UNIVERSITY OF CALIFORNIA,
IRVINE

Sequential and Temporal Analysis of Human-generated Data

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Computer Science

by

Jihyun Park

Dissertation Committee:
Chancellor's Professor Padhraic Smyth, Chair
Assistant Professor Sameer Singh
Professor Mark Warschauer

2019

DEDICATION

To my family.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and foremost, I would like to thank my advisor Prof. Padhraic Smyth for providing me with trust and guidance, and supporting me and my work in many ways throughout my whole PhD years. He has helped me build confidence, have initiative, and showed me the way to think critically. I deeply appreciate his time meeting with students every week despite his busy schedule, and the effort to encourage and compliment each and every student. It meant a lot to me and has really made me get back up again when I was struggling. I have learned a lot from him; not only the insight and knowledge about research in machine learning, but how to be well organized, how to be patient, how to talk and treat people. I am grateful to be a student of his and deeply thank him for always being a sincere advisor and teacher.

I would also like to thank the rest of my committee members for spending time on my dissertation and providing helpful feedback. The education projects, which take up about half of my thesis, were only possible because of Prof. Mark Warschauer. Being part of the IVLE project group was one of the greatest experiences during my PhD. It also made me have more interest in education and socially-relevant problems. Prof. Sameer Singh gave helpful advice on the second part of my thesis. I thank him for coming up with the idea of our pretty color sequence plots that are highly complimented by many people. I also appreciate him for always being friendly and willing to help.

I also thank all my teachers who inspired and guided me prior to UCI: Prof. Jungtae Kim, Prof. Deog-Kyoon Jeong, Prof. Hyunggon Park, Dr. Homayoon Beigi, Dr. Dan Ellis, and Prof. Truong-Thao Nguyen. Some of them may not remember me very well, but their lectures, guidance, advice and assistance have led me to the place where I am right now. I am glad that I had an opportunity to meet and learn from them.

My PhD was all about collaborating with people in other fields, and I was very fortunate to meet such a nice group of people for all of my projects. All the IVLE or the school of education people, including Prof. Mark Warschauer, Prof. Fernando Rodriguez, Renzhe Yu, Prof. Rachel Baker; I thank them for always being friendly and helpful throughout the collaboration. I believe the team has taken a huge step forward, and I am grateful that I have contributed at least some part of it. PCORI folks, including Prof. Zac Imel, Prof. Ming Tai-Seale, Abhishek Jindal, Patty Kuo, and Dr. Mike Tanana; all the work I have done with the group, I truly enjoyed it, and I really appreciate their help, feedback, and discussions. Special thanks to Robert Logan IV and Dimitrios Kotzias for taking over the PCORI project during the summer and helping me with many related things. Finally, it has been a real pleasure meeting and working with all the SAP team: Prof. Kai Zheng, Brian Tran, Dr. Tobias Schimmer, Dr. Hans-Martin Will, and Lucy Li.

During my time at UC Irvine, I was extremely fortunate to make lasting friendships with amazing people. First of all, Ellie Lee, who is the most energetic and funniest person that I know of, has given me so much power to survive here in graduate school since 2014. Ellie and her family provided help, care, and support whenever I needed, and made me feel like I have a family here in Irvine. I will never forget their generosity and thoughtfulness. Dohyun

Lee would make crazy jokes which made me laugh, but at the same time he gave me sincere advice as a person who had gone through the process. Hyungik Oh has been a colleague that I needed so much who I can share the struggles of PhD life. I am going to miss our fun chats and hope we can all meet at Java City again. I also thank Minhaeng Lee for answering all my random questions and helping me with so many different things, and Eun Jeong Shin for her caring and support.

And to our #4202 crew; it was very cool to have such awesome labmates who I can hang out outside of the office. Dimitrios Kotzias always made me feel better and gave me confidence whenever I was struggling or feeling sad. Disi Ji, a.k.a. DiSenior Disu, has been someone who I can always talk to, about literally anything, and brought me out for so many good things: food, boba, yoga, and movies, which presented me with lots of refreshing moments. Irvine could have been a very boring place without the weekend brunch, movie, and baking sessions with them. Also, Robby Logan and Casey Graff were always eager to help and would give me good feedback and comments on my work. I appreciate their time and effort to do that, and also thank for filling #4202 with bright energy. All the other past and current #UCIDataLab and #4202 members: Moshe Lichman, Eric Nalisnick, Zach Butler, Chris Galbraith, Lars Hertel, Homer Strong, Efi Karra Taniskidou, Abhishek Jindal, and Preston Putzel. I am lucky that I had a smart and friendly group of people around me throughout my PhD. I will always remember our board full of hashtags and memories. #ZotZotZot

I thank all other people who have helped and supported me during my PhD from Irvine and all over the world: Sun Young Park, JungYeon Ku, Hyesin Cho, Soohee Park, Sooyeon Kim, and Jiayang Chen.

Lastly, I would like to thank my parents Jong Jin Park and Sungjae Kim, my sister Soohyun Park, and my husband Daniel Wonjoon Song for being my mentors, advisors, therapists, and best friends for the entire time. Getting a PhD would not have been something that I could imagine without them in my life. I thank for their endless love and trust, and for always being there for me. I also thank my other family members: my in-laws including my parents-in-law, Joo Seok Song and Yoonjung Choi, my grandparents, great uncle and aunt, aunts and uncles, cousins, and even my nephews and nieces. I especially thank my cousin Bitnarae Kim for listening to me when I was going through hard times and her daughter Yeseo Woo for always making me smile. I am too fortunate to have such a huge but close family with full of love and support. I dedicate my dissertation to all of them.

This dissertation was supported by National Science Foundation under Grant Number 1535300, Patient-Centered Outcomes Research Institute® (PCORI®) Award (ME-1602-34167), and by funding support from SAP.

CURRICULUM VITAE

Jihyun Park

EDUCATION

Doctor of Philosophy in Computer Science **2019**
University of California, Irvine *Irvine, California*

Master of Science in Electrical Engineering **2013**
Columbia University *New York, New York*

Bachelor of Science in Electronic Engineering **2011**
Ewha Womans University *Seoul, South Korea*

RESEARCH EXPERIENCE

Graduate Research Assistant **2014–2019**
DataLab *Irvine, California*
University of California, Irvine

REFEREED JOURNAL PUBLICATIONS

Jihyun Park, Dimitrios Kotzias, Patty Kuo, Robert L Logan IV, Kritzia Merced, Sameer Singh, Michael Tanana, Efi Karra Taniskidou, Jennifer Elston Lafata, David C Atkins, Ming Tai-Seale, Zac E Imel, and Padhraic Smyth. **Detecting Conversation Topics in Primary Care Office Visits from Transcripts of Patient-Provider Interactions**. *Journal of the American Medical Informatics Association (JAMIA)*, 2019.

REFEREED CONFERENCE PUBLICATIONS

Jihyun Park, Kameryn Denaro, Fernando Rodriguez, Padhraic Smyth, and Mark Warschauer. **Detecting Changes in Student Behavior from Clickstream Data**. *In Proceedings of the Seventh International Learning Analytics and Knowledge Conference (LAK17)*, Mar 2017. – *Best Paper: Honorable Mention*

Jihyun Park, Renzhe Yu, Fernando Rodriguez, Rachel Baker, Padhraic Smyth, and Mark Warschauer. **Understanding Student Procrastination via Mixture Models**. *In Proceedings of the Eleventh International Conference on Educational Data Mining (EDM 2018)*, July 2018. – *Best Paper Award*

Fernando Rodriguez, Renzhe Yu, **Jihyun Park**, Mariela Janet Rivas, Mark Warschauer, and Brian K. Sato. **Utilizing Learning Analytics to Map Students' Self-Reported Study Strategies to Click Behaviors in STEM Courses**. *In Proceedings of the Nineth International Learning Analytics and Knowledge Conference (LAK19)*, Mar 2019.

TEACHING EXPERIENCE

Teaching Assistant - Project in Artificial Intelligence **Winter 2016**
University of California, Irvine *Irvine, California*

Teaching Assistant - Programming in C/C++ **Fall 2015**
University of California, Irvine *Irvine, California*

Grader - Advanced Digital Signal Processing **Fall 2012**
Columbia University *New York, New York*

PROFESSIONAL EXPERIENCE

Data Science Intern **2019**
Obsidian Security, Inc. *Newport Beach, California*

Research Intern **2018**
Baidu Silicon Valley A.I. Lab *Sunnyvale, California*

Research Staff **2013 – 2014**
Recognition Technologies, Inc. *White Plains, New York*

ACADEMIC REVIEWING

Conference on Artificial Intelligence (AAAI) **2020**
Women in Machine Learning (WiML) Workshop **2017**

ABSTRACT OF THE DISSERTATION

Sequential and Temporal Analysis of Human-generated Data

By

Jihyun Park

Doctor of Philosophy in Computer Science

University of California, Irvine, 2019

Chancellor's Professor Padhraic Smyth, Chair

Machine learning and data mining have the potential to provide meaningful solutions to many real-world problems that could impact society. This dissertation explores machine learning methods that utilize the sequential or temporal aspect of data to solve socially-relevant problems in two domains: one in education, the other in medicine.

In the domain of education and learning analytics, I present two different methods that help better understand students' online learning behaviors based on their clickstream data. First, statistical change detection is used to detect when and how students change their behavior during a course relative to the student population as a whole. I group the students depending on the type of changes and examine how the changes can be related to the course outcomes. The second method I explore is the use of probabilistic clustering to allow different types of temporal behavioral patterns to emerge from clickstream data. The resulting patterns are analyzed in the context of procrastination and time-management, demonstrating that the procrastinating behavior and course outcomes are highly correlated.

Secondly, in the medical domain, different materials and methods are discussed to test the feasibility of using machine learning models to obtain structured information from patient-physician conversations in primary care visits, as a means to potentially address the physician burnout problem. Several hundred dialog transcripts of doctor-patient conversations are

used to predict topic labels for talk-turns. Different machine learning models are trained to operate on single or multiple talk-turns (logistic classifiers, support vector machines, gated recurrent units [GRUs]) as well as sequential models that integrate information across talk-turn sequences (conditional random fields, hidden Markov models, and hierarchical GRUs). From the results, I show that incorporating sequential information across talk-turns improves the accuracy of topic prediction in dialogs. Moreover, I examine the degree to which topic classification accuracy drops by adding an automatic speech recognition (ASR) system for transcription to the pipeline. A systematic evaluation is carried out by measuring the performance of ASR (with word error rate) and the downstream classification accuracy.

In the second part of my work in the medical domain, I investigate the use of machine learning methods for predicting emotional valence of both doctors and patients at the utterance level, based on transcripts of doctor-patient visits. This is one of the first large-scale investigations of emotional valence prediction at the utterance level in long medical dialogs. Using a variety of evaluation metrics, I show that current machine learning methods can achieve accuracies for this task that are close to that of human performance.

Chapter 1

Introduction

We live in a time where we have become highly dependent on science and technology. Machine learning and data mining, in particular, has been employed in many domains that are intimately related to our everyday lives. It is becoming more important to communicate with society and provide benefits to human beings, rather than developing technology just for its own sake [54, 101].

Historically, machine learning research has not placed much emphasis on applications that benefit society [130] with relatively little incentive to tackle social problems [1]. To raise awareness of the importance of socially-relevant applications, there has been attempts to introduce and involve such applications within the community [1, 9, 130]. The main areas of interest include education, health, environmental protection, cyber security, assistive technology, etc.

For the data that are generated by humans, it is crucial to think ahead of the right purpose of a study so that it can eventually be used for human well-being. Examples of human-generated data are human speech and text data, online activity data, sensor data from mobile and wearable devices, to name a few.

One aspect of the human-generated data is that the data often involve information over time. This is because time cannot be separated from human life, and the data are generated as we live through time. For the examples mentioned above, speech is an analog time-series data that is an inherently dynamic process [62], and text has to follow certain ordering to deliver contextual information [89]. Also, with the advent of the internet and associated technological advancements, activity data from the internet, internet-of-things, mobile and wearable devices are a common type of sequential data that are easily accessible today. This includes website clicking [109], post or comment threads in social network systems [174], GPS or location data [30], human sensing data such as heart-rate signals [32], etc. Therefore, utilizing sequential or temporal properties of human data could be helpful to capture more meaningful information.

In this dissertation, I touch two domains in applications that can directly influence society and improve human life, with the use of sequential or temporal characteristics in human-generated data. First, I analyze problems in the domain of education, to provide valuable insights about student activities in an online learning environment, which is a challenging task without the help of machine learning. The second domain I focus on is in medicine, to alleviate physician burnout in doctor visits by providing structured information of the conversation, and to assist the doctors in having a more empathetic communication with their patients. The methodologies used in the dissertation can be explained in many different ways, but here I introduce them by categorizing them into unsupervised and supervised approaches in a broader context.

1.1 Unsupervised Approaches

Unsupervised learning in machine learning refers to the learning methods using data without labels. It is usually used to find out hidden patterns that are innate to the data. Since

there is no true label that can be compared with the output to calculate error metrics, unsupervised learning is a less well-defined problem than supervised learning. However, because the problem definition is flexible, there are a potentially broad range of applications [113], and it has the merit of not involving an expensive labeling process. Also, regardless of the cost of labeling, labels are not always enough to explain what we are looking for, and important information can be obtained from the data itself. Especially for human data, one of the major interests is detecting, analyzing, and understanding patterns in human behavior, which is something that can be rather described given the data than defined a priori.

Since human behavior can often be represented as a composition of recurrent patterns over time, observing the temporal patterns of activities help understanding human behaviors [26]. By capturing the repeated patterns or routines and modeling them, various studies have found ways to detect certain behaviors [99] and discover causal relationships in routines [8].

Clustering analysis and mixture modeling techniques are canonical methods for unsupervised learning. In Chapter 3, I use a Poisson mixture model to find temporal behavioral patterns of students in university classes, that could inform the instructors and students whether the students are procrastinating or not. Clustering methods have also been applied to detect outliers in behaviors to find anomalies, since it is a common challenge to solve problems without labels in anomaly detection [61].

Another unsupervised method used in this dissertation is changepoint detection, which is a method that identifies the times when change occurs in time series data. Changepoint detection techniques can be used to detect cyber attacks [168] and to evaluate articulatory disorders [34]. One of the simple approaches is finding the changepoints offline after retrieving all the data, by fitting multiple piecewise regression models instead of a single model. Chapter 2 uses such a method for the student click activity data, to find whether students have increased or decreased their online activity during the course.

1.2 Supervised Approaches

When the output labels exist in the dataset, supervised learning approaches are used to learn a mapping from the input data to the output. The output label can have either continuous or categorical value. Supervised learning methods that are independent of time or do not have a sequential structure could work well as a baseline, even when ignoring such information in the data. However, using a model with a sequential structure can in principle add much more information, particularly when predicting labels in sequences (e.g., part of speech tagging and named entity recognition).

Here I briefly review some of the commonly used methods for sequential human-generated data that are also mentioned in the later chapters of the dissertation. Hidden Markov model (HMM) [11] and conditional random fields (CRF) [88] are the two well-known examples of structured learning methods for sequential data. HMMs are good for integrating predictions over time, with applications to problems such as automatic speech recognition [79], part of speech tagging [87], activity recognition from videos [147], etc.[1]

A recurrent neural network (RNN) [45] is a type of neural network model that has feedback connections between the nodes, allowing to have sequential inputs and learns the structured properties of them. Two examples of RNNs with the gated architectures are long short-term memory networks (LSTMs) [69] and gated recurrent units (GRUs) [31]. These are two of the most powerful and popular models, especially in natural language processing.

Dialog transcripts are sequential human-generated data, with sequential structure not only between the words, but also between the utterances or talk-turns. For a task that predicts a label for each utterance, the information from the previous or the next (for offline analysis) spoken utterances is very useful. In Chapters 4, 5, and 6, I apply GRUs with hierarchical structure to doctor-patient dialog data to predict topical and emotional content of the con-

[1]The details of the models are well explained in chapters 17.3 and 19.6 of Murphy's book [113].

versation. For both topic and emotion predictions (Chapters 4 and 6), I demonstrate that the models using sequential information perform better than independent models.

1.3 Notation

The following notation is used throughout the dissertation unless specified otherwise. Upper-case and bold-faced letters (e.g., $\boldsymbol{X}$) are used for matrices, lower-case and bold-faced letters (e.g., $\boldsymbol{x}$) for vectors, and lower-case letters without any bolding (e.g., x) are used to denote scalars. The upper-case and unbolded letter (X) is used to describe the size of a vector, or to represent a set or a collection.

For clickstream data in Chapters 2 and 3, the data is represented as a matrix, where each row ($\boldsymbol{x}_i$) is a data vector for each student i. The student index i ranges from 1 to N (the total number of students) and is often used as a subscript.

In Chapters from 4 to 6, subscript i, j indicates talk-turn or utterance j in visit i.

1.4 Outline of the Dissertation

This dissertation mainly consists of two socially-relevant applications using human data. The first two chapters (Chapter 2 and Chapter 3) study applications in education using students' clickstream data. Both chapters utilize unsupervised methods to better explain the behaviors of students at the individual level from the noisy click data.

In Chapter 2, statistical change detection technique is used to investigate students online behaviors. Using clickstream data from two large university courses, one face-to-face and one online, I illustrate how this methodology can be used to detect when students change their

previewing and reviewing behavior, and how these changes can be related to other aspects of students activity and performance.

Chapter 3 uses a probabilistic mixture model to unveil latent types of behavior from clickstream data and analyzes the resulting temporal patterns in the context of procrastination. Two sets of clickstream data from the same online course are used, where the course has a weekly repeating structure. The mixture model discovers two distinctive behaviors—procrastination and non-procrastination. Overall, the results show that the students identified as non-procrastinators tend to perform significantly better than procrastinators.

The other chapters (Chapters 4, 5, and 6) focus on developing methods to improve the quality of patient-physician interactions. Using patient-provider dialog transcripts, I first investigate the effectiveness of different supervised machine learning methods for automated annotation of medical topics in Chapter 4. In this chapter, I show the importance of employing sequential information in dialog texts for predicting labels at the talk-turn level.

As an extension to Chapter 4, Chapter 5 questions how much degradation in classification performance can be observed when automatic speech recognition (ASR) is involved, compared to the case of using the original human transcribed texts. Different ASR systems from major cloud software companies are compared by measuring the performance of ASR (by word error rate) and the final annotation (by the same metrics used in Chapter 4).

Chapter 6 also includes analyzing patient-physician dialog transcripts, but focuses on the emotional valence of speaker utterances. To the best of my knowledge, this is the first large-scale study of the effectiveness of machine learning for predicting emotional valence in primary-care patient doctor visits. For model evaluation, I analyze the overall emotional flow within a visit (or a session) of each speaker by calculating the Pearson correlation coefficient, and the R-precision scores for each positive, neutral, and negative class. Rather than focusing on classification accuracies, the evaluation using correlation and R-precision

metrics provide a better sense of model performance in the context of medical dialog. I also evaluate human inter-rater reliability, by evaluating a set of labels by one labeler and treating the others as the ground truth, and find that model-human and human-human performance is comparable on this task.

Chapter 2

Detecting Changes in Student Behavior from Clickstream Data

One of the major goals in educational data mining (EDM) is to use student clickstream data to describe and understand students' behavioral patterns. While past findings have advanced our understanding of what we can learn from clickstream data, a significant challenge involves devising statistical techniques that help us identify students who are changing behavior in the middle of a term. There are a number of reasons motivating this problem; one is to identify students who are in need of assistance during the course, another is to identify reasons that students are changing their behavior so that a course could be improved overall. The analysis of clickstream data within a course can also provide invaluable information to course instructors and to education researchers, and there is a need to be able to both summarize and visualize the results in a straightforward manner.

In this chapter, I will focus on clickstream data from two courses at a large university: one

The material in this chapter is based on the paper "Detecting Changes in Student Behavior from Clickstream Data" by Park, J., Denaro, K., Rodriguez, F., Smyth, P., and Warschauer, M., published in *the Proceedings of the Seventh International Learning Analytics and Knowledge Conference (LAK17)*, 2017

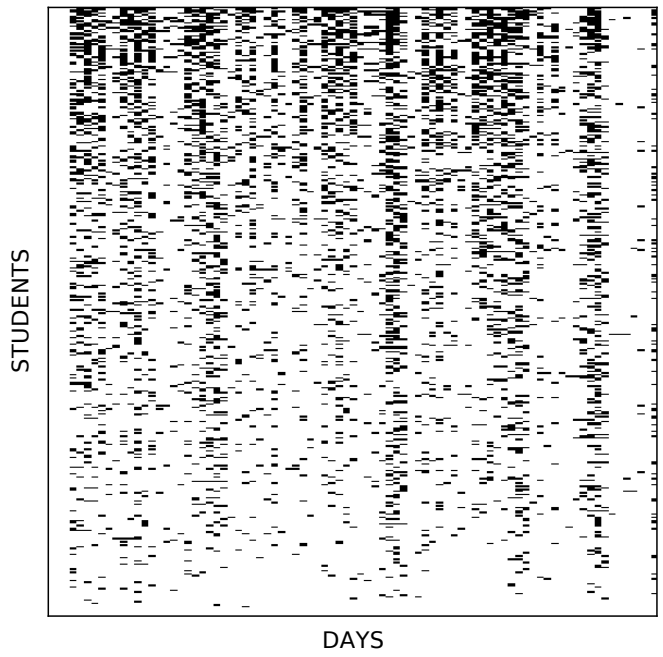


Figure 2.1: A plot of student clickstream activity in the 10-week face-to-face course over time, where each row represents an individual student, and each column represents a day. A black marker in cell i, t indicates clickstream activity for student i on day t.

face-to-face course and one online course, both from the 2015–2016 academic year. For each course, clickstream data is obtained through a course management system in the form of {student ID, time stamp, activity}. The types of activities recorded correspond to broad categories of student behavior, such as previewing lecture notes, submitting assignments, or posting and responding to discussion board questions. For instance, one of the courses examined in this chapter had 377 registered students who generated approximately 380,000 click events over a 10-week period. Figure 2.1 displays each of the individual student clickstreams over the 85 days of the course, with each row corresponding to a student. While the plot shows some general increases in click activities around quiz and exam dates, it is not easy to see much else, nor to understand how individual student behaviors are related to the overall population due to significant variability in students' click patterns. Furthermore, it is difficult to determine whether students change their click behaviors in any significant way, or whether or not these behaviors are correlated with course performance.

