2.1: Fragment of the annotated transcript of a dialogue between a counselor and an adolescent

Annotation	Description	Speaker	Text
331	Open-ended question, elicit change talk positive	Counselor	do you feel like making healthier choices for your snacks and your meals is something you would be able to do ? mm-hmm meaning is that food available for you ?
117	Low Uptake, positive	Adolescent	Yes
301	Structure Session	Counselor	okay and thats an important thing for us to think about cause i would not want to help you come up with a plan that you would not be able to do without somebody else help so the last part of your plan is how somebody could be supportive to you meaning how they can help you be successful and so we should choose somebody who you feel like is around often enough
112	Change Talk positive	Adolescent	my um aunt
301	Structure Session	Counselor	okay so lets stick something my aunt can do
112	Change Talk positive	Adolescent	she could when i am doing when i am eating something that i should i could not be eating but so i can choose something healthy she could tell me not to eat it
309	Affirm, low	Counselor	okay that sounds like a really great suggestion

To conduct a detailed analysis of performance of classification methods, we used the following two-stage process to create the codebooks with different number of codes for adolescent and caregiver sessions. In the first stage, we merged conceptually similar behavior codes as well as the codes with similar data distributions, while in the second stage, we eliminated the codes with insufficient data samples. In case of the adolescent sessions, we started with 55 adolescent session-specific codes and, after

merging the codes with subtle differences (e.g. converting valiances of change talk CHT+1, CHT+2 and CHT+3 into CHT+), obtained a codebook with 41 classes. We further reduced this codebook to 20 classes after merging 21 classes with similar sample distributions. After eliminating the codes that had less than 10 data samples (to ensure that there can be at least one sample of each class in each fold when using 10-fold cross validation experimental design), we obtained a third codebook with 17 codes. Using the same approach, we created the codebooks containing 58, 19, and 16 caregiver session-specific codes. Table 2.2 shows the distribution of utterances over 16 classes in the caregiver session transcripts. As follows from Table 2.2, the distribution of utterances over classes is highly imbalanced even for the codebook of the smallest size, which is fairly common for clinical text.

Table 2.2: Distribution of utterances over 16 classes in the caregiver dataset

Code	**Description**	**Utterance**	**%**
209	Caregiver Change Talk, negative	297	6.82
212	Caregiver Change Talk, positive	1107	25.40
232	Low Uptake, positive	518	11.89
235	High uptake	231	5.30
301	Structure Session	206	4.73
302	General Information, positive	309	7.09
305	Emphasize Autonomy	148	3.40
306	Closed question, Elicit Feedback	50	1.15
307	Support	108	2.48
308	Affirm	289	6.63
315	Reflect, change talk positive, about caregiver	659	15.12
329	Self-disclose	44	1.01
330	Statement, other	121	2.78
331	Open-ended question, elicit change talk positive	200	4.59
343	Open-ended question, target behavior neutral	33	0.76
344	Open-ended question, elicit barriers	38	0.87

After creating the codebooks, we pre-processed the dataset using the Snowball stemmer available as part of the Weka [59] machine learning toolkit[1]. We also found out that stopword removal decreased the performance of classification models for our

[1]http://www.cs.waikato.ac.nz/ml/weka/

Table 2.3: Feature representation of each utterance in machine learning pipeline

Feature Type	Description	Purpose
Lexical features	One feature per each distinct word in the set of training interview transcripts. The value of each lexical feature is the number of times that the corresponding word appears in the utterance.	To capture the vocabulary that is indicative of each label.
Contextual features	One feature per each codebook label. The value of the feature is set to 1, if the previous utterance in the dialog was annotated with the corresponding label, and to 0, otherwise.	Context changes the likelihood of observing speech acts. for example, if the previous speaker was requesting information, then the next speech act is more likely to be providing the requested information.
Semantic features	One feature per each of the sixty-eight LIWC lexicons. The value of each semantic feature is the number of times a word from the corresponding dictionary appears in the utterance.	To capture psycho-linguistic clues related to the thought processes, emotional states, intentions and motivations of the speaker.

task (e.g., in case of the codebook consisting of 17 classes, the accuracy of Naïve Bayes decreased from 67% to 47.10%, while the accuracy of SVM decreased from 70.76% to 55.26%). A likely reason is that, although negations are typically considered as stopwords, they are fairly important clues for inferring certain behavior types (e.g., removing the stopword "not" completely transforms the meaning of a phrase "not great").

2.3.2 Features

Different feature types used in experiments are summarized in Table 2.3, while Figure 2.1 illustrates the process of extracting these features from a sample interview fragment. First, we compared the performance of all classification models using only lexical features, which were derived from the unigram bag-of-words representation of utterances. According to this approach, a set of unique terms (vocabulary) of

size N is first determined for a given collection of textual fragments (in our case, interview transcripts) and then each textual fragment f (in our case, adolescent or caregiver utterance) is represented as a feature vector $[nw_{1,f}, ..., nw_{N,f}]$, where $nw_{n,f}$ is a feature representing the number of times an n^{th} word from the collection vocabulary occurred in f. For example, the vocabulary of a collection consisting of only one textual fragment "what you think about your weight right now and your health" would be ("about", "and", "health", "now", "right", "think", "weight", "what", "you", "your") and the unigram bag-of-words feature vector for this fragment based on the representation would be [1,1,1,1,1,1,1,1,1,2]. Since the question mark (?) is an important indicator of some communication types, it was also used as a feature.

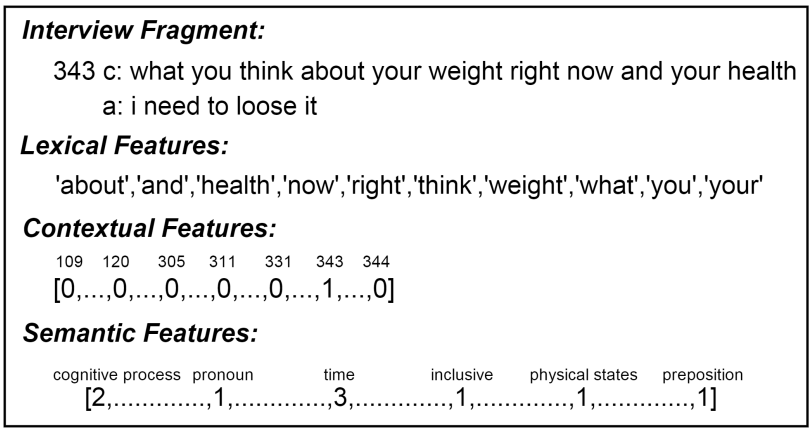

Figure 2.1: Features extracted from a sample interview fragment

Second, we expanded lexical features with the features derived from Linguistic Inquiry and Word Count (LIWC) lexicon [126]. LIWC lexicon consists of the dictionaries, which had been manually compiled and validated for over a decade by psychologists, sociologists and linguists. Dictionaries are organized around sixty-eight psychological and social dimensions, which are structured as an ontology-like hierarchy and may overlap. Each dictionary corresponds to a well-defined concept or psychological construct (e.g. social, positive emotions, negative emotions, money). Social dictionary consists of the nouns and pronouns that refer to other people (e.g., "they", "she", "us", "friends") as well as the verbs that indicate interaction (e.g.,

"talking", "sharing"). Dictionaries of positive (e.g. "happy", "love", "good") and negative (e.g. "sad", "kill", "afraid") words cover the entire spectrum of corresponding emotions from happiness to anxiety. We use the vector of counts of terms in the utterance across LIWC dictionaries as additional semantic features. For example, the sentence "what you think about your weight right now and your health" is represented as a vector [2,...,1,...,3,...,1,...,1,...,1], in which each element is the number of counts of words that fall under each of the sixty eight categories [cognitive process,...,pronoun,...,time,...,inclusive,...,physical states,..., preposition]. LIWC has been applied to successfully predict the onset of depression in individuals based on the text from social media [37] and characterize the emotional variability of pregnant mothers from Twitter posts [36]. In case of annotation of clinical interview transcripts, LIWC features provide important psychological clues related to thought processes, emotional states, intentions, and motivations of patients.

Finally, in addition to lexical features, we also considered the context of interview utterances in the form of the label of the preceding utterance. We hypothesize that contextual features of an utterance play an important role during annotation process since the interviews proceed in sequential manner with participants asking or responding to questions of the previous speaker. Therefore, we use the automatically assigned category of the preceding counselor (adolescent or caregiver) utterance as an additional contextual feature when annotating adolescent or caregiver session transcripts, and vice versa. For example, if the set of codes specific to the counselor utterances is [109,...,120,...,305,...,311,...,331,...,343,...,344], then the additional contextual feature vector for the adolescent utterance "i need to lose it", which is preceded by the counselor utterance annotated with the code 343, is [0,...,0,...,0,...,0,...,0,...,1,...,0].

2.3.3 Classification models

We first describe a general architecture of the classification system used in experiments and, then provide a brief overview of each evaluated machine learning method.

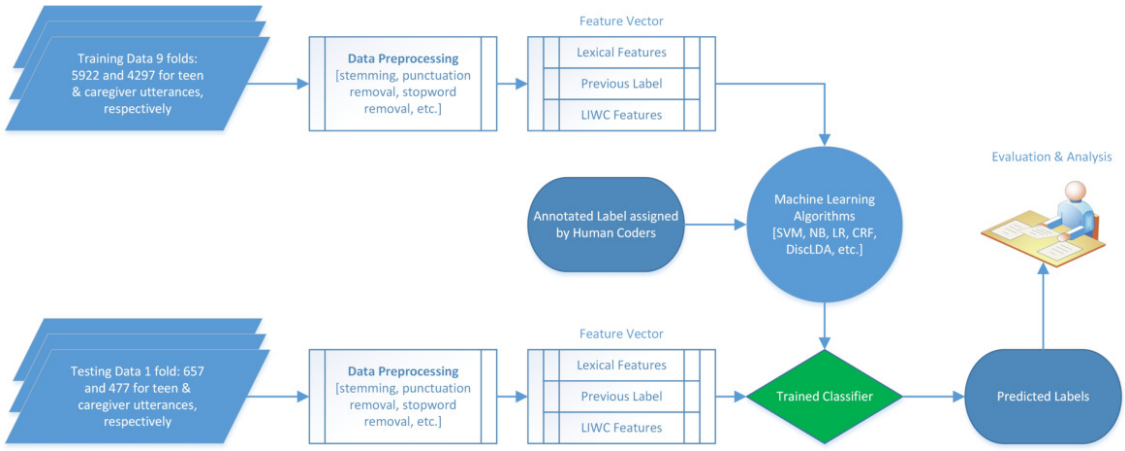

Figure 2.2: Architecture of the pipeline for automatic annotation of clinical interview fragments using different supervised machine learning methods

Figure 2.2 shows the architecture of the pipeline used for the classification of medical interview transcripts.

The pipeline consists of two stages: training and testing. Prior to the training stage, we preprocess the collected clinical interview transcripts by performing stemming, punctuation removal, word segmentation and tokenization. Features are then extracted from the preprocessed data. During this stage, previous label and LIWC features are used in conjunction with the lexical features to create the feature vectors. After that, classifiers are trained on the feature vectors extracted from the training samples and their associated annotations. In the testing stage, after creation of feature vectors, the previously trained classifiers predict the label of each utterance in the testing sample. Finally, performance of different classifiers is evaluated by calculating standard metrics such as precision, recall, F-score (F1), kappa measure and accuracy. Specifically, we evaluated the performance of the following state-of-the-art supervised machine learning methods.

Naïve Bayes (NB)

Naïve Bayes (NB) is as a popular probabilistic method [68, 122] for text classification due to its robustness and relative simplicity. Experimental results reported in

this project were obtained using standard implementations of binomial Naïve Bayes (NB) and multinomial Naïve Bayes (NB-M) algorithms [92] provided by the Weka toolkit.

Support Vector Machine (SVM)

Support Vector Machine (SVM) [34, 42] belongs to a family of generalized linear binary classifiers, which map an input feature vector into a higher dimensional space and finds a hyperplane that separates the samples into two classes in such a way that the margin between the closest samples in each class is maximized. Open-source implementation of SVM with different kernels in publicly available LibSVM[2] [26] package was used for the experiments reported in this work. The parameters of each kernel have been empirically optimized using cross-validation. Figures 2.4, 2.5 and 2.3 illustrate the variance in performance of SVM with different setting of parameters for RBF, polynomial and sigmoid kernels, respectively. As follows from Figure 2.5, when the number of classes is large, SVM has optimal performance when quadratic polynomial kernel is used or when γ is set to 0.1 for a sigmoid kernel. The best performance of SVM among all kernels, however, is achieved when it is used with a Radial Basis Function kernel (RBF) with the parameters C and γ set to 4.0 and 0.1, respectively. We also found that L1 loss function performs better than L2 loss function for Linear SVM.

Conditional Random Fields (CRF)

Conditional Random Fields (CRF) [79] is a probabilistic model, which is different from all other classifiers in that, in addition to lexical features, it also considers the dependencies between the labels of consecutive data samples. We also explain more about the CRF model in Chapter 4. We used linear chain CRF provided by MALLET [93], a publicly available machine learning toolkit[3].

[2]https://www.csie.ntu.edu.tw/ cjlin/libsvm/
[3]http://mallet.cs.umass.edu/

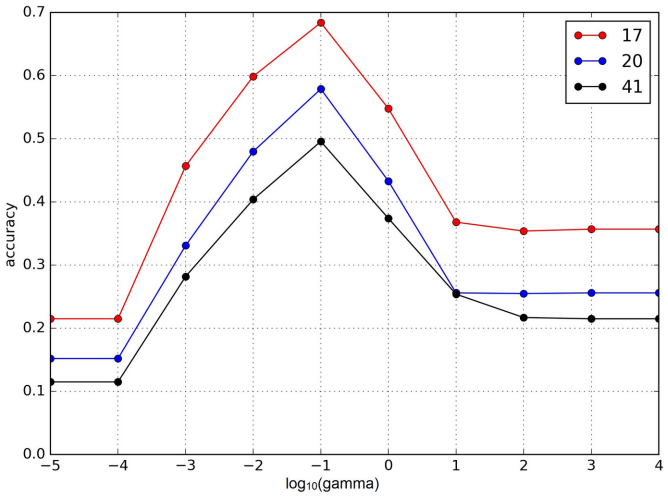

Figure 2.3: Performance of SVM with sigmoid kernel by varying γ

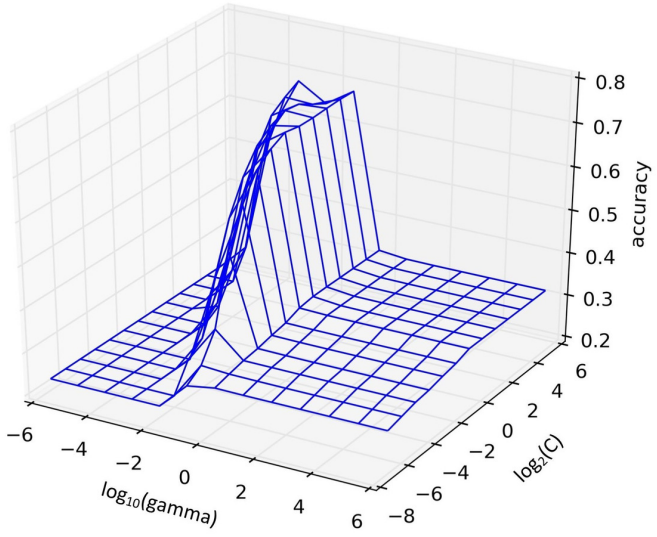

Figure 2.4: Performance of SVM with RBF kernel by varying kernel parameters C and γ

Decision tree (J48)

J48 [119] is an open source implementation of the C4.5 decision tree classification algorithm provided by Weka. Decision trees are interpretable classifiers, which model the classification process as a tree traversal.

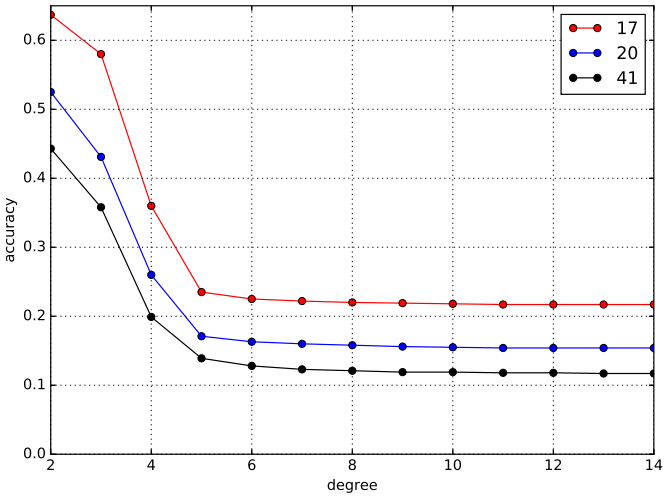

Figure 2.5: Performance of SVM with polynomial kernel by varying the degree

AdaBoost

AdaBoost [47] (short for "Adaptive Boosting") is one of the most widely used and studied machine learning meta-algorithms. Boosting algorithms belong to a group of voting techniques [46], which produce classification decision as a linear combination of the output of other classifiers (also called "base" or "weak" classifiers) [58]. In particular, we used J48 decision tree classifier as a weak learner for AdaBoost.

Random Forest (RF)

Random Forest [21] is an ensemble method that uses bagging to improve classification performance by combining the output of several classifiers. The main idea behind ensemble methods is that a large number of "weak learners" can be used to create a "strong learner". In case of Random Forest, a "weak learner" is a decision tree. Figure 2.6 illustrates the performance of Random Forest by varying the number of individual decision trees. From Figure 2.6, it follows that increasing the number of trees beyond 150 results in minor performance improvement. We used 300 trees for RF, which we empirically determined to result in the best performance of this classifier the codebooks of different size.

DiscLDA

DiscLDA [58] is a dimensionality reduction method that incorporates supervision in the form of class labels into Latent Dirichlet Allocation (LDA) [20] to uncover the latent structure in document collections and leverage this structure to improve the accuracy of classification. Experimental results reported in this project were obtained by setting alpha to $50/T$ [57] where T is a number of topics and β to 0.1 and running the model for 150 iterations. Figure 2.7 shows the performance of DiscLDA depending on the number of topics. From Figure 2.7, it follows that the accuracy of DiscLDA is maximized when 250 topics are used.

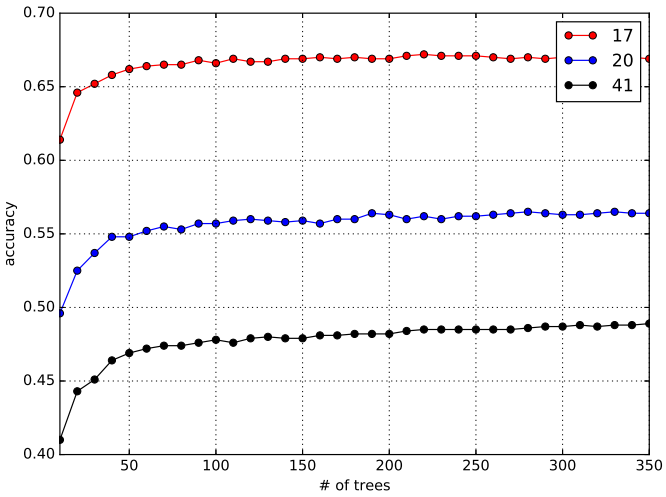

Figure 2.6: Performance of random forest by varying the number of decision trees

Deep Learning (DL)

Deep Learning (DL) exploits the idea of a hierarchy of explanatory factors, in which higher level learned more abstract concepts from the lower level ones. A greedy layer-by-layer method is often used to construct these architectures. Deep learning helps to disentangle these abstractions and select the features that are useful for learning. For supervised learning tasks, instead of extracting manually designed features from the data, deep learning methods translate the data into a compact intermedi-

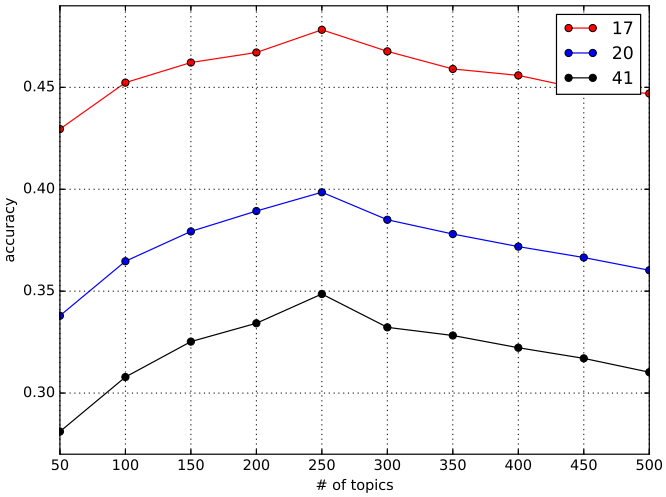

Figure 2.7: Performance of DiscLDA by varying the number of topics

ate representation, similar to other dimensionality reduction techniques, and derive layered structures, which eliminate redundancy in feature representation. We used a convolutional neural network (CNN) with one layer of convolution [71] on top of the latent dimensional representation of each word in an interview fragment using the publicly available WORD2VEC[4] vectors, which were obtained from an unsupervised neural language model [95] estimated on 100 billion word corpus from Google News. If a WORD2VEC vector was not available for a particular word, we used random initialization for its latent dimensional representation. In the architecture of this CNN, shown in Figure 2.8, an interview fragment consisting of n words is represented by n 300 dimensional WORD2VEC vectors, which were fine-tuned for our dataset through backpropagation. A convolution operation using multiple filters corresponding to the windows of size 3, 4 and 5 words was then applied to produce new features. After that, a max-over-time pooling [33] is used to capture the most important feature for each particular filter. These features form the penultimate layer and are then passed to a fully connected softmax layer whose output is a probability distribution over category assignments for a given interview fragment. Based on emperical analysis

[4] `https://code.google.com/p/word2vec/`

in [139], we tuned two important parameters to improve the performance of CNN: dropout rate and number of featuremaps. The effect of dropout rate and the number of featuremaps on performance of CNN is shown in Figures 2.10 and 2.9, respectively. As follows from Figure 2.9, the number of faturemaps does not have a significant effect on the performance of CNN, when the number of classes is large.