As discussed in more detail in the next section, student clickstream data has been the subject of a number of prior studies, such as the investigation of potential predictive relationships between online student activity and student outcomes (such as course grades). Here I focus instead on detecting changes in individual student activity over time, relative to the activity of the class of a whole. In particular, I investigate the use of statistical change detection techniques (e.g., Kirch and Tajduidje Kamgaing [84]) to automatically detect changes in activity over time for each student. I model the activity of each student relative to the aggregate activity of all students in the class and compare two models on a per-student basis; a model where there is no change in student activity versus a model where there is a significant change in activity at some unknown point during the period of the course. Likelihood-based techniques are used to fit both models on a per-student basis, and model selection criteria are implemented in order to determine whether each student is best modeled under the "change" or "no-change" model.

2.1 Related Work

Clickstream data analysis in an educational setting has focused on what the clickstream can say about the students in terms of learning behavior through a variety of features derived from the clickstream. Much of the prior work on clickstream data analysis for understanding student behavior has occurred in the context of Massive Open Online Courses (MOOC) setting. Many of these analyses have focused on using the clickstream data to predict MOOC completion (for example in Crossley et al. [37]) and to predict learning outcomes within a MOOC. For example, the relationship between the number of posts and the learning gains of the students has been investigated [170], as well as how discussion forum views are potentially related to learning outcomes [14]. There has also been research focused on improving predictions of learning outcomes by incorporating clickstream events as well as

summaries of the clickstream [20].

A secondary research topic has focused on describing students with similar clickstreams (e.g., Wang et al. [167]), the activities that the students are engaging in, and in understanding the student's typical online interaction within a class. As an example, clickstream data analysis was used to better understand whether or not students were following a defined learning path [39]. In other work, students' clickstreams were grouped into similar plans of action to better understand: learning pathways [116], how discussion forums and other activities in the MOOC were related to country and culture [96], and examined whether engagement on discussion forums increased based on the type of video a student watched [16]. All of these clickstream analyses have an underlying goal of describing student behaviors through the clickstream and to draw meaningful conclusions about those students.

MOOCs are typically used by people as a way to learn new skills or keep up to date with current ones. Because most MOOCs do not offer formal degrees, there are no serious consequences for doing poorly or dropping out. In contrast, college course grades determine whether students succeed or fail (whether they advance to the next course, remain in their intended major, or graduate). Thus, findings from MOOC clickstream studies cannot offer broad explanations about student learning experiences in higher education settings. So while MOOCs and college courses share some similarities, in terms of course management systems and clickstream data, investigating college courses may require a different set of goals and statistical techniques.

For instance, one important area of higher education research focuses on student engagement. Studies find that students who are not engaged with the learning process—that is, students who do not put in the time and energy into purposeful learning—are at greater risk for failing courses and dropping out of college [86]. While this finding is not new, understanding how to quickly identify these students, especially at the course-level, remains a significant challenge.

Clickstream data has the potential to address this since the data is obtained in real-time. Researchers can provide instructors with immediate insights on how students are engaging with the course management system. This is especially important in courses with large enrollments, where problems with student engagement can often go unnoticed [111]. Some recent work has found that student engagement with the course management system, as indexed by the number of days students visited the site relative to their peers, was positively related to course outcomes [91]. My work, as described in this chapter, adds to this area of research by using statistical change detection techniques to further understand course engagement.

More broadly, changepoint detection techniques for event time-series are a widely studied topic, and a variety of statistical methodologies have been developed (e.g., [44, 84]), with much of this work focused on single (univariate) time-series. Web user behavior has been analyzed to detect changes in an individual's behavior, to report "interesting" sessions, and to detect changes in user activity [71]. There has not been any prior work (to my knowledge) on change detection applied to multiple clickstreams of students in an educational setting.

Thus far, previous work in the analysis of clickstream data in an educational setting has focused on grouping students into similar groups, understanding possible dropouts, predicting student success in a course, and defining learning pathways. The goal of the work in this chapter is to add to the current body of research in a meaningful way by using changepoint detection techniques as a proxy for understanding student engagement. By detecting whether student behavior changes in a significant manner over the time period of a particular term, I will demonstrate how to identify students who increase, decrease, or show no change in their clickstream activities, and whether these changes relate to course performance.

2.2 Methods

I discuss below the approach for modeling and change detection of student activity. I begin by defining some general notation and then introduce two different models: a Bernoulli model for binary data and a Poisson model for count data. The section concludes with a description of changepoint detection for both of these models.

2.2.1 Notation

Let N be the number of individual students in a course and let i be an index that refers to an individual student in the class, $i = 1, \ldots, N$. I assume below that time is discrete[1] with T discrete time-points and $t = 1, \ldots, T$ being an index running from the first to the last time-period of clickstream logging for the course. Below I refer to t on a daily time-scale for convenience, but in general other time-periods–such as days or weeks–could be used.

Let $\boldsymbol{X}$ be the observed data for a course, represented as an $N \times T$ array whose entries are counts $x_{it} \in \{0, 1, 2, ...\}$. Note that x_{it} represents the number of click events for student i on day t, where $1 \leq i \leq N$ and $1 \leq t \leq T$. A binarized version of the data x'_{it} is also considered, where $x'_{it} = I(x_{it} > 0)$, and $I()$ is an indicator function (as in Figure 2.1 for example). The number of clicks x_{it} (counts) by student i on a given day t in principle contains more information than the binarized version x'_{it}, but could also be quite noisy in the sense that more clicks might not necessarily correlate well with relevant student activity. I explore both options since the choice of looking at a count versus the binarized version in practice will depend on the context of a particular analysis.

[1] A changepoint methodology using a continuous-time model could in principle also be developed in a manner similar to the discrete-time methodology described in this chapter.

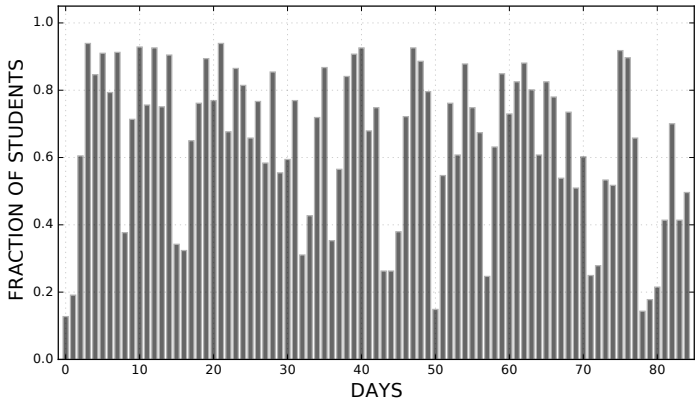


Figure 2.2: Proportion of students who click each day during a 10-week course.

2.2.2 Bernoulli Models for Binary Data

For the binary data, x'_{it}, let π_{it} be the probability that each student i is active on day t (i.e., the probability that student i generates one or more clicks on day t). The log-odds of π_{it} is modeled as:

$$\log \frac{\pi_{it}}{1-\pi_{it}} = \mu_t + \alpha_i, \tag{2.1}$$

where $\mu_t, t = 1, \ldots, T$ can be viewed as a time-varying population mean for the log-odds and $\alpha_i, 1 \leq i \leq N$ is a student-dependent offset to account for individual-level variation in student behavior.

The role of α_i in this model is to modulate the time-varying population mean μ_t in a student-specific manner. A positive value of α_i for student i will increase the log-odds above the population mean μ_t, which in turn means that student i tends to click more than the mean student as represented by μ_t. A negative value of α_i has the opposite effect; student i has a lower probability of clicking compared to the average student. μ_t represents time-varying population behavior on a log-odds scale.

The approach to change detection described in this chapter relies on modeling each student's activity relative to that of the overall student population in the class. The population

(or background) rate μ_t typically varies significantly as a function of time t since student behavior is strongly affected by temporal effects such as days of lectures, weekday versus weekend effects, assignment deadlines, exams, and etc. As an example, Figure 2.2 shows the proportion of students who clicked on a file each day, aggregating the data shown earlier in Figure 2.1.

Modeling the log-odds as a linear function is a standard technique in generalized linear modeling and ensures that the resulting probability π_{it} above lies between 0 and 1, i.e., Equation 2.1 above can be rewritten as the following:[2]

$$\pi_{it} = \frac{1}{1 + e^{-(\mu_t + \alpha_i)}}. \tag{2.2}$$

2.2.3 Estimation of Model Parameters

The parameters $\mu = \{\mu_1, \ldots, \mu_T\}$ and $\alpha = \{\alpha_1, \ldots, \alpha_N\}$ are estimated from the $N \times T$ data array $\boldsymbol{X}'$ with entries $x'_{it} \in \{0, 1\}, 1 \leq i \leq N, 1 \leq t \leq T$. Since the x'_{it}'s are binary the likelihood for each individual data point x'_{it} can be written as:

$$L(\mu, \sigma | x'_{it}) = \pi_{it}^{x'_{it}} (1 - \pi_{it})^{(1 - x'_{it})}, \tag{2.3}$$

where π_{it} is defined in Equation 2.2. The likelihood of the full dataset $\boldsymbol{X}'$ is then defined as:

$$\begin{aligned} L(\mu, \alpha | \boldsymbol{X}') &= P(\boldsymbol{X}' | \mu, \alpha) \\ &= \prod_{i=1}^{N} \prod_{t=1}^{T} \pi_{it}^{x'_{it}} (1 - \pi_{it})^{(1 - x'_{it})}. \end{aligned} \tag{2.4}$$

[2]From an ML perspective, we could also think of this as a logistic model, although the parameterization via μ_t and α_i is somewhat different to how logistic models are used in classification problem.

Here I make the assumption that the observed data for each student on each day is conditionally independent of all other observations (for students and for days) given the parameters μ and α. This is a simplification since it ignores (for example) possible time-varying trends in student behavior. Nonetheless, as seen later in the experimental results, it provides a useful basis for change detection.

A two-stage procedure is used for parameter estimation.[3] First, an estimate $\hat{\mu}_t$ for the population mean is generated as follows:

$$\hat{\mu}_t = \log \frac{\hat{q}_t}{1 - \hat{q}_t}, \quad 1 \leq t \leq T, \tag{2.5}$$

where $\hat{q}_t = \frac{1}{N} \sum_{i=1}^{N} x'_{it}$, which is the proportion of students (across all students) that generated a click on day t.

In the second step, a regression model is fitted for each student i in Equation 2.1 with the population mean $\hat{\mu}_t$ set as an offset, by maximizing the likelihood. α_i can be thought of as a student-specific intercept term for each student i.

2.2.4 Poisson Models for Count Data

The counts x_{it} can be modeled directly as well, where x_{it} can have values $\{0, 1, 2, ...\}$. A natural model in this context is the Poisson model.

I develop the count model in a manner similar to that of the binary case earlier. In particular, the logarithm of the mean of the Poisson distribution, $\log \lambda_{it}$, is modeled as a linear function of a time-varying population rate μ_t and an individual student effect α_i:

$$\log \lambda_{it} = \mu_t + \alpha_i. \tag{2.6}$$

[3]The estimation could be done in a single step; I would expect similar results to what are obtained in the two-step approach.

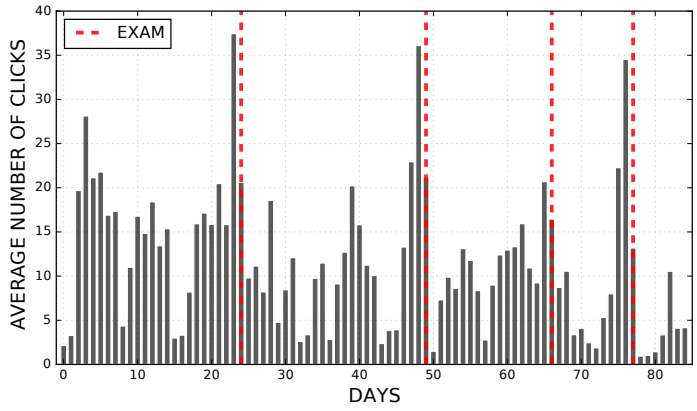


Figure 2.3: Average number of click events per student each day during a 10-week course.

Note that although for convenience the same notation is used for the two sets of parameters, μ and α, and they play an analogous role as their "namesake" parameters in the binary model, these parameters are different from those in the binary model described earlier.

Figure 2.3 shows the average number of click events for each student per day, reflecting the type of time-varying population behavior that μ_t is intended to capture. The red dashed lines are the dates for the three midterms and the final, and we can see much more click activity right before the exam dates.

The likelihood function for a single count x_{it} can be written as

$$P(x_{it}|\mu_t, \alpha_i) = \frac{\lambda_{it}^{x_{it}} e^{-\lambda_{it}}}{x_{it}!}, \tag{2.7}$$

where λ_{it} is defined in Equation 2.6. As with the binary case, assuming that the observations x_{it} are conditionally independent given the parameters, the full likelihood can be written as:

$$\begin{aligned} L(\mu, \alpha|\boldsymbol{X}) &= P(\boldsymbol{X}|\mu, \alpha) \\ &= \prod_{i=1}^{N} \prod_{t=1}^{T} \frac{\lambda_{it}^{x_{it}} e^{-\lambda_{it}}}{x_{it}!}. \end{aligned} \tag{2.8}$$

A conditional independence assumption is made here again for the Poisson model. A two-stage parameter estimation process is carried out as before. In the first step, $\hat{\mu}_t$ is estimated as follows:

$$\hat{\mu}_t = \log \hat{m}_t, \quad 1 \leq t \leq T, \tag{2.9}$$

where $\hat{m}_t = \frac{1}{N} \sum_{i=1}^{N} x_{it}$, representing the average number of click events across the population that were generated on day t. In the second step, I fit a Poisson regression model (by maximizing the likelihood) for each student i as in Equation 2.6 with an offset $\hat{\mu}_t$ to get an estimate for each α_i.

2.2.5 Detecting Changes in Activity

To detect changes in activity, the method allows for the possibility that each student's activity rate changes at some unknown time point during the course. For simplicity, the Bernoulli model is used to describe the method below, but note that the proposed approach works in the same manner for both the Bernoulli binary model and the Poisson count model. The only difference is in how the likelihood is defined and the parameters are estimated for each (as described earlier), and the issue is whether to fit a model with a change or with no change.

For each student i, two different models are considered. The first model is the one that assumes the student's rate of activity α_i, defined relative to the background activity μ_t, does not change over time. The second model, the changepoint model, assumes that a student's activity rate switches at some unknown changepoint. Both models are used to fit the data for each student and use a data-driven model selection technique to select which model is justified given the observed data.

In the changepoint model, an assumption is made that there is one activity rate α_{i1} for student i before changepoint τ_i and a different activity rate α_{i2} after the changepoint τ_i. The changepoint model for binary data (for example) can be written as follows, where $I()$ is an indicator function:

$$\log \frac{\pi_{it}}{1-\pi_{it}} = \mu_t + \alpha_{i1} I(t < \tau_i) + \alpha_{i2} I(t > \tau_i), \tag{2.10}$$

with a similar definition for the Poisson model. This model can be interpreted as fitting two regression models with different means on either side of the changepoint.

The value of the changepoint τ_i for each student is unknown. Since time t is discrete the values of τ_i can take one of $T-1$ possible values, corresponding to the $T-1$ boundaries between the T observation times.

In effect, this changepoint model has 3 parameters (assuming μ_t is known): the two activity rates and the changepoint. Maximum likelihood estimates of the parameters for the ith student are generated by maximizing the log-likelihood defined as follows:

$$\begin{aligned} & l_i(\alpha_{i1}, \alpha_{i2}, \tau_i, \mu) \\ & = \sum_{t<\tau_i} \log P(x'_{it}|\alpha_{i1}, \mu_t) + \sum_{t>\tau_i} \log P(x'_{it}|\alpha_{i2}, \mu_t). \end{aligned} \tag{2.11}$$

A similar equation is used for counts x_{it} and the Poisson model.

To fit this model, a similar two-stage approach is used for the model with no-change described earlier. In the first stage, I fit the background rate μ_t using the data across all students, in the same manner as for the no-change model. In the second stage, I find the values $\alpha_{i1}, \alpha_{i2}, \tau_i$, for each student i, that maximize the log-likelihood defined above. Since τ_i is discrete, the optimization problem can be reduced to finding the values of α_{i1} and α_{i2} for a fixed τ_i and then iterate over the $T-1$ possible values of τ_i. For each fixed value of τ_i,

the log-likelihood splits into the two parts on the right-hand side of Equation 2.11 above, a log-likelihood term containing α_{i1} and a second log-likelihood term containing α_{i2}. Each can be optimized independently using the same procedure described earlier for estimating α_i for the no-change model.

For each student i, once the parameters of both the no-change and the changepoint models have been estimated, the best model from the two candidate models is selected. The likelihood (or log-likelihood), evaluated at the maximum likelihood values of the parameters, is not useful for model selection since the changepoint model will always have a likelihood value that is at least as high as the no-change model (this is because the changepoint model contains the no-change model as a special case).

There are a variety of model selection techniques in the statistical literature to handle the issue of how to fairly compare models (in the case where models have different numbers of parameters), including techniques such as penalized likelihood, Bayesian criteria, and cross-validation [33]. For the results in this chapter, I use the Bayesian Information Criterion (BIC), which is a well-established and easily interpretable method for model selection. The BIC score is defined for each student as:

$$BIC_{iM} = -2l_{iM} + p_M \log T, \tag{2.12}$$

where M indicates a particular model ($M = 1$ corresponds to the no-change model, and $M = 2$ corresponds to the changepoint model), l_{iM} is the log-likelihood for model M for student i's data evaluated at the maximum likelihood values of the parameters, p_M is the number of parameters in each model ($p_1 = 1$, $p_2 = 3$, for the no-change and changepoint models respectively)[4], and T is the number of observations per student. The second term in Equation 2.12, $p_M \log T$, can be interpreted as a penalty for having additional parameters in

[4]Technically, the background model parameter μ should be counted as well, but since this is the same for both models we can omit it.

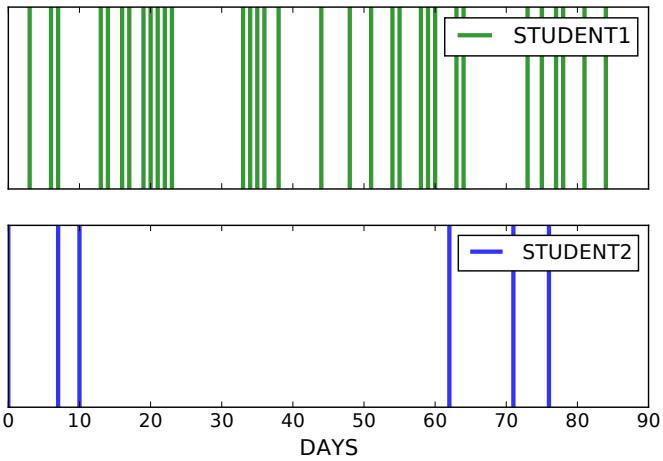


Figure 2.4: Simulated activity data for two students.

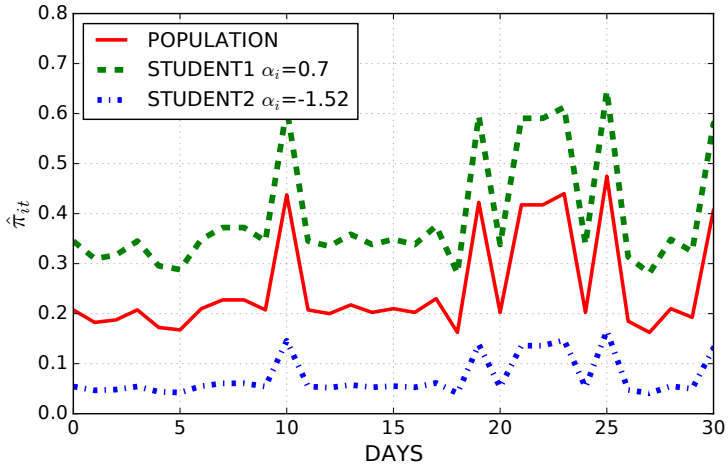


Figure 2.5: Estimated activity probabilities $\hat{\pi}_{it}$ for the two simulated students and for the population.

a model.

The BIC method selects the model with the lowest BIC score for each student. In particular, in the context of this changepoint application, BIC can be used to detect if there is evidence that a student's rate of activity changed, i.e., if $BIC_{i2} < BIC_{i1}$ then the evidence supports the changepoint model over the no-change model for student i.

2.3 Results for Simulated Data

To illustrate how the change-detection methods work, I simulated daily binary time series of student click activity for 400 students over 85 days (numbers that are roughly similar to the larger of the two classes that are analyzed later in the chapter). The true population rate μ_t switched back and forth between two different values over time, one with a high rate and one with a low rate. The variability in the simulation roughly corresponds to what was observed in the real student data. The offsets, α_{ij}, for each student were sampled independently from a normal distribution; $\alpha_{ij} \sim Normal(0, 1.5)$. Half of the students were simulated with one α_{i1}, i.e. no change in behavior over time. The other half of the students had two different offsets sampled, α_{i1} and α_{i2}, on either side of a changepoint τ_i which was sampled independently from a uniform distribution; $\tau_i \sim Uniform(15, 70)$.

Figure 2.4 is a plot of binary data for two simulated students who did not have changepoints. Student 1 is much more active than Student 2, and therefore Student 1 is going to have a larger estimated value for α_i. The estimated π_{it}'s for these students over the first 30 days (time $t = 1, ..., 30$) is shown in Figure 2.5. The plot illustrates how the estimated activity varies relative to the population probability $\hat{\mu}_t$ (the solid red curve). The more active student (green dashed line) has higher probabilities of clicking over time, while the less active student (blue dotted line) has lower probabilities, and both probabilities rise and fall relative to the behavior of the population. For example, when student activity on average rises on a particular day such as day 10 (e.g., due to an assignment), the click probability for both students rises.

Next, I show the results of two different simulated students, one with a changepoint and the other without a changepoint. Figure 2.6 (a) is a plot from a student with a changepoint, with the raw data in the lower plot and the fitted model (plotted on a log-odds scale) in the upper plot. There is a clear change in student behavior around day 59, and this is visible

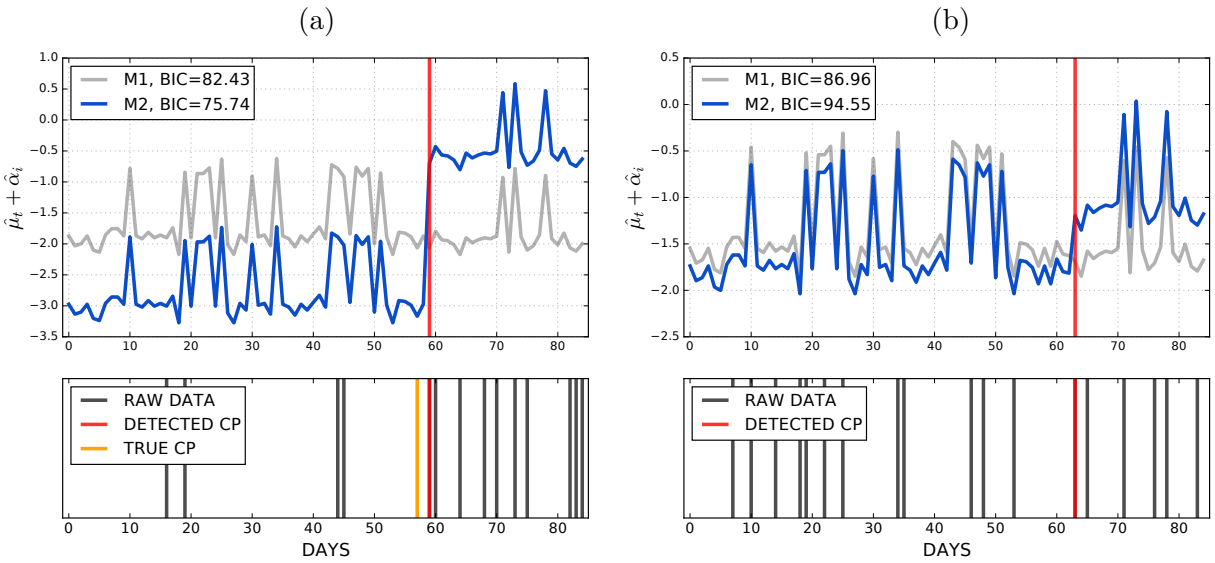


Figure 2.6: Log-odds of $\hat{\pi}_{it}$, for M1 and M2 (top), and simulated data of a student (bottom). The left plot (a) shows an example of a student with a changepoint at $t = 57$, and the right plot (b) shows an example of a student without a changepoint.

both in the raw data (lower plot) and the fitted changepoint model (top plot). The BIC for the model with the changepoint ($M = 2$) is significantly smaller than that of a model without the changepoint ($M = 1$) for this simulated student's data, i.e., the BIC method was able to successfully detect that a model with a changepoint is preferred over a model with no changepoint for this data. In contrast, Figure 2.6 (b) displays the results for a simulated student with no changepoint. Both plots show the results of fitting the changepoint model M2. The changepoint model puts a change at day 63, but the BIC method selects the no-changepoint model since the BIC for no change $M = 1$ is much smaller than that of the changepoint model $M = 2$.

The BIC method for binary simulated data reliably detected changepoints when the magnitude of the change in α_{i1} and α_{i2} (before and after the changepoint τ_i) was relatively large, but as the change became smaller it became more conservative. Out of the 400 simulated cases, BIC detected a change in 100 cases, with a precision of 91% (91 out of the 100 detected were true changes) and a recall of 46% (91 out of the 200 true changes were detected). The remaining 54% of true changes had much lower magnitude changes (0.96 on average)

compared to the detected cases (magnitude 2.61 change on average).

2.4 Clickstream Datasets

The clickstream data that used in this chapter was recorded via the Canvas learning management system (LMS). Canvas is an open-source LMS that serves as a supplemental instructional technology for students. It has been adopted as the campus-wide LMS system by a number of US universities, including UC Irvine. Students use Canvas to download course content, take online quizzes, watch videos, and submit assignments. The most common data available are clickstream data; every time a student clicks on a URL within the Canvas LMS, the click is recorded and logged with student ID, URL, and time-stamp.

Two types of student behavior are considered for the students described in this chapter. The first is a student's previewing behavior. An event is defined as a "Preview" event when a student views or downloads a file prior to the event start date. This could indicate how well a student is performing in terms of being prepared for the course. The second type of behavior is related to the students' reviewing activities. A "Review" event is defined as an event when a student views or downloads a file after the event end date, e.g., a student downloading a lecture file after the class in which the material was covered. Focusing on these two types of events can help to screen out less relevant information in the clickstream data and extract more meaningful information about students' activities. While in this chapter I focus on events related to my definitions of previewing and reviewing activity, the methodology for change detection is applicable to arbitrary sets of clickstream events.

Extracting each of the previewing and reviewing events results in an activity matrix of size $N \times T$, where the cell i, t indicates that the number of previewing or reviewing events by student i on day t. The data can be binarized to create a binary representation for the

Bernoulli model described in Subsection 2.2.2.

I used datasets from two courses at UC Irvine in the study; both offered during the 2015–2016 academic year. The first is a face-to-face 10-week course with 377 enrolled students. Lectures were held three times a week, and there were three midterms and one final exam. Figure 2.3 shows the average number of click events on each day per student. There is a significant variation in students' clicking activity over time. For example, students tended to be much more active during days close to the exams (shown as red dashed lines).

The second dataset is somewhat different from the first in that it was an online course offered for 5 weeks. There were 176 enrolled students in this course. This dataset is significantly smaller than that for the first course both in terms of the number of students N and the number of days T. There were 25 video lectures in total, and students were supposed to watch one lecture per day from Monday through Friday. The final exam was held on campus after the 5 lecture weeks.

2.5 Experimental Results

In this section, I discuss the application of the proposed change detection methodology to the two clickstream datasets described in the previous section.

2.5.1 Example 1: 10-Week Face-to-Face Course

The clickstream data spanned 85 days, which included 10 weeks of instruction as well as activity before and after the 10 weeks. The change-detection methodology was applied to 4 different versions of the $N \times T$ data matrices: for preview and review events, in binary and count form. The changepoints were restricted to be in the range from day 10 to day

Event Type	$N_{Increase}$	$N_{Decrease}$	$N_{NoChange}$
Preview, binary	7	9	361
Preview, count	112	96	169
Review, binary	39	23	315
Review, count	121	159	97

Table 2.1: Number of students who showed increase, decrease, or no change in their activities for each activity data type for the 10-week face-to-face course.

75, since changepoint detection at the beginning or end of the sequences (i.e., outside of this range) tends to be unreliable due to small sample sizes and not so meaningful in terms of interpreting actual student behavior.

The students that were considered to have changed by the BIC scores were categorized into two groups: students who increased their click activity and students who decreased their click activity. I refer to these groups as *Increased* and *Decreased*, respectively. Note that these terms should be interpreted in a relative sense, since increase and decrease are from the α_i coefficient for each student *relative to the background rate* μ_t. Thus, a detected increase for student i means in effect that the student is ranked higher in the class in terms of activity relative to other students after the changepoint τ_i, compared to their rank before τ_i (and conversely for a decrease). The students without detected change are referred to as the *NoChange* group.

The numbers of students detected as belonging to each group, for each event type, are shown in Table 2.1. The Poisson count model detects significantly more student changes than the Bernoulli binary model, for both preview and review event types. This is to be expected since the Poisson model has more information to work with (and thus has better sensitivity) compared to the Bernoulli model, which only sees a binarized version of the daily counts (and thus has less information per day about student activity). In the discussion below, I focus primarily on the Poisson results with counts given its better sensitivity.

Figure 2.7 shows the click data for each of the students for which a change was detected,

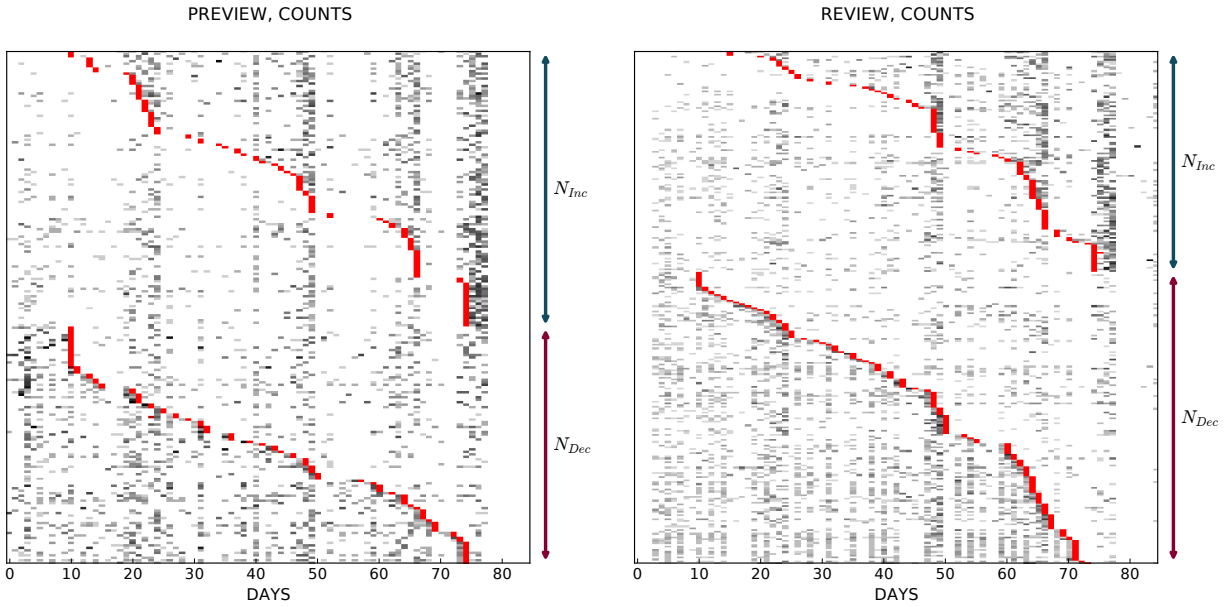


Figure 2.7: Student preview and review activity data over time, for the students who increased or decreased their behavior in the 10-week face-to-face course. The gray marker at t-th column in each row means that there was click activity on day t for that student, with darker colors reflecting larger counts (more clicks).

with one student per row, and one plot per type of event (Preview and Review). The students are split into two groups within each plot depending on whether their detected changes were increases or decreases, and rows were then ordered within each group based on the chronological location of the changepoint per student. The changepoint locations are marked in red, and the plots show a clear distinction between the days with more activity and the days with less activity.

Figure 2.8 provides a week-by-week summary of the information in Figure 2.7, showing the number of detected student changes per week, for each type of event. The vertical lines that are visible in the two count matrices are the exam dates. There are some obvious temporal patterns in this data. For example, the upper plot (preview events) shows that more than a quarter of the students increased their previewing activities in the third week, which is the week before the first midterm. This agrees with the intuition that prior to the first major exam in a class, we would expect to see some significant shifts in student activity. The lower

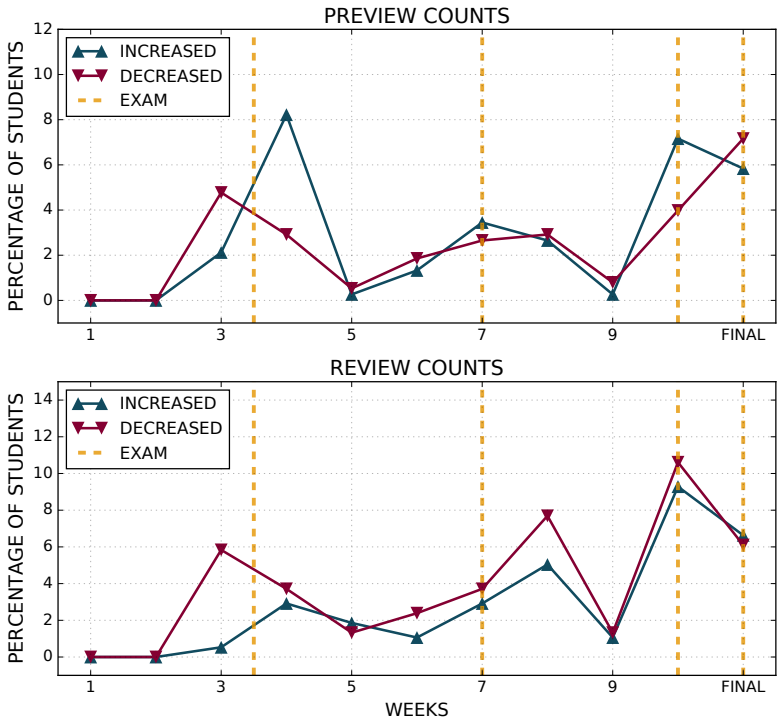


Figure 2.8: Percentage of 10-week face-to-face course students who increased or decreased within each week.

	$P(Pass\|Inc)$	$P(Pass\|Dec)$	$P(Pass)$
Probability	0.93	0.76	0.83
$\Delta Pass$ (%)	12.10	-7.44	0
p-value	0.0025	0.0458	-

Table 2.2: Probability of a student getting a passing grade (A, B, C) depending on which group the student is in for the review data.

plot shows that most of the changes in reviewing activity happened towards the end of the quarter, particularly during week 10 before the final exam. Again it makes sense that there are significant changes across students in their relative rates of reviewing activity prior to the final exam. We can also see in both plots that the number of detected changes per week, for increases and for decreases, are strongly correlated. As mentioned earlier, this is to be expected with this model since increase and decrease for this model are defined relative to overall mean population behavior.

I also investigated how detected changes in preview and review activities were correlated

with student outcomes in terms of the students' final grades in the class. I calculated the probability of a student getting a passing grade given that the student is in the *Increased* group, $P(Pass|Increase)$, or in the *Decreased* group, $P(Pass|Decrease)$, and compared these numbers with the marginal (unconditional) probability of a student passing $P(Pass)$. For both preview and review count events, I used a two-sided binomial test with $P(Pass)$ as the null hypothesis to compute p-values for $P(Pass|Increase)$ and $P(Pass|Decrease)$.

Table 2.2 shows the results for review count data. $\Delta Pass(\%)$ in Table 2.2 is the percent change of the probability of passing relative to the marginal probability $P(Pass)$, depending on what group the student is in. It is calculated by the following equation.

$$\Delta Pass = \Big(\frac{P(Pass|G)}{P(Pass)} - 1\Big) \times 100, \tag{2.13}$$

where G is the behavior group. At the 0.01 level of significance, $P(Pass|Increase)$ is significant and $P(Pass|Decrease)$ is significant at the 0.05 level.[5] Students in the *Increased* group have a higher probability of passing the course, while the students in the *Decreased* group have a higher probability of failing. This means that students who increased their reviewing behavior (relative to all of the students in the course), at some point during the quarter, ended up getting better grades on average than those that did not.

For preview counts, the probabilities were also in the direction of increases in previewing, leading to better outcomes on average (and vice versa), but these changes were not statistically significant. This may suggest, for this particular course, that changes in review activities are better predictors of student outcomes than preview activities.

Finally, for the 10-week course, I analyzed in more detail the results for two specific students (using their Review data) to illustrate how the model can be used to interpret clickstream activity at the individual student level. Figure 2.9 (a) illustrates the results for a student

[5]Students in *NoChange* group had a passing probability of 0.8, and it was not statistically significant in terms of difference from the students as a whole, i.e., the marginal probability $p(Pass)$.

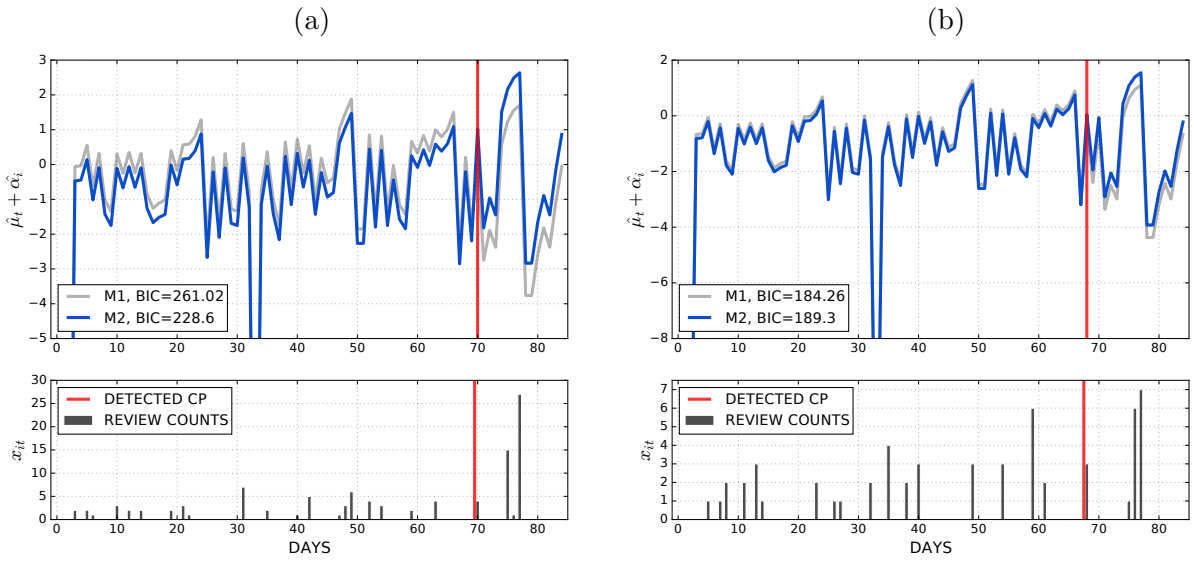


Figure 2.9: Log of $\hat{\lambda}_{it}$ from M1 and M2 (top), and the raw data of a student from the 10-week face-to-face course (bottom). For (a), the model with changepoint (M2) was selected by the BIC method, and for (b), the no-changepoint model (M1) was selected by the BIC method.

where the lower plot shows the observed daily review clicks, and the upper plot shows the Poisson models for the no-change model and the changepoint model (with a detected change at day 70). For this student, the BIC method preferred the changepoint model over the no-change model, with $BIC_2 < BIC_1$ by a large margin. This is reflected in the observed data in the lower plot where the number of counts for this student increases significantly after the changepoint.

Figure 2.9 (b) shows the same type of plot for a student where the BIC method selected the model without the changepoint. From the raw counts (lower plot), it looks like the student's activity level could have changed (increased) after day 68. However, relative to the background activity (particularly around days 76 to 78, leading up to the final exam) this student's activity level is not sufficiently different to the mean population behavior to justify the additional parameters in the changepoint model, as reflected in the BIC scores ($BIC_1 < BIC_2$).

Data Type	$N_{Increase}$	$N_{Decrease}$	$N_{NoChange}$
Preview, binary	6	8	162
Preview, counts	41	40	95
Review, binary	11	6	159
Review, counts	47	66	63

Table 2.3: Number of students who showed increase, decrease, or no change in their activities for each activity data type for the online course.