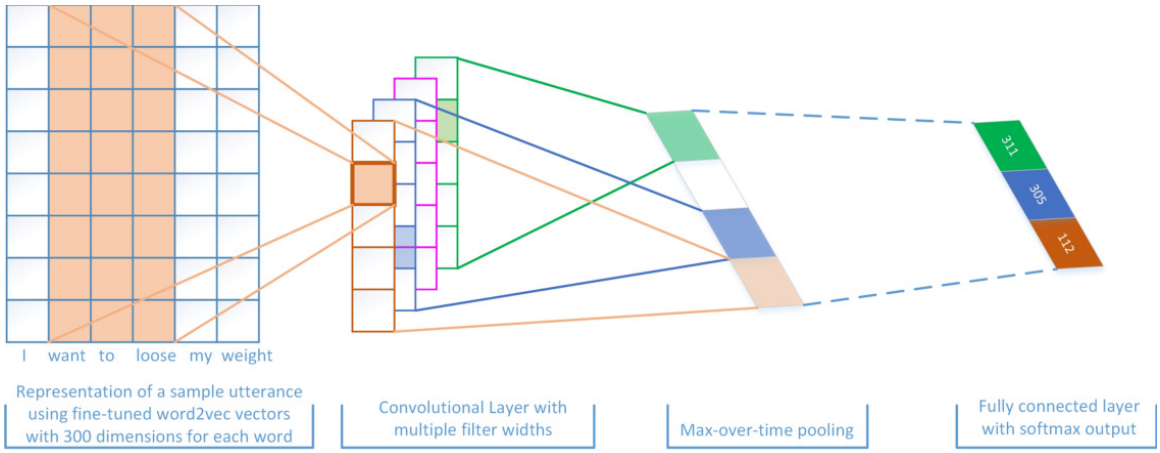

Figure 2.8: Architecture of convolutional neural network for automatic annotation of clinical interview transcripts

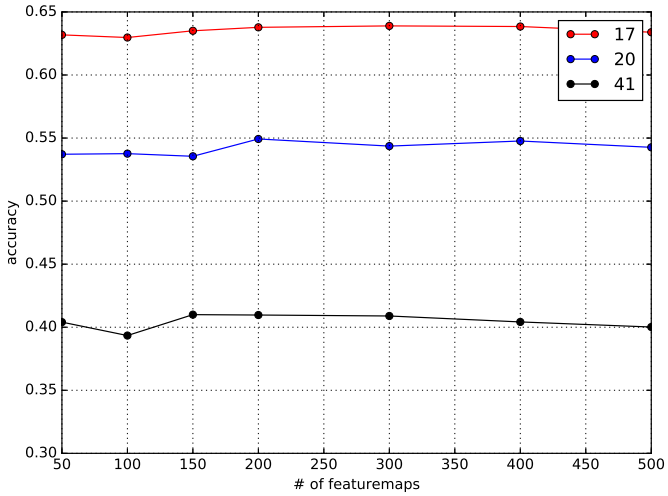

Figure 2.9: Performance of CNN by varying the number of featuremaps

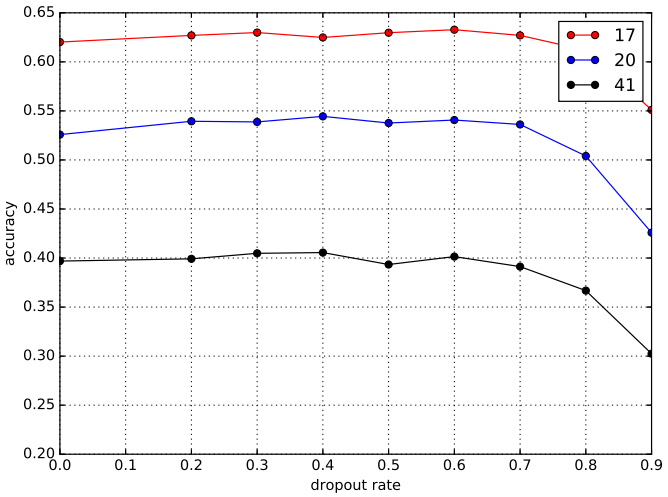

Figure 2.10: Performance of CNN by varying the dropout rate

2.3.4 Evaluation

To ensure the robustness of performance estimates, we used 10-fold cross validation [73] as an experimental design. The performance of different classifiers and feature sets was evaluated in terms of precision, recall, F1 score (F1), kappa measure and accuracy using weighted macro-averaging over 10 folds.

2.4 Results

Experimental evaluation of automatic annotation using machine learning included several dimensions:

- determining the performance of classifiers on the codebooks of different size;
- determining the effectiveness of the proposed contextual and semantic features.

Since clinical researchers typically annotate caregiver and adolescent sessions separately, we first created two experimental datasets consisting of only adolescent and only caregiver session transcripts. Second, besides evaluating the accuracy of annotating adolescent and caregiver transcripts with the codebooks containing an entire set of codes, we also conducted a series of experiments with the codebooks of smaller sizes created as outlined above. Third, besides training and testing NB, SVM, CRF,

Decision Tree, Boosting, DiscLDA, Random Forest and CNN classifiers using only lexical features, we also evaluated the effectiveness of the proposed contextual and semantic features.

2.4.1 Quality of automatic annotation using only lexical features

Standard performance metrics[5] of different classification models using only lexical features for the task of annotating adolescent and caregiver session transcripts are summarized in Tables 2.4 and 2.5, respectively.

Table 2.4: Performance of classification models using only lexical features according to different evaluation metrics for the task of annotating adolescent interview session transcripts. Highest value for each metric and codebook size across all models is highlighted in boldface.

Cls.	Model	Acc.	Prec.	Rec.	F1	Kappa
17	NB	0.544	0.603	0.544	0.552	0.497
	NB-M	0.670	0.662	0.670	0.643	0.622
	J48	0.595	0.573	0.595	0.580	0.539
	AdaBoost	0.627	0.600	0.627	0.609	0.574
	RF	0.670	0.662	0.670	0.625	0.616
	DiscLDA	0.477	0.454	0.477	0.431	0.388
	SVM	**0.708**	**0.705**	**0.708**	**0.680**	**0.663**
	CNN	0.678	0.633	0.678	0.670	0.509
20	NB	0.487	0.509	0.487	0.482	0.448
	NB-M	0.579	0.582	0.579	0.559	0.537
	J48	0.479	0.467	0.479	0.470	0.431
	AdaBoost	0.504	0.488	0.504	0.493	0.458
	RF	0.563	0.564	0.563	0.519	0.514
	DiscLDA	0.400	0.410	0.400	0.356	0.330
	SVM	**0.610**	**0.611**	**0.610**	**0.592**	**0.571**
	CNN	0.586	0.588	0.586	0.587	0.476
41	NB	0.406	0.434	0.406	0.405	0.375
	NB-M	0.513	0.479	0.513	0.484	0.478
	J48	0.396	0.375	0.396	0.382	0.356
	AdaBoost	0.436	0.412	0.436	0.421	0.398
	RF	0.495	0.487	0.495	0.453	0.455
	DiscLDA	0.362	0.387	0.362	0.301	0.304
	SVM	**0.537**	**0.513**	**0.537**	**0.504**	**0.502**
	CNN	0.396	0.369	0.396	0.382	0.170

[5]Cls.: # of classes, Acc.: Accuracy, Prec.: Precision, Rec.: Recall

Table 2.5: Performance of classification models using only lexical features according to different evaluation metrics for the task of annotating caregiver interview session transcripts. Highest value for each metric and codebook size across all models is highlighted in boldface.

Cls.	Model	Acc.	Prec.	Rec.	F1	Kappa
16	NB	0.571	0.608	0.571	0.575	0.518
	NB-M	0.633	0.629	0.633	0.604	0.573
	J48	0.578	0.563	0.578	0.567	0.514
	AdaBoost	0.602	0.582	0.602	0.588	0.539
	RF	0.640	0.631	0.640	0.596	0.574
	DiscLDA	0.482	0.442	0.482	0.421	0.362
	SVM	**0.664**	**0.653**	**0.664**	**0.639**	**0.606**
	CNN	0.657	0.641	0.657	0.648	0.512
19	NB	0.477	0.504	0.477	0.467	0.434
	NB-M	0.536	0.539	0.536	0.512	0.487
	J48	0.436	0.431	0.436	0.432	0.382
	AdaBoost	0.467	0.457	0.467	0.460	0.415
	RF	0.507	0.508	0.507	0.467	0.450
	DiscLDA	0.374	0.370	0.374	0.333	0.287
	SVM	**0.545**	**0.547**	**0.545**	**0.535**	**0.497**
	CNN	0.510	0.498	0.510	0.504	0.401
58	NB	0.379	0.392	0.379	0.370	0.350
	NB-M	0.442	0.404	0.442	0.386	0.401
	J48	0.340	0.321	0.340	0.328	0.302
	AdaBoost	0.381	0.359	0.381	0.366	0.344
	RF	0.402	0.358	0.402	0.352	0.358
	DiscLDA	0.288	0.258	0.288	0.234	0.229
	SVM	**0.451**	**0.420**	**0.451**	**0.418**	**0.414**
	CNN	0.118	0.102	0.118	0.109	0.032

Several observations can be made based on Tables 2.4 and 2.5. First, SVM consistently demonstrates the best performance while DiscLDA and J48 consistently have the worst performance in terms of all metrics and for the codebooks of all sizes on both adolescent and caregiver interview session transcripts. In case of DiscLDA, this indicates that dimensionality reduction is less effective when the number of classes is large. In case of J48, this indicates that decisions trees are not effective in case of sparse high-dimensional feature vectors and large number of classes. Furthermore, the difference in performance between SVM and other classifiers keeps increasing with the number of classes in the codebook. For example, in case of adolescent interview

transcripts, the difference in accuracy between SVM and CNN (the best and second best) is 3% when the codebook with 17 labels is used, 2.4% when the codebook with 20 labels in used and 14.1% when the codebook with 41 labels is used. This indicates superior robustness of SVM compared to other machine learning methods. Second, although boosting with AdaBoost consistently improves the performance of J48 in terms of all metrics and for the codebooks of all sizes and on both adolescent and caregiver interview session transcripts, SVM and, in many cases, multinomial NB, outperformed AdaBoost, particularly in case of the codebooks with large number of codes (41 labels in case of the adolescent and 58 labels in case of caregiver session-specific codebooks), which indicates that boosting is less effective for classification tasks involving large number of classes. Third, CNN outperforms all other classifiers except CRF and SVM in all codebook sizes except 41 and 58. The differences in accuracy between SVM and CNN are 0.7%, 3%, 3.5%, 2.4%, 14.1% and 33.3% when the codebooks of size 16, 17, 19, 20, 41 and 58 are used, respectively. These results indicate that CNN is less effective for classification problems when the number of classes is large. Fourth, the performance of all classification models is consistently lower on caregiver utterances than on adolescent utterances, which can be explained by the relative simplicity of the language used by the adolescents.

ROC curves in Figures 2.11, 2.12 and 2.13 illustrate the relative performance of different classifiers for the codebooks of different size.

2.4.2 Quality of automatic annotation using lexical and non-lexical features

Summary of performance[6] of CRF and SVM using the combinations of lexical and contextual (SVM-PL), lexical and semantic (SVM-LIWC) and all features (SVM-AF) on adolescent and caregiver session transcripts is provided in Tables 2.6 and 2.7, respectively.

[6]Cls.: # of classes, Acc.: Accuracy, Prec.: Precision, Rec.: Recall

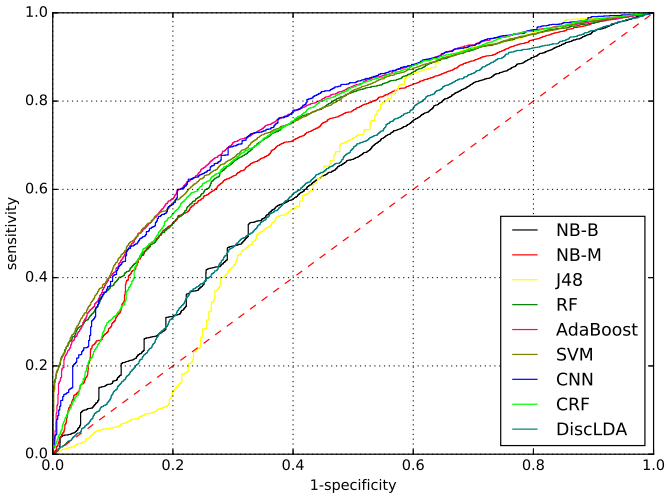

Figure 2.11: ROC curves for all classifiers when the codebook with 17 classes is used

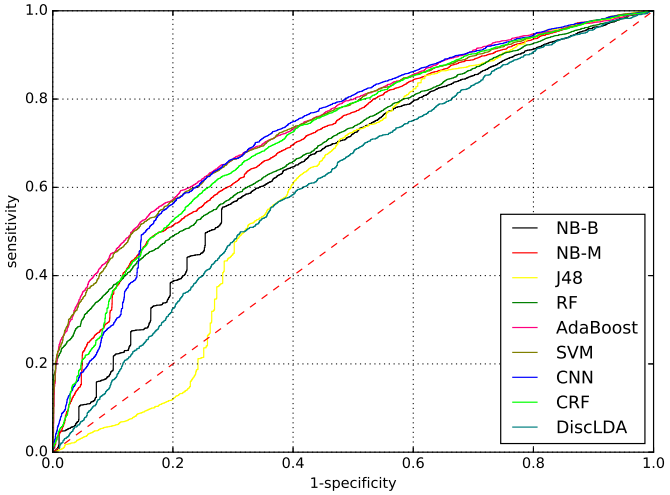

Figure 2.12: ROC curves for all classifiers when the codebook with 20 classes is used

Several important conclusions can be made by comparing the experimental results in Tables 2.6 and 2.7 with Tables 2.4 and 2.5. First, CRF outperformed multinomial NB, achieving 1.2% and 0.2% higher accuracy and 3.4% and 2.1% higher F1 score when the codebooks with 17 and 20 labels, respectively, were used to annotate the adolescent transcripts and 2.1% and 0.3%higher accuracy and 4.9% and 2.8% higher F1 score when the codebooks with 16 and 19 labels, respectively, were used to annotate the caregiver transcripts. However, CRF provides 2% and 2.7% lower accuracy with 2% and 2.7% lower F1 score and 2% and 0.4% lower accuracy with 2.7% and

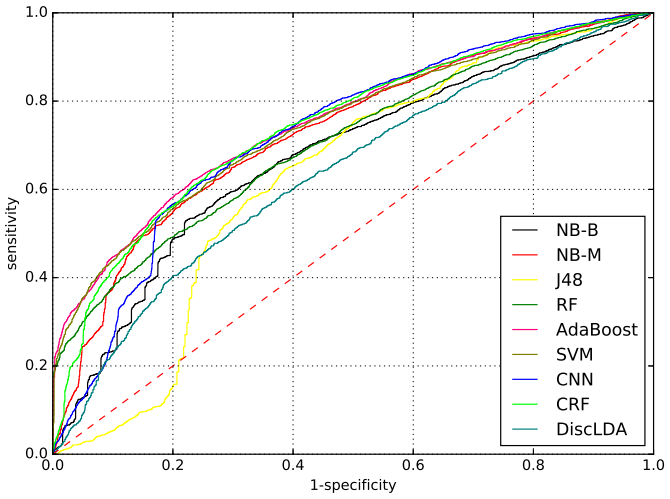

Figure 2.13: ROC curves for all classifiers when the codebook with 41 classes is used

3.7% lower F1 score when 41 and 58 labels are used respectively. On the other hand, the accuracy of CRF is worse than the accuracy of SVM using lexical features by 2.6% , 2.9% and 4.4% with codebook size 17, 20 and 41, respectively, on adolescent transcripts and by 7.4% , 0.6% and 1.3% with codebook size 16, 19 and 58, respectively, on caregiver transcripts. Nevertheless, since CRF considers both lexical features as well as the labels of previous utterances, these results highlight the importance of accounting for context when annotating the utterances in clinical interview transcripts.

Second, the performance of SVM improves in terms of all metrics on both adolescent and caregiver datasets and for the codebooks of all sizes when either contextual (SVM-PL) or semantic (SVM-LIWC) features are used in addition to the lexical ones. When both of these features are used together (SVM-AF), the annotation performance of SVM improves even further achieving the best performance in terms of all metrics using the codebooks of all sizes on both adolescent and caregiver transcripts. In particular, by using contextual and semantic features in addition to the lexical ones, the accuracy of SVM improves by 4.3%, 7.2%, and 3.1%, while it's F1 score improves by 5.9%, 8.2%, and 4.2%, when the codebooks with 17, 20, and 41 labels,

Table 2.6: Performance of classification models using a combination of lexical and different types of non-lexical features according to standard metrics for the task of annotating adolescent interview session transcripts

Cls.	Model	Acc.	Prec.	Rec.	F1	Kappa
17	CRF	0.682	0.673	0.682	0.677	0.636
	SVM	0.708	0.705	0.708	0.680	0.663
	SVM-PL	0.715	0.711	0.715	0.696	0.673
	SVM-LIWC	0.742	0.740	0.742	0.727	0.704
	SVM-AF	**0.751**	**0.750**	**0.751**	**0.739**	**0.715**
20	CRF	0.581	0.579	0.581	0.580	0.540
	SVM	0.610	0.611	0.610	0.592	0.571
	SVM-PL	0.639	0.642	0.639	0.630	0.604
	SVM-LIWC	0.653	0.653	0.653	0.657	0.619
	SVM-AF	**0.682**	**0.685**	**0.682**	**0.674**	**0.651**
41	CRF	0.493	0.485	0.493	0.457	0.502
	SVM	0.537	0.513	0.537	0.504	0.502
	SVM-PL	0.565	0.543	0.565	0.542	0.535
	SVM-LIWC	0.538	0.518	0.538	0.507	0.503
	SVM-AF	**0.568**	**0.549**	**0.568**	**0.546**	**0.538**

respectively, are used to annotate the adolescent transcripts. When contextual and semantic features are used, the accuracy of SVM improves by 7.4%, 9.3%, and 3.7% and its F1 score improves by 8.8%, 9.6%, and 4.4% when the codebooks with 16, 19, and 58 labels, respectively, are used to annotate the caregiver transcripts.

Comparison of performance of different classification models

The accuracy of NB-M, SVM, CNN, CRF, SVM-AF, J48 decision tree, Random Forest, AdaBoost, and DiscLDA classification models for the task of annotating adolescent and caregiver datasets is compared across the codebooks of different sizes in Figures 2.14 and 2.15.

From Figure 2.14 and 2.15, it follows that SVM and CRF achieve around 52%, 60%, and 70% accuracy when using the codebooks consisting of 41, 20, and 17 labels, respectively, to annotate adolescent session transcripts and 45%, 55%, and 66% ac-

Table 2.7: Performance of classification models using a combination of lexical and different types of non-lexical features according to standard metrics for the task of annotating caregiver interview session transcripts

Cls.	Model	Acc.	Prec.	Rec.	F1	Kappa
16	CRF	0.654	0.652	0.654	0.653	0.603
	SVM	0.664	0.653	0.664	0.639	0.606
	SVM-PL	0.670	0.658	0.670	0.651	0.614
	SVM-LIWC	0.730	0.730	0.730	0.717	0.686
	SVM-AF	**0.738**	**0.733**	**0.738**	**0.727**	**0.696**
19	CRF	0.539	0.541	0.539	0.540	0.492
	SVM	0.545	0.547	0.545	0.535	0.497
	SVM-PL	0.566	0.570	0.566	0.559	0.522
	SVM-LIWC	0.620	0.625	0.620	0.613	0.581
	SVM-AF	**0.638**	**0.639**	**0.638**	**0.631**	**0.601**
58	CRF	0.438	0.409	0.438	0.423	0.385
	SVM	0.451	0.420	0.451	0.418	0.414
	SVM-PL	0.480	0.462	0.480	0.456	0.446
	SVM-LIWC	0.459	0.445	0.459	0.429	0.422
	SVM-AF	**0.488**	**0.466**	**0.488**	**0.462**	**0.454**

curacy when using the codebooks consisting of 58, 19, and 16 labels, respectively, to annotate caregiver session transcripts. CNN also has approximately the same performance as SVM and CRF, when the codebooks consisting of 16, 17 and 20 labels are used. However, CNN has significantly lower performance compared to SVM and CRF in terms of all metrics when the codebook of size 41 and 58 labels are used. SVM-AF consistently outperforms all other methods across the codebooks of all sizes on both datasets, achieving the highest accuracy of 75.1% (which is close to human accuracy), when the codebook consisting of 17 classes is used for annotating adolescent interview session transcripts, and of 73.8%, when the codebook consisting of 16 classes is used for annotating caregiver interview session transcripts.

Depending on the type of the interview transcript and the codebook size, SVM-AF achieves 3%–9% higher accuracy and 4%–10% higher F1 score than SVM and 4%–10% higher accuracy and 4%–11% higher F1 score than CRF, which highlights

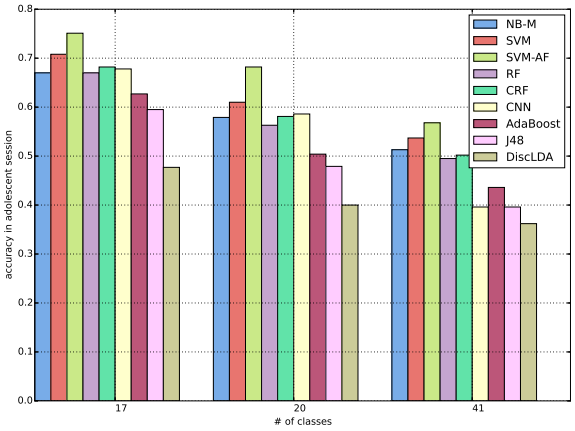

Figure 2.14: Comparison of annotation accuracy of adolescent interview fragments with different machine learning methods and feature sets

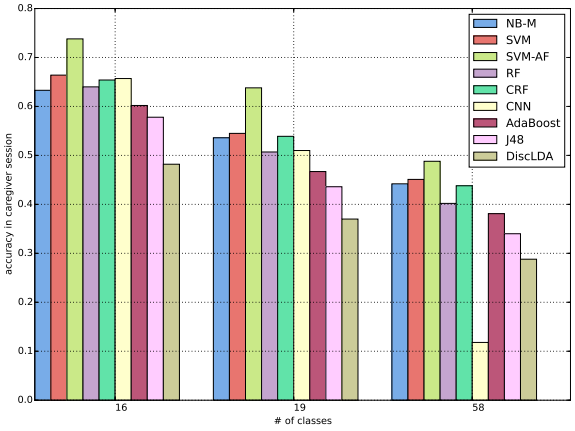

Figure 2.15: Comparison of annotation accuracy of caregiver interview fragments with different machine learning methods and feature sets

the importance of contextual and semantic features.

Reliability of the best classifier in other study data

We tested the accuracy and reliability of the best machine learning classification model developed in the above work in a new treatment setting, HIV medical care. The training dataset for this study was composed of 80 patient-provider clinical interactions during routine HIV clinic visits previously coded with the MY-SCOPE coding instrument. We also tested the robustness of our SVM-AF model with 49 eCoaching

sessions and 129 Obesity sessions. Our working hypothesis was that the classification model developed in the above study would demonstrate the transferability of knowledge by achieving a high level of coding accuracy. Table 2.8 shows the performance of the SVM-AF model on HIV, eCoaching and Obesity datasets.

Table 2.8: Performance of SVM model using a combination of lexical and different types of non-lexical features according to standard metrics for the task of annotating MI interview session transcripts in HIV, eCoaching and Obesity studies, respectively

Dataset	Accuracy	Precision	Recall	F1-Measure
HIV	0.720	0.701	0.720	0.696
e-Coaching	**0.798**	**0.793**	**0.798**	**0.782**
Obesity	0.751	0.750	0.751	0.739

The SVM-AF model, with no modifications from the above study, achieved 69.6% F1-score with 70.1% precision and 72.0% recall for the task of automatic annotation of utterances in patient-provider encounters in HIV clinic. The SVM-AF model also demonstrated good performance in both datasets, achieved 79.8% F1-score with 79.3% precision and 79.8% recall in eCoaching sessions and 73.9% F1-score with 75% precision and 75.1% recall in Obesity sessions. These results illustrate the effectiveness of transfer learning strategies or applying machine learning models trained on one clinical context (e.g., weight loss) to another clinical context (e.g., HIV patient visits). Effective transfer of machine learning models can significantly reduce the time and resources needed to develop the training datasets for different types of clinical discourse.

2.5 Discussion

Our experimental evaluation of supervised machine learning methods for the task of automatic annotation of clinical interview transcripts resulted in several important observations and conclusions. First, although CNN has comparable performance to SVM when the number of classes is relatively small, its performance drastically decreases when the number of classes gets large. Remarkably, for very large number

of classes (41 and 56, in our case) Deep Learning is less effective-than a random guess. Second, multinomial and binomial Naïve Bayes, AdaBoost, Random Forest, and DiscLDA have been consistently outperformed on both datasets and codebook sizes by CNN, CRF and SVM, when all models use only lexical features. Superior generalization ability of SVM even in case of a large number of classes and features (which is the case when lexical features are used) can be attributed to its ability to learn the classification model independent of the dimensionality of feature space.

We also observed a consistent trend of performance improvement for SVM when adding non-lexical features, such as the label of the preceding utterance and the features derived from LIWC dictionaries, to the lexical ones. The first result indicates that the context of an utterance in clinical interview transcripts in the form of the label of the preceding utterance plays an important role in the classification process, besides the content of the utterance itself. The second result indicates that, for the purpose of classification, the semantics of an utterance in clinical interviews can be approximated with a distribution of its words across LIWC dictionaries.

2.6 Summary

In this chapter, we propose novel features and report the results of an extensive experimental evaluation of state-of-the-art supervised machine learning methods for text classification using those features, to help clinical researchers and practitioners assess the feasibility of using these methods for the task of automatic annotation of clinical text using the codebooks of realistic size. We found out that Support Vector Machine using only lexical features consistently outperforms all other classifiers on caregiver and adolescent datasets according to most metrics. Adding contextual and semantic features further improves the performance of SVM on both datasets, achieving close to human accuracy when the codebooks consisting of 16 and 17 classes are used to annotate caregiver and adolescent transcripts, respectively.

This work has important practical implications. First, it can facilitate researchers

to establish causal relationship between different communication strategies and desired behavioral outcomes without having to repeatedly wade through pages of interview transcripts. Second, since automatic annotation is significantly faster than manual, it can dramatically accelerate the pace of research in behavioral sciences. Although all experiments were conducted on interview transcripts, the proposed methods and features are not specific to a particular domain of Motivational Interviewing, and thus there is also no prima facie reason to believe that they will not be effective for annotation of any other type of clinical conversation.

CHAPTER 3 SEQUENTIAL ANALYSIS OF CLINICAL CONVERSATION

In this chapter, we describe the sequential analysis of annotated clinical conversation to inform best clinical practice by facilitating the use of more effective and tailored counselor communication. We first explain the problem of classifying patient-counselor communication sequences in the context of the clinical conversation. Specifically, we focus on predicting the success (i.e. eliciting a particular type of patient behavioral response) of motivational interviews with obese adolescents and their caregivers based on an observed sequence of coded patient-counselor communication exchanges during those interviews. We then move on sequential analysis of pre-coded clinical conversation to identify patterns of patient-counselor communication in successful and unsuccessful sequences in MI sessions.

3.1 Classification of communication sequences

3.1.1 Introduction

Temporally ordered sequences of discrete or continuous observations generated by molecular, psychological or psychological process(es) arise in many different areas of biology and medicine (e.g., DNA base-pairs, protein sequences, ECG measurements, laboratory results, diagnostic codes, utterances in the clinical dialog). Classification (or categorization) is a type of analysis of those sequences that has a broad range of important practical applications, from protein function [137] or structure [39] prediction to detecting individuals with a heart disease [134]. Taking into account both the entire set of observations in a sequence, as well as the temporal order and potential dependencies between observations, makes sequence classification a more challenging task than a classification of independent observations. Predicting the outcome of those sequences (e.g. physiological or behavioral response) can also be viewed as a sequence classification problem.

In this work, we address the problem of predicting the outcome of coded patient-

provider communication (PPC) sequences in the context of the clinical dialog. Specifically, we focus on predicting the success (i.e. eliciting a particular type of patient behavioral response) of motivational interviews with obese adolescents and their caregivers based on an observed sequence of coded PPC exchanges during those interviews. Childhood obesity is a serious public health concern in the United States. Recent estimates indicate that approximately one-third (31.8%) of U.S. children 2-19 years of age are overweight and 16.9% are obese [107]. Adolescents, who are obese, are likely to be obese in adulthood and have a greater risk of heart disease, type 2 diabetes, stroke, cancer, and osteoarthritis [52]. One approach to effective obesity intervention is Motivational Interviewing (MI), an evidence-based counseling technique to increase intrinsic motivation and self-efficacy for health-related behavior change. The goal of MI is to encourage patients to explore their own desires, ability, reasons, need for and commitment to the targeted behavior change. These statements collectively referred to as "change talk" (CHT), consistently predict the actual behavior change[10] that can be sustained for as long as 34 months[132] after an interview. However, the ability of providers to consistently elicit this type of patient communication requires knowledge of effective communication strategies for a variety of patients, which can only be obtained through analysis of a large number of annotated interviews. Since manual examination and analysis of MI interview transcripts is a very time-consuming process, designing effective MI interventions and tailoring them to particular populations can take years. Therefore, there is a need for informatics-based methods to facilitate the development of effective behavioral interventions, in general, and theoretically-grounded computational models to explore the mechanisms of MI's efficacy, in particular.

We compared the accuracy of probabilistic models, such as MC and HMM, and deep learning methods, such as LSTM and GRU, for the task of predicting the success of clinical interviews (i.e. eliciting a particular type of patient behavioral response,

such as CHT) at any point during a clinical interview based on a sequence of coded previous PPC exchanges in the same interview. This study was a continuation of our previous work [74, 62], in which we explored several machine learning methods for automatic annotation of clinical interview fragments with a large number of patient and provider behavior codes from a specialized codebook [24]. While there have been some previous qualitative studies of patient-provider dialog in a clinical setting [43], this is the first work explored the applicability of state-of-the-art methods for sequence modeling to the analysis of PPC exchanges, in general, and predicting the desired patient behavioral response in the context of motivational interviews, in particular.

3.1.2 Related work

In general, sequence classification methods fall into one of three major classes: feature-based, distance-based and model-based. Feature-based methods transform a sequence into a feature vector and apply a standard supervised machine learning method, such as Support Vector Machine [83] or Decision Tree [32] to arrive at classification decision. The methods in this class have had limited success since traditional feature representation methods cannot easily account for the order of and dependencies between observations in a sequence. For an example, behavioral codes could be represented as a bag of codes (features) disregarding the order of its codes but keeping counts. Distance-based methods classify a sequence by finding the most similar sequences with known classes based on a distance metric. The most commonly used distance metric is Euclidean distance, the similarity between two sequences of the same length can be computed by taking the sum of the ordered point-to-point distance between them. Another metric Dynamic Time Wrapping (DTW) [69] makes distance comparisons more robust because it supports a variable length sequence and insensitive with respect to signal shifting and scaling. However, these distance metrics are primarily designed for time series data, in which the observations are discretized by timestamps.

The third type of sequence classification methods first creates a probabilistic model, such as the Markov Chain (MC) or Hidden Markov Model [114] (HMM), for sequences in each class based on the training data and then, classifies new sequences by applying the created models. While MCs and HMMs can capture first- and second-order dependencies between adjacent observations in a sequence, learning higher-order dependencies with these models requires prohibitively large amounts of data and utilized as a baseline for our sequence classification study. [63] By encoding sequences into low-dimensional representations, Recurrent Neural Networks (RNNs) are able to capture both short- and long-term dependencies and were shown to be effective at modeling different types of sequential data [84]. Long Short-Term Memory (LSTM) [65] is a variant of RNNs, which successfully addressed the vanishing gradient problem [17] of traditional RNN. LSTM demonstrated excellent performance in different domains, from speech [55] and handwriting recognition[106] to health informatics [85, 30]. LSTM was also effectively used for predicting the diagnosis and medication codes, given a sequence of codes from the previous patient visits [30]. A further simplification and improvement of LSTM model, called the Gated Recurrent Unit (GRU)[31], was later proposed. LSTM and GRU demonstrated markedly better performance among all other RNN variants for a variety of tasks in different domains.

3.1.3 Methods

Dataset

The experimental dataset for this work was constructed from the transcripts of 129 motivational interviews, which consist of a total of 50,239 segmented and annotated utterances. Each transcript corresponds to an MI interview session, which typically involves a counselor, an adolescent and a caregiver. The utterances were annotated based on the MYSCOPE codebook [24], in which the behavior codes are grouped into the patient (adolescent and caregiver) codes and the counselor codes. Annotated utterances were divided into successful and unsuccessful communication sequences.

Table 3.1: Fragment of the annotated transcript of a dialogue between a counselor and an adolescent. MYSCOPE codes assigned to the utterances and their meaning are shown in the first two columns.

Code	**Behavior**	**Speaker**	**Utterance**
SS	Structure Session	Counselor	Okay. Can I meet with Xxxx alone for a few minutes?
OQO	Open-ended question, other	Counselor	So, Xxxx, how you doing?
HUPO	High uptake, other	Adolescent	Fine
OQTBN	Open-ended question, target behavior neutral	Counselor	That's good. So, tell me how do you feel about your weight?
CHT+	Change talk positive	Adolescent	It's not the best.
CQECHT+	Closed question, elicit change talk positive	Counselor	It's not the best?
CHT+	Change talk positive	Adolescent	Yeah
CQTBN	Closed question, target behavior neutral	Counselor	Okay, so have you tried to lose weight before?
HUPW	High uptake, weight	Adolescent	Yes

Successful communication sequences are the ones, which resulted in positive change talk (CHT+) or commitment language (CML+) statements by an adolescent or a caregiver, while unsuccessful sequences are the ones, which resulted in negative change talk (CHT-) or commitment language (CML-), or the ones, in which no change talk or commitment language statements were made.

A fragment of an adolescent session transcript is presented in Table 3.1. In this example, $SS \rightarrow OQO \rightarrow HUPO \rightarrow OQTBN \rightarrow CHT+$ is a successful sequence, in which a counselor starts with an open-ended question and ultimately is able to elicit a positive change talk statement. As follows from this example, similar utterances, such as "Yeah" and "Yes", can be assigned different behavior codes (CHT+ and HUPW), depending on the context.

The resulting experimental dataset was highly imbalanced. Out of 5143 observed sequences, 4225 or 82.15% were positive and only 918 or 17.85% were negative. No major differences were observed in the average length of successful (9.79 utterances) and unsuccessful (9.65 utterances) sequences.

Since severely imbalanced datasets often distort the true performance of a classification method relative to a simple "majority vote" baseline (e.g. simply classifying every communication sequence as successful would result in 82.15% accuracy on our dataset), it is important to properly address the class imbalance. We evaluated the performance of probabilistic and deep learning methods using both under-sampling and over-sampling for balancing the number of samples in different classes. Synthetic Minority Over Sampling Technique (SMOTE) [27] is a widely used oversampling method for imbalanced datasets, in which new synthetic examples are generated for minority classes. Specifically, we generated synthetic examples at the borderline between the majority and minority classes [105]. On the other hand, the under-sampling method reduces the number of samples in majority class by replacing the clusters of samples identified by the k-means clustering algorithm with the cluster centroids.

Sequence classification methods

In general, a sequence can be viewed as a temporally ordered set of observations. An observation corresponds to a behavior code, which has a symbolic representation, such as $LUP+$ (low uptake, positive), $OQECHT+$ (open-ended question, elicit change talk positive), etc. Given a sequence of behavior codes $S_i = \{c_1, c_2, ..., c_n\}$ representing PPC exchanges during some part of a motivational interview, the task of predicting interview success can be considered as sequence classification. Given a set of class labels $L = \{l_1, l_2, ..., l_m\}$ (in our case, the labels are "successful" and "unsuccessful" motivational interview), a sequence classifier C learns a function $S_i \rightarrow l_i, l_i \in L$ that maps a sequence S_i into a class label $l_i \in L$.

Our designed baseline prediction method consists of two steps. In the first step,

we model successful and unsuccessful patient-provider interactions using first and second-order Markov Chain and Hidden Markov Model, which are popular probabilistic models for discrete observation sequences with finite vocabulary. In the second step, we classify each test sequence based on the maximum likelihood of generating that sequence from each model. Although HMM was originally developed for speech recognition [114], it is one of the most widely used methods for sequence modeling [103, 135]. However, the latest advances in deep learning suggest that RNNs may provide better results than conventional machine learning methods for the task of sequence classification. To verify this hypothesis, we employed two state-of-the-art variants of RNN in our experiments: Long Short-Term Memory (LSTM) [65] and Gated Recurrent Unit (GRU) [31].

Markov Chain (MC) is a probabilistic model that conditions each observation in a sequence only on preceding observation and not on any other past observation. First, we estimated two Markov models M and $\overline{M}$, summarizing counselor strategies and patient responses, in the cases of successful (M) and unsuccessful ($\overline{M}$) motivational interviews. A Markov model M can be represented as a weighted directed graph $G = (V, E, p)$, in which:

- $V = \{CML+, CHT+, CHT-, AMB-, LUP+, LUP-, HUPW, CQECHT+, ...\}$ is a set of vertices, consisting of adolescent, caregiver and counselor MI behavior codes;

- $E \subseteq V \times V$ is a set of edges corresponding to possible transitions from one MI behavior code to the other in a sequence;

- $p_M : E \rightarrow [0...1]$ is a function that assigns probability $p(c_i|c_j)$ to an edge between the MI behavior codes c_i and c_j based on the maximum likelihood estimation:

$$P_M(c_j|c_i) = \frac{n_{c_i,c_j}}{n_{c_i}} \tag{3.1}$$

where n_{c_i,c_j} and n_{c_i} are the number of times a transition between the MI behavior codes c_i and c_j and the number of times the code c_i have been observed in the training data, respectively. Given a Markov model M (such that $S \subseteq V$), the probability that a sequence of MI behavior codes $S = \{C_1, ..., C_N\}$ has been generated from a Markov model M is:

$$P_M(S) = \prod_{i=2}^{N} p_M(c_i|c_1, \ldots, c_{i-1}) = \prod_{i=2}^{N} p_M(c_i|c_{i-1}) \tag{3.2}$$

In the second step, we quantify the likelihood of success of a given motivational interview at a certain time point given a sequence of MI behavior codes S observed prior to that point as:

$$p(S \rightarrow successful) = \log\left(\frac{P_M(S)}{P_{\overline{M}}(S)}\right) = \sum_{i=2}^{N} \log p_M(c_i|c_{i-1}) - \sum_{i=2}^{N} \log p_{\overline{M}}(c_i|c_{i-1}) \tag{3.3}$$

If $p(S \rightarrow successful) > 0$, a communication sequence is predicted to be successful (i.e. result in positive change talk or commitment language). Otherwise, it is predicted to be unsuccessful.

The above model is also referred as first-order MC, since it only considers immediately preceding behavior code, when computing the state transition probabilities. In our experiment, we also considered second-order Markov model, which conditions each observation on the preceding two observations.

Hidden Markov Model (HMM) is another probabilistic model used for modeling processes varying in time. HMMs are widely used for sequence analysis because of their ability to identify hidden states, corresponding to clusters of observations. Mathematically, HMM can be defined as $\lambda = (A, B, \pi)$, where:

- A is an $N \times N$ state transition probability distribution matrix $A = \{a_{ij}\}$
- B is an $N \times M$ matrix $B = \{b_j(k)\}$ with observation symbol probability distribution for each state

- π is the initial state distribution vector $\pi = \{\pi_i\}$

Hence, N is a number of hidden states in the model and M is a number of distinct observations per hidden state, i.e. the discrete vocabulary size. The key difference between HMM and MC is that HMM requires specifying the number of hidden states as a model parameter. HMM deduces a sequence of hidden states that best explains the observations along with the state transition probabilities and the distributions of observations (emission probabilities) per each hidden state. The Baum-Welch algorithm [114] is used to estimate the parameters of HMMs for successful and unsuccessful interviews using the corresponding training set, while the Viterbi algorithm [114] is used to determine the most likely sequence of hidden states for a given sequence of observations. After assignment of hidden states, the log-likelihood of success for an interview can be estimated using Eq. 3.3 as well.

Behavior code embeddings. Representation of behavior codes was inspired by the recent success of word embeddings[16, 95, 111]. Embedding is a representation of an object in low-dimensional space using a real-valued vector. In our study, embeddings of behavior codes were obtained as a by-product of training LSTM and GRU after feeding one-hot vectors as a representation of behavior codes as input to these RNNs. Behavior code embeddings have the property of representing similar codes with the vectors that are close to each other in low-dimensional space. Figure 3.1 illustrates the MYSCOPE code embeddings visualized in 2-dimensional space by t-SNE [88]. It can be seen that positive behavior codes such as OQECHT+, OQECML+, AF, AFL, SUP, RCML+S, CQECML+, etc. formed a cluster in the left part of Figure 3.1. The nearest neighbors of CQECML+ are highlighted by different color intensity (i.e. OQECML+ being more purple indicates that it is more similar to CQECML+). The right part of the figure demonstrates another cluster formed with negative behavior codes including CQECML-, AMB-, RCHT-C, OQECHT-, GINFO-, RBAC, LUP-, RCHT-S, RPTBC, RAMBC, AMB-, RCML-S, etc. It is interesting

that the behaviors intended to elicit CHT+/CML+ group together, whereas the ones intended to elicit CHT-/CML- also group together and are located in the opposite regions of semantic space.

Figure 3.1: 2-D representation of behavior code embeddings

Recurrent Neural Networks (RNN) are a class of neural networks that have an internal memory, which makes them particularly suitable for processing sequences of observations. The ability of RNNs to capture long-term dependencies and remember past observations for predicting future observations is their main advantage over MCs and HMMs. These features are very useful in the analysis of motivational interviews, in which any behavior observed at a particular point in the interview may be indicative of other behaviors that are observed later. In order to mitigate the vanishing gradient problem of earlier versions of RNN [17], Hochreiter et al.[65] proposed Long Short Term Memory networks (LSTM). There are several variants of LSTM model, among which the most notable one is the Gated Recurrent Unit[28] (GRU). GRUs are simpler than LSTMs and have been shown to be effective for a variety of

Natural Language Processing tasks [28]. GRU is formally defined as follows:

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \tag{3.4}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \tag{3.5}$$

$$\tilde{h}_t = tanh(W_h x_t + r_t \odot U_h h_{t-1} + b_h) \tag{3.6}$$

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \tag{3.7}$$

In Eq. 3.4-3.7, σ corresponds to sigmoid function and $\odot$ designates an element-wise product. The update gate z_t and reset gate r_t at time step t are computed by the Eq. (3.4) and (3.5), where W_z, W_r, W_h, U_z, U_r, U_h are the weight matrices and b_z, b_h and b_r are bias vectors. The activation h_t of the GRU at time t is a linear combination of the previous activation h_{t-1} and the candidate activation $\tilde{h}_t$, which is represented by Eq. (3.7) and (3.6).