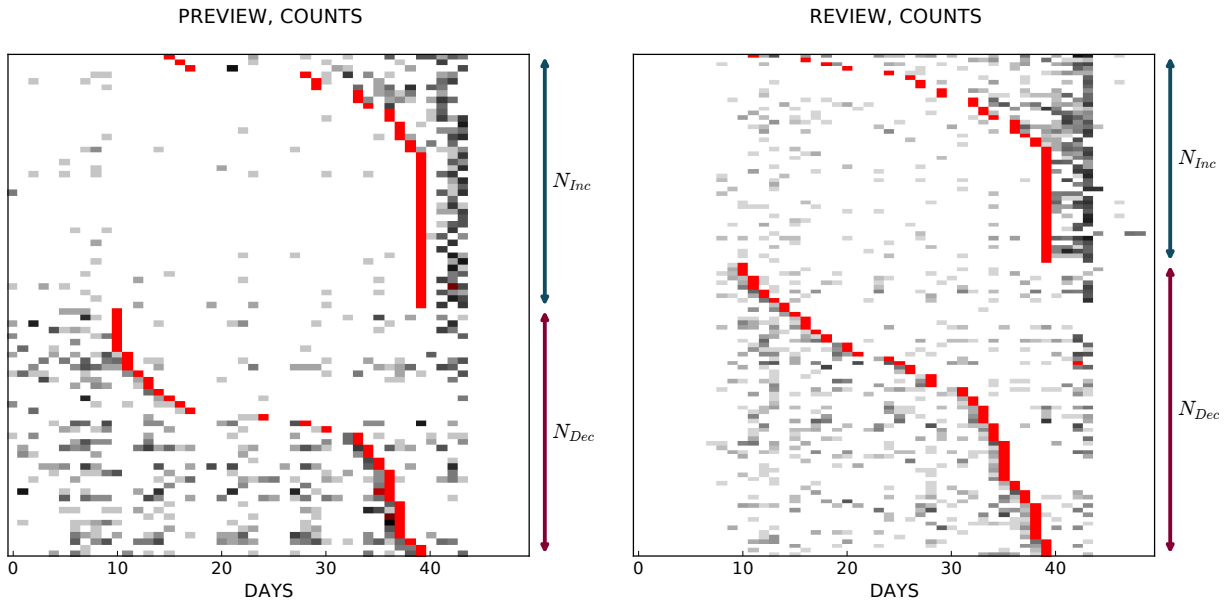


Figure 2.10: Student preview and review activity data over time, for the students who increased or decreased their behavior in the 5-week online course. The gray marker at t-th column in each row means that there was click activity on day t for that student, with darker colors reflecting larger counts (more clicks).

2.5.2 Example 2: Online 5-Week Course

The second course was a 5-week summer online course. The event dataset used for this course was smaller than the first in terms of both the number of students ($N = 176$) and the number of days with clickstream activity ($T = 50$). The course was offered online and the students were expected to watch a lecture video on every weekday over the 5 weeks, leading to more uniformity and less variability in student clickstream activity over time. In addition, the 10-week class had 3 midterm exams and a final exam, while the 5-week online class only had a single final exam at the end of the course.

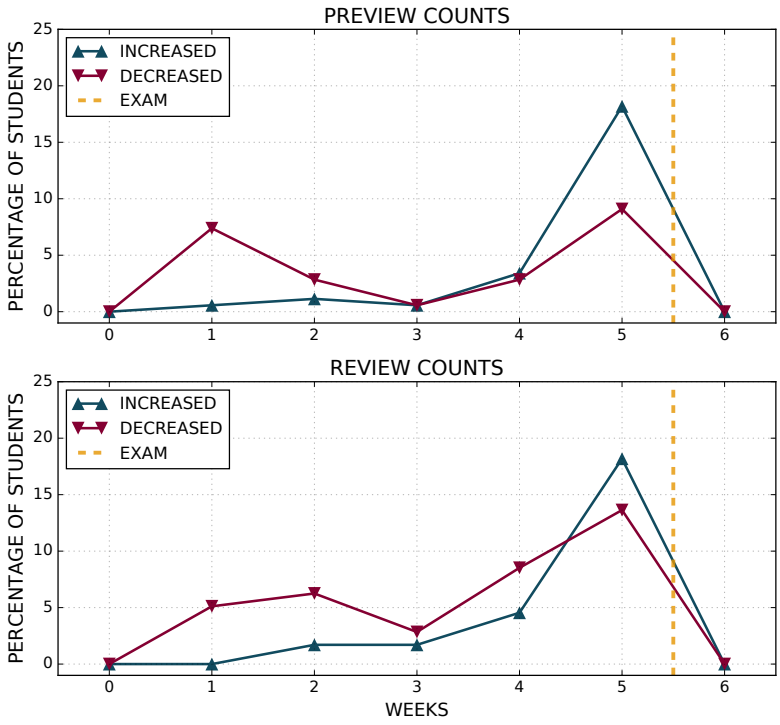


Figure 2.11: Percentage of students with detected increase or decrease in activity for each week in the online 5-week course.

The numbers of students detected for each of the *Increased* and *Decreased* groups, for both preview and review events, are shown in Table 2.3. We see a similar overall pattern to that for the 10-week class, namely that the Poisson model using counts detects considerably more changes than the Bernoulli method using binary data. The overall proportions of changes detected are roughly similar across both classes, with about 50% of students having increased or decreased count activity relative to the population, for each of the two types of events. One difference I found between the two courses was the proportion of students who exhibited no change at all, for either preview or review events: 13% of students in the 10-week course and 25% in the 5-week courses. This difference might be due to the intermediate exams (3 midterms) in the 10-week course, leading to more variability in student behavior compared to the 5-week course, which only had a final exam.

The clickstreams for the students with detected changes are shown in Figure 2.10. There are very high activities at the end of the course session for students in the *Increased* group,

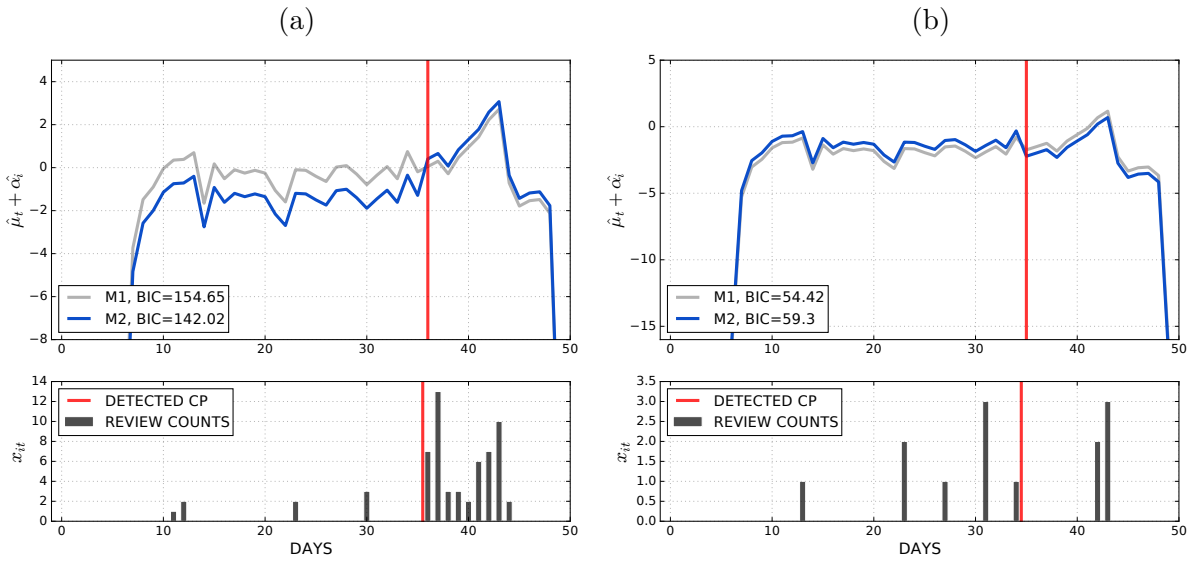


Figure 2.12: Log of $\hat{\lambda}_{it}$ from M1 and M2 (top), and the raw data of a student from the 5-week online course (bottom). For (a), the BIC method selected the changepoint model (M2), and for (b), the BIC method selected the no-changepoint model (M1).

for both Preview and Review event types. Also, the majority of the changepoints occur just before the darker area of the plot. Figure 2.11 shows that, among the students who had an increased change, most of them had a changepoint in the fifth week, which is the last week of the course before the final. The relationship of click activity and course outcomes for this course is not analyzed since fewer than 5% of the students received grades of D or F in the class, resulting in a sample size that is too small for reliable inferences.

As with the 10-week class, the results for review events for two specific students are examined to illustrate the methodology at the level of individual students. Figure 2.12 (a) shows the results for a student where the method detected a change in activity at day 35. Figure 2.12 (b) shows the results for a student where the no-change model was preferred by BIC. Both students exhibited increases in their review activities after day 40. However, the magnitude of change for the first student is significantly greater than that for the second student (as can be seen in the lower panels of both plots)—relative to the student population as a whole, the second student did not exhibit a significant change in activity.

2.6 Conclusions

Student clickstream data is inherently difficult to work with given its complex and noisy nature. This chapter described a statistical methodology for detecting changepoints in such data and illustrated the potential of the approach by applying the methodology to two large university courses. The proposed approach is relatively simple and allows for a number of possible extensions; the development of more flexible changepoint models (such as systematic drifts in student activity levels), allowing for more than a single changepoint, post hoc adjustments for multiple testing, and using robust estimation techniques for parameters and their respective standard errors. Bayesian methods could also be potentially useful in this context for both parameter estimation and model selection to more fully reflect uncertainty in inferences at the individual student level. A useful extension for educators would be to develop an online detection variant of the offline approach proposed here, potentially allowing for identification of at-risk students, instructor feedback, or interventions while a course is in session.

While the results in this chapter are promising and there are interesting methodological avenues to pursue, the most important future direction from an education research perspective will involve more in-depth investigation of the utility of these types of methods in terms of providing actionable insights that are relevant to the practice of education.

The primary contributions of this chapter include the following:

- I developed regression models for binary and count clickstream data to fit each individual student's preview and review activity relative to the whole class.
- I utilized statistical change detection techniques to investigate the change in student behaviors.
- I applied the method to two large university courses, one face-to-face and one online,

and validated the method.

- I analyzed the results by grouping the students into three behavioral groups (*Increased*, *Decreased*, and *NoChange*), and found the relationship between the behavioral groups and their course outcomes, with the conclusion for the face-to-face course that increased activity led to positive change in course outcome (the probability of passing the class).

Chapter 3

Understanding Student Procrastination via Mixture Models

As colleges and universities continue to increase the number of online course offerings, these classes are becoming a normal part of students' learning experiences. While online courses have made learning more accessible to students, prior work suggests that students enrolled in online courses have worse learning outcomes when compared to students enrolled in face-to-face courses [15]. One important reason for this is that the online learning environment requires a higher degree of self-regulation than the face-to-face environment [21]. Students must effectively plan and regulate their learning time, and monitor their own progress in order to meet important deadlines [175], but students may lack these important skills. Moreover, online courses have a high degree of anonymity. Students are not physically present in a classroom, and their activity on Learning Management Systems (LMS) is not made public

The material in this chapter is based on the paper "Understanding Student Procrastination via Mixture Models" by Park, J., Yu, R., Rodriguez, F., Baker, R., Smyth, P., and Warschauer, M., published in *the Proceedings of the Eleventh International Conference on Educational Data Mining (EDM 2018)*, 2018. The paper received the best paper award.

to the rest of the class. This absence of face-to-face accountability may cause students to disengage with the course much more than they would in traditional classrooms. The lack of structure and anonymity may lead students to procrastinate, putting off work until close to important deadlines. Therefore, understanding students' learning behaviors relating to time management, especially procrastination, could be one important mechanism for improving online learning.

Clickstream datasets have provided rich resources for analyzing students' time management behaviors. Procrastination has been measured using the specific time points at which students take certain actions within an online course, such as accessing content pages, watching lectures, and submitting quizzes. A common way to measure procrastination is to calculate the amount of time a student is engaged with the LMS prior to an important course deadline. Studies that use these types of measures as indicators of procrastination find that the indicators are negatively correlated with course outcomes [73, 81, 173].

Motivated by these previous studies, I utilize clickstream data to further understand procrastination using two online classes offered at a large public university. These two classes were designed so that the students are expected to space out their studies on a daily basis and to set weekly deadlines. In this chapter, I investigate the use of probabilistic mixture modeling to analyze time-stamped logs of student activity in the context of these two online classes. The mixture model identifies different behavioral patterns in the data, where the patterns can be clearly identified as reflecting procrastinating and non-procrastinating behavior among the students. The developed methodology enables finer-grained analyses of procrastination and its relationship with learning outcomes, which can inform more effective instructional reforms in online learning.

3.1 Related Work

3.1.1 Self-Regulation, Procrastination, and Academic Success

Self-regulated learning refers to the process of directing one's own learning experience [175], and these processes encompass several attitudes and behaviors. For instance, models of self-regulated learning generally distinguish between motivational beliefs about learning, goal setting and planning behaviors, specific learning strategies, and metacognitive monitoring processes [125]. While each of these facets plays an important role in the learning process, research on online learning finds that students' planning and time management behaviors are important indicators of course success [46, 173]. Procrastination behaviors, which refer to delaying coursework until major deadlines, reflect poor planning and time management.

Several studies have focused on procrastinating as a major barrier that hinders students from succeeding in online courses [46, 172]. Using online course analytic data, one recent study found that students who did not begin working on assignments until hours before a deadline received lower course grades when compared to students who began their work earlier [43]. Other studies have found similar results, where students who delay working on assignments are more likely to perform poorly [172, 173]. These results confirm the undesirable nature of procrastination as well as the importance of regular learning behaviors.

Another extensive body of work has shown that students from underrepresented backgrounds, such as those who come from low-income households, or who are first to attend college, are at greater risk for leaving STEM majors [28]. This problem may be additionally exacerbated in online coursework. There are many important factors that explain issues surrounding underrepresented student success, such as lack of mentoring, financial concerns, and feelings of exclusion [146]. With regard to self-regulatory behaviors such as procrastination, prior work has also shown an increased tendency for underrepresented groups to

engage in more procrastination than their counterparts [141]. However, this study was not conducted in an educational context, and procrastination was measured subjectively using surveys. With this in mind, a side aim of my work in this chapter is to explore the relationship between individual differences in procrastination (time management behavior, in general) and students' external background characteristics, specifically for the students taking online courses.

3.1.2 Measuring Procrastination

Measures of procrastination are relatively straightforward and similar across various learning environments. In the most common measures, researchers capture the time that students finish a certain task and calculate the difference between this time and either the release time [18] or the deadline [81, 73] of the task. This type of measure has the merit of being very interpretable, but a limitation is that it only captures the average degree of procrastination without depicting nuanced patterns in these behaviors.

3.1.3 Cluster Analysis and Mixture Modeling

Clustering in general is a widely used technique in data analysis for automated data-driven discovery of groups or clusters in data. In the context of analyzing education data, clustering algorithms have found broad application as a technique for grouping students based on their behavioral patterns. For example, Toth et al. [158] cluster students based on their problem-solving interaction patterns using the X-means algorithm (a variation of the well-known K-means clustering algorithm) for a better understanding of complex problem-solving behaviors and identifying levels of problem-solving proficiency. Ng, Liu, and Wang [115] use survey scores of motivated strategies for learning questionnaires to cluster students into multiple groups. The resulting groups obtained by hierarchical clustering with Ward's method exhibit

distinct learning profiles of motivational beliefs and self-regulatory strategies.

The clustering approach I follow in this chapter is probabilistic model-based clustering [52, 107]. In this framework, each cluster corresponds to a probability distribution (also known as a "component") in a mixture model and the entities being clustered are assumed to have been generated by one of the component distributions. This probabilistic framework for clustering has a number of advantages over non-probabilistic techniques such as K-means clustering or hierarchical clustering. For example, as I describe later in the chapter, the framework can model count data in a natural manner using Poisson distributions as components in the mixture model, where each component (or cluster) represents a different Poisson distribution over count outcomes. The Poisson mixture model has been applied to a number of different fields including marketing [19], finance [78], biology and bioinformatics [23, 50], document analysis [95], and so on. However, to my knowledge, there had not been any work prior to this on the development of Poisson mixtures in an education context, particularly for the problem of clustering students based on their observed activity in online classes.

3.2 Methods

3.2.1 Student Activity Counts

For courses where time-stamped student-generated events are tracked via logs of clickstream data, we can count these events on a daily basis. Thus, we can get a set of *daily activity counts* for each student throughout a course, where the activity can correspond to specific types of tasks of interest (such as video-watching, quiz submission, and so on). Figure 3.1 shows the daily activity count data for two students from one of the course datasets used in this chapter. The data for each student is displayed as a matrix, where the grayscale indicates

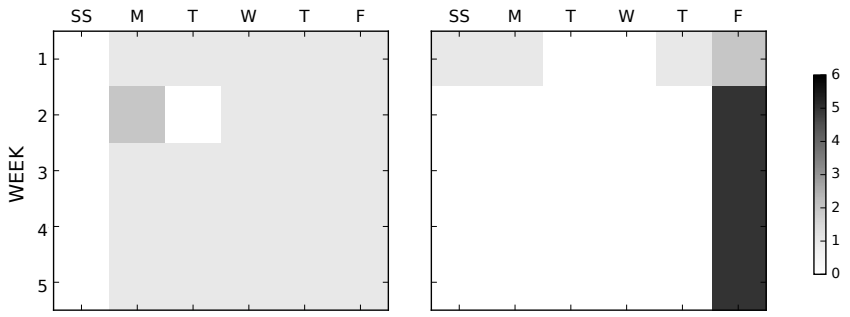


Figure 3.1: Examples of student *daily activity counts* (specifically, the number of video watching tasks per day) displayed as a matrix of week × day. SS indicates Saturday and Sunday.

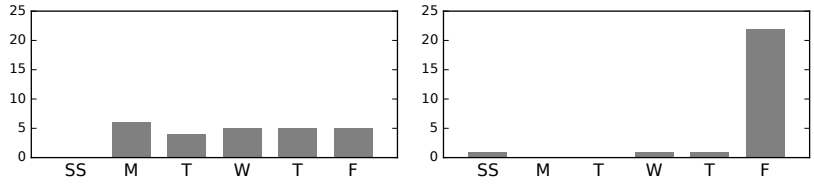


Figure 3.2: *Aggregated daily task counts* across weeks ($\boldsymbol{x}_i$) for the two students shown in Figure 3.1.

the number of tasks performed by each student on each day over the 5 week duration of the course. This type of display is useful in terms of capturing the temporal aspect of when a student is engaged in a particular activity, such as watching a lecture video or submitting a quiz. It also indicates that one of the students (on the right) may be procrastinating each week—I discuss these types of patterns in more detail later in the chapter.

It is also possible to compute the sum over weeks to get the "aggregated" daily counts assuming that there is a structure in the course that repeats every week (which is the case for the two classes studied in this chapter). Examples of the aggregated daily counts are shown in Figure 3.2 as bar plots, computed by aggregating across the weekly rows of data for each student in Figure 3.1.

3.2.2 Mixture Model with Gamma Priors

In this section, I discuss the use of a Poisson mixture model to cluster students based on their activity counts, focusing on the aggregated daily counts as in Figure 3.1. In terms of notation, let $\boldsymbol{x}_i$ be the vector of aggregated daily counts for student i, where $i = 1, \ldots, N$. The dimensionality D of each vector is the number of days ($D = 6$ in this case since Saturday and Sunday are collapsed into one). Thus, the data consists of N students, each with a D-dimensional vector of aggregated daily counts.

To model this data, I use a probabilistic mixture model with Poisson components. Let K be the number of components (or clusters) with an index $k = 1, 2, \ldots, K$. The unobserved latent variable z_i takes values from the set $\{1, \ldots, K\}$ and corresponds to the latent component or cluster that student i is presumed to belong to. Each of the k components consists of a vector of Poisson rate parameters, $\boldsymbol{\lambda}_k = [\lambda_{k1}, \ldots, \lambda_{kd}, \ldots, \lambda_{k6}]$, where d from $\{1, \ldots, 6\}$ represents a specific day of the week. For example, one component could have very low values for all the λ_{kd}'s, for students with low daily activity, and another component could have high values for all the λ_{kd}'s, for students with high daily activity.

When fitting this mixture model to data, I take a Bayesian approach [57] and use Gamma prior distributions for the rate parameters λ_{kd}. The primary reason for doing this is to encourage the model to avoid degenerate solutions with a small component that has one or more rate parameters λ_{kd} at or near a value of 0. This can produce a high-likelihood solution but one that is not useful. In the experimental results later in the chapter I used hyperparameters of $\alpha = 1.1$ and $\beta = 0.1$ for the Gamma distribution. These hyperparameter choices have the effect of making the Gamma prior behave like a step function, putting zero probability mass at $\lambda_{kd} = 0$ ($k = 1, \ldots, K, d = 1, \ldots, 6$), and a relatively flat uninformative prior distribution over positive rate parameter values. Figure 3.3 depicts a graphical model representation of the Poisson mixture model with a Gamma prior on the $\boldsymbol{\lambda}$ parameters for

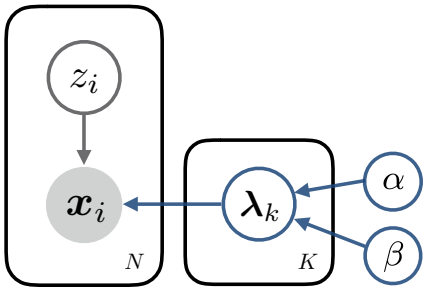


Figure 3.3: Graphical representation of the Poisson mixture model with Gamma prior. $\boldsymbol{x}_i$ and $\boldsymbol{\lambda}_k$ are 6 dimensional vectors. N is the number of students, and K is the number of mixture components.

each component.