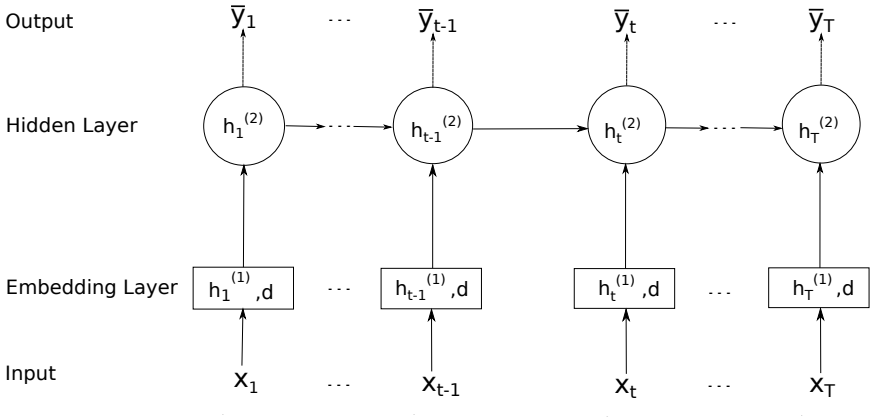

Figure 3.2: Proposed RNN model with target replication (TR)

The RNN architecture employed for sequence classification is shown in Figure 3.2. As can be seen from Figure 3.2, softmax is used at each time step to predict the class of a sequence observed so far. Since the sequence label is predicted at each observation, the proposed architecture is referred to as Recurrent Neural Network with Target Replication (TR). It was trained by minimizing the following hybrid loss

function:

$$\tilde{\mathcal{L}} = \alpha \cdot \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}(\bar{y}^{(t)}, y^{(t)}) + (1 - \alpha) \cdot \mathcal{L}(\bar{y}^{(T)}, y^{(T)}) \tag{3.8}$$

As follows from Eq. 3.8, the total loss $\tilde{\mathcal{L}}$ is a convex combination of the final loss $\mathcal{L}(\bar{y}^{(T)}, y^{(T)})$ and the average loss over all observations in a sequence, where T is the total number of observations, $\bar{y}^{(t)}$ is the output at step t, and $\alpha \ \epsilon \ [0, 1]$ is a hyper-parameter controlling the relative importance of each loss type. We experimentally determined that the best performance is achieved when $\alpha = 0.5$. Our model also contains several other hyperparameters, such as the number of embedding dimensions, the number of hidden units, learning rate, batch size, etc., which were optimized on the validation set. We implemented our models in Tensorflow with Adam optimizer as well as early stopping based on the validation loss and observed that our model converges after 100 epochs.

Evaluation

Performance of probabilistic and deep learning methods was evaluated in terms of precision, recall, and F-measure using 10 folds cross-validation and weighted macro-averaging of these metrics over the folds. However, LSTM and GRU were trained on 80% of the data and validated on 10%, with the remaining 10% of the data used for testing.

3.1.4 Results

All sequence classification methods were evaluated in the case of both under and over-sampling. Predictive performance summary of all methods is summarized in Table 3.2[1].

[1]Prec.: Precision, Reca.: Recall

Table 3.2: Performance of MC, HMM, LSTM and GRU with and without target replication (TR) for predicting the success of patient-provider communication sequences when under- and over-sampling were used to balance the dataset. The highest value for each performance metric is highlighted in bold.

Method	Under-sampling			Over-sampling		
	Prec.	**Reca.**	**F1**	**Prec.**	**Reca.**	**F1**
Markov Chain 1^{st} Order	0.7060	0.7044	0.7038	0.7932	0.7799	0.7775
Markov Chain 2^{nd} Order	0.6395	0.6385	0.6379	0.7111	0.7029	0.7000
Hidden Markov Model	0.6244	0.6143	0.6067	0.7775	0.7567	0.7520
LSTM	0.8672	0.8626	0.8622	0.8411	0.8372	0.8368
LSTM-TR	**0.8733**	**0.8681**	**0.8677**	**0.8424**	**0.8385**	**0.8381**
GRU	0.8674	0.8648	0.8646	0.8379	0.8342	0.8337
GRU-TR	0.8705	0.8676	0.8673	0.8412	0.8377	0.8373

Predictive performance in the case of under-sampling

We used a small learning rate of 0.00005 and the batch size of 8 along with early stopping strategy for training deep learning models on the dataset balanced with under-sampling. Five major conclusions can be drawn from the results in Table 3.2. First, recurrent neural networks outperform probabilistic models and achieve 16.39%-26.1% higher F1-score. Second, LSTM with target replication has the best performance over all other RNN methods, and achieved F1-score 0.8677 with precision 0.8733 and recall 0.8681. Third, target replication strategy improves the performance of GRU and LSTM, with conventional GRU showing better performance than traditional LSTM. Fourth, among probabilistic models, the MC based method generally outperforms HMM across all metrics for under-sampled sequences. Fifth, second-order MC has lower precision, recall, and F-measure than first-order MC. In particular, precision, recall and F-measure decrease by 9.42%, 9.36% and 9.36%, when going from first to second-order MC model.

Predictive performance in the case of over-sampling

Similar to the under-sampling scenario, early stopping strategy was also employed for training deep learning models on the dataset balanced with over-sampling. How-

Table 3.3: Most likely communication sequences in successful and unsuccessful motivational interviews.

Type	Most likely communication sequences
successful	GINFO+: General information, positive → LUP+: Low uptake, positive → OQTBN: Open-ended question, target behavior neutral
successful	SS: Structure session → GINFO+: General information, positive → CQECHT+: Closed-ended question, elicit change talk positive
successful	SO: Statement, other → LUP+: Low uptake, positive → AF: Affirm → HUPW: High uptake, weight → OQECML+: Open-ended question, elicit commitment language positive.
unsuccessful	ADV+: Advise, positive → AMB-: Ambivalence negative → OQECHT-: Open-ended question, elicit change talk negative
unsuccessful	CQECHT+: Open-ended question, elicit change talk positive → RCHT-S: Reflect, change talk negative → OQECHT-: Open-ended question, elicit change talk negative
unsuccessful	SUP: Support → AF: Affirm → CQTBN: Closed-ended question, target behavior neutral → OQECHT-: Open-ended question, elicit change talk negative → AMB-: Ambivalence negative

ever, in this case, RNN models were trained with the learning rate of 0.00010 and the batch size of 55. Experimental results indicate that HMM had better performance than second-order MC, achieving 9.34%, 7.65%, and 7.43% higher precision, recall, and F-measure, while HMM still had 1.98%, 2.97%, and 3.28% lower precision, recall, and F-measure than first-order MC. Also similar to the under-sampling scenario, target replication improves the performance of RNN models and LSTM with target replication has the highest F1-score among all models. However, the predictive performance of LSTM and RNN decreases when over-sampling is used, while the performance of probabilistic models increases.

Most likely communication sequences

Table 3.3 provides examples of typical patient-provider communication sequences that frequently appear in successful and unsuccessful motivational interviews. We observed that in successful motivational interviews information is frequently provided using patient-centered communication (GINFO+) and structure session (SS) utterances, in which the counselor either explains the therapeutic agenda or attempts to

transition to a new topic or session content. Sometimes, counselors also acknowledge the clients' communication or an off topic comment (SO). We also observed that affirmations (AF) and open-ended questions (OQECML+) have a strong effect on eliciting positive change talk or commitment language, which is consistent with MI theory. It can also be seen that providing advice using non-patient centered strategies (ADV-) leads to negative ambivalence (AMB-), which results in the interview heading in therapeutically wrong direction. Questions posed to elicit negative change talk or commitment language lead to CHT-, CML- or AMB-, which is consistent with the manual analysis by clinicians.

3.1.5 Discussion

We made the following conclusions after analyzing the experimental results of different communication sequence outcome prediction methods. First, the overall predictive performance of RNN based methods is substantially higher than that of probabilistic models. In particular, the RNN-based methods achieve near-human accuracy for predicting the success of motivational interviews. This indicates that RNN is able to capture the structure of discourse in motivational interviews by preserving long-term dependencies among the behavior codes, which reflect the overall progression of the interviews. This provides evidence that RNNs are able to successfully replicate human cognitive processes to integrate previous information when making decisions. In addition to that, embeddings allow to reduce the dimensionality of codes in PPC sequences and consequently improve both precision and recall of prediction.

Second, using target replication to compute the loss at each time step results in better performance for all configurations of the proposed RNN-based methods. This indicates that the average of the losses over all steps emphasizes the dependencies between the pairs of patient and provider codes, which results in more accurate estimates of the model parameters. Better estimates of parameters in RNN models of motivational interviews are propagated to the next step based on the relative im-

portance of intermediate output, where they are aggregated into predictions for the entire sequence. This allows to achieve an improvement in prediction accuracy.

Third, using first-order Markov model results in better prediction accuracy compared to higher-order Markov models, which we attribute to the fact that the number of states in higher-order Markov models may grow exponentially with their order. As a result, accurate estimation of transition probabilities requires much larger training data. Using smaller datasets, which is the case when under-sampling is employed, will result in a sparsity problem, when many transitions are either not observed in the training set at all or observed only a few times, leading to missing or potentially inaccurate probability estimates. Obtaining large training sets cannot be easily accomplished in many domains, including motivational interviewing. In this project, we found out that using first-order Markov models is a reasonable trade-off between efficiency and accuracy.

Fourth, similar to traditional Markov model, HMM achieves a dramatic improvement in the prediction accuracy when larger training set is used. This indicates that sufficient training data is required to find the optimal settings of hyperparameters, such as the number of hidden states, initial state distribution, transition probabilities, and emission probabilities.

Fifth, the proposed method can be used to identify the most effective communication strategies for eliciting a particular type of behavioral response. Awareness of these strategies by researchers can significantly decrease the time and effort required to develop effective interventions to address many public health conditions, such as childhood obesity, and tailor these interventions to particular patient cohorts. Awareness of these strategies by the counselors can lead to a greater success rate of motivational interviews.

3.1.6 Summary

In the first section of this chapter, we compared the accuracy of Recurrent Neural Networks with Markov Chain and Hidden Markov Model for the task of predicting the success of motivational interviews. We found out that individual PPC exchanges are highly indicative of the overall progression and future trajectory of clinical interviews and can be used to predict their overall success. Our methods can facilitate motivational interviewing researchers in establishing causal relationships between different communication strategies and the desired behavioral outcomes during the interviews without resource-intensive manual qualitative analysis of interview transcripts, which can significantly decrease the time and effort required to develop behavioral interventions. These methods can also help to identify the most likely sequences in successful and unsuccessful motivational interviews, which can directly inform clinical practice and increase the effectiveness of behavioral interventions. Our experimental results also indicate that our methods can be used for real-time monitoring of the progression of clinical interviews. This work also has broad implications for public health research by providing a theoretically-grounded computational approach to qualitative data analysis.

3.2 Sequential patterns in clinical conversation

3.2.1 Introduction

Motivational Interviewing (MI) is an evidence-based strategy for communicating with patients about behavior change [97]. The theory underlying MI's clinical efficacy posits that behavior change is triggered by fostering an atmosphere of change, which is accomplished through the exercise of relational and technical skills [97]. The relational hypothesis suggests that counselors' use of accurate empathy, positive regard and congruence create the "spirit of MI", an optimal therapeutic state to explore behavior change. MI's technical hypothesis [98] states that counselors' use of communication techniques consistent with the MI framework ("MI-consistent" or MICO;

e.g., open-ended questions, reflections, advise with permission, affirmations, emphasize control, reframe and support) will lead to patient "change talk". Change talk is patient statements during clinical encounters that express their internal desire, ability, reasons, need for and/or commitment to behavior change [7]. Previous studies [10] have shown that change talk expressed during treatment sessions consistently predicts behavior change with results persisting as long as 34 months post-intervention [132]. In contrast, MI-inconsistent communication behaviors (MIIN; e.g., advising without permission, warning about behavioral consequences and confronting) are hypothesized to lead to arguments against behavioral change and/or to maintain the status quo (referred to as counter change talk or sustain talk). Multiple studies have linked high rates of MICO to the expression of change talk and MIIN to sustain talk [89]. These studies have relied on session-level behavior counts and correlational analyses, which ignore the temporal order of utterances in patient-counselor communication, thereby limiting researchers' ability to test MI's technical hypothesis.

In this project, we focused on computational methods to facilitate the sequential analysis of pre-coded MI transcripts to identify patterns of patient-counselor communication in successful and unsuccessful sequences in MI sessions. Analysis of these patterns provides empirical support for the specific counselor communication strategies that are effective at eliciting patient change talk. This knowledge will inform MI theory by providing additional evidence to support MI's technical hypothesis. It will also inform clinical practice by facilitating the use of more effective and tailored counselor communication. This study was the first empirical evaluation of the effectiveness of closed frequent pattern mining to analyze patient-counselor communication sequences during MI sessions. Bertholet et al. [18] used HMM to identify hidden states in a brief motivational intervention. Limiting their HMM model to three hidden states which were characterized as "towards change", "away from change" and "non-determined", these states were used to predict drinking outcomes 12 months

post-intervention. In this project, we identified the optimal number of hidden states using HMM modeling of successful and unsuccessful sequences of patient-counselor communication. The goal of this study was to evaluate the utility of using HMM and frequent pattern mining to better understand the specific counselor communication strategies leading to patient change talk and sustain talk during Motivational Interviewing sessions. These two approaches offer the following advantages over the first-order Markov Chain-based methods most typically used in MI research. First-order Markov Chain models identify the likely transitions between individual behaviors. In contrast, HMM summarizes transitions between clusters of related behavior codes (i.e., hidden states) allowing the identification of clusters of behaviors antecedent to change talk in successful patient-counselor communications and sustain talk in unsuccessful patient-counselor communications. Frequent pattern mining can identify patterns involving long-range dependencies between patient and counselor behaviors. Accounting for such long-range dependencies is important, since human behaviors, such as patient-counselor communications during MI sessions, are informed by all the antecedent behaviors and not just the immediately preceding behavior.

3.2.2 Related work

Sequential analysis is an analytic approach to examine temporally ordered sequences of events or observations [13, 14]. Moyers and Martin [100] were the first to apply sequential analysis approach in a study of adults in treatment for alcohol abuse and found that change talk was significantly more likely after MICO or "MI-consistent", counselors use of communication techniques consistent with the MI framework and sustain talk more likely after MIIN or "MI-inconsistent", communication behaviors inconsistent with the MI framework. A follow-up study with the same population found that change talk was more likely after two MICO behaviors, counselor questions about the positive and negative aspects of drinking and reflections of change talk, but these behaviors also led to sustain talk [101]. Surprisingly,

MIIN was unrelated to sustain talk, but decreased the likelihood of change talk. Gaume et al. [50] used sequential analysis to study communication patterns during brief motivational interviewing for hazardous alcohol consumption with young adults conscripted into military service. They found that MICO led to both change talk and sustain talk but the MIIN-to-sustain talk pattern was not observed. A second study with the same population confirmed that MICO leads to significantly more change talk and sustain talk [49]. In this sample, MIIN led to greater sustain talk, but was unrelated to change talk. Further analyses revealed that reflections were the only MICO behavior linked to increased change talk; reflections and other MICO behaviors, excluding questions, were related to increased sustain talk. Glynn and colleagues [53] linked reflections of change talk to the elicitation of change talk and reflections of sustain talk to the elicitation of sustain talk among incarcerated adolescents with high rates of alcohol and marijuana use. In a study of adolescents engaged in weight loss treatment, Carcone et al. [24] used sequential analysis to identify three counselor behaviors likely to result in change talk: open-ended questions phrased to elicit change talk, reflections of change talk and statements emphasizing decision-making autonomy. A parallel study of the adolescents' caregivers [67] drew a similar conclusion that asking questions phrased to elicit change talk, reflections of change talk and autonomy-supportive statements were the counselor behaviors, which led to the elicitation of change talk. Across these studies, counselors' use of reflections was consistently linked to change talk; other MICO behaviors, however, led to change talk in some treatment contexts, but not others, suggesting a need for additional research to understand the treatment contexts, in which various MICO strategies are effective. Our sequential analysis contributes to existing knowledge by examining African American adolescents in weight loss treatment.

The sequential analysis procedure used in the above MI process studies [100, 25, 129, 94] is based on the first-order Markov Chain model [100, 101, 49]. The Markov

Chain model is a discrete-time stochastic process built on the assumption that the state of a system or condition changes over time and only depends on the previous event. Hence, Markov Chain models have two main drawbacks. The first is their inability to preserve the long-range dependencies between observations in a sequence. In MI, an observed behavior can be influenced by any of the preceding behaviors. The second drawback is their inability to consider similarities between behavior codes and, consequently, first-order Markov chain models are unable to identify multiple similar behaviors that lead to the same outcome. Thus, first-order Markov models may be insufficient to fully understand the associations between behaviors in patient-counselor communication sequences. There is a need for more powerful computational methods, which consider clusters of similar behavior codes and long-range dependencies between behaviors, to identify causal relationships. To achieve this goal, we tested the applicability of data mining and machine learning methods to identify effective patterns of patient-counselor communication. The current work builds on our recent work [63] by examining the efficacy of Hidden Markov Models (HMMs) and frequent pattern mining for the identification of the counselor communication strategies leading to patient change talk.

HMMs are widely used for the analysis of sequence data due to their ability to model long-range dependencies between clusters of discrete observations in a sequence. The HMM associates each observation in a sequence with a "hidden" state, which corresponds to a distribution over all distinct observations in a sequence (i.e., probabilities associated with each observation, when HMM is in this hidden state), such that each "hidden state" corresponds to a different distribution. Sequences of observations are modeled as transitions between different hidden states and sampling observations from distributions corresponding to each hidden state. HMMs were originally proposed for speech recognition [114], in which the states were used to represent all English language sounds. In biomedical informatics, HMMs were employed for

the diagnosis of diseases and biological sequence modeling [131, 8]. For example, an HMM-based classifier was applied to Doppler ultrasound imaging data to extract features from the images that were then used to distinguish healthy patients from those with heart disease [131]. In another study, HMM was used to capture important characteristics of protein families [8]. In the application of HMM to patient-counselor communication, hidden states and the sets of related behavior codes associated with the hidden states may correspond to patients' underlying motivational state during a patient-counselor encounter.

Although the MI literature has established patient change talk and commitment language (a special class of change talk where patients express their intentions, plans and action steps toward behavior change [15]) as the antecedents of patients' behavior change [10], there is less clarity regarding which counselor communication strategies influence the articulation of change talk. Modeling successful and unsuccessful communication sequences during MI sessions with HMM can provide additional evidence to identify the counselor communication strategies that are likely to lead to patient change talk and commitment language.

Frequent pattern mining [4] is a class of data mining methods to identify sets of items (or observations, referred to as itemsets) which frequently appear together. Agrawal and Srikant [5] first introduced frequent pattern mining with the Apriori algorithm, developed to identify customer purchasing patterns. Since its introduction, frequent pattern mining has been applied to several other domains, including health informatics [2, 108, 19, 136], medical imaging [108], chemical and biological analysis [40, 82, 138], web mining [120] and outlier analysis [3]. Now, our new published study was the first to use this approach for studying patient-counselor communication. A major challenge in applying frequent pattern mining methods to patient-counselor communication sequences is the large number of resulting patterns, which include redundant patterns. To address this problem, we utilized the closed frequent itemset

mining method [110], which produces fewer patterns in a more compact form that are easier to interpret. In this project, we leveraged FPClose [54], an efficient state-of-the-art closed frequent pattern mining method, to identify the counselor behaviors that frequently lead to patient change talk. FPClose is a state-of-the-art closed frequent itemset mining algorithm, which has demonstrated good performance in terms of running time and memory consumption.

3.2.3 Methods

Dataset

This project utilized the same dataset annotated with MYSCOPE codebook described in chapter 1 excluding conversations that correspond to greetings, farewell and interview setups such as table and camera settings. The experimental dataset consists of 7,192 patient, caregiver and counselor utterances segmented and annotated with the MYSCOPE behavior codes, illustrated in Table 3.4.

Table 3.4: MYSCOPE codebook

Annotation	Behavior	Description	Example
Counselor			
AF	Affirmation	Positive or complimentary statements that express appreciation, confidence, or reinforce the patient's strengths or efforts.	"You guys, as a family, are already doing a lot of really positive things."
AR	Action reflection	Statements that reflect back the patient's statement(s) while at the same time embedding a solution to a barrier or an action plan.	"If you decide to follow a meal plan, it has to include occasional dessert."

Table3.4 (continued)

Annotation	Behavior	Description	Example
EA	Emphasize autonomy	Statements that directly acknowledge, honor, or emphasize the patient's freedom of choice, autonomy, personal responsibility and so forth.	"Okay. Well, it's your plan, so whatever works best for you. If you feel like you want one that's written down that you can refer back to, then let's write it and if not then that's fine."
GINFON	General information negative	The counselor gives advice, makes a suggestion, offers a solution/possible action, gives feedback, or offers educational information in a non-patient-centered manner.	"Healthy weight loss is about one to two pounds a week and once we get you set up and actually into the program you can look for that to happen for about one to two pounds a week to get you on that goal."
GINFOP	General information positive	The counselor gives advice, makes a suggestion, offers a solution/possible action, gives feedback, expresses a concern, or offers educational information in a patient-centered manner (i.e., asking permission, using the third person, giving the opportunity to reject the information and offering a menu of options).	"Okay. Alright so I just wanted to tell you that I will be asking you a lot of questions. It may get redundant. So, if at any point in time you need a break or I'm asking too much go ahead and let me know."

Table3.4 (continued)

Annotation	Behavior	Description	Example
QEB	Question to elicit barriers	Questions designed to initiate a discussion of barriers to change.	"Alright. So, are these ideas you feel you can put in place for this week?"
QECHTP	Question to elicit change talk positive	Questions that ask about the patient's desire, ability, reasons, or need for change or that reference past action toward behavior change or barriers to change.	"Okay. And tell me a little bit more about that. Like what do you foresee your goal in this program? Like what do you want to happen out of this program?"
QECMLP	Question to elicit commitment language positive	Questions that ask about current or future action toward behavior change or reference barriers to change.	"Okay. Is there something else that you could do eat maybe instead of a Pop-Tart that's a little bit healthier?"
QEF	Question to elicit feedback	Statements that solicit the patient's thoughts, ideas, or feelings about a specific recommendation or piece of information.	"So, do you have any questions about that?"
QEST	Question to elicit sustain talk	Questions designed to elicit negative change talk or negative commitment language.	"And about how many hours would you say you watched TV for today? Or played video games or YouTube?"

Table3.4 (continued)

Annotation	Behavior	Description	Example
QO	Question other	Open- or close-ended questions unrelated to the target behavior.	"Yup. What do you think might get in your way of being able to provide that kind of support for [your daughter]?"
RCHTP	Reflect change talk positive	A reflective listening statement that captures and returns a patient's statement or behavior from the current or a previous session that describes the patient's desire, ability, reasons, or need for change or past action or barriers to change.	"So, it sounds like you just want to be healthy and you want to be stylish. You want to fit into some different types of clothes."
RCMLP	Reflect commitment language positive	A reflective listening statement that captures and returns a patient's statement or behavior from the current or a previous session that describes current or future action or references barriers to changing with the goal of problem-solving.	"You are ready to start this plan today."

Table3.4 (continued)

Annotation	Behavior	Description	Example
RO	Reflect other	A reflective listening statement that captures and returns a patient's utterance or behavior from the current or previous session that is unrelated to the target behavior.	"You are having a hard time at work."
RST	Reflect sustain talk	These statements reflect negative change talk or negative commitment language made by the patient.	"Oh okay. So, money influences your environment."
SO	Statement other	An utterance eliciting feedback, offering support, self-disclosure, or of some other form besides a strategy or reflection	"You're being pulled in a million directions"
SPT	Support	These are generally supportive, understanding comments. They have the quality of commenting on a situation, or of agreeing or siding with the patient in a genuine way.	"I'm concerned about you, given all these difficulties you've been having."

Table3.4 (continued)

Annotation	Behavior	Description	Example
SS	Structure session	A communication strategy that suggests an attempt to describe what will happen in the session or to refocus a meandering conversation back to the target behaviors	"Maybe when the three of us come together in a few minutes, that's something that we could just clarify with her, like is that really what she wants."
SUM	Summary	A reflective listening statement that captures and returns at least 2 different ideas from a patient's utterance or behavior from the current session	"You have thought a lot about this. Sometimes it feels like losing weight is just too hard. Yet you have lots of reasons to lose weight. If you could find a program you could stick to, a program that would not have too many changes at once, you would consider it."
Patient			
CT	Change Talk	Statements that express the patient's desire, ability, reasons, need for, or commitment to (intentions, plans and action steps) changing their behavior	"I will try to buy less junk food."

Table3.4 (continued)

Annotation	Behavior	Description	Example
ST	Sustain Talk	Statements that express the patient's desire, ability, reasons, need for, or commitment to (intentions, plans and action steps) to maintain the status quo or not change their behavior	"I didn't get to the gym this week."
HUPW	High uptake weight	A turn that does develop the topic of the conversation. High Uptake statements include: weight-related statements about actions of commitment, change talk and ambivalence that occurred in the past, patient questions to the counselor and session interruptions by persons who are not an active part of the treatment session.	"Support is always good. You know that's a key factor. Mm-hmm."
HUPO	High uptake other	An utterance that develops the topic of the conversation but is about non-target behaviors or interruptions	"Yeah because the mentor comes and they take off and they go someplace for a little while."

Table3.4 (continued)

Annotation	Behavior	Description	Example
LUP	Low uptake	An utterance that does not develop the topic of conversation but still allows it to continue	"Mm-hmm. Right."

Data preprocessing

Utterances in MI session transcripts were segmented into successful and unsuccessful communication sequences which is shown in section 4.1. For each MI transcript, the stream of behavior codes from the beginning of a session to the end of the session was analyzed. Successful sequences were defined as those that resulted in a patient change talk or commitment language statement. Unsuccessful sequences were similarly created for sequences resulting in sustain talk. A total of 1,360 sequences were generated using this approach. The majority of the sequences (n=1,102) were successful, which is expected for a treatment-seeking population, in which patients initial motivation for behavior change is typically high. Successful sequences had an average length of 5.28 utterances, while unsuccessful sequences had on average 5.29 utterances.

Data modeling

Hidden Markov Model: We applied the Hidden Markov Model (HMM)[2] to identify clusters of behavior codes corresponding to successful and unsuccessful communication sequences and to describe the relationships (transitions) between these clusters. Given a set of behavior code sequences, the posterior inference of HMM parameters involves the deduction of a temporal sequence of hidden states that best

[2]we used the implementation in the hmmlearn package publicly available at http://hmmlearn.readthedocs.io/

explains observations in each sequence. The rows in the emission probability matrix correspond to the distribution of observation symbols (i.e., the MYSCOPE behaviors displayed) for each hidden state and the transition probability matrix describes the transitions between the hidden states. Training an HMM with a given number of hidden states (N) involves estimating the following parameters using the Baum-Welch algorithm:

- M is the number of distinct observations symbols per state, i.e. the discrete codebook size (Table 3.4)
- T is an $N \times N$ state transition probability matrix, in which t_{ij} is the probability of HMM transitioning from state i to state j
- E is an $N \times M$ emission probability matrix, in which e_{jk} is the probability of observing symbol k, when HMM is in state j
- π is the initial state distribution vector where π_i is the probability of the i^{th} state to be the first state

We trained two HMM models, one using all successful sequences and the other one using all unsuccessful sequences. Each model was trained with the objective of maximizing the log-likelihood of all observations in the corresponding set of sequences. The optimal number of hidden states was determined by estimating the Bayesian information criterion (BIC) of HMM models with a different number of hidden states and selecting the model with the smallest value of BIC, which takes into account both log-likelihood and a penalty term for the number of parameters in the model to avoid overfitting. Experiments with a different number of hidden states in HMMs estimated on successful and unsuccessful sequences indicated that 5 hidden states were optimal for successful sequences and 2 hidden states were optimal for unsuccessful sequences (Figure 3.3).

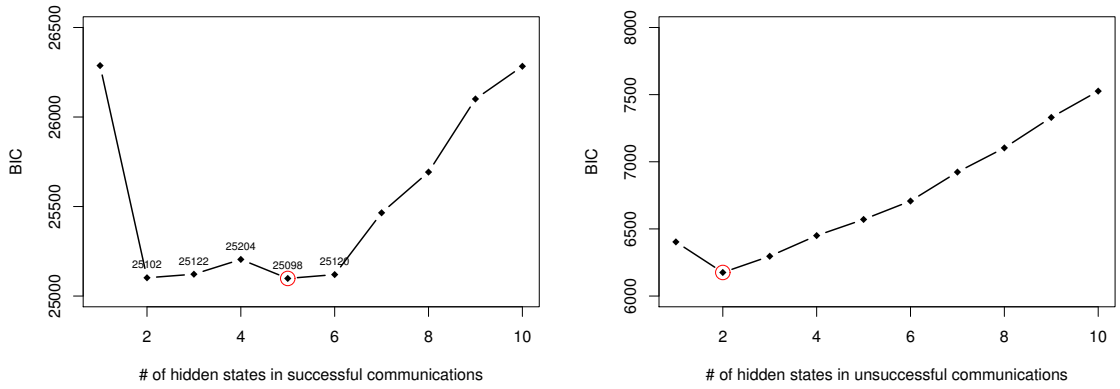

Figure 3.3: Bayesian information criterion (BIC) of HMM models of successful (left) and unsuccessful (right) interviews by varying the number of hidden states

Frequent Pattern Mining: We applied frequent pattern mining to identify frequently occurring patterns of patient-counselor behavior codes in successful and unsuccessful communication sequences. Behavior codes in these patterns may be separated by one or more other codes. For this purpose, we utilized FPClose [54], an efficient state-of-the-art closed frequent pattern mining algorithm implemented in SPMF [44, 45], to identify frequent patterns of patient-counselor communication behaviors in successful and unsuccessful MI communication sequences. SPMF is an open-source library providing more than 150 data mining algorithms. Popular non-closed frequent pattern mining algorithms include Apriori [5] and FP-Growth [60]. A *frequent pattern* is defined as a pattern of observations, which appears in a given set of sequences more often than a user-specified threshold called the *minimum support count*. For example, {A}, {C}, {D} and {C, D} are frequent patterns in the example set of sequences in Figure 3.4, since these patterns appear at least 2 times, which is the minimum support count in this example. In this work, we identified and analyzed *closed frequent patterns* among all sequences of behavior codes in successful and unsuccessful communication sequences. A *frequent pattern* is *closed* if none of its supersets have the same support count [110], where a set X is a superset of another

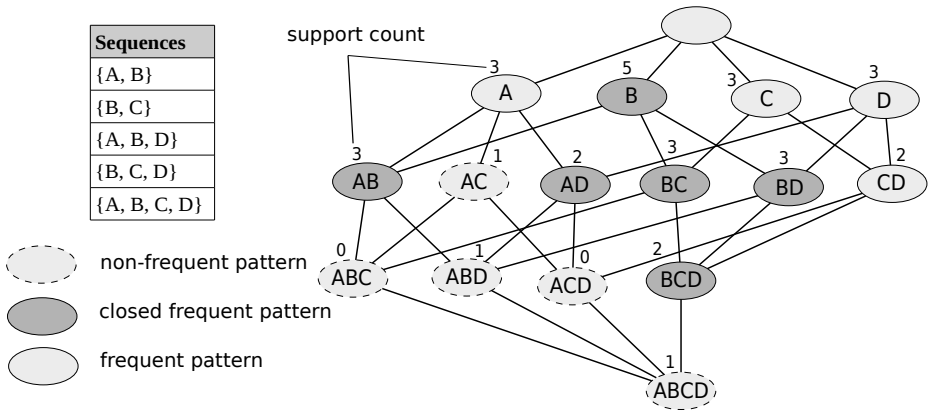

Figure 3.4: A sample collection of sequences and different types of frequent patterns obtained by a frequent pattern mining method with the minimum support of 2

set Y, if X contains all the elements of the set Y. For example, the itemsets {A}, {C}, {D} and {C, D} in Figure 3.4 are not closed frequent patterns since their supersets {A, B}, {B, C}, {B, D} and {B, C, D} have the same support count. Therefore, {B}, {A, B}, {A, D}, {B, C}, {B, D} and {B, C, D} are closed frequent patterns since none of their supersets have the same support count. On the other hand, itemsets {A, C}, {A, B, C}, {A, B, D}, {A, C, D} and {A, B, C, D} have support counts of 1, 0, 1, 0 and 1, respectively, which is less than the minimum support count and thus are identified as non-frequent itemsets. Since the threshold for minimum support count depends on a task and is typically determined by the domain expert, we followed prior work [99, 86] and set the minimum support count as 10% of the total number of all communication sequences, which is 110 for successful and 25 for unsuccessful communication sequences. For each pattern, the statistical significance of the difference between successful and unsuccessful sequences was computed with Pearson's chi-square test.

3.2.4 Results

The transition and emission probability matrices of the HMM models are reported in Tables 3.5 and 3.6. Three behaviors represented 45-69% of each state's emission probability mass and, thus, were used to interpret the emission matrix and label the

Table 3.5: Hidden Markov Model Emission Matrices

State	AF	AR	EA	GIN-FON	GIN-FOP	QEB	QEC-HTP	QEC-MLP	QEF	QE-ST	QO	RC-HTP	RC-MLP	RO	RST	SO	SPT	SS	SUM	HU-PO	HU-PW	LUP
Successful Sequences																						
High Motivation	12%	5%	4%	0%	3%	0%	0%	1%	0%	0%	1%	29%	17%	1%	7%	2%	5%	6%	6%	0%	0%	0%
High Receptivity	15%	2%	5%	0%	18%	1%	16%	5%	3%	1%	0%	1%	0%	6%	5%	3%	3%	7%	4%	4%	1%	1%
Moderate Receptivity	2%	1%	14%	0%	8%	1%	20%	9%	5%	1%	2%	2%	0%	1%	0%	1%	0%	5%	4%	5%	2%	16%
Low Receptivity	4%	1%	17%	0%	3%	1%	11%	0%	2%	0%	0%	5%	3%	3%	3%	0%	1%	7%	1%	6%	0%	28%
Active Feedback	1%	1%	3%	0%	7%	1%	0%	9%	3%	0%	1%	0%	0%	0%	3%	1%	0%	1%	1%	10%	12%	47%
Unsuccessful Sequences																						
Ambivalent	13%	5%	4%	1%	6%	0%	1%	5%	0%	2%	2%	20%	7%	2%	12%	3%	4%	5%	6%	0%	2%	0%
Avoidant	3%	1%	6%	0%	11%	7%	8%	1%	3%	2%	2%	1%	1%	5%	4%	0%	1%	2%	0%	11%	4%	29%

Table 3.6: Hidden Markov Model Transition Matrices

State	High Motivation	High Receptivity	Moderate Receptivity	Low Receptivity	Active Feedback
Successful Sequences					
High Motivation	2%	9%	24%	24%	41%
High Receptivity	7%	9%	20%	25%	39%
Moderate Receptivity	8%	19%	24%	23%	27%
Low Receptivity	11%	26%	28%	18%	17%
Active Feedback	25%	34%	20%	15%	6%
Unsuccessful Sequences					
	Ambivalent	Avoidant			
Ambivalent	16%	84%			
Avoidant	39%	61%			

hidden states. We observed 4,724 state transitions in the successful sequences; 4,679 and 45 state transitions occurred between different and same states, respectively. On an average, 5 state transitions occurred within a sequence for both successful and unsuccessful sequences. Most (84%) successful sequences began in a state characterized as *"high motivation"* as evidenced by a greater proportion of three counselor behaviors: reflections of change talk (29%), reflections of commitment language (17%) and affirmations (12%). Successful sequences began in a state of *"high receptivity"* 11% of the time. *"High receptivity"* sequences were characterized by nearly equal proportions of information offered using patient-centered strategies (18%), questions to elicit change talk (16%) and affirmations (15%). Few successful sequences began in states of *"moderate receptivity"* and *"low receptivity"* (2% and 3% of the time, respectively). These two states were characterized by different proportions of the same behaviors. *"Moderate receptivity"* sequences were distinguished from *"low receptivity"* sequences by a greater proportion of counselor questions to elicit change talk (20% versus 11%) and a lower proportion of patient low uptake statements (16% versus 28%); counselor statements emphasizing the patient's autonomy were about the same (14% versus 17%). No (0%) successful sequence began in the *"active feedback"* state, which was characterized by three patient behaviors, low uptake (47%), weight-related high uptake (12%) and other-related high uptake (10%). Successful sequences transitioned from *"high motivation"* to *"active feedback"* most often (41%). *"Active feedback"*, in turn, most frequently transitioned to *"high receptivity"* (39%). The *"moderate receptivity"* state most often transitioned to *"active feedback"* (27%) and back to *"moderate receptivity"* (24%) or to *"low receptivity"* (23%) with similar frequency. The full transition matrix is presented in Table 3.6.

In contrast, 1,106 state transitions occurred within the unsuccessful sequences; 697 and 409 state transitions happened between different and same states, respectively. The majority of unsuccessful sequences (98%) began in a state of *"ambivalence"*

as indicated by the greater proportion of counselor reflections of both change talk (20%) and sustain talk (12%) as well as affirmations (13%). About 2% of the time unsuccessful sequences started in a state of *"avoidance"*. Higher rates of patient low uptake (29%) and other-related high uptake (11%) statements and counselor patient-centered information (18%) distinguished *"avoidant"* sequences. Both *"ambivalent"* (84%) and *"avoidant"* (61%) states most frequently transitioned to the *"avoidant"* state.

Results from frequent pattern mining analysis are presented in Table 3.7. Reflections of change talk were the most frequent counselor communication behavior in both successful (36.1%) and unsuccessful sequences (33.7%). Successful sequences were distinguished from unsuccessful sequences by a higher frequency of counselor questions phrased to elicit change talk (30.8% versus 17.4%, Pearson's chi-square test $p < 0.001$), statements emphasizing the patient's decision-making autonomy (28.5% versus 18.6%, p=0.001), questions phrased to elicit commitment language (18.1% versus 11.6%, p=0.011) and reflections of commitment language (20.7% versus 15.1%, p=0.042). In contrast, unsuccessful sequences were characterized by greater frequency of questions to elicit perceived barriers (14.7% versus 0%, $p < 0.001$), reflections of sustain talk (27.1% versus 15.8%, $p < 0.001$), providing information (28.7% versus 22.1%, p=0.025) and other reflections (11.6% versus 0%, $p < 0.001$). In 14.0% of the successful sequences, reflections of change talk were paired with a question phrased to elicit change talk; this pattern did not appear in >10% of the unsuccessful sequences. In contrast, in 10.5% of the unsuccessful sequences, reflections of change talk were paired with information; this pattern did not appear in >10% of the successful sequences.

3.2.5 Discussion

We applied HMM and frequent pattern mining to test the fundamental hypothesis guiding Motivational Interviewing, which posits that counselors use of "MI-

Table 3.7: Frequent communication patterns in successful and unsuccessful patient-counselor communication sequences

Successful			Unsuccessful		
LUP	573	52.0%	LUP	118	45.7%
RCHTP	398	36.1%	RCHTP	87	33.7%
LUP, RCHTP	224	20.3%	LUP, RCHTP	45	17.4%
QECHTP	339	30.8%	GINFOP	74	28.7%
LUP, QECHTP	184	16.7%	LUP, GINFOP	44	17.1%
AF	314	28.5%	RST	70	27.1%
LUP, AF	166	15.1%	LUP, RST	30	11.6%
EA	314	28.5%	AF	68	26.4%
LUP, EA	188	17.1%	EA	48	18.6%
GINFOP	244	22.1%	LUP, EA	28	10.9%
LUP, GINFOP	143	13.0%	QECHTP	45	17.4%
RCMLP	228	20.7%	RCMLP	39	15.1%
LUP, RCMLP	121	11.0%	QEB	38	14.7%
QECMLP	200	18.1%	RO	30	11.6%
RST	174	15.8%	QECMLP	30	11.6%
LUP, RST	114	10.3%	SUM	29	11.2%
RCHTP, QECHTP	154	14.0%	SS	28	10.9%
SUM	138	12.5%	RCHTP, GINFOP	27	10.5%
LUP, SS	112	10.2%	HUPO	47	18.2%
HUPO	173	15.7%			

Note: Patterns that are aligned to the right are included in the immediately preceding pattern count. In these patterns, a counselor behavior was paired with a patient low uptake/facilitative comment, which is a marker of patient attention to the conversation and feedback suggesting the line of discussion may continue.

consistent" communication strategies (MICO) will lead to patient change talk [98]. Previous studies have empirically linked counselors' use of MICO communication strategies to higher rates of patient change talk in first-order Markov Chain models [100, 101, 49]. Our study leveraged data mining methods to provide an even stronger evidence for MI's fundamental hypothesis by considering longer-range dependencies in the data. Unlike simple first-order Markov Chain models, frequent pattern mining considers behavioral antecedents beyond the counselor behavior immediately preceding a patient change talk statement, while HMM identifies groups of communication behaviors occurring in successful and unsuccessful communication

sequences. The ability of HMM and frequent pattern mining to identify critical patterns in patient-counselor communication sequences advances research in the field of Motivational Interviewing, which has previously relied upon simple Markov Chain models [100, 101, 50, 49, 53, 24, 67].

In both analyses, MICO communication strategies were characteristic of successful sequences (i.e., those resulting in a change talk statement). In HMM, the majority of successful sequences began in the *"high motivation"* state, when counselors frequently use reflections of change talk or commitment language as well as affirmations. Other high-frequency counselor behaviors observed in successful sequences included statements emphasizing patients' decision-making autonomy, questions phrased to elicit change talk and the provision of information using patient-centered strategies. The frequent pattern mining results were similar. Reflections of change talk was the most frequent counselor communication strategy in successful sequences, followed by open questions phrased to elicit change talk, affirmations, statements emphasizing the patient's decision-making autonomy and sensitively provided information. Previous studies of MI behavior code sequences, which relied on first-order Markov Chain models to analyze communication sequences, have linked patients' expression of change talk to counselor reflections of change talk, [101, 49, 53, 24, 67] open questions phrased to elicit change talk, [101, 24, 67] and statements emphasizing the patient's decision-making autonomy [24, 67]. However, these studies did not find a link between change talk and counselors' use of affirmations or the provision of information, when examining specifically which of the MICO communication strategies were empirically linked to the elicitation of change talk. Thus, this publication is the first to provide empirical evidence for these causal linkages. One reason for this unique finding may be the treatment context, adolescent patients engaged in a voluntary weight loss trial. Adaptations of MI for the healthcare setting suggest that asking questions, demonstrating active listening through reflections and the provision of information

are critical communication skills for encouraging health-related behavior change [41]. Thus, providing information in a patient-centered manner in the context of health care treatment may be necessary to ensure patients have the requisite knowledge of their health care problem and its treatment.