The likelihood for the data $\boldsymbol{x}_i$ for each student i under this mixture model can be written as:

$$p(\boldsymbol{x}_i|\boldsymbol{\lambda}) = \sum_{k=1}^{K} p(\boldsymbol{x}_i|z_i = k, \boldsymbol{\lambda}_k)p(z_i = k), \tag{3.1}$$

where $p(z_i = k)$ is the marginal mixing weight for each component, and each component distribution can be written as:

$$p(\boldsymbol{x}_i|z_i = k, \boldsymbol{\lambda}_k) = \prod_{d=1}^{D} p(x_{kd}|\lambda_{kd}, z_i = k), \tag{3.2}$$

assuming conditional independence of the daily counts x_{kd} given component k. λ_{kd} is the rate for day d for component k and each distribution $p(x_{kd}|\lambda_{kd}, z_i = k)$ is a Poisson distribution. The prior distribution is defined as a product over independent Gamma priors, one for each λ_{kd}, each with parameters $\alpha = 1.1$ and $\beta = 0.1$.

3.2.3 Learning Parameters with the EM Algorithm

To estimate the parameters $\boldsymbol{\lambda}_k$ of the model, the Expectation-Maximization (EM) algorithm is used. It is an iterative algorithm that is widely used in fitting mixture models to data [41, 106]. More specifically, I use the EM algorithm to maximize the product of the data likelihood times the prior (both defined above). This results in both (a) maximum a posteriori (MAP) parameter estimates for the Poisson components in the model, and (b) membership weights w_{ik} that reflect the probability (under the fitted model) that each student i belongs to component (or cluster) k.

Each iteration of the EM algorithm consists of two steps, the E (expectation) step and the M (maximization) step. In the E-step, conditioned on some fixed (current) values of the parameters, the probability of membership w_{ik} is computed for each component $k = 1, \ldots, K$, for each student $i = 1, \ldots, N$.

$$\begin{aligned} w_{ik} &= p(z_i = k | \boldsymbol{y_i}, \boldsymbol{\lambda}, \alpha, \beta) \\ &\propto p(\boldsymbol{x}_i, z_i = k, \boldsymbol{\lambda}_k | \alpha, \beta) \\ &\propto p(\boldsymbol{x}_i | z_i = k, \boldsymbol{\lambda}_k) p(\boldsymbol{\lambda}_k | \alpha, \beta) p(z_i = k). \end{aligned} \tag{3.3}$$

These membership weights are important in the later analyses, since they provide information of how likely it is that each data point i (student i) was generated by component k. In the M-step, conditioned on the set of membership probabilities w_{ik}, a point estimate of each parameter is estimated via MAP estimation.

$$\hat{\boldsymbol{\lambda}}_k = \frac{\sum_i w_{ik}(\boldsymbol{x}_i + \alpha - 1)}{\sum_i w_{ik}(1 + \beta)} \tag{3.4}$$

$$\hat{p}(z_i = k) = \frac{\sum_i^N w_{ik}}{N}. \tag{3.5}$$

These MAP parameter estimates provide the input for the next E-step, and thus, the cycle

of E and M-steps continue iteratively.

The EM algorithm as a whole consists of randomly initializing the parameters of the model, followed by repeated computation of pairs of E and M steps, until the log-likelihood is judged to have converged (i.e., when the improvement in log-likelihood from one iteration to the next is less than some small value ϵ, or when the average membership probability value is not changing significantly from one iteration to the next).[1]

3.3 Datasets

Two datasets from the same undergraduate online course were used in this study: one from summer 2016, and the other from summer 2017. The course was offered for 5 weeks in both years. While each class was taught by two different instructors, the class content, such as the lecture videos, resources, and assignments, were the same. In both classes, students were assigned 5 video lectures every week, and each lecture video had a corresponding quiz. The instructors encouraged students to watch one lecture video and complete the corresponding quiz each day, from Monday through Friday.

Although students were encouraged to follow this schedule, the actual deadline for watching the 5 lecture videos and completing the quizzes was on Fridays at midnight. While this structure gave students freedom to watch the lecture videos when they wanted, this flexibility also allowed them to procrastinate.

Most of the students' activities were recorded through the Canvas Learning Management System (LMS). These activities included downloading course content, watching lecture videos, taking online quizzes, submitting assignments, etc. Every time a student clicked on a URL within the Canvas system, the click event was logged with the student ID, URL, and time-

[1]Python code for the EM algorithm on Poisson mixture model is available online at `https://github.com/ucidatalab/student_poisson_mixture`.

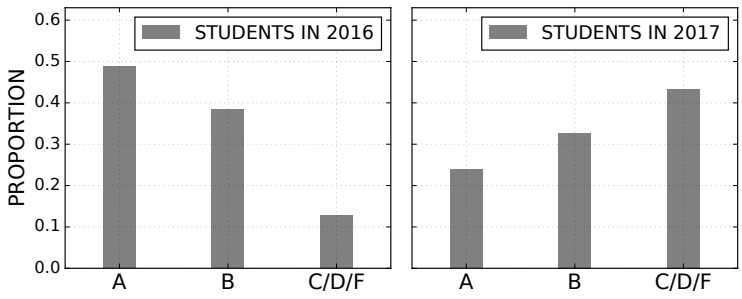


Figure 3.4: Grade distributions of students in 2016 class (left) and in 2017 class (right). Two classes show very different grade distributions. Almost half of the students received an A for the class in 2016, whereas more students got lower grades in 2017.

stamp. The clickstream data was processed so that it only focused on the activities of daily tasks, resulting in daily activity counts, as mentioned in the previous section (Figure 3.1 and Figure 3.2). Only one event per task was counted, and thus the sum of the matrix for each student was 25 or less (for 5 video lectures $\times$ 5 weeks). I chose to count only the first attempt (first click event) for each task.

In addition to the clickstream data, student demographic data was available through the university's institutional research office. It included both demographic information (gender, ethnicity, first generation status, low income status, and full-time status) and prior academic achievement (total SAT score[2]). Some students did not agree to provide this demographic data, although most did. For this reason, the later analyses based on demographic information are based on the subset of students who agreed to share this information.

Although both classes used the same materials and implemented the same deadlines, there were some notable differences in how the click events were recorded, as well as how each instructor structured the course. These differences are described in the following sections.

[2]A standardized test widely used for college admissions in the United States.

3.3.1 Class in 2016

Online lectures and daily quizzes were offered outside the Canvas LMS for this class. Each lecture video was embedded on a separate web page on the server that was accessible, and the links to the web pages were provided via the Canvas weekly module.

Logs for the daily quiz attempts were not accessible, so instead I used the first "video clicks," which are from the logs of HTTP GET requests of the video embedded web pages. For each student, the IP addresses of the video logs (from the server) were matched with the IPs recorded on the Canvas LMS.

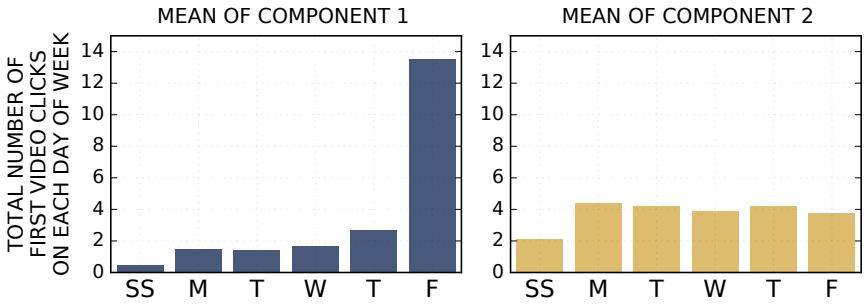
After removing 4 students with very low activity (0 or 1 video clicks in total), there were 172 students with activity counts available for analysis. More than 90% of the students received a passing grade, and half of the students received an A (Figure 3.4). Completing the daily tasks (watching videos and solving quizzes) counted as 15% towards the overall grade for each student.

3.3.2 Class in 2017

The video click logs for this class were not available since the videos were uploaded on a third-party server. However, the daily quizzes that students took after watching the lecture videos were recorded through the Canvas system, and it was possible to obtain students' quiz submission time-stamps via the corresponding clickstream data. Therefore, for this class I focused on the first clicks for daily "quizzes." Note that this is different from the 2016 class data, which used the first clicks for each video-watching event.

There were 145 students in the class—I used data for 140 students after dropping 5 students with very low activity (as with the 2016 class). As previously noted, a different instructor taught the class in 2017 than in 2016. The instructor for the 2017 class changed the contri-

(a) Class in 2016



(b) Class in 2017

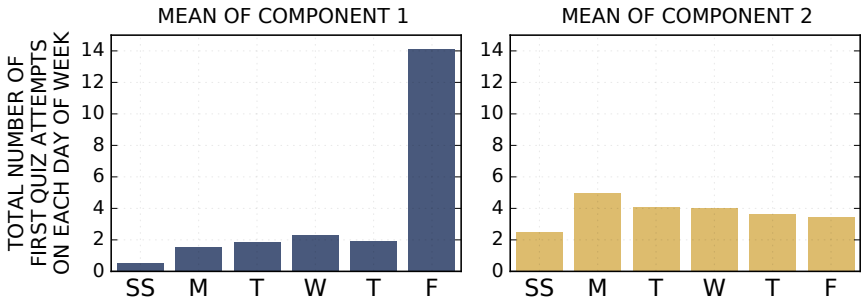


Figure 3.5: Poisson mixture component means ($\boldsymbol{\lambda}_k$'s) from modeling aggregated daily task counts ($\boldsymbol{x}_i$) for the class in 2016 (upper) and 2017 (lower).

bution to 8% of the total grade for watching and completing the lecture videos, significantly less than in the 2016 class (15%). The grade distribution of the 2017 class in Figure 3.4 is also significantly different from that in 2016—there are significantly fewer students who received A's or B's in 2017 compared to the 2016 class.

3.4 Procrastination as a Mixture Component

Below, I present and discuss the results of fitting a two-component ($K = 2$) Poisson mixture model to the aggregated daily task counts for the two classes described in Section 3.3. I also explored models with more components, $K = 3, 4, \ldots$, but found that the $K = 2$ model broadly captured the primary modes of student behavior and that higher values of K tended to split the two main modes into further subgroups without providing any significant

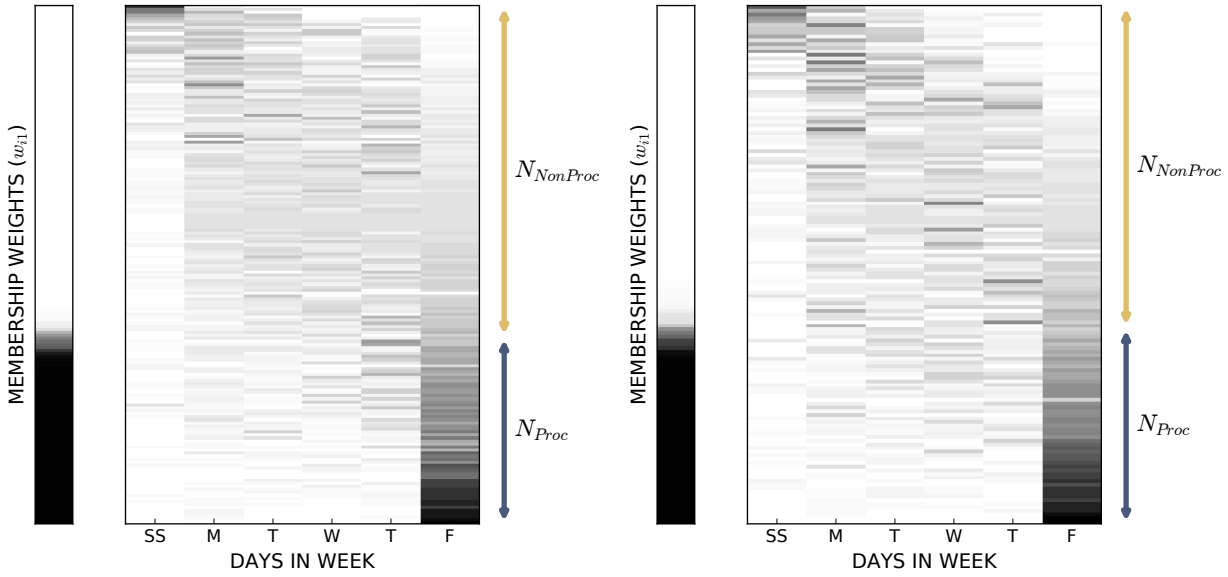


Figure 3.6: Aggregated daily task counts shown along with the membership weights. Each row represents a student, and the students are sorted by the membership weight for the procrastination group w_{i1}. The left figure is for the class in 2016, and the right figure is for the class in 2017.

additional insight.

Figure 3.5 shows the expected number of counts on each weekday per group, i.e., the rate parameters, $\boldsymbol{\lambda}_k$'s. The two group-dependent rate patterns across the days of the week, for both 2016 and 2017, show two very distinct behavioral patterns. One of the mixture components has a very high rate on Friday and low rates on the other days of the week. The other component has low and relatively flat rates from Monday to Friday. Considering the fact that the deadline for daily tasks in these courses is on Fridays, these two patterns clearly reflect two different types of student behaviors: *procrastination* and *non-procrastination*.

3.4.1 Characteristics of the Two Behavioral Groups

The membership weights can be thresholded at 0.5 to classify each student $i = 1, \ldots, N$ into one of the two groups, i.e., if $w_{i1} > 0.5$ then student i is assigned to the *procrastination* group (where $k = 1$ corresponds to the *procrastination* group). About 36–37% of the students were

assigned to the *procrastination* group in each of the two years.

The two plots in Figure 3.6 illustrate the students' week-aggregated activities along with the students' membership weights. Each row in each plot represents a student, and the wider matrix plot shows the aggregated daily counts, sorted by their membership weight w_{i1}. The values in the matrix range from 0 to 25, and a darker color means that there are more task activities on that day of the week. The two plots from different years look almost identical, and they clearly show the two types of behavior. The students (rows) at the bottom of each plot have more counts (darker colors) on Fridays and belong to the *procrastination* group. There is also a small group of students at the top of both plots who tend to be more active over the weekend. The size of this group of students is relatively small and their behavior pattern is effectively that of non-procrastinators since they are the "early birds" who check out the lecture videos or the quizzes early in the week.

The membership weights are shown on the narrower bar plot (left of each year's plot), where a darker color represents a higher membership weight of belonging to the procrastination group (with a weight close to 1). We can observe that there is a relatively small amount of grey area in the bar plot (for both years), which means that the majority of the students have a very high probability of being assigned to one group or the other.

3.4.2 Association between Behaviors and Grades

I further analyzed the relationship between the two different behavioral groups and the grades. The grade distribution in each group is shown in Figure 3.7. Results from the two classes are shown side by side. It is obvious from the figure that the non-procrastinators tend to get significantly more A grades than the procrastinators, whereas the procrastinators get more C, D, and F's. Even though the overall grade distributions were quite different in the two classes (see Figure 3.4), a strong correlation between the behavioral groups and course

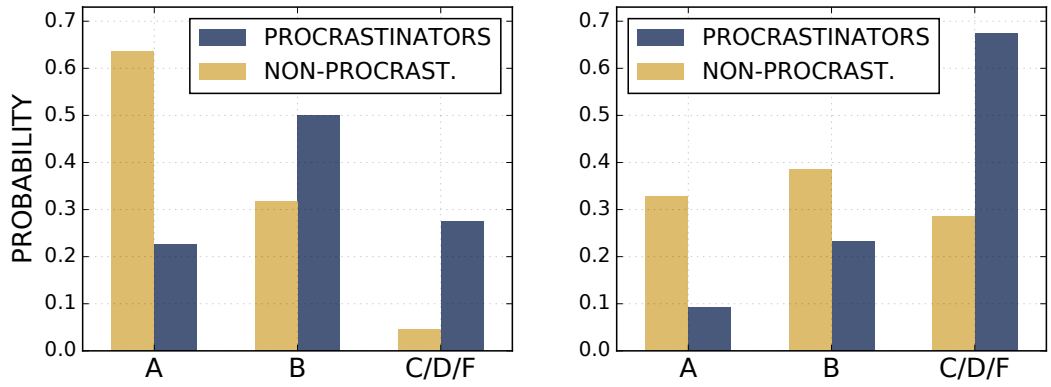


Figure 3.7: Probability of receiving each grade given that the student is in the *procrastination* group or in the *non-procrastination* group in 2016 (left) and in 2017 (right).

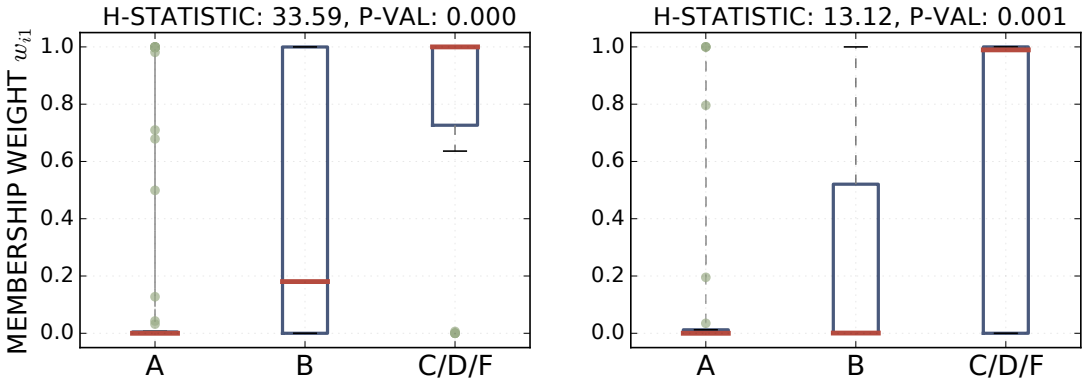


Figure 3.8: Distribution of procrastinating group membership weights w_{i1} in different grade groups of class in 2016 (left) and in 2017 (right). H-statistic comes from a Kruskal-Wallis test.

outcomes can be found. In both classes, the non-procrastinating students are about three times more likely to get an A grade than the procrastinating students. These probabilities were significant at the 0.01 level using a chi-squared test.

Another analysis on finding the relationship between procrastination behavior and grade outcomes can be done by grouping students by their grade (rather than by the behavioral group) and looking at the patterns of behavior for each grade group.

As shown in Figure 3.4, a majority of the students got a passing grade in 2016. The number of students who received A, B, and the others (C, D, or F) were 84, 66, and 22, respectively. The left boxplot in Figure 3.8 informs that "A students" have very low membership weight

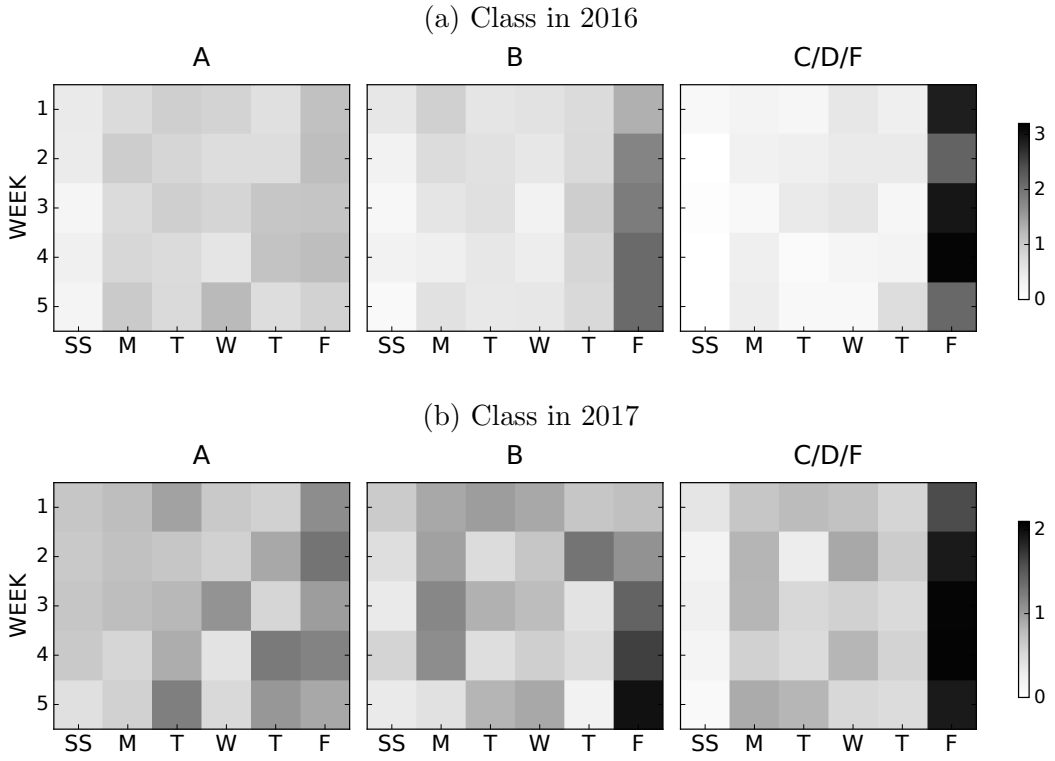


Figure 3.9: The number of task counts per day, for each of the 5 weeks, averaged over the students in each grade group. Left: students who received A, middle: students who received B, right: students who received C, D, or F.

(w_{i1}) values, but the "C, D, F students" have very high membership weight values. This can be interpreted as saying that the students who received lower grades (C, D, or F) have higher probabilities of being procrastinators.

A similar result can be observed for the class in 2017 from the right side of the plot in Figure 3.8. There were 27, 37, 49 students in each of the grade groups (there were 27 students whose grade information was unavailable). The broader distribution of weights in the C, D, F group may be due to the fact that there were many more students with lower grades than higher grades in this year of the course.