The analysis of unsuccessful sequences, i.e., those resulting in a patient sustain talk statement, was typified by a combination of MICO and MI-inconsistent communication strategies (MIIN). Specifically, the majority of unsuccessful sequences in the HMM analysis began in a state of *"ambivalence"* which was characterized by large proportions of counselor reflections of both change talk and sustain talk. Similarly, in the frequent pattern mining analysis of unsuccessful sequences, reflections of change talk and sustain talk were two of the three most frequent counselor behaviors observed. These results are consistent with those of Gaume et al. [49] who found both MICO and MIIN were linked to the elicitation of sustain talk in a sample of at-risk young adult drinkers enlisted into the military. Specifically, counselors' use of simple and complex reflections and "other MICO" behaviors (an index of affirmations, statements emphasizing patient control, reframing and support) were empirically linked to the elicitation of sustain talk; neither open or closed questions were related to the elicitation of sustain talk. Carcone et al. [24] found counselors' questions and reflections specifically phrased to elicit patient sustain talk were the counselor behaviors most likely to elicit sustain talk among adolescents engaged in a weight loss trial. In contrast, Moyers et al. [101] found questions about the positive and negative aspects of the target behavior and reflections of sustain talk were empirically linked to the elicitation of sustain talk but MIIN was not. These variable findings suggest a need to tailor the MI communication strategies to the treatment context.

The task presented in this section is part of a line of research to develop machine-learning models to annotate (code) and analyze patient-counselor communication patterns. We have previously reported on the development of probabilistic genera-

tive models [75, 74] and application of novel features for maximum margin and deep learning classifiers [62] with the goal of automated annotation of MI session transcripts. Experiments applying the annotation model to novel datasets are underway to assess the generalizability of the model to more diverse types of clinical encounters (e.g, email coaching to increase fruit and vegetable intake, HIV clinical care visits [23]). We also developed and evaluated probabilistic and deep learning methods for the task of predicting the change talk at any point during the motivational interview [63]. The above work built on this past work to automatically annotate clinical encounters, specifically, this study presented two approaches for the sequential analysis of patient-counselor communication data for the purpose of identifying the counselor communication strategies linked to the elicitation of change talk and sustain talk. We are planning to examine the performance of the HMM and frequent pattern mining models in diverse data sets representing different populations and behavioral problems. Annotation and sequential analysis models together form the basis of a complete system to automatically code and analyze patient-counselor interactions. An automated system for behavioral coding and analysis could substantially accelerate the pace of research on the causal mechanisms of Motivational Interviewing and inform both the theory and clinical practice by providing clinicians with information about how to best tailor their communication strategies to different patient populations.

The above study was limited by the use of one dataset composed of 37 Motivational Interviewing transcripts of counseling sessions with African American adolescents in weight loss treatment. Thus, there is a need to replicate these findings with larger and more diverse data samples as the findings may not be representative of communication patterns in other contexts employing the Motivational Interviewing framework. In fact, when interpreted in light of the published literature, the results obtained in these experiments suggest that communication patterns are likely to vary

given the treatment context. There are, however, consistencies with previous Motivational Interviewing process studies providing support for the validity of our findings and suggesting some counselor communication strategies may cut across treatment contexts. Another limitation of this work was the fact that successful and unsuccessful sequences were analyzed independently. One implication of this approach is that the utility of a counselor behavior, such as the provision of information, to shift an interaction destined for failure to success, cannot be determined from these analyses.

3.2.6 Summary

Experimental results reported in this section, add to the growing evidence base examining the mechanisms of effect in Motivational Interviewing using modeling approaches that overcome critical shortcomings of previous methods. While counselors' use of "MI-consistent" communication behaviors has been previously linked to higher rates of change talk in correlational studies [100, 25, 129, 94] and simple Markov Chain models [100, 101, 49], the use of HMM and frequent pattern mining analyses improves upon these approaches by considering long-range dependencies in the data. The results of this pattern mining work suggest a more complex pattern between counselor communication behaviors and patient talk that varies depending on the context in which Motivational Interviewing is being used.

CHAPTER 4 SEGMENTATION OF CLINICAL CONVERSATION

In the previous two chapters, we examined the utility of machine learning methods for automated annotation [62, 74] and sequential analysis [63, 61] of in-person MI sessions. Experimental data utilized in those studies were transcribed auto recordings of in-person MI sessions with a counselor, which were segmented into counselor and client utterances during the transcription process. In this chapter, we focus on email-based clinical conversation to automate the segmentation of clinical conversation into groups of codable MI behaviors.

4.1 Introduction

The emergence of e-Health technologies has greatly expanded the reach of behavioral interventions. One such intervention is email-delivered Motivational Interviewing (MI). In this project, we focus on the analysis of email-delivered MI, or e-Coaching, to promote healthy eating among young adults. The e-Coaching dataset is composed of email correspondence between an MI counselor and the young adult patient. Unlike transcribed in-person exchanges, email correspondence is not clearly segmented into codable speech acts (i.e., utterances). Thus, the unstructured nature of e-Coaching exchanges poses a unique set of analytic challenges. Segmentation of e-Coaching exchanges into textual fragments that correspond to distinct e-Coach and patient communication behaviors is a significant barrier to qualitative analysis of this type of clinical conversation. Automating this task is a unique and challenging problem due to the following reasons:

1. Emails are unstructured text containing informal information exchange in a non-traditional format. For example, an e-Coach usually responds to several previous patient statements in one email. In contrast, in a traditional, in-person MI session, each utterance is assumed to be a response to an immediately preceding utterance.

2. Discourse segments in e-Coaching do not have a clear breakpoint, such as the end of a sentence or a paragraph. One sentence may be divided into fragments corresponding to multiple MI behaviors. On the other hand, an MI behavior may comprise several sentences.

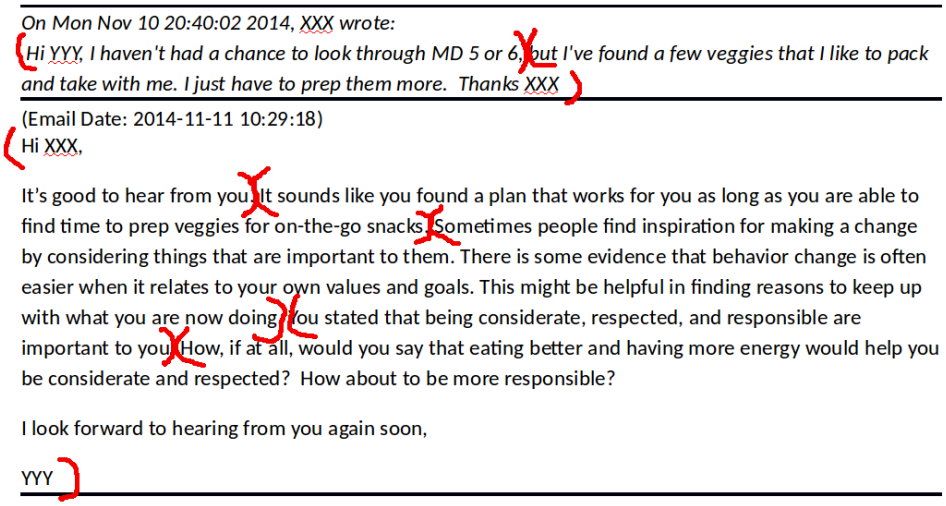

Figure 4.1: Example of an e-Coaching exchange segmented into fragments corresponding to MI behaviors of an e-Coach and a patient

Figure 4.1 illustrates a segmentation of an e-Coaching exchange, in which the first sentence is segmented into 2 MI behavior fragments, while the fourth and fifth MI behavior fragments comprise one and three sentences, respectively. Segmentation of e-Coaching exchanges constitutes a special case of clinical discourse analysis [133] aimed at better understanding the effective communication strategies specific to this type of behavioral interventions.

The goal of this project is to assess the effectiveness of deep learning methods for the task of automated segmentation of e-Coaching emails into textual fragments corresponding to individual patient and provider behaviors. For this study, we utilized the data from MENU GenY (Making Effective Nutrition Choices for Generation Y) [6], a web-delivered public health intervention with email-based coaching to encourage increased fruit and vegetable intake among young adults, aged 2130. A secondary

goal of the MENU GenY project was to identify the specific communication strategies used by e-Coaches to elicit change talk for healthier eating among young adult patients. Segmentation of clinical conversation in the context of electronically delivered interventions into groups of MI behaviors is traditionally performed manually by MI researchers, which significantly slows down its qualitative analysis. *This work is the first work to evaluate the empirical effectiveness of deep learning architectures in addressing the problem of discourse segmentation in the context of email-based behavioral interventions.*

Specifically, we evaluate the effectiveness of distributed representations (i.e. embeddings) of words and punctuation marks as well as part-of-speech (POS) features in conjunction with both traditional supervised machine learning methods, such as linear-chain Conditional Random Fields (CRF) [79] and deep learning methods, such as Multi-Layer Perceptron (MLP) [117], Bidirectional Recurrent Neural Network (BRNN) [118] and Convolutional Recurrent Neural Network (CRNN) [130], to determine the best performing method and feature combination for the task of segmentation of e-Coaching emails into MI behaviors.

4.2 Related work

Prior work on textual segmentation in the biomedical domain primarily focused on sentence boundary detection [56, 76, 130] and segmentation of clinical documents in patients' electronic health records (EHR) into sections and headers. [11, 38, 128, 29] In particular, maximum entropy models [128] and Support Vector Machine (SVM) along with word vector similarity metrics and several heuristics [11] have been applied to identify specific sections in EHR, such as general patient information, medical history, procedures, findings, etc. Denny et al. [38] proposed SecTag algorithm, which combined natural language processing techniques, terminology-based rules and a Naïve Bayes classifier to identify sections and headers in EHR. Segmentation of e-Coaching emails, however, is different from segmentation of other clinical documents,

since the focus is on dialog acts in clinical conversation.

SVM in conjunction with prosodic and part-of-speech features [76] and recurrent convolutional neural networks [56] have also been utilized for *sentence boundary detection* in general text. Liu et al. [87] demonstrated that a linear-chain CRF outperforms Hidden Markov and maximum entropy models for this task.

Segmentation of e-Coaching emails is also different from traditional shallow discourse analysis [48], which besides identification of speech acts, also aims to determine the types of transitions between speech acts and label speech acts with the speakers who performed them in a multi-speaker conversation. The proposed methods will automate the process of segmenting clinical exchanges into MI behaviors, which will significantly reduce the time and resources required to perform such segmentation manually. Furthermore, these methods can be integrated with the automated MI behavior coding methods [62, 74] to create a software pipeline for fully automated analysis of email-delivered behavioral interventions.

4.3 Methods

4.3.1 Dataset

The experimental dataset for this work was constructed from 49 e-Coaching sessions, which include 330 and 281 emails by e-Coaches and patients, respectively. Various statistics of the experimental dataset are provided in Table 4.1. Each e-Coaching session represents an MI intervention delivered via email. Emails were segmented into 3,138 text fragments and annotated with MY-SCOPE codebook. Email segmentation can be considered as sequence tagging, which can be framed as a binary classification problem, in which each word or punctuation mark is annotated with one of the two class labels ("new segment" or "same segment") to indicate whether it is a beginning of a new MI behavior segment or not. In total, the dataset consists of 95,777 words and 7,140 punctuation marks and includes 3,138 "new segment" and 99,779 "same

segment" instances, illustrated in Table 4.1[1]. In this study, we experimented with traditional machine learning methods, such as Conditional Random Fields (CRF) [79] and deep learning methods, such as Multi-Layer Perceptron (MLP) [117], Bi-directional Recurrent Neural Network (BRNN) [118] and Convolutional Recurrent Neural Network (CRNN) [130]. In the case of MLP, training and testing samples were created based on a sliding window of $2n$ words or punctuation marks over each position (which could be a word or a punctuation mark) in a given input sequence, such that each sample consists of the n words or punctuation marks after the current position and n words or punctuation marks prior to the current position, including the position itself. In the case of CRF, BRNN and CRNN models, an e-Coaching email was taken as an input sequence, POS tags and embeddings of each word or punctuation mark were used as input and binary labels corresponding to "new segment" and "same segment" classification decisions were considered as the model output. In the gold standard, words or punctuations within the same segment were assigned the label of 0 and the last word or punctuation mark of a segment were assigned the label of 1.

Table 4.1: Summary of statistics of the experimental dataset and example of a segmented sequence

Instances	**Class labels**		**Tokens**		**Emails**		**Annotation**	
	new	**same**	**words**	**punc.**	**pat.**	**prov.**	**method**	**codes**
102,917	3,138	99,779	95,777	7,140	281	330	MYSCOPE	115

4.3.2 Features

We utilized three types of features in conjunction with CRF, MLP, BRNN and CRNN: word embeddings as lexical features, punctuation and POS features. Syntactic abstractions of individual words, such as POS tags, have been previously shown to be effective features for similar natural language processing tasks [87, 130]. To extract POS features, we pre-processed e-Coaching emails using the NLTK POS tag-

[1]punc.: punctuation marks, pat.: patient, prov.: provider

ger[2]. Punctuation marks, which correspond to one of the symbols {'.', ',', '!', '?', ':', ';'} between a pair of words, were also used as a feature, since punctuation marks designate the boundary of a sentence, clause or a phrase and often also correspond to a segment boundary [29]. For natural language processing (NLP) tasks, inputs are received as textual fragments, in which individual words are as the basic lexico-semantic units. Therefore, it is important to represent a word in such a way that preserves all relevant lexical and semantic information. Embedding is a form of distributed representation, when each word is associated with a dense real-valued vector in low-dimensional space. Embeddings have been previously shown to effectively capture semantic, syntactic and morphological properties of words [111, 95]. For experiments reported in this study, we utilized word embeddings pre-trained on Google News corpus consisting of 1.6 billion words using word2vec software package.[3] For words or punctuation marks, which do not have pre-trained embeddings, we utilized the embeddings of the same dimensionality trained on the experimental dataset. CRF utilized lexical features, POS tags and the preceding label.

4.3.3 Segmentation models

We experimented with 4 different classifiers, including one traditional machine learning model (CRF) and three deep learning methods (MLP, BRNN and CRNN). Since deep learning architectures provide a flexible mechanism for constructing complex models, we take advantage of this flexibility to test different variations of MLP, BRNN and CRNN models for the task of segmentation of e-Coaching emails.

Conditional Random Fields (CRF): CRF has been widely used in various NLP tasks that involve sequence annotation, such as part-of-speech tagging.[79, 64] Unlike the maximum-entropy Markov model, which uses per-state exponential models for conditional probability of the next state given a current state, CRF model directly estimates a distribution of the entire output sequence conditioned on the ob-

[2] https://www.nltk.org/

[3] https://code.google.com/p/word2vec/

servation sequence. A linear-chain CRF model is defined as a conditional probability distribution $p(y|x)$ of output sequence y, given input sequence x:

$$p(y|x) = \frac{1}{Z_x} \exp\left(\sum_{t=1}^{T}\sum_{k} \lambda_k f_k(y_{t-1}, y_t, x, t)\right) \tag{4.1}$$

where Z_x is a normalization factor, $f_k(y_{t-1}, y_t, x, t)$ is a feature function, and λ_k is a learned weight associated with feature f_k. The optimal output sequence y^* for input sequence x, $y^* = arg\,max_y p(y|x)$, is obtained efficiently using the Viterbi algorithm. In our experiments, the following features were utilized in conjunction with CRF: i) current word or punctuation ii) next and previous 3 words or punctuations iii) binary feature indicating whether a word or punctuation is a special character (';', '?', '.', ',', '!', ':', etc.) or not iv) binary feature indicating whether a word is a title word or not (e.g. "The" is a title word but "the" is not) v) POS tags.

Multi-Layer Perceptron (MLP): MLP is a neural network, which consists of multiple fully connected layers that map an input to one or several outputs [117]. Figure 4.2 illustrates a multi-layer perceptron with a single hidden layer. MLPs have no cycles or loops. Information in them flows only forward, from the input layer through the hidden layer(s) to the output layer. The MLP in this study utilizes one hidden layer consisting of 128 neurons and rectified linear unit (ReLU) as a nonlinear activation function. In order to prevent over-fitting, we applied dropout (random masking of neurons [121] to fully connected layers during training. Dropout was also applied to a fully connected layer in CRNN.

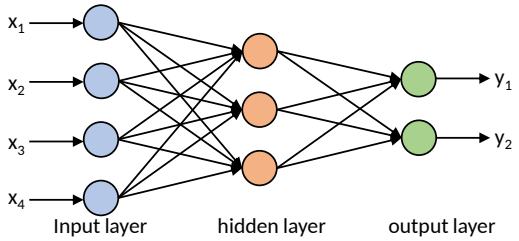

Figure 4.2: Multi-layer perceptron with a single hidden layer

Bi-directional Recurrent Neural Network (BRNN): BRNN is a neural network designed to capture sequential patterns by considering both past and future inputs as well as complex relationships between input features and output labels [118]. The hidden state of BRNN is an aggregation of the hidden states of a forward and backward recurrent neural networks (RNNs). Gated Recurrent Units (GRU) [31] capable of handling variable size input sequence and having internal memory, which can be reset, were utilized as an RNN in this work.

Convolutional Recurrent Neural Network (CRNN): CRNN[130] shown in Figure 4.3 is a deep neural architecture, which combines convolutional and recurrent layers. Our implementation of CRNN consists of 5 layers: 1) input layer 2) embedding layer 3) convolution layer with max pooling 4) BRNN layer 5) fully connected layer with dropout and sigmoid output. E-coaching email exchanges are represented as a sequence of m words and punctuations, which are fed into the input and embedding layers to produce a $m \times n_e$ matrix after fetching the embeddings for words and punctuations in the input sequence. This matrix is a distributed representation of an input email exchange, which contains rich morpho-syntactic information that can be utilized for its segmentation. When POS tags are utilized along with word embeddings, they are represented with a 10-dimensional vector, which is concatenated with 300-dimensional word embeddings to obtain new embedding vectors $n_e = [n_w; n_p]$ of size 310. The primary purpose of a convolution layer is to extract new features for each word or punctuation mark based on the neighboring words or punctuation marks. A one-dimensional (1D) convolution operation is utilized in this layer in our implementation of BRNN. In 1D convolution, one filter is responsible for the extraction of one feature. After applying n_f different filters with zero-padding on both sides of the input text, n_f features are produced by the convolution layer for each word. A max pooling over time operation is then applied to find the most significant features in a textual fragment. The bi-directional recurrent layer receives new features

extracted from the convolution layer. Unidirectional RNNs are typically utilized to capture long-range dependencies in a sequence of observations. Bi-directional RNNs, on the other hand, are capable of capturing both past and future contexts through forward and backward traversals of a sequence. The purpose of the fully connected layer in CRNN is to use the output of the bidirectional RNN layer for classifying each word or punctuation into "new segment" or "same segment" classes. Since a fully connected layer has a larger number of parameters, they are more likely to excessively co-adapt to other parameters in the network and result in over-fitting. To prevent this, we utilized dropout by randomly ignoring 50% of the connections in the fully connected layer of CRNN. Finally, logistic sigmoid outputs the probability of classifying or labeling each word or punctuation mark with "same segment" class. We experimentally determined the optimal parameters using 5-fold cross-validation and found out that the best performance is achieved when filter length in the convolution layer is 7, number of filters is 100, max pooling size is 3, ReLU is used as an activation function in the convolution layer, hyperbolic tangent is used as an activation function in the bi-directional RNN layer and the number of dimensions in the hidden state of RNNs is 200. Adam [72] with 50 epochs, the batch size of 32 and learning rate of 0.001 was used for optimization and the early stopping strategy was applied.[4]

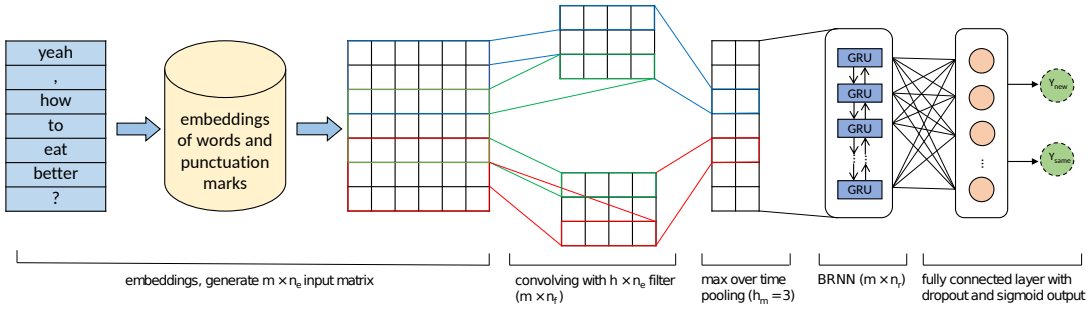

Figure 4.3: Architecture of a convolutional recurrent neural network for automated segmentation of e-Coaching emails into fragments corresponding to MI behaviors

[4]source code of all methods is available at `https://github.com/teanalab/eCoaching-Text-Segmentation`

4.3.4 Evaluation

We report standard metrics of precision, recall and F1-measure to evaluate the performance of the classifiers [1]. Accuracy is not reported as a performance metric, since it is highly sensitive to the distribution of prior class probabilities, which is skewed when datasets with unbalanced classes are involved. The results are reported based on 5-fold cross-validation (one fold was used as a test set and the remaining 4 folds were used as a training set) and weighted macro-averaging over the folds. We also report the area under the precision-recall curve (AUPR), due to its effectiveness in measuring the performance of binary classifiers in the case of the datasets with imbalanced class distribution [35].

4.4 Results

Experimental results of this work spanned three dimensions. First, we determined the optimal sizes of word embedding vectors and the sliding window of MLP. Second, we evaluated the performance of different methods with respect to detecting "new segment" as well as the weighted average over "new segment" and "same segment" classes. Third, we assessed the impact of different types of features as well as their combination on the performance of different machine learning methods on the e-Coaching email segmentation task.

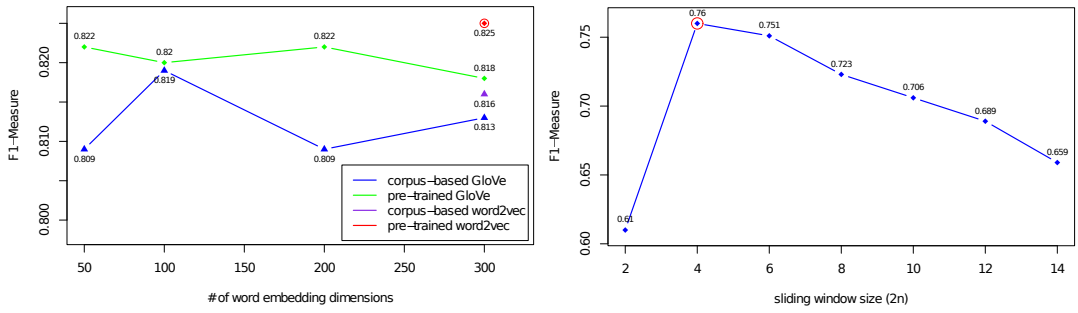

Figure 4.4: F1-measure of CRNN on the task of e-Coaching email segmentation by varying the number of dimensions in pre-trained and corpus-based GloVe and word2vec embeddings (left). F1-measure of MLP on the task of e-Coaching email segmentation by varying the size of the sliding window (right).

Figure 4.4 (left) illustrates the performance of CRNN on the task of e-Coaching email segmentation by varying the number of dimensions in pre-trained and corpus-based GloVe [5] and word2vec embeddings. We observed that the best performance is achieved with pre-trained 300-dimensional word2vec word vectors, when three types of features are used together. Therefore, we report the results for other deep learning models used in this study when 300-dimensional word2vec embedding vectors are utilized. The input layer of MLP consists of a sum of embeddings of n words or punctuation marks before and the sum of embeddings of n words or punctuation marks after the word or punctuation mark, which is the center of a sliding window of $2n$ words or punctuation marks. Figure 4.4 (right) demonstrates the performance of MLP on e-Coaching email segmentation by varying the size of the sliding window. It can be observed that the best performance of MLP is achieved when the size of the sliding window is 4 (or $n = 2$). Therefore, MLP results in the remaining experiments are reported when n is set to 2.

Table 4.2: Performance of CRF, MLP, BRNN and CRNN on "new segment" detection as well as the weighted average over "new segment" and "same segment" classes when only lexical features are used. The highest value for each performance metric is highlighted in boldface.

Method	**New Segment**			**Overall**			**AUPR**
	Prec.	**Reca.**	**F1**	**Prec.**	**Reca.**	**F1**	
CRF	0.782	0.691	0.733	0.983	0.984	0.984	0.780
MLP	**0.836**	0.593	0.694	0.982	0.983	0.982	0.736
BRNN	0.606	0.680	0.641	0.977	0.976	0.976	0.655
CRNN	0.775	**0.797**	**0.785**	**0.986**	**0.986**	**0.986**	**0.818**

As follows from Table 4.2[6], CRNN outperforms all other methods in terms of recall and F1-measure achieving 0.797 recall and 0.785 F1-measure for new segment detection. CRNN also shows superior performance according to all performance metrics calculated as a weighted average over "new segment" and "same segment" classes. BRNN had the lowest performance among all models in terms of precision and F1-

[5] `https://nlp.stanford.edu/projects/glove/`

[6] Prec.: Precision, Reca.: Recall

measure. On the other hand, MLP had the highest precision of 0.836 when lexical features are used to identify "new segment". CRF achieves 0.733 F1-measure, the second highest in identifying "new segment". CRF also demonstrated the second best performance among all models according to all metrics calculated as a weighted average over both classes. Experimental results indicate that the performance of all classifiers according to all metrics calculated as a weighted average over both classes is significantly higher than their performance on "new segment" detection, which is expected since 96.95% of instances belong to the "same segment" class and 99.3% of them are correctly classified. For example, CRNN achieves 27.23%, 23.71% and 25.61% higher precision, recall and F1-measure calculated as a weighted average over "new segment" and "same segment" classes, compared to the "new segment" detection.

Table 4.3: Performance of CRF, MLP, BRNN and CRNN on "new segment" detection as well as the weighted average over "new segment" and "same segment" classes when all types of features are used together. The highest value for each performance metric is highlighted in boldface.

Method	**New Segment**			**Overall**			**AUPR**
	Prec.	**Reca.**	**F1**	**Prec.**	**Reca.**	**F1**	
CRF	0.813	0.772	0.792	0.988	0.988	0.988	**0.877**
MLP	**0.817**	0.710	0.760	0.986	0.987	0.986	0.842
BRNN	0.683	0.820	0.745	0.985	0.983	0.984	0.770
CRNN	0.789	**0.864**	**0.825**	**0.990**	**0.989**	**0.989**	0.867

Table 4.3[7] summarizes the results of all models on the task of segmentation of e-Coaching emails when word embeddings or lexical features are used in combination with punctuation and POS features. Similar to results in Tables 4.2, CRNN demonstrates the best performance among all methods achieving 0.864 recall with 0.825 F1-measure for "new segment" detection and 0.990 precision with 0.989 recall and F1-measure overall. BRNN and CRF demonstrated the lowest and second highest performance on the task of email segmentation among all methods, respectively. We

[7]Prec.: Precision, Reca.: Recall

observed that classification performance significantly improved for "new segment" detection when lexical features are used in combination with punctuation and POS features. Specifically, precision increases by 3.96%, -2.27%, 12.71% and 1.81%; recall increases by 11.72%, 19.73%, 20.59% and 8.41%; and F1-measure increases by 8.05%, 9.51%, 16.22% and 5.1% for CRF, MLP, BRNN and CRNN methods, respectively, on new segment detection when all types of features are utilized together. Similarly, precision increases by 0.51%, 0.41%, 0.82% and 0.41%; recall increases by 0.41%, 0.41%, 0.72% and 0.3%; and F1-measure increases by 0.41%, 0.41%, 0.82% and 0.3% for CRF, MLP, BRNN and CRNN methods, respectively, as a weighted average over "new segment" and "same segment" classes when lexical features are used in combination with punctuation and POS features.

Table 4.4: Area under the precision-recall curve (AUPR) values of all classifiers demonstrating the impact of different types of features on e-Coaching email segmentation performance. Highest AUPR value for each feature set across all models is highlighted in boldface.

Features	**CRF**	**MLP**	**BRNN**	**CRNN**
word embeddings only	0.780	0.736	0.655	**0.818**
word embeddings + POS	0.797 (+2.18%)	0.746 (+1.36%)	0.647 (-1.22%)	**0.798 (-2.44%)**
word embeddings + punctuation	**0.876 (+12.31%)**	0.835 (+13.45%)	0.774 (+18.17%)	0.874 (+6.85%)
all features	**0.877 (+12.44%)**	0.842 (+14.4%)	0.770 (+17.56%)	0.867 (+6%)

Table 4.4 illustrates the impact of different types of features as well as their combination on e-Coaching email segmentation performance. Punctuation and POS features have similar effect measured by the AUPR, which increases by 12.44%, 14.4%, 17.56% and 6% for CRF, MLP, BRNN and CRNN, respectively, when all features are used together. Individually, although punctuation features improve the performance of all classifiers, POS features improve the performance of only CRF and MLP. CRF achieved the highest AUPR when all types of features are used together. On the other hand, POS features degraded the AUPR of BRNN and CRNN.

4.5 Discussion

This study is the first effort to design and evaluate machine learning methods for automated segmentation of e-Coaching sessions. Experimental results indicate that CRNN is the best model among all machine learning methods considered for this study. CRNN achieved 0.989 F1-measure overall and 0.825 F1-measure for detecting "new segment". The robust performance of CRNN provides an evidence that deep learning models are capable of detecting the boundaries of patient and provider behaviors in email delivered behavioral interventions. Our experiments also highlight the importance of punctuation and POS features along with word embeddings for all machine learning methods employed this study. Although the domain of this study was intentionally focused, we believe that the proposed methods are not limited to e-Coaching sessions and our conclusions can be generalized to other domains, which require discourse segmentation.

Punctuation marks and POS features resulted in significant improvement in the performance of traditional machine learning and deep learning methods. Punctuation features had a stronger individual impact on model performance than POS features. In all cases, CRF and MLP performed better, when word embeddings were used in conjunction with punctuations and POS features. Considering punctuations improved the performance of BRNN and CRNN measured by precision, recall and F1-measure, while POS features lowered their AUPR.

The convolution layer made a significant difference between the performance of CRNN and BRNN in MI session discourse segmentation. CRNN had 22.46% and 10.74% higher F1-measure in "new segment" detection and 1.02% and 0.51% higher F1-measure overall compared to BRNN, when word embeddings and all other features were used, respectively. In CRNN, a convolution layer performs a series of convolution and pooling operations, which produce a number of important high-level features from input embeddings. These high-level features are then utilized by the bidirectional

RNN layer in CRNN, which translates to a significant increase in performance. In contrast, BRNN utilizes the input embeddings directly as features.

Although punctuation marks play an important role in segmentation boundary detection, a few errors were triggered by the presence of punctuation marks. For example, a text segment from an e-Coaching email *"A typical day in regards to fruit and vegetable has me eating about a serving at breakfast (our cafe has cut up fruit) and then maybe a piece of fruit later in the day or as a snack. Vegetable tends to be a side serving at lunch and dinner and I get celery or carrot cuts with dressing for a snack a lot of times. I could probably add some sort of vegetable into my breakfast (like spinach in an omelet) and snack on another piece of fruit when I am hungry rather than the junk food I tend to eat."* was incorrectly segmented after the first sentence, when period was encountered. Similarly, additional information is a common cause for misclassification of an email segment into multiple segments. For instance, although the first sentence in the above email segment represents a positive commitment to behavior change, the next two sentences provide additional information to support the patient's commitment.

4.6 Summary

Segmentation is the first step of qualitative analysis of unstructured clinical communications, such as e-Coaching. Although several studies have focused on the segmentation problem in biomedical context, they are limited to segmenting clinical text in EHR into sections and sentences. No previous studies considered the task of automated segmentation of clinical communications into groups of MI behaviors in the context of unstructured MI sessions. By comparing the performance of machine learning methods for the task of segmentation of e-Coaching emails, we found out that convolutional recurrent neural networks demonstrate the best performance in terms of most performance metrics. Manual segmentation of e-Coaching sessions is a very resource-intensive and time-consuming task, which can significantly decrease

the time and effort required to develop effective behavioral interventions. Our proposed methods can help to identify textual segments corresponding to MI behaviors in unstructured clinical dialog, which can then be annotated with MI behavior annotation methods in a pipeline setting. Automated segmentation and annotation of e-Coaching emails can significantly decrease the time to identify effective communication strategies in email-based MI.

CHAPTER 5 CONCLUSION AND FUTURE WORK

5.1 Conclusion

In this dissertation, we presented our research accomplishments to fully automate the analysis of patient-provider counseling and understand the MI mechanism of effect.

First, we propose novel features and report the results of an extensive experimental evaluation of state-of-the-art supervised machine learning methods for text classification using those features, to help clinical researchers and practitioners assess the feasibility of using these methods for the task of automatic annotation of clinical text using the codebooks of realistic size. We found out that Support Vector Machine using only lexical features consistently outperforms all other classifiers on caregiver and adolescent datasets according to most metrics. Adding contextual and semantic features further improves the performance of SVM on both datasets, achieving close to human accuracy when the codebooks consisting of 16 and 17 classes are used to annotate caregiver and adolescent transcripts, respectively.

Second, we perform two sequential analysis of pre-coded MI transcripts. In the first experiment, we compared the accuracy of Recurrent Neural Networks with Markov Chain and Hidden Markov Model for the task of predicting the success of motivational interviews. We found out that individual PPC exchanges are highly indicative of the overall progression and future trajectory of clinical interviews and can be used to predict their overall success. Our methods can facilitate motivational interviewing researchers to identify the most likely sequences in successful and unsuccessful motivational interviews, which can directly inform clinical practice and increase the effectiveness of behavioral interventions. In our second experiment, we overcome the critical shortcomings of previous methods. While counselors' use of "MI-consistent" communication behaviors has been previously linked to higher rates of change talk in correlational studies [100, 25, 129, 94] and simple Markov Chain

models [100, 101, 49], the use of HMM and frequent pattern mining analyses improves upon these approaches by considering long-range dependencies in the data. The results of this pattern mining work suggest a more complex pattern between counselor communication behaviors and patient talk that varies depending on the context in which Motivational Interviewing is being used.

Finally, we propose various segmentation models because segmentation is the first step of qualitative analysis of unstructured clinical communications, such as e-Coaching. Although several studies have focused on the segmentation problem in a biomedical context, they are limited to segmenting clinical text in EHR into sections and sentences. By comparing the performance of machine learning methods for the task of segmentation of e-Coaching emails, we found out that convolutional recurrent neural networks demonstrate the best performance in terms of most performance metrics. Our proposed methods can help to identify textual segments corresponding to MI behaviors in unstructured clinical dialog, which can then be annotated with MI behavior annotation methods in a pipeline setting.

5.2 Future research directions

We plan to explore the following possible future research directions.

First, our study in this dissertation has focused on manual feature extraction methods. An interesting automated feature extraction method can be considered to improve the performance of utilized machine learning models.

Second, Attention-based models are increasingly popular because information is lost by compressing variable-length long sequences into a fixed-size vector in RNN. Therefore, we would like to consider attention-based neural networks in order to improve the performance of our annotation, segmentation and sequence models.

Third, our experimental results indicate that ML methods can be used for real-time monitoring of the progression of clinical interviews. We plan to integrate the sequential model with segmentation and auto-coding classifiers to develop a fully

automated e-Coaching.

Finally, the limitation of our study is that our dissertation data is collected from a single medical institute; formatting, style and email segment can be different in other settings. Therefore, there is a need to replicate the experiments with different data sets. As our future work, we plan to evaluate our approach on the datasets from other behavioral interventions.

APPENDIX

Gold Standard: a term used to describe a collection of a labeled dataset which has been manually labeled by the experts.

State-of-the-art: the most recent or latest version of a particular technology. State-of-the-art machine learning methods refer to the best available machine learning methods developed using modern techniques and technologies.

REFERENCES

[1] Aas, K., and Eikvil, L. Text categorisation: A survey, 1999.

[2] Abdullah, U., Ahmad, J., and Ahmed, A. Analysis of effectiveness of apriori algorithm in medical billing data mining. In *Emerging Technologies, 2008. ICET 2008. 4th International Conference on* (2008), IEEE, pp. 327–331.

[3] Aggarwal, C. C. Outlier analysis. In *Data mining* (2015), Springer, pp. 237–263.

[4] Aggarwal, C. C., and Han, J. *Frequent pattern mining*. Springer, 2014.

[5] Agrawal, R., Srikant, R., et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB* (1994), vol. 1215, pp. 487–499.

[6] Alexander, G. L., Lindberg, N., Firemark, A. L., Rukstalis, M. R., and McMullen, C. Motivations of young adults for improving dietary choices: Focus group findings prior to the menu geny dietary change trial. *Health Education & Behavior* (2017), 1090198117736347.

[7] Amrhein, P. C. How does motivational interviewing work? what client talk reveals. *Journal of Cognitive Psychotherapy 18*, 4 (2004), 323–336.

[8] Amrhein, P. C., Miller, W. R., Yahne, C. E., Palmer, M., and Fulcher, L. Client commitment language during motivational interviewing predicts drug use outcomes. *Journal of consulting and clinical psychology 71*, 5 (2003), 862.

[9] Apodaca, T., Manuel, J. K., Moyers, T., and Amrhein, P. Motivational interviewing with significant others (miso) coding manual. *Unpublished manuscript* (2007).

[10] Apodaca, T. R., and Longabaugh, R. Mechanisms of change in motivational interviewing: a review and preliminary evaluation of the evidence. *Addiction 104*, 5 (2009), 705–715.

[11] Apostolova, E., Channin, D. S., Demner-Fushman, D., Furst, J., Lytinen, S., and Raicu, D. Automatic segmentation of clinical texts. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE* (2009), IEEE, pp. 5905–5908.

[12] Atkins, D. C., Steyvers, M., Imel, Z. E., and Smyth, P. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science 9*, 1 (2014), 49.

[13] Bakeman, R., and Gottman, J. M. *Observing interaction: An introduction to sequential analysis*. Cambridge university press, 1997.

[14] Bakeman, R., and Quera, V. *Sequential analysis and observational methods for the behavioral sciences*. Cambridge University Press, 2011.

[15] Baldi, P., Chauvin, Y., Hunkapiller, T., and McClure, M. A. Hidden markov models of biological primary sequence information. *Proceedings of the National Academy of Sciences 91*, 3 (1994), 1059–1063.

[16] Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. A neural probabilistic language model. *Journal of machine learning research 3*, Feb (2003), 1137–1155.

[17] Bengio, Y., Frasconi, P., and Simard, P. The problem of learning long-term dependencies in recurrent networks. In *Neural Networks, 1993., IEEE International Conference on* (1993), IEEE, pp. 1183–1188.

[18] Bertholet, N., Faouzi, M., Gmel, G., Gaume, J., and Daeppen, J.-B. Change talk sequence during brief motivational intervention, towards or away from drinking. *Addiction 105*, 12 (2010), 2106–2112.

[19] Bethel, C. L., Hall, L. O., and Goldgof, D. Mining for implications in medical data. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on* (2006), vol. 1, IEEE, pp. 1212–1215.

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

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

[22] Can, D., Marín, R. A., Georgiou, P. G., Imel, Z. E., Atkins, D. C., and Narayanan, S. S. it sounds like...: A natural language processing approach to detecting counselor reflections in motivational interviewing. *Journal of counseling psychology 63*, 3 (2016), 343.

[23] Carcone, A., Kotov, A., Hasan, M., Dong, M., Eggly, S., Hartlieb, K., Alexander, G., Lu, M., and Naar, S. Using natural language processing to understand the antecedents of behavior change. In *Annals Of Behavioral Medicine* (2018), vol. 52, pp. S422–S422.

[24] Carcone, A. I., Naar-King, S., Brogan, K., Albrecht, T., Barton, E., Foster, T., Martin, T., and Marshall, S. Provider communication behaviors that predict motivation to change in black adolescents with obesity. *Journal of developmental and behavioral pediatrics: JDBP 34*, 8 (2013), 599.

[25] Catley, D., Harris, K. J., Mayo, M. S., Hall, S., Okuyemi, K. S., Boardman, T., and Ahluwalia, J. S. Adherence to principles of motivational interviewing and client within-session behavior. *Behavioural and Cognitive Psychotherapy 34*, 1 (2006), 43–56.

[26] Chang, C.-C., and Lin, C.-J. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST) 2*, 3 (2011), 27.

[27] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research 16* (2002), 321–357.

[28] Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).

[29] Cho, P. S., Taira, R. K., and Kangarloo, H. Text boundary detection of medical reports. In *Proceedings of the AMIA Symposium* (2002), American Medical Informatics Association, p. 998.

[30] Choi, E., Bahadori, M. T., Schuetz, A., Stewart, W. F., and Sun, J. Doctor ai: Predicting clinical events via recurrent neural networks. In *Proceedings of Machine Learning for Healthcare Conference* (2016), pp. 301–318.

[31] Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[32] Chuzhanova, N. A., Jones, A. J., and Margetts, S. Feature selection for genetic sequence classification. *Bioinformatics 14*, 2 (1998), 139–143.

[33] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., and Kuksa, P. Natural language processing (almost) from scratch. *Journal of Machine Learning Research 12*, Aug (2011), 2493–2537.

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

[35] Davis, J., and Goadrich, M. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning* (2006), ACM, pp. 233–240.

[36] De Choudhury, M., Counts, S., and Horvitz, E. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2013), ACM, pp. 3267–3276.

[37] De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. Predicting depression via social media. *ICWSM 13* (2013), 1–10.

[38] Denny, J. C., Spickard III, A., Johnson, K. B., Peterson, N. B., Peterson, J. F., and Miller, R. A. Evaluation of a method to identify and categorize section headers in clinical documents. *Journal of the American Medical Informatics Association 16*, 6 (2009), 806–815.

[39] Deshpande, M., and Karypis, G. Evaluation of techniques for classifying biological sequences. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining* (2002), Springer, pp. 417–431.

[40] Deshpande, M., Kuramochi, M., Wale, N., and Karypis, G. Frequent substructure-based approaches for classifying chemical compounds. *IEEE Transactions on Knowledge and Data Engineering 17*, 8 (2005), 1036–1050.

[41] Douaihy, A., Kelly, T. M., and Gold, M. A. *Motivational interviewing: A guide for medical trainees*. Oxford University Press, USA, 2015.

[42] Durgesh, K. S., and Lekha, B. Data classification using support vector machine. *Journal of Theoretical and Applied Information Technology 12*, 1 (2010), 1–7.

[43] Eide, H., Quera, V., Graugaard, P., and Finset, A. Physician–patient dialogue surrounding patients expression of concern: applying sequence analysis to rias. *Social Science & Medicine 59*, 1 (2004), 145–155.

[44] Fournier-Viger, P., Gomariz, A., Gueniche, T., Soltani, A., Wu, C.-W., and Tseng, V. S. Spmf: a java open-source pattern mining library. *The Journal of Machine Learning Research 15*, 1 (2014), 3389–3393.

[45] Fournier-Viger, P., Lin, J. C.-W., Gomariz, A., Gueniche, T., Soltani, A., Deng, Z., and Lam, H. T. The spmf open-source data mining library version 2. In *Joint European conference on machine learning and knowledge discovery in databases* (2016), Springer, pp. 36–40.

[46] Freund, Y. Boosting a weak learning algorithm by majority. *Information and computation 121*, 2 (1995), 256–285.