The association between behavior and grade outcome is also clearly visible in the raw data, i.e., the task activity counts, for both years. Figure 3.9 clearly illustrates the behavior patterns for students with different grades. Much darker colors are seen on Fridays on the

(a) Class in 2016

Demographics	N	Test	p-value
$FirstGen$	144	χ^2-test	0.566
$LowIncome$	151		0.672
SAT	147	K-W test	0.238

(b) Class in 2017

Demographics	N	Test	p-value
$FirstGen$	120	χ^2-test	0.218
$LowIncome$	128		0.955
SAT	125	K-W test	0.802

Table 3.1: Statistical test results for understanding the relationship between demographic variables and behavioral group assignment for two classes.

rightmost matrices (students who received C, D, or F grades), and more evenly distributed colors on the left matrix plots, which shows the activities of students who received A grades.

3.5 Relationship with Student Background

Having explored the fine-grained differences in students' procrastinating behaviors and their relationship with outcomes, I further examined if these variations can be explained by students' background characteristics. The goal of this analysis is to understand whether there exists a potential risk factor among underrepresented students, or if instead, the observed behavioral differences are more at the individual level in nature. I also sought to explore whether prior academic achievement could explain differences in procrastinating behaviors.

From a comprehensive list of demographic variables, I chose three that are of general interest in education research: *Low Income Status*, *First Generation* and *Total SAT Score*. The first two binary variables represent a student's socioeconomic status, and the last continuous variable is a proxy for prior academic achievement.

I separately tested the relationships between these three variables and the behavioral group assignment (as in Section 3.4.1). The type of statistical tests and their results are reported

in Table 3.1 (a) for the class in 2016, and Table 3.1 (b) for the class in 2017. Because there are missing values in the demographic information, only the students who have relevant information are included in each of the tests (the number of students, N, is reported in the table).

The results show that for both classes none of these demographic variables have any significant relationship with procrastination. This suggests that failures in time management may arise more from students' inherent factors than specific background characteristics, and that effective instructional interventions are less likely to be hampered by students' underrepresented backgrounds. However, due to the limited class sizes, this inference still needs to be further explored at scale.

3.6 Conclusions

In this chapter, I introduced a data-driven methodology for characterizing student procrastination in online courses. Based on Poisson mixture modeling, the proposed approach can be applied to courses where tasks with clear deadlines are regularly assigned and students' time-stamped activities related to those tasks are recorded. In the experiments with two undergraduate online classes, this method identified two distinct patterns in students' weekly planning behavior, which can be further utilized to measure procrastination. This measure was found to be strongly correlated with course outcomes for both classes. Interestingly, while the procrastination behavior was a strong predictor of course outcomes, it was not significantly related to students' demographics or prior academic achievement. These results suggest that, as a whole, procrastination behaviors seem to be more of an inherent characteristic.

These types of clickstream data and analyses allow for rich complements to other types

of educational research. For example, the proposed behavioral grouping can be combined with survey data to examine how accurate students' perceptions of their skills are, and to identify students who might be especially prone to benefit from support. From a practical perspective, these data-driven approaches can be incorporated into learning management systems and work in real time. This would potentially facilitate automated assessment and intervention regarding time management skills.

There are also a number of potentially useful extensions to the methodological approach proposed here. For example, the mixture components in the two classes that were analyzed are straightforward to interpret with regard to procrastination, but this might not be the case for different course designs and structures. In these broader scenarios, it may be useful to incorporate informative Gamma prior distributions into the mixture model, with, for instance, three prior components for procrastination behavior, non-procrastination behavior, and mixed behavior, respectively.

The primary contributions of this chapter include the following:

- I developed a general data-driven method for identifying procrastination. The method analyzed counts of student activity and can work with any online course with periodic deadlines and that has corresponding time-stamped clickstream data.
- I validated the method using two online university classes, and identified two distinct behavioral patterns which can be used to measure an individual student's degree of procrastination.
- For the two classes that were analyzed, I discovered that behavioral groups related to procrastination and course outcomes were highly correlated, lending support to prior theories of self-regulated learning and procrastination while also providing new insights.

Chapter 4

Detecting Conversation Topics in Primary Care Office Visits from Transcripts

Patient-physician conversations are complex, multi-dimensional, and multifunctional [13, 63, 72, 108, 160]. Patients present multiple issues during an office visit requiring clinicians to divide time and effort during a visit to address competing demands [51, 151] (e.g., a patient could be concerned about blood pressure, knee pain, and blurry vision in a single appointment). Moreover, visit content does not solely focus on biomedical issues, but also on psychosocial matters, personal habits, mental health [150], patient-physician relationship [49], and small talk [27, 150]. Health communications researchers analyze the content of patient-physician communication by directly labeling the interaction using trained raters

The material in this chapter is based on the paper "Detecting conversation topics in primary care office visits from transcripts of patient-provider interactions" by Park, J. et al., published in *Journal of the American Medical Informatics Association (JAMIA)*, 2019.

to label the topical content of clinical interactions. For example, during periodic health examinations with their primary care physicians, only one-third of patients with mental health needs had a discussion about mental health [149, 150]. These findings suggest that quality improvement efforts that evaluate the content of clinical interactions might help address gaps in service delivery.

Labeling each talk-turn, which is the utterances in each speaker's turn, using labeling systems designed to capture the content of medical visits (e.g., the Multi-Dimensional Interaction Analysis system [27, 151, 149]) is labor-intensive, and costly. It requires training a team to label the text, and establish and maintain reliability. Depending on the extensiveness of the labeling system, it can take several hours to label one patient-physician interaction [24, 110]. This level of effort means that even in research settings with resources for detailed evaluation, studies of patient-physician interaction are often limited in scale [151]. As a result, the direct evaluation of clinical interactions is not feasible in routine care settings for quality improvement purposes [70, 93]. Automated methods capable of extracting the topical content of clinical encounters could support physicians who overlook asking patients about critical issues (e.g., suicide, blood pressure medication) when it is clinically indicated. The methods can help reduce the burden of documentation currently placed on physicians, and facilitate large-scale research on the quality of patient-physician communication.

4.1 Related Work

The past decade has seen an explosion of interest in machine learning and natural language processing (NLP) in medical contexts [114], targeting problems such as automated extraction of information from clinical notes and electronic health records [40, 42, 129]. Of direct relevance to the present chapter is a growing body of work dedicated to applying methods from machine learning and NLP to the automatic annotation of conversations between physicians

and patients. The typical approach in such studies begins with labeling a corpus of transcript data, e.g., providing human-generated labels for each utterance or talk-turn in each visit in the corpus. Machine learning techniques are then used to learn a classification model from a subset of the corpus, the training data, and the model's predictions are evaluated by comparing its predictions with the known human labels on unseen test transcripts. For example, Mayfield et al. [102] analyzed patient-physician transcripts from the Enhancing Communication and HIV Outcomes (ECHO) study [12] and employed a logistic regression model to classify utterances into the categories "information-giving" or "information-requesting" [90]. More recently, Kotov et al. [85] analyzed transcripts from motivational interviewing related to pediatric obesity and developed probabilistic machine learning models for classifying patient utterances into classes of behavioral responses. In later work on the same dataset, Hasan et al. [65] compared probabilistic models with more recent recurrent neural network approaches and found the latter to be generally more accurate on that dataset.

There has been less prior work on the problem of automated classification of topical content in patient-physician dialog. Wallace et al. [166] developed machine learning models to classify dialog utterances into one of six high-level discussion topics: Biomedical, Logistics, Psychosocial, ARV, Missing/other, and Socializing. Using the same ECHO dataset as used in the Mayfield et al. study [12], they evaluated the performance of conditional random fields (CRFs) for this prediction task and concluded that the results showed promise for automated classification of patient-physician interactions into clinically relevant topics. Gaut et al. [56] proposed the use of labeled topic models for classifying psychotherapy sessions with 161 possible topic labels, using a dataset published by Alexander Street Press, and again found that these models showed promise in terms of predictive ability.

These earlier studies demonstrate that machine learning systems can generate plausible annotations of medical dialogs—my work in this chapter pursues this line of research further. It differs from earlier work on topic classification in a number of aspects. For example, I

Topic Label	Transcript Text
MusSkePain	MD: Good. Good. Alright . Yeah, the function, uh, the muscle function seems good.
MusSkePain	PT: Mm-hmm.
MusSkePain	MD: We'll see what this shows, okay ?
MusSkePain	PT: Okay.
PhysicalExam	MD: Let's have you stand up. I'm going to do a, uh, excuse me, I'm going to do a, uh, hernia check and prostate exam and well be about done today.
PhysicalExam	PT: Okay. Mm-hmm.
PhysicalExam	MD: And as you may recall, I'm sorry, this is going to be uncomfortable.
PhysicalExam	PT: Yeah. Probably.
PhysicalExam	MD: Please bear with me.
PhysicalExam	PT: Mm-hmm.
WorkLeisure	MD: I'm sorry. So, keeping you busy at work?
WorkLeisure	PT: Yeah. They've been doing that. Actually filming the life of -name-.
WorkLeisure	MD: Oh, really ?
WorkLeisure	PT: They're doing it right now. -name- is doing the, uh, lead part.

Figure 4.1: A short excerpt from an annotated dialog transcript. Topic labels are assigned to each talk-turn. MD and PT indicate the speaker for each talk-turn, where MD stands for "Medical Doctor," "Physician," or "Medical Provider," and PT stands for "Patient."

compare both probabilistic and neural network classification methods for topic classification of talk-turns, whereas Wallace et al. [166] and Gaut et al. [56] only focused on probabilistic approaches. I also evaluate performance for a more detailed set of 27 topics compared to the 6 high-level topics used in the Wallace et al. study. The Gaut et al. study also differs from this work in that it primarily focused on session-level labels and did not investigate the use of sequential information across talk-turns for talk-turn level predictions as it is done here. I believe that this is the first study that systematically compares sequential and non-sequential classification methods, for both probabilistic and neural network models, on the problem of talk-turn topic classification from transcripts of patient-physician dialog.

Short Topic Name	Brief Description	Percentage of Talk-Turns Assigned
BiomedHistory	Biomedical history, symptoms, and medical condition	29.85
PreventiveCare	Preventive medical measures	14.67
MusSkePain	Musculoskeletal pain	8.30
VisitFlowMgmt	Agenda setting, opening of visit, closing of visit	6.38
GynGenitoUrinary	Gynecological and genitourinary problem	4.72
PhysicalExam	Physical exam	3.41
Family	Family and significant other	3.10
HealthCareSystem	Health care system	2.89
WorkLeisure	Work and leisure activities	2.73
Tests	Tests and diagnostic procedures	2.59
Cigarette	Cigarette	2.43
Weight	Weight	2.38
DizzyDentHearVision	Dizzyness, vision, hearing, dental issues	2.03
Other	Other (various rare topics)	1.94
Exercise	Exercise	1.89
Depression	Depression	1.86
Medication	Medications	1.84
SmallTalk	Small talk	1.72
GeneralAnxieties	General anxieties and worries	1.38
MDLife	Physician personal life	1.04
Diet	Diet, food (exclude supplements)	0.69
Alcohol	Alcohol	0.57
Sleep	Sleep	0.53
TherapeuticIntervention	Therapeutic intervention	0.33
RiskyBehavior	Risky behaviors (e.g., international travel, weapons at home) and risk avoidance preventive practices (e.g., safe sex, wearing seatbelt or bike helmet)	0.32
OtherAddictictions	Caffeine, or other addictions	0.21
Age	Age	0.17

Table 4.1: Name and brief description of each topic ordered by the percentage of each topic in talk-turns.

4.2 Dataset

The source data include transcripts of audio-recordings of primary care office visits from the Mental Health Discussion (MHD) Study [149]. Each transcript corresponds to a visit between a patient and a physician—a small fraction of the dialog corresponds to other participants in the conversation (such as a nurse and family member). Data collection occurred from 2007 to 2009 in a health system in Michigan with 26 ambulatory care clinics. Patients were 50 to 80 years old, all had insurance, and were due for a colorectal cancer screening at the time of appointment. All aspects of the research protocol were approved by relevant organizations' institutional review boards (IRB).

Each visit is comprised of a series of talk-turns, with 122,083 talk-turns in total across 279 visits (median/mean of 408/438 talk-turns per visit, with lower/upper quartiles of 312/522) from 59 physicians. Each talk-turn was manually assigned by a human labeler to one of 39 different topic labels [149] that were modified from the Multi-dimensional Interaction Analysis (MDIA) coding system [27]. A topic is defined as an issue raised in a conversation that required a response from the other member of the dyad and had at least two exchanges between the dyad. A small number of talk-turns were split into two if the turn straddled two topics. This resulted in a few of the original talk-turns being represented as two separate talk-turns after coding. Figure 4.1 illustrates how different topics were assigned to talk-turns during a short portion of a particular visit. Topic labels that occurred in less than 20 visits were merged into a single topic denoted as *Other*, resulting in a total of 27 unique topics in the corpus. Table 4.1 provides the names of the topic labels, a brief description of each, as well as the percentage of talk-turns assigned to each by the labelers across the corpus. The topic label distribution is skewed towards topics relevant to periodic health exams—the three most frequent topics (*BiomedHistory*, *PreventiveCare*, and *MusSkePain*) account for more than half of all talk-turns.

a-ha	hm-mm	mm-hmm	uh-hum
able	hmm	mmm	uhhuh
ah	huh	mmm-hmm	uhm
ahh	huh-uh	oh	uhmm
alright	hum	ok	uhuh
blah	isn	okay	um
did	kind	ooh	um-hmm
didn	kinda	really	um-hum
does	know	right	umhum
doesn	let	said	ve
doing	like	say	want
don	little	sort	wasn
done	ll	thing	way
eh	lot	things	yea
feel	m-hmm	think	yeah
going	maybe	u-hum	yep
good	mean	uh	yes
got	mhm	uh-hmm	yknow
guess	mm	uh-huh	yup
hm	mm-hm		

Table 4.2: List of additional stopwords used for generating bag-of-words features.

4.3 Methods

4.3.1 Text Preprocessing

A number of preprocessing steps are applied to convert the dialog text into a set of tokenized words. I first replaced the patient names and numbers with `-NAME-` and `-NUMBER-` tokens to remove potentially identifiable information. After removing symbols, other than a set of punctuation symbols, such as ".", "?", "-", the sentences in each talk-turn were tokenized into a list of words using the standard Python NLTK tokenizer [17]. For the models that use bag-of-words encoding, the vocabulary included all unigrams and bigram noun phrases that occurred at least 5 times in the corpus, except for two sets of stopwords. In addition to the standard stopwords list from the NLTK library, a customized list of stopwords shown in Table 4.2 is used. The list includes the words that are common in a dialog (such as "uh", "umm", etc.) that was adapted from [56]. It also includes variants of standard stopwords that are

due to typos and incorrect punctuation that exist in the data. For the data described in this chapter, removing stopwords for models using bag-of-words encoding typically resulted in a roughly 1% increase in performance on different metrics. The size of the final vocabulary was $V = 14800$. For the neural network models, the vocabulary consisted of all unigrams, with neither set of stopwords removed (as is customary in neural network models) with a final vocabulary size of 5073 including an unknown token.

Different preprocessing pipelines were used for the neural-network models and non-neural models because this is standard practice for both sets of the models in the literature. In particular, for bag-of-words encoding with models such as logistic regression, feature engineering (such as including bigrams and removing stopwords) is widely acknowledged in the literature as useful for improving accuracy [5, 159]. On the other hand, recurrent neural networks learn their own features sequentially in an end-to-end fashion, and the standard approach in recurrent neural models for text is that all word tokens in the sequence are retained (e.g., Collobert and Weston [35]).

4.3.2 Representing Talk-turns for Classification Models

The data for each visit i, $1 \leq i \leq 279$, is represented as a sequence of labeled talk-turns j, with L_i talk-turns in the i-th visit, $1 \leq j \leq L_i$. Let $W_{i,j}$ and $y_{i,j}$ represent the list of word tokens and the topic label, respectively, for the j-th talk-turn in the i-th visit. As mentioned earlier, $y_{i,j}$ can take values from 1 to 27, corresponding to each of the 27 unique topics. Each word in $W_{i,j}$ is encoded as a binary vector ("one-hot encoding") of length V, where V is the vocabulary size. For example, if a word occurs in a talk-turn and the ID in the vocabulary for the word is 10, then the binary/one-hot-encoded vector becomes a vector of length V, where the 10th entry of the binary word vector is set to 1 and all other entries are set to zero.

For the majority of the classification models that are evaluated in this chapter, the binary word vectors in each talk-turn were aggregated into a single talk-turn vector $\boldsymbol{e}_{i,j}$ by adding the individual binary word vectors (also known as a bag-of-words encoding) and re-weighting using tf-idf weights,[1] a common text preprocessing step which downweights common and uninformative words. The vector $\boldsymbol{e}_{i,j}$ represents each talk-turn and has dimensionality equal to the size of the vocabulary $V = 14800$.

For the neural network models, a different representation is used. One vector $\boldsymbol{e}_{i,j}$ is generated per talk-turn using a network composed of an embedding layer and a bidirectional set of gated recurrent units (GRUs). The embedding layer, initialized with pre-trained GloVe [122] vectors, takes each binary word vector in the talk-turn and maps it to a dense embedded vector representation. The GRU component takes the sequence of embedded vectors within a talk-turn (one embedded vector per word) and produces a single fixed-dimensional vector $\boldsymbol{e}_{i,j}$ to represent the talk-turn. This approach has been shown to be useful in NLP applications for encoding variable-length sequential information from words in a sentence (talk-turn) into a fixed-dimensional vector that can be used as a feature vector for downstream classification [60]. The GRU unit size was 128, and the output talk-turn vector $\boldsymbol{e}_{i,j}$ had size 256 since the GRU outputs in both directions are concatenated.

Given the talk-turn vectors $\boldsymbol{e}_{i,j}$, each talk-turn is classified either independently or by using sequential information across talk-turns. Figure 4.2 provides a high-level overview of the three primary different types of models are explored: (i) independent models that classify each talk-turn j only using the words in talk-turn j, (ii) window-based models that also use words from a window of talk-turns both before and after talk-turn j, and (iii) fully sequential models that also consider the topic labels (or predictions) of talk-turns before and after j when predicting a topic for talk-turn j. In addition, another type of sequential model is considered, that uses the talk-turn level GRUs on top of word level GRU outputs. The

[1] "Tf" stands for term frequency, and "idf" stands for inverse document frequency. Chapter 6.5 in Jurafsky and Martin [80] provide more detailed information.

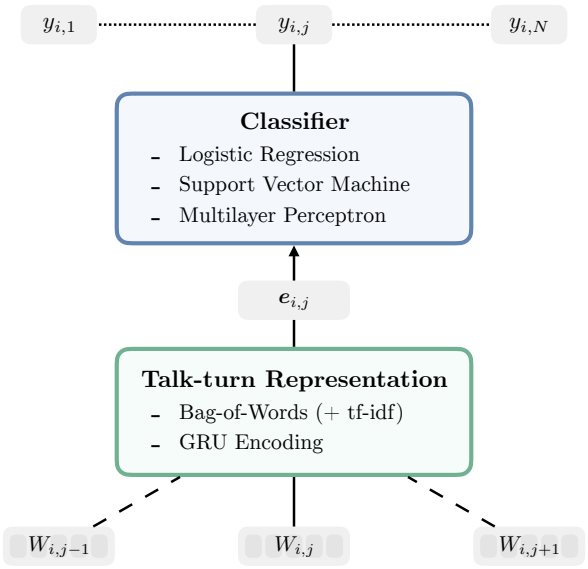


Figure 4.2: High-level diagram of the various models discussed in the chapter. i, j stands for talk-turn j in visit i. $W_{i,j}$ is the list of tokenized words in talk-turn j. For each talk-turn j, I first generate the vectorized talk-turn representation $\boldsymbol{e}_{i,j}$, and the talk-turn representation $\boldsymbol{e}_{i,j}$ is used as an input to different classifiers to predict the topic label $y_{i,j}$, which is the topic label for talk-turn j. Windowed models use adjacent talk-turns to create the talk-turn level representation, and the fully sequential models make use of the sequential dependencies between the topic labels.

model is depicted in Figure 4.3. I describe each of the three approaches in more detail below.

4.3.3 Independent Models

Independent models classify each talk-turn j by using only the words in that talk-turn, $W_{i,j}$, independently of the other talk-turns. The independent models used in my study were the logistic regression (LR) classifiers, support vector machines (SVM), and feedforward neural networks with a single hidden layer. The LR and SVM classifiers used the bag-of-words vectors with tf-idf weights as input to predict the topic label $y_{i,j}$. For the feedforward neural network, the output talk-turn vectors $\boldsymbol{e}_{i,j}$ of the bidirectional GRU units were used as inputs, and the softmax function was used for the activation function in the perceptron. The parameters (weights) of the embedding layer, the GRU, and the feedforward neural network

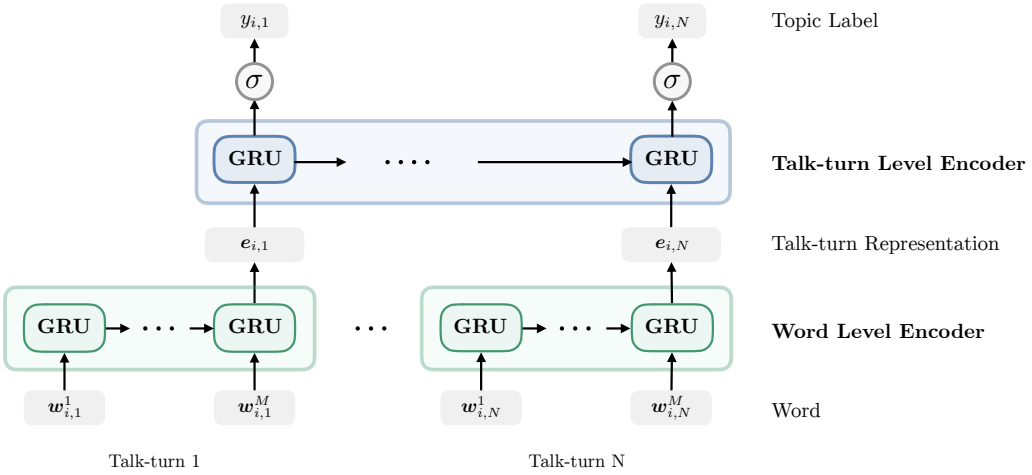


Figure 4.3: Simplified diagram of the Hierarchical GRU (Hier-GRU). Each word in talk-turn j in visit i is fed into the word level encoder to get a talk-turn representation $\boldsymbol{e}_{i,j}$, which becomes the input to the talk-turn level encoder. The embedding layer is omitted for brevity and the embedding of the k-th word in talk-turn j is shown as $\boldsymbol{w}^k_{i,j}$. The model has dependencies at the hidden state of talk-turn level GRUs. Both encoders were bidirectional in the experiments.

were all trained together as a single network, and I refer to this model as "Independent GRU."

4.3.4 Window-based Models

Instead of using a single talk-turn vector $\boldsymbol{e}_{i,j}$ as an input (as in the independent models), for the window-based models, the adjacent M vectors before and after each talk-turn are concatenated. The input vector was defined as $[\boldsymbol{e}_{i,j-M}, \ldots, \boldsymbol{e}_{i,j}, \ldots, \boldsymbol{e}_{i,j+M}]$, with dimension $V \times (2M+1)$, where V is the size of the vocabulary for bag-of-words representations. These windowed input vectors were then used as input to either the LR or SVM classifier. In Figure 4.2, the window-based approach is indicated with dashed lines at the input level.

The window-based representation captures potentially useful sequential information across talk-turns in a manner not available to the independent models—but at the cost of having on the order of $2M$ times as many parameters. The approach is particularly useful when

making topic predictions for short talk-turns with very little information—information from neighboring talk-turns can be used to help to make predictions in such cases.

4.3.5 Fully Sequential Models

To model the sequential dependencies between the topic labels $y_{i,j}$, I used linear chain conditional random fields (CRFs) and hidden Markov models (HMMs). The linear chain CRF is widely used for sequence labeling tasks such as named-entity recognition or part-of-speech tagging [166, 88]—here I applied it to the problem of predicting the topic label of a talk-turn, given a sequence of talk-turns. The HMMs are constructed by using the output class label probabilities from each of the independent models discussed above. I converted the class probabilities from the classifiers, $p(y_{i,j} = k|W_{i,j})$, to emission probabilities, $p(W_{i,j}|y_{i,j} = k)$ (needed by the HMM), by using the fact that $p(W_{i,j}|y_{i,j} = k) \propto p(y_{i,j} = k|W_{i,j})/p(y_{i,j} = k)$, given that the talk-turn probabilities $p(W_{i,j})$ do not depend on a particular topic k. The emission probabilities are combined with the Markov transition probabilities (which can be directly estimated from label sequences in the training data) via the Viterbi algorithm to compute the sequence of topic labels that has the highest joint probability across talk-turns, conditioned on the observed talk-turn words.

I also found that using speaker-specific transition matrices improved the accuracy of the sequential models. Not surprisingly, physicians tend to start new topics during a conversation more than the patients do. Figure 4.4 shows the percentage of time that a particular speaker (physician, patient, or other) starts each topic. To incorporate speaker information in the HMM approach, I augmented the standard HMM to use two topic transition matrices, one for the physician and one for the patient or other speakers. Each transition matrix corresponds to the speaker of the state that the HMM is transitioning to (e.g., one transition matrix for transitioning to physician, and the other for transitioning to all others). The decoding

process in the Viterbi algorithm is modified so that it uses the appropriate matrix depending on the speaker.

Another type of sequential model that is entirely neural-network-based does not have direct dependencies between the topic labels, but has bidirectional connections between the hidden states of the talk-turn level GRUs. The model, which is referred to as "Hier-GRU," has a hierarchical structure having two different GRUs, one at the word level to generate talk-turn level representation, and the other which takes the talk-turn vectors as inputs to predict the output label $y_{i,j}$ for each talk-turn j.[2] Similar to Independent GRU, the talk-turn level GRU output is connected to a fully-connected layer and then a softmax to make prediction. Cross-entropy loss is used for end-to-end training.

4.4 Experiments and Results

4.4.1 Experimental Details

All models are evaluated using 10-fold cross-validation, and the evaluation metrics are computed both at the talk-turn level and at the visit level. At the talk-turn level, I computed the classification accuracy by comparing (a) the predicted topic from a model with (b) the human-generated topic for the talk-turn. To obtain results at the visit level, I aggregated the predictions from the individual talk-turns within each visit to generate a visit-level binary-valued prediction vector of dimension 27 (the number of topic labels) with 1 for topic label k if the model predicted topic k for one or more talk-turns in the visit, and 0 otherwise. Using such a vector for each visit, I calculated the accuracy, precision, recall, and F1 scores. The metrics were micro-averaged by globally counting the true positives, false positives, etc. for

[2]This type of structure has been used for dialog generation (e.g., Serban et al. [132]) as well as query suggestion (e.g., Sordoni et al. [140]).

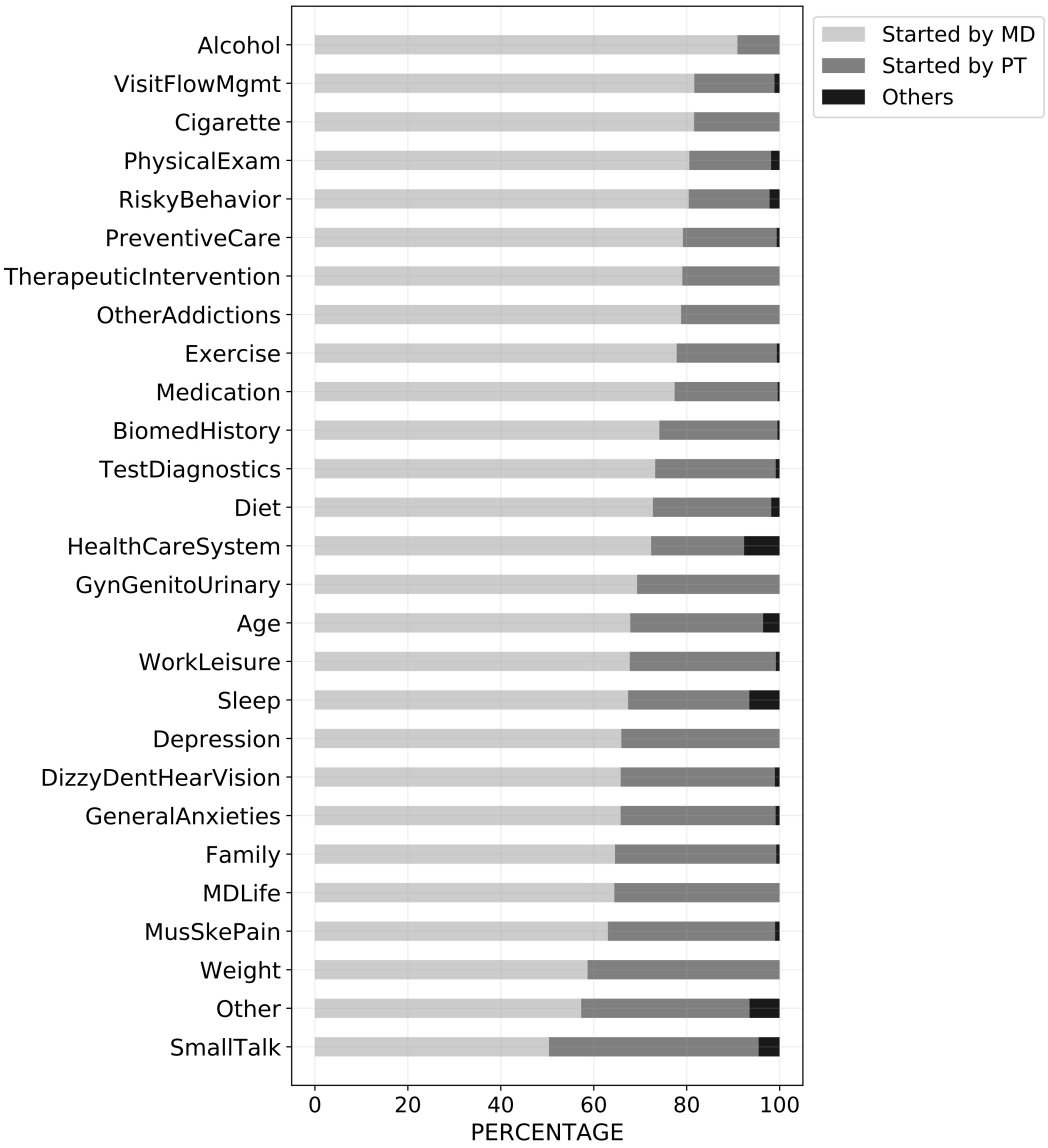


Figure 4.4: Percentage of time that a particular speaker generates the first talk-turn in a sequence of talk-turns for each topic. Physicians (MD, medical doctors) tend to start a new topic more often than patients (PT) in general—49.9% and 48.2% of the talk-turns were generated by physicians and patients, respectively. In particular, more than 80% of the conversations about *Alcohol* and *Cigarette* were started by the physicians, whereas the percentages of both speakers starting casual conversations (*SmallTalk*) were almost equal.

all the topic labels.

To evaluate the performance of the classifiers, I compared the results with those of simple baseline models that predict the most common topics. The baseline at the talk-turn level just predicts the most common topic in the corpus, *BiomedHistory* (as shown in Table 4.1). At the visit level, the baseline always predicts the set of topic labels that occur in 50% or more of all visits, irrespective of the words within each visit.

Implementation Details

- **LR and SVM**

  For the logistic regression, the Python Scikit-learn library [121] is used with the one-versus-rest option for multi-class classification. The inverse of the L2 regularization strength (`C`) was set to 1.6, which was the best performing value selected for a single fold in the 10-fold cross-validation data. Scikit-learn library is again used for the support vector machine (SVM), with RBF kernel and the one-versus-one decision function. The inverse regularization was selected to have the number 100, also optimized using the same fold.

- **Windowed-based Models**

  The M parameter for the windowed model was set to $M = 2$ in the experiments based on the knowledge that the transcripts were labeled with the convention that each topic label must span at least two exchanges (four talk-turns).

- **CRF**

  The java-based package MALLET [104] with default parameter settings was used. The parameters used in Wallace et al. [166] produced less accurate predictions for this dataset.

- **HMM**

Python code[3] is implemented for both parameter estimation (transition probabilities, marginal probabilities, and etc.) using the training data and for generating talk-turn topic predictions using Viterbi decoding with the test data.

- **GRU (Independent model)**

  The neural network models are implemented using AllenNLP library [55], which is based on PyTorch. The word embedding layer had size 100, which was initialized using 100-dimensional GloVe[122] vectors that were pre-trained with Wikipedia and Gigaword. I selected the size of the bidirectional GRU, the learning rate, and the dropout rate for each layer by conducting a grid-search over those parameters using one of the 10 folds used for cross-validation. The value I used for GRU size was 128, word embedding dropout rate was 0.3, GRU dropout rate was 0.5 and the final softmax layer dropout was 0.3. I used Adam [83] to optimize the network with an initial learning rate of 0.001.

- **Hier-GRU**

  Similar to the independent GRU model, both the word level and the talk-turn level GRU encoders used size 128 with dropout 0.5, and were both bidirectional. All other dropout rates including the dropout at word embedding layer were 0.2, and the model was also trained with Adam optimizer starting with a learning rate of 0.001.

4.4.2 Summary of Experimental Results

Table 4.3 shows the classification accuracies at the talk-turn level for independent models, windowed models, and sequential models. The most accurate independent model is the GRU and the most accurate windowed model is the Windowed LR. The models with sequential information clearly outperform independent models with the Hier-GRU yielding the highest

[3]Available at `https://github.com/UCIDataLab/PP_dialog_models`.

Model		Talk-turn Level Accuracy (%)
Baseline		29.85
Independent Models	LR	37.00
	SVM	36.64
	GRU	**38.85**
Window-based Models	Windowed LR	**51.12**
	Windowed SVM	50.46
Fully Sequential Models	CRF	48.37
	HMM-LR	56.89
	HMM-SVM	51.52
	HMM-GRU	**57.60**
	Hier-GRU	**61.77**

Table 4.3: Accuracies for topic prediction at the level of talk-turns for different prediction models. Micro-averaged precision and recall scores are the same as accuracy.

Model		Visit Level (%)			
		Accuracy	Precision	Recall	F1
Baseline		72.29	73.79	62.22	67.42
Independent Models	LR	75.19	67.91	84.25	**75.15**
	SVM	72.45	63.40	90.40	74.50
	GRU	72.47	64.19	86.59	73.68
Window-based Models	Windowed LR	77.28	69.82	86.47	77.21
	Windowed SVM	79.81	75.06	82.13	**78.37**
Fully Sequential Models	CRF	74.19	80.43	58.42	67.64
	HMM-LR	80.00	80.16	73.31	76.55
	HMM-SVM	75.21	78.70	60.90	68.63
	HMM-GRU	9.00	74.98	79.35	**77.06**
	Hier-GRU	78.96	73.69	82.43	**77.78**

Table 4.4: Micro-averaged accuracy, precision, recall, and F1 scores, at the visit level, for different prediction models.

Label	Hier-GRU			HMM-GRU			Assigned
	Precision	Recall	F1	Precision	Recall	F1	Topic (%)
BiomedHistory	65.80	76.61	70.79	74.24	56.26	64.01	29.85
PreventiveCare	73.34	83.41	78.05	77.98	72.49	75.13	14.67
MusSkePain	67.48	69.14	68.30	67.68	64.27	65.93	8.30
VisitFlowMgmt	63.64	63.58	63.61	64.97	64.54	64.76	6.38
GynGenitoUrinary	67.62	54.76	60.51	72.15	57.66	64.09	4.72
PhysicalExam	48.91	52.33	50.56	50.21	63.20	55.96	3.41
Family	49.31	47.56	48.42	42.79	54.21	47.83	3.10
HealthCareSystem	37.81	28.85	32.73	46.10	39.49	42.54	2.89
WorkLeisureActivity	47.26	46.20	46.72	52.90	53.14	53.02	2.73
TestDiagnostics	49.67	32.43	39.24	37.24	52.88	43.70	2.59
Cigarette	71.38	85.84	77.94	81.38	72.73	76.81	2.43
Weight	49.54	52.60	51.02	47.20	46.77	46.98	2.38
Other	21.42	9.59	13.25	14.45	15.20	14.81	2.03
Depression	50.82	64.27	56.76	45.71	44.24	44.96	1.94
DizzyDentHearVision	23.08	14.01	17.44	46.96	50.83	48.82	1.89
Medication	38.97	27.09	31.96	22.46	64.29	33.29	1.86
Exercise	56.94	59.23	58.06	45.42	67.01	54.14	1.84
SmallTalk	20.61	18.70	19.61	32.79	31.09	31.92	1.72
GeneralAnxieties	32.51	18.63	23.68	16.36	32.86	21.84	1.38
MDLife	33.81	14.19	19.99	12.79	11.69	12.22	1.04
Diet	27.15	19.62	22.78	27.52	57.10	37.14	0.69
Alcohol	56.08	36.12	43.94	52.19	64.91	57.86	0.57
Sleep	13.40	2.33	3.97	29.80	40.32	34.27	0.53
TherapeuticIntervention	0.00	0.00	0.00	2.04	5.30	2.95	0.33
RiskyBehavior	33.90	5.87	10.00	44.87	41.06	42.88	0.32
OtherAddictions	0.00	0.00	0.00	23.55	41.40	30.02	0.21
Age	0.00	0.00	0.00	13.23	20.00	15.93	0.17

Table 4.5: Precision, recall, and F1 scores of each topic, calculated at the talk-turn level using Hier-GRU and HMM-GRU prediction results. The rows are sorted by the percentage of talk-turns of each topic. In general, the more frequently discussed topics have higher F1 scores. The five highest F1 scores for each model are highlighted in bold-faced text.

accuracy of 61.77% over all the models. The improvement in accuracy of Hier-GRU and HMM-GRU are both statistically significant, with $p < 0.01$, relative to each of the Independent GRU and Windowed LR models (using dependent t-tests for paired samples across the 10 folds of cross-validation). The visit-level evaluation scores are shown in Table 4.4. Interestingly, the gap in performance (as measured by the micro-averaged F1 score) between independent, windowed, and sequential models is much smaller in the visit-level scores. The primary reason for this is that the independent models tend to have relatively high recall scores, whereas sequential models have relatively high precision scores. The prediction models also can be evaluated at the level of individual topics to understand variability in prediction accuracy across topics. Precision, recall, and F1 scores were calculated by treating each topic label separately at the talk-turn level. Scores from the two best performing models (Hier-GRU and HMM-GRU) are shown in Table 4.5, sorted by the percentage of each topic. The more common topics (that occur for example in at least 5% of the talk-turns) generally have higher F1 scores. However, there are some less common but highly-specific topics, such as the *Cigarette* topic (in 2.43% of talk-turns), that also have relatively high F1 scores.

4.5 Discussion

Using sequential information across talk-turns in predictive models systematically leads to more accurate predictions, particularly when predicting topic labels for talk-turns. To illustrate this point in more detail, Figure 4.5 shows sequences of predicted and human-assigned topic labels, for one particular visit, where different colors represent different topic labels. The top plot is from the Independent GRU model. The lack of sequential information in the model leads to predictions that are noisy and lack the sequential smoothness of the human-assigned labels (bottom plot). The second plot is from the Hier-GRU model, and the third

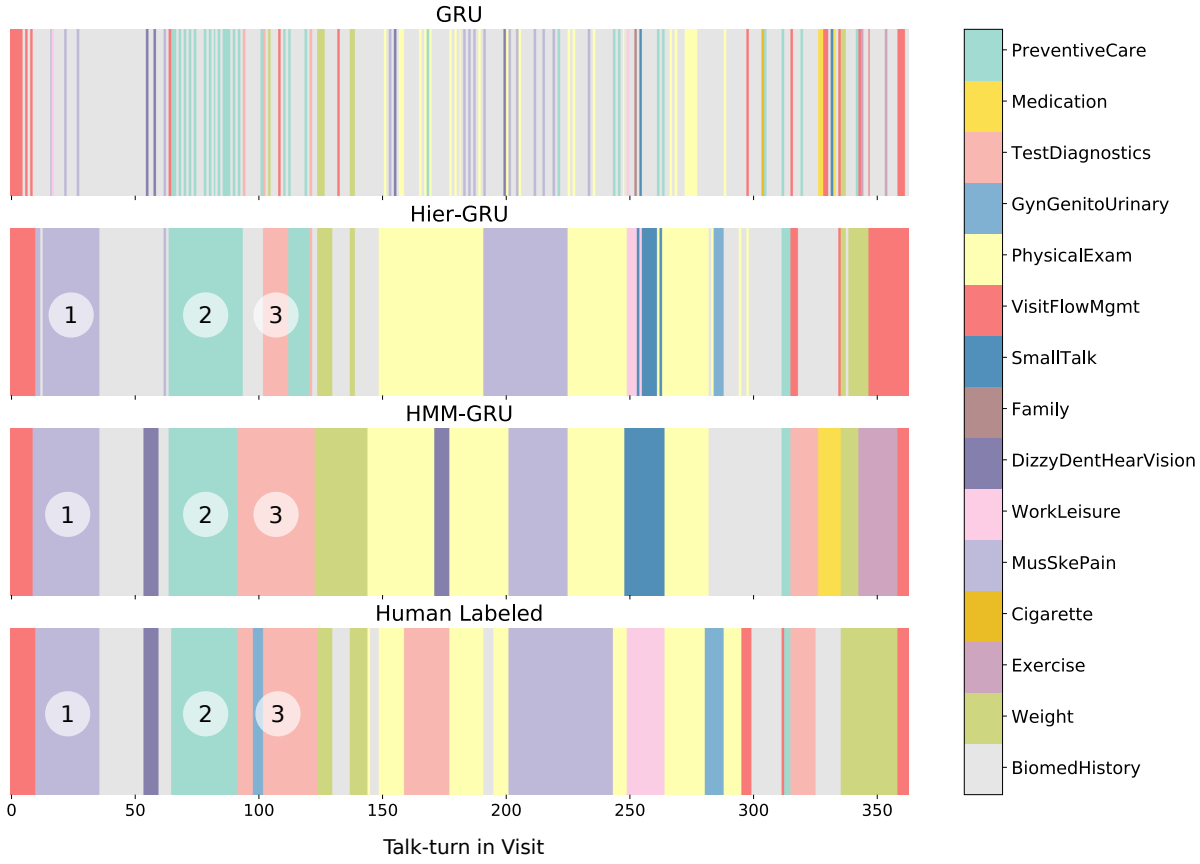


Figure 4.5: Sequences of color-coded topic labels for one of the visits in the dataset. The upper plot shows the predicted topic labels from an independent model, and the center two plots show those from fully sequential models. The lower plot corresponds to the human-generated labels. The segments for the *MusSkePain* topic marked as (1) had lengths of 23 talk-turns for Hier-GRU, 27 for HMM-GRU, and 26 for Human labeled. Similarly, the *PreventiveCare* segments (2) had lengths 30, 28, and 27, and the *TestDiagnostics* segment (3) had lengths 10, 31, and 22 in talk-turns, respectively for Hier-GRU, HMM-GRU, and Human labeled.