[47] Freund, Y., Schapire, R., and Abe, N. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence 14*, 771-780 (1999), 1612.

[48] Galley, M., McKeown, K. R., Fosler-Lussier, E., and Jing, H. Discourse segmentation of multi-party conversation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics* (2003).

[49] Gaume, J., Bertholet, N., Faouzi, M., Gmel, G., and Daeppen, J.-B. Counselor motivational interviewing skills and young adult change talk articulation during brief motivational interventions. *Journal of substance abuse treatment 39*, 3 (2010), 272–281.

[50] Gaume, J., Gmel, G., Faouzi, M., and Daeppen, J.-B. Counsellor behaviours and patient language during brief motivational interventions: A sequential analysis of speech. *Addiction 103*, 11 (2008), 1793–1800.

[51] Gaut, G., Steyvers, M., Imel, Z. E., Atkins, D. C., and Smyth, P. Content coding of psychotherapy transcripts using labeled topic models. *IEEE journal of biomedical and health informatics 21*, 2 (2017), 476–487.

[52] General, U. S. Surgeon generals vision for a healthy and fit nation. *Washington, DC: HHS* (2010).

[53] Glynn, L., Houck, J., Moyers, T., Bryan, A., and Montanaro, E. Are change talk and sustain talk contagious in groups? sequential probabilities and safer-sex outcomes in alcohol-and marijuana-using adolescents. *Alcoholism: Clinical & Experimental Research 38* (2014), 328A.

[54] Grahne, G., and Zhu, J. Fast algorithms for frequent itemset mining using fp-trees. *IEEE transactions on knowledge and data engineering 17*, 10 (2005), 1347–1362.

[55] Graves, A., Mohamed, A.-R., and Hinton, G. Speech recognition with deep recurrent neural networks. In *Proceedings of the 2013 IEEE International Conference on Acoustics, Speech and Signal Processing* (2013), IEEE, pp. 6645–6649.

[56] Griffis, D., Shivade, C., Fosler-Lussier, E., and Lai, A. M. A quantitative and qualitative evaluation of sentence boundary detection for the clinical domain. *AMIA Summits on Translational Science Proceedings 2016* (2016), 88.

[57] Griffiths, T. L., and Steyvers, M. Finding scientific topics. *Proceedings of the National academy of Sciences 101*, suppl 1 (2004), 5228–5235.

[58] Günter, S., and Bunke, H. New boosting algorithms for classification problems with large number of classes applied to a handwritten word recognition task. In *International Workshop on Multiple Classifier Systems* (2003), Springer, pp. 326–335.

[59] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. The weka data mining software: an update. *ACM SIGKDD explorations newsletter 11*, 1 (2009), 10–18.

[60] Han, J., Pei, J., and Yin, Y. Mining frequent patterns without candidate generation. In *ACM sigmod record* (2000), vol. 29, ACM, pp. 1–12.

[61] Hasan, M., Carcone, A. I., Naar, S., Eggly, S., Alexander, G. L., Hartlieb, K. E. B., and Kotov, A. Identifying effective motivational interviewing communication sequences using automated pattern analysis. *Journal of Healthcare Informatics Research*, 1–21.

[62] Hasan, M., Kotov, A., Carcone, A. I., Dong, M., Naar, S., and Hartlieb, K. B. A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories. *Journal of biomedical informatics 62* (2016), 21–31.

[63] Hasan, M., Kotov, A., Carcone, A. I., Dong, M., and Naar-King, S. Predicting the outcome of patient-provider communication sequences using recurrent neural networks and probabilistic models. In *Proceedings of the 2018 AMIA Informatics Summit* (2018), American Medical Informatics Association.

[64] Hirohata, K., Okazaki, N., Ananiadou, S., and Ishizuka, M. Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-I* (2008).

[65] Hochreiter, S., and Schmidhuber, J. Long short-term memory. *Neural computation 9*, 8 (1997), 1735–1780.

[66] Imel, Z. E., Steyvers, M., and Atkins, D. C. Computational psychotherapy research: Scaling up the evaluation of patient–provider interactions. *Psychotherapy 52*, 1 (2015), 19.

[67] Jacques-Tiura, A. J., Carcone, A. I., Naar, S., Brogan Hartlieb, K., Albrecht, T. L., and Barton, E. Building motivation in african american caregivers of adolescents with obesity: application of sequential analysis. *Journal of pediatric psychology 42*, 2 (2016), 131–141.

[68] John, G. H., and Langley, P. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence* (1995), Morgan Kaufmann Publishers Inc., pp. 338–345.

[69] Keogh, E. J., and Pazzani, M. J. Scaling up dynamic time warping for datamining applications. In *Proceedings of the 6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2000), ACM, pp. 285–289.

[70] Kibriya, A. M., Frank, E., Pfahringer, B., and Holmes, G. Multinomial naive bayes for text categorization revisited. In *Australasian Joint Conference on Artificial Intelligence* (2004), Springer, pp. 488–499.

[71] Kim, Y. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* (2014).

[72] Kingma, D. P., and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[73] Kohavi, R., et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai* (1995), vol. 14, Montreal, Canada, pp. 1137–1145.

[74] Kotov, A., Hasan, M., Carcone, A., Dong, M., Naar-King, S., and BroganHartlieb, K. Interpretable probabilistic latent variable models for automatic annotation of clinical text. In *Proceedings of the 2015 Annual Symposium of the American Medical Informatics Association* (2015), pp. 785–794.

[75] Kotov, A., Idalski Carcone, A., Dong, M., et al. Towards automatic coding of interview transcripts for public health research. In *Proceedings of the Big Data Analytic Technology For Bioinformatics and Health Informatics Workshop (KDD-BHI) in conjunction with ACM SIGKDD Conference on Knowledge Discovery and Data Mining. New York, NY* (2014).

[76] Kreuzthaler, M., and Schulz, S. Detection of sentence boundaries and abbreviations in clinical narratives. In *BMC medical informatics and decision making* (2015), vol. 15, BioMed Central, p. S4.

[77] Lacoste-Julien, S., Sha, F., and Jordan, M. I. Disclda: Discriminative learning for dimensionality reduction and classification. In *Advances in neural information processing systems* (2009), pp. 897–904.

[78] Lacson, R., and Barzilay, R. Automatic processing of spoken dialogue in the home hemodialysis domain. In *AMIA Annual Symposium Proceedings* (2005), vol. 2005, American Medical Informatics Association, p. 420.

[79] Lafferty, J. D., McCallum, A., and Pereira, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning* (2001), pp. 282–289.

[80] Laws, M. B., Beach, M. C., Lee, Y., Rogers, W. H., Saha, S., Korthuis, P. T., Sharp, V., and Wilson, I. B. Provider-patient adherence dialogue in hiv care: results of a multisite study. *AIDS and Behavior 17*, 1 (2013), 148–159.

[81] Laws, M. B., Taubin, T., Bezreh, T., Lee, Y., Beach, M. C., and Wilson, I. B. Problems and processes in medical encounters: the cases method of dialogue analysis. *Patient education and counseling 91*, 2 (2013), 192–199.

[82] Leach, A. R., and Gillet, V. J. *An introduction to chemoinformatics*. Springer Science & Business Media, 2007.

[83] Leslie, C., and Kuang, R. Fast string kernels using inexact matching for protein sequences. *Journal of Machine Learning Research 5*, Nov (2004), 1435–1455.

[84] Lipton, Z. C., Berkowitz, J., and Elkan, C. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019* (2015).

[85] Lipton, Z. C., Kale, D. C., Elkan, C., and Wetzell, R. Learning to diagnose with lstm recurrent neural networks. *arXiv preprint arXiv:1511.03677* (2015).

[86] Liu, B., Hsu, W., and Ma, Y. Mining association rules with multiple minimum supports. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (1999), ACM, pp. 337–341.

[87] Liu, Y., Stolcke, A., Shriberg, E., and Harper, M. Using conditional random fields for sentence boundary detection in speech. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL05)* (2005), pp. 451–458.

[88] Maaten, L. v. d., and Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research 9*, Nov (2008), 2579–2605.

[89] Magill, M., Gaume, J., Apodaca, T. R., Walthers, J., Mastroleo, N. R., Borsari, B., and Longabaugh, R. The technical hypothesis of motivational interviewing: A meta-analysis of mi's key causal model. *Journal of consulting and clinical psychology 82*, 6 (2014), 973.

[90] Martin, T., Moyers, T. B., Houck, J., Christopher, P., and Miller, W. R. Motivational interviewing sequential code for observing process exchanges (mi-scope) coder's manual. *Retrieved March 16* (2005), 2009.

[91] Mayfield, E., Laws, M. B., Wilson, I. B., and Penstein Rosé, C. Automating annotation of information-giving for analysis of clinical conversation. *Journal of the American Medical Informatics Association 21*, e1 (2013), e122–e128.

[92] McCallum, A., Nigam, K., et al. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization* (1998), vol. 752, Citeseer, pp. 41–48.

[93] McCallum, A. K. Mallet: A machine learning for language toolkit.

[94] McCambridge, J., Day, M., Thomas, B. A., and Strang, J. Fidelity to motivational interviewing and subsequent cannabis cessation among adolescents. *Addictive behaviors 36*, 7 (2011), 749–754.

[95] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (2013), pp. 3111–3119.

[96] Miller, W. R., and Rollnick, S. *Motivational interviewing: Helping people change*. Guilford press, 2012.

[97] Miller, W. R., and Rollnick, S. Motivational interviewing: Helping people change. 3. *New York: Guilford publications* (2013).

[98] Miller, W. R., and Rose, G. S. Toward a theory of motivational interviewing. *American psychologist 64*, 6 (2009), 527.

[99] Moradi, M., and Ghadiri, N. Quantifying the informativeness for biomedical literature summarization: An itemset mining method. *Computer methods and programs in biomedicine 146* (2017), 77–89.

[100] Moyers, T. B., and Martin, T. Therapist influence on client language during motivational interviewing sessions. *Journal of substance abuse treatment 30*, 3 (2006), 245–251.

[101] Moyers, T. B., Martin, T., Houck, J. M., Christopher, P. J., and Tonigan, J. S. From in-session behaviors to drinking outcomes: a causal chain for motivational interviewing. *Journal of consulting and clinical psychology 77*, 6 (2009), 1113.

[102] Moyers, T. B., Martin, T., Manuel, J. K., Hendrickson, S. M., and Miller, W. R. Assessing competence in the use of motivational interviewing. *Journal of substance abuse treatment 28*, 1 (2005), 19–26.

[103] Mutsam, N., and Pernkopf, F. Maximum margin hidden markov models for sequence classification. *Pattern Recognition Letters 77* (2016), 14–20.

[104] Nguyen, A., Moore, D., McCowan, I., and Courage, M.-J. Multi-class classification of cancer stages from free-text histology reports using support vector machines. In *Engineering in Medicine and Biology Society, 2007.*

EMBS 2007. 29th Annual International Conference of the IEEE (2007), IEEE, pp. 5140–5143.

[105] Nguyen, H. M., Cooper, E. W., and Kamei, K. Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms 3*, 1 (2011), 4–21.

[106] Nion, T., Menasri, F., Louradour, J., Sibade, C., Retornaz, T., Métaireau, P.-Y., and Kermorvant, C. Handwritten information extraction from historical census documents. In *Proceedings of the 12th International Conference on Document Analysis and Recognition* (2013), pp. 822–826.

[107] Ogden, C. L., Carroll, M. D., Kit, B. K., and Flegal, K. M. Prevalence of obesity and trends in body mass index among us children and adolescents, 1999-2010. *Jama 307*, 5 (2012), 483–490.

[108] Olukunle, A., and Ehikioya, S. A fast algorithm for mining association rules in medical image data. In *Electrical and Computer Engineering, 2002. IEEE CCECE 2002. Canadian Conference on* (2002), vol. 2, IEEE, pp. 1181–1187.

[109] Pang, B., Lee, L., et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval 2*, 1–2 (2008), 1–135.

[110] Pasquier, N., Bastide, Y., Taouil, R., and Lakhal, L. Discovering frequent closed itemsets for association rules. In *International Conference on Database Theory* (1999), Springer, pp. 398–416.

[111] Pennington, J., Socher, R., and Manning, C. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (2014), pp. 1532–1543.

[112] Pérez-Rosas, V., Mihalcea, R., Resnicow, K., Singh, S., and An, L. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2017), vol. 1, pp. 1426–1435.

[113] Pérez-Rosas, V., Mihalcea, R., Resnicow, K., Singh, S., Ann, L., Goggin, K. J., and Catley, D. Predicting counselor behaviors in motivational interviewing encounters. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (2017), vol. 1, pp. 1128–1137.

[114] Rabiner, L. R. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*, 2 (1989), 257–286.

[115] Rish, I., et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (2001), vol. 3, IBM New York, pp. 41–46.

[116] Roberts, K., and Harabagiu, S. M. A flexible framework for deriving assertions from electronic medical records. *Journal of the American Medical Informatics Association 18*, 5 (2011), 568–573.

[117] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *nature 323*, 6088 (1986), 533.

[118] Schuster, M., and Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing 45*, 11 (1997), 2673–2681.

[119] Sharma, A. K., and Sahni, S. A comparative study of classification algorithms for spam email data analysis. *International Journal on Computer Science and Engineering 3*, 5 (2011), 1890–1895.

[120] Srivastava, J., Cooley, R., Deshpande, M., and Tan, P.-N. Web usage mining: Discovery and applications of usage patterns from web data. *Acm Sigkdd Explorations Newsletter 1*, 2 (2000), 12–23.

[121] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research 15*, 1 (2014), 1929–1958.

[122] St, K. N. t. A. M. Learning to classify text from labeled and unlabeled documents.

[123] Sutton, C., and McCallum, A. *An introduction to conditional random fields for relational learning*, vol. 2. Introduction to statistical relational learning. MIT Press, 2006.

[124] Tanana, M., Hallgren, K., Imel, Z., Atkins, D., Smyth, P., and Srikumar, V. Recursive neural networks for coding therapist and patient behavior in motivational interviewing. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality* (2015), pp. 71–79.

[125] Tanana, M., Hallgren, K. A., Imel, Z. E., Atkins, D. C., and Srikumar, V. A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of substance abuse treatment 65* (2016), 43–50.

[126] Tausczik, Y. R., and Pennebaker, J. W. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology 29*, 1 (2010), 24–54.

[127] Teal, C. R., and Street, R. L. Critical elements of culturally competent communication in the medical encounter: a review and model. *Social science & medicine 68*, 3 (2009), 533–543.

[128] Tepper, M., Capurro, D., Xia, F., Vanderwende, L., and Yetisgen-Yildiz, M. Statistical section segmentation in free-text clinical records. In *LREC* (2012), pp. 2001–2008.

[129] Thrasher, A. D., Golin, C. E., Earp, J. A. L., Tien, H., Porter, C., and Howie, L. Motivational interviewing to support antiretroviral therapy adherence: The role of quality counseling. *Patient Education and Counseling 62*, 1 (2006), 64–71.

[130] Treviso, M., Shulby, C., and Aluísio, S. Sentence segmentation in narrative transcripts from neuropsychological tests using recurrent convolutional neural networks. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers* (2017), vol. 1, pp. 315–325.

[131] Uğuz, H., Arslan, A., and Türkoğlu, İ. A biomedical system based on hidden markov model for diagnosis of the heart valve diseases. *Pattern Recognition Letters 28*, 4 (2007), 395–404.

[132] Walker, D., Stephens, R., Rowland, J., and Roffman, R. The influence of client behavior during motivational interviewing on marijuana treatment outcome. *Addictive Behaviors 36*, 6 (2011), 669–673.

[133] Webber, B., Egg, M., and Kordoni, V. Discourse structure and language technology. *Natural Language Engineering 18*, 4 (2012), 437–490.

[134] Wei, L., and Keogh, E. Semi-supervised time series classification. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2006), ACM, pp. 748–753.

[135] Won, K.-J., Prügel-Bennett, A., and Krogh, A. Training hmm structure with genetic algorithm for biological sequence analysis. *Bioinformatics 20*, 18 (2004), 3613–3619.

[136] Wright, A. P., Wright, A. T., McCoy, A. B., and Sittig, D. F. The use of sequential pattern mining to predict next prescribed medications. *Journal of biomedical informatics 53* (2015), 73–80.

[137] Yakhnenko, O., Silvescu, A., and Honavar, V. Discriminatively trained markov model for sequence classification. In *Proceedings of the 5th IEEE International Conference on Data Mining* (2005), IEEE.

[138] Zhang, S., and Wang, J. T. Discovering frequent agreement subtrees from phylogenetic data. *IEEE Transactions on Knowledge and Data Engineering 20*, 1 (2008), 68–82.

[139] Zhang, Y., and Wallace, B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820* (2015).

ABSTRACT

MACHINE LEARNING METHODS FOR THE ANALYSIS OF CLINICAL CONVERSATION

by

MD MEHEDI HASAN

August 2019

Advisor: Dr. Alexander Kotov

Major: Computer Science

Degree: Doctor of Philosophy

Motivational Interviewing (MI) is an evidence-based communication technique to increase intrinsic motivation and self-efficacy for behavior change. This goal is achieved through the exploration of the patient's own desires, ability, reasons, need for and commitment to the targeted behavior change. However, communication science approaches to understanding the efficacy of MI are inherently limited by traditional qualitative coding methods which is a time-consuming and resource-intensive process. Thus, an efficient method is required to automate the coding process which will accelerate the pace of communication research in behavioral science. The specific provider behaviors responsible for the elicitation of change talk, are also less clear and may vary by treatment context. Therefore, new design objective and perspective are necessary to understand which provider behaviors and in which contexts lead to patient change talk. In this dissertation, we deal with two types of clinical conversation, one that involves a face to face dialogue between patient and counselor and another one which involves an email-based conversation between patient and an ecoach.

First, we leverage eight supervised machine learning models to automatically annotate counseling sessions with 37 African American adolescents with obesity and their caregivers. We examine the performance of classifiers using lexical, contextual, and semantic features, to predict the behavioral codes in the previously coded data.

Second, understanding motivational interviewing mechanisms of effect, we focus on deep learning and probabilistic models and analyze the sequencing of patient-provider communication. The goal of these experiments is to identify the communication behaviors leading to the elicitation of client change talk, a marker of success in MI, and counter change talk, a marker of unsuccessful communication. Two approaches, recurrent neural networks and Markov models, were tested. As a continuation of our sequential analysis, we analyze pre-coded MI transcripts to identify the specific counselor communication behaviors effective for eliciting patient change talk. We evaluate the empirical effectiveness of the hidden Markov model and closed frequent pattern mining to inform MI practice.

Finally, we propose various segmentation models for the analysis of email-based counseling sessions since segmentation is a necessary and critical step to process email-based conversation for developing autocoding and sequence analysis models. We formulate the segmentation task as a classification problem and utilizes word and punctuation mark embeddings in conjunction with part-of-speech features to address it. We evaluate the performance of conditional random fields as well as a multi-layer perceptron, bi-directional recurrent neural network and convolutional recurrent neural network for the task of clinical text segmentation.

Experimental results indicate that machine learning models achieve performance near human coders for the segmentation and annotation of clinical conversation, which will significantly increase the pace of communication research in behavioral science. Our methods can facilitate motivational interviewing researchers to identify the most likely sequences in successful and unsuccessful motivational interviews, which can directly inform clinical practice and increase the effectiveness of behavioral interventions. We can integrate the sequential model with segmentation and auto-coding classifiers to develop a fully automated system for the analysis of clinical conversation.

AUTOBIOGRAPHICAL STATEMENT

MD MEHEDI HASAN

EDUCATION

- Master of Science (Computer Science), 2019
  Wayne State University, Detroit, MI, USA
- Bachelor of Science (Computer Science and Engineering), 2009
  Bangladesh University of Engineering and Technology, Dhaka

PUBLICATIONS

1. **Hasan, M.**, Kotov, A., Naar, S., Alexander, G.L. and Carcone, A.I. "Deep neural architectures for discourse segmentation in email-based behavioral intervention". In Proceedings of the AMIA Informatics Summit (2019). American Medical Informatics Association, pp. 443–452.

2. Carcone, A.I., **Hasan, M.**, Alexander, G.L., Dong, M., Eggly, S., Brogan Hartlieb, K., Naar, S., MacDonell, K. and Kotov, A. "Developing machine learning models for behavioral coding". Journal of pediatric psychology (2019), 44(3), pp. 289–299.

3. **Hasan, M.**, Carcone, A.I., Naar, S., Eggly, S., Alexander, G., BroganHartlieb, K. and Kotov, A. "Identifying Effective Motivational Interviewing Communication Sequences Using Automated Pattern Analysis". Journal of Healthcare Informatics Research (JHIR) (2018), 3(1), pp. 86–106.

4. **Hasan, M.**, Kotov, A., Carcone, A.I., Dong, M. and Naar, S. "Predicting the Outcome of Patient-Provider Communication Sequences using Recurrent Neural Networks and Probabilistic Models". In Proceedings of the AMIA Informatics Summit (2018). American Medical Informatics Association, pp. 64–73.

5. **Hasan, M.**, Kotov, A., Carcone, A.I., Dong, M., Naar, S. and Hartlieb, K.B. "A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories". Journal of biomedical informatics (2016), 62, pp. 21–31.

6. **Hasan, M.**, Kotov, A., Mohan, A., Lu, S. and Stieg, P.M. "Feedback or Research: Separating Pre-purchase from Post-purchase Consumer Reviews". In European Conference on Information Retrieval (2016). Springer International Publishing, pp. 682–688.

7. Kotov, A., **Hasan, M.**, Carcone, A.I., Dong, M., Naar-King, S. and Brogan-Hartlieb, K. "Interpretable probabilistic latent variable models for automatic annotation of clinical text". In AMIA Annual Symposium Proceedings (2015). American Medical Informatics Association, pp. 785–794.