Talk turn ID	GRU Label	H-GRU Label	HMM + GRU Label	Human Label	Transcript Text
2	VisitFlowMgmt	VisitFlowMgmt	VisitFlowMgmt	VisitFlowMgmt	MD: Hello, -name-. How are you today?
3	VisitFlowMgmt	VisitFlowMgmt	VisitFlowMgmt	VisitFlowMgmt	PT: Good, thanks. Good.
4	VisitFlowMgmt	VisitFlowMgmt	VisitFlowMgmt	VisitFlowMgmt	MD: Good, good, good.
5	BiomedHistory	VisitFlowMgmt	VisitFlowMgmt	VisitFlowMgmt	PT: Yeah.
6	VisitFlowMgmt	VisitFlowMgmt	VisitFlowMgmt	VisitFlowMgmt	MD: Doing your, uh, your physical today?
7	BiomedHistory	VisitFlowMgmt	VisitFlowMgmt	VisitFlowMgmt	PT: yeah.
8	VisitFlowMgmt	VisitFlowMgmt	VisitFlowMgmt	VisitFlowMgmt	MD: Okay. Very good. Um, well, uh, let's go over things, then, um, have some specific things you wanted to go over today.
9	BiomedHistory	VisitFlowMgmt	MusSkePain	VisitFlowMgmt	PT: Uh-uh.
10	BiomedHistory	MusSkePain	MusSkePain	MusSkePain	MD: And well, so a problem with your foot there?
11	BiomedHistory	MusSkePain	MusSkePain	MusSkePain	PT: Right, the left one.
12	BiomedHistory	BiomedHistory	MusSkePain	MusSkePain	MD: Okay. What's been happening?
13	BiomedHistory	MusSkePain	MusSkePain	MusSkePain	PT: Not much. I think I had a little fracture in it, and then, but the, uh, little toe and the one next to it still feel a little, little numbness in it.
14	BiomedHistory	MusSkePain	MusSkePain	MusSkePain	MD: Really?
15	BiomedHistory	MusSkePain	MusSkePain	MusSkePain	PT: Right. But that's been, like, over a couple months ago.
16	MusSkePain	MusSkePain	MusSkePain	MusSkePain	MD: How did you injure your foot?

Figure 4.6: The beginning part of the visit shown in Figure 4.5. Each talk-turn is presented with predicted labels from three different models (Independent GRU, Hier-GRU, and HMM-GRU) and the human-generated labels. For the short talk-turns the *BiomedHistory* topic label is predicted quite often by the Independent GRU, while the two other models produce label sequences that are more similar to human-generated labels.

Talk turn ID	GRU Label	H-GRU Label	HMM + GRU Label	Human Label	Transcript Text
221	PreventiveCare	MusSkePain	MusSkePain	MusSkePain	MD: But it's not a comfortable test because these needles electric shocks.
222	BiomedHistory	MusSkePain	MusSkePain	MusSkePain	PT: Yeah.
223	BiomedHistory	MusSkePain	MusSkePain	MusSkePain	MD: But um, well see if theres evidence of nerve damage there, okay?
224	BiomedHistory	MusSkePain	MusSkePain	MusSkePain	PT: Okay.
225	PhysicalExam	PhysicalExam	PhysicalExam	MusSkePain	MD: Um, let's see. Let's have you lift your knee up off the table, please.
...					
233	MusSkePain	PhysicalExam	PhysicalExam	MusSkePain	MD: Your feet up. Bend your feet up at the ankle. Bend your feet up at the ankles, like this.
234	BiomedHistory	PhysicalExam	PhysicalExam	MusSkePain	PT: Oh, like that?
235	PhysicalExam	PhysicalExam	PhysicalExam	MusSkePain	MD: Yeah. Push your feet down.
236	BiomedHistory	PhysicalExam	PhysicalExam	MusSkePain	PT: Down ?
237	BiomedHistory	PhysicalExam	PhysicalExam	MusSkePain	MD: Like you're stepping on the gas.
238	BiomedHistory	PhysicalExam	PhysicalExam	MusSkePain	PT: Oh, okay.
239	BiomedHistory	PhysicalExam	PhysicalExam	MusSkePain	MD: Good. Good. Alright. Yeah, the function, uh, the muscle function seems good.
240	BiomedHistory	PhysicalExam	PhysicalExam	MusSkePain	PT: Mm-hmm.
241	BiomedHistory	PhysicalExam	PhysicalExam	MusSkePain	MD: Well see what the shows, okay?
242	BiomedHistory	PhysicalExam	PhysicalExam	MusSkePain	PT: Okay.
243	PreventiveCare	PhysicalExam	PhysicalExam	PhysicalExam	MD: Let's have you stand up. I'm going to do a, uh, excuse me, I'm going to do a, uh, hernia check and prostate exam and well be about done today.
244	BiomedHistory	PhysicalExam	PhysicalExam	PhysicalExam	PT: Okay. Mm-hmm.
245	PreventiveCare	PhysicalExam	PhysicalExam	PhysicalExam	MD: And as you may recall, I'm sorry, this is going to be uncomfortable.
246	BiomedHistory	PhysicalExam	PhysicalExam	PhysicalExam	PT: Yeah. Probably.
247	PhysicalExam	PhysicalExam	PhysicalExam	PhysicalExam	MD: Please bear with me.
248	BiomedHistory	PhysicalExam	SmallTalk	PhysicalExam	PT: Mm-hmm.
249	WorkLeisure	WorkLeisure	SmallTalk	WorkLeisure	MD: I'm sorry. So, keeping you busy at work?
250	BiomedHistory	WorkLeisure	SmallTalk	WorkLeisure	PT: Yeah. They've been doing that. Actually filming the life of -name-.
251	BiomedHistory	WorkLeisure	SmallTalk	WorkLeisure	MD: Oh, really?
252	Family	WorkLeisure	SmallTalk	WorkLeisure	PT: They're doing it right now. -name- is doing the, uh, lead part.
253	BiomedHistory	SmallTalk	SmallTalk	WorkLeisure	MD: Oh, really?
254	SmallTalk	WorkLeisure	SmallTalk	WorkLeisure	PT: They restructured the whole hospital and put it back in the -num- s and it's really nice. That's what's going on now.

Figure 4.7: Another excerpt from the same visit in Figure 4.5. Topics that are semantically similar are confusable (*PhysicalExam* and *MusSkePain* in talk-turns 233–242, and *SmallTalk* and *WorkLeisure* in talk-turns 249–254).

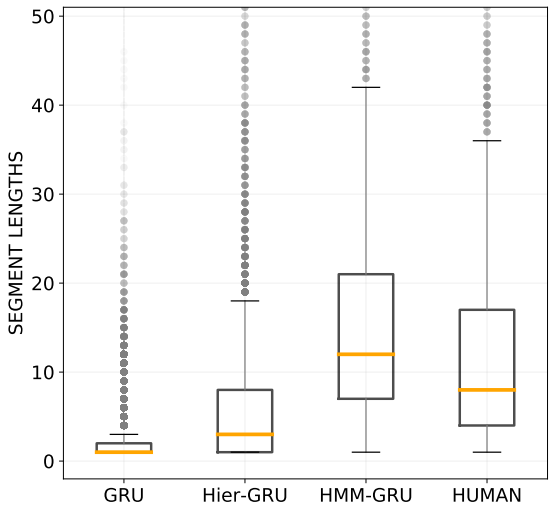


Figure 4.8: Boxplots of segment lengths from four different models for all 279 sessions.

plot shows the predicted sequence of topics from the Viterbi parse of the HMM-GRU model, i.e., the probabilistic predictions from the same GRU model in the top plot but which are now sequentially smoothed by the HMM transition matrices. It is visually apparent (not only for this visit but for all visits) that the sequential models (second and third) are much more similar to the human labeling (bottom) than the independent model (top).

Figures 4.6 and 4.7 provide a more detailed look at portions of the transcript corresponding to Figure 4.5. For example, there are quite a few short talk-turns that have no words with topic-relevant information, such as "Yeah" (talk-turns 5 and 7 in Figure 4.6 and 222 in Figure 4.7) and "Okay" (talk-turn 224 in Figure 4.7). The sequential models are able to use the context information to assign these talk-turns to the same topic as the human labeler. The GRU independent model, however, does not have any context and assigns these talk-turns by default to the topic with the highest marginal probability (*BiomedHistory*).

While the smoothing in sequential models helps to improve prediction accuracy, it can also produce errors due to over-smoothing. In particular, I found that HMM-GRU model tends to predict longer topic segments relative to the human-labeled results, as can be seen in talk-turns 100 to 200 of the visit in Figure 4.5. The human labels contain short bursts

of topics *GynGenitoUrinary* and *BiomedHistory* that are not detected by the HMM-GRU model. Figure 4.8 shows the boxplots of the topic segment lengths for the four sequences of labels presented in Figure 4.5. Overall, the independent GRU tends to generate the shortest segment lengths, and the HMM-GRU tends to generate the longest segment lengths. The median segment length of the human-labeled topic sequences lies between those from Hierarchical GRU (Hier-GRU) and HMM-GRU. The overall distribution of the topic segment lengths of HMM-GRU predictions is better matched, than other mtehods, to the length distribution of human-generated topic labels. This is further quantified by the visit-level results in Table 4.4, where the recall scores of the fully sequential models are systematically lower than the independent models, and the reverse for the precision scores.

I also observed that some topics are semantically similar and easily confusable. For example, in Figure 4.7, from talk-turn 233 to 242, the two sequential models predict the topic *PhysicalExam*, while the human labeled *MusSkePain*—from the corresponding transcript text either prediction seems reasonable. Similarly, from talk-turn 249 to 254 the HMM-GRU predicts *SmallTalk*, while the human labeled *WorkLeisure*—from the text the corresponding talk-turns appear to be a mixture of both. Other examples were found across the corpus where the model frequently gets confused among small groups of related topics (e.g., *GeneralAnxieties* and *Depression*; *Weight* and *Diet*). The full confusion matrices are shown in Figure 6. There is inevitably a subjective aspect to the human labeling, suggesting that there is likely to be a performance ceiling in terms of the accuracy of any algorithm relative to human labels on this data.

From the topic-specific results in Table 4.5, we can see that while the predictions are relatively accurate for some topics, for others (e.g., *Age*, *TherapeuticIntervention*, *OtherAddictions*, *Other*, *MDLife*, etc.), the scores are quite low. The broad nature of these topics is a likely contributor to the low accuracies, but the relative lack of training data per topic may also be another contributing factor. These topics account for roughly 1% (or less) of talk-turns in the

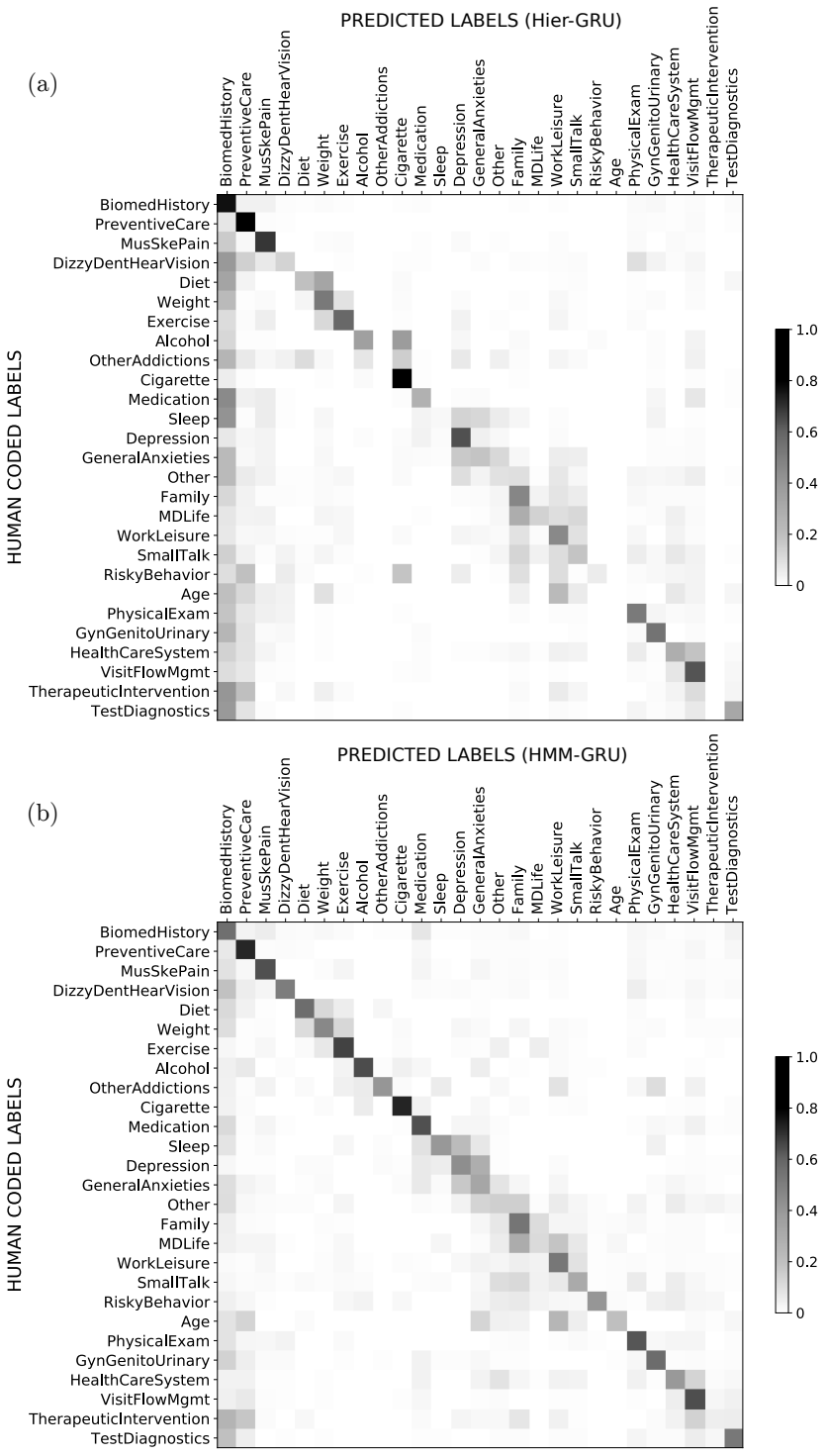


Figure 4.9: Confusion matrices generated by (a) Hier-GRU and (b) HMM-GRU, where the intensity of each cell shows the conditional probabilities of $p($ predicted label | human-generated label $)$ and each row sums to 1. A number of subsets of topics have high confusion probabilities, including *Diet/Weight/Exercise*, *Depression/GeneralAnxieties*, and *Family/MDLife/WorkLeisure*.

corpus and many of these talk-turns are relatively short with little topical content, leading to relatively less signal, particularly for training neural network models. One possible approach to improve accuracy would be to incorporate additional external information relevant to these topics, such as incorporating lists of relevant words from ontological sources such as UMLS (Unified Medical Language System) into the training of prediction models, and/or adding relevant information from sources such as physician or specialist notes.

4.6 Conclusions

Patient-physician communication is an essential component of healthcare. In this context, prediction models that annotate patient-physician transcripts can in principle provide both useful information about the nature of topics discussed in specific conversations as well as contribute to a broader understanding of patient-physician communication. The results demonstrate that machine learning methods show promise for building models that can automatically predict discussion topics in dialog at the talk-turn and visit level. In particular, using a large real-world patient-physician dialog corpus, I investigated the performance of a variety of classification models including logistic regression, support vector machines, feed-forward neural network model with GRU, hierarchical GRU, conditional random fields, and hidden Markov models. I found that sequential models (such as Hier-GRU and HMM-GRU) are more accurate compared to non-sequential models for predicting topic labels for talk-turns. In addition, I found that semantic similarity of discussion topics can be a significant contributor to prediction error.

While additional research and model improvement is needed, my results show promise for a number of medical topics that are critical quality indicators in primary care (e.g., cigarette smoking, pain). Potential applications might include exploring systems that incorporate prior information from a list of problem areas or prior diagnoses found in the medical record.

For example, the presence or absence of smoking cessation counseling during primary care encounters may inform population health management programs aimed at helping patients quit smoking. Deployment of systems like this in real-world primary care may also be useful for obtaining the scale of data needed to improve model performance.

The primary contributions of this chapter include the following:

- I trained different types of machine learning models for automated annotation of medical topics in patient-provider dialog transcripts by categorizing the models into independent, window-based, and fully sequential models.
- I evaluated and compared the models' predicted results both at the visit-level and at the talk-turn level with the baseline scores.
- I provided evidence that the sequential or contextual information is crucial in talk-turn level topic classification in conversation data.
- I suggested challenges and possible future directions in detecting topical content in medical dialogs.

Chapter 5

Evaluating Cloud Speech Recognition Engines for Topical Analysis of Clinical Conversations

Appropriate documentation of the clinical visit is critical for communication among medical professionals [67, 137], enabling quality assurance [38], and accurate billing and reimbursement [75]. The traditional way of documenting a clinical visit in the Electronic Health Record (EHR), namely physicians' notes, provide a source of valuable information on what occurred during the interaction and what physicians consider to be important. EHRs have improved the accessibility of medical information [171], but patients are demanding access to information at a greater scale [138]. Pressure to quickly document the medical visit may lead physicians to type into the record during the medical visit, which can negatively impact patient-physician communication [133, 162, 139]. Primary care physicians spend about half of their time working on computers [6, 152], which may be partly responsible for growing concerns of physician burnout across a wide range of physician specialties [53, 135, 148]. In addition, physician-generated notes do not always provide an accurate representation of

what occurred during the visits [64, 157].

Working with medical scribes could alleviate the problem. Medical scribes are individuals who primarily enter notes during the patient encounter on behalf of the physician, allowing physicians to sidestep much interaction with their EHRs [136]. However, employing medical scribes are associated with costs, often limiting their use to high-volume specialty clinics [7, 165].

Technologies that could reduce the burden of documentation on physicians, and provide more accurate documentation of clinical interactions are greatly needed. Natural language processing (NLP) technologies combined with advances in automatic speech recognition (ASR) [29, 128] offer potentially promising solutions [10, 47]. Information extraction and summarization technologies built on top of resulting transcripts will be needed to take the next step in reducing the burden of documentation on physicians and in providing clinical decision support to both patients and physicians [128]. If successful, ASR with machine-learning-enabled charting system could free up valuable time for physicians to talk to their patients rather than typing extensively during clinical encounters [68, 163].

The work discussed in this chapter takes a step forward in the automation of documenting the patient-physician conversations, by evaluating the topic classification models with two types of input—the human-transcribed and the ASR-transcribed—and examine the drop in performance at the output of the ASR and the output of the model. Moreover, I compare the results with different modalities of the recording devices and readily available speech recognition systems to understand the variability in performance.

5.1 Materials and Methods

5.1.1 Dataset

The same data used in Chapter 4 is used in this chapter. The dataset consists of 279 transcripts from the Mental Health Discussion (MHD) study [149], where one of 27 topic labels is assigned to each talk-turn. Detailed information can be found in Section 4.2. Out of 279 transcripts of visits, 36 visits are randomly selected as a test set.

Since the actual audio or video files from the original study are not available, additional steps for obtaining the audio files are necessary to create ASR-generated transcripts. Therefore, two readers, who are affiliated with the University of California, Irvine, read the transcripts from the test set with a set of rules and settings. I describe the details below.

Pre-processing and De-identification

To ensure Protected Health Information (PHI) is not transferred to the selected cloud-based ASR engines, additional pre-processing steps were completed in accordance with the HIPAA (Health Insurance Portability and Accountability Act) privacy rules. The transcripts were pre-processed in two stages: 1) automatic processing for all the transcripts and 2) additional manual replacements for the test set (36 visits).

First, I replaced a list of names, locations, and patient related information that are identifiable in the dataset with generic tokens (e.g. `-NAME-`, `-LOCATION-`, `-PATIENT-ID-`). Some of the names and numbers were already de-identified in the original transcripts, however, many parts of the transcripts still included sensitive information. Including the names of the 50 states, about 500 cities in the United States, and the 400 most popular baby names in 1950s[1]

[1] `https://www.ssa.gov/oact/babynames/decades/names1950s.html`

are searched and replaced with non-PHI tokens. Also, there existed some other information in the original transcripts that can be helpful for the readers, such as door closing sound, laughing, and etc. These are detected using regular expressions and also converted to tokens enclosed with hyphens, such as `-DOOR-SOUND-` and `-LAUGH-`. The list of all the tokens are presented in Appendix A. Note that the non-verbal tokens are only to help the readers, and are removed for training and testing.

Secondly, the readers of the transcripts replaced the tokens with the words that are contextually appropriate. For example, `-NAME-` to "John Doe" and `-LOCATION-` to "downtown." Since not all the information was captured by the first automated pre-processing step, the readers manually examined the processed transcripts, overwriting any identifiers not previously replaced with generic tokens, based on the Safe Harbor Method [117]. Note that the "human-generated transcripts" in the later sections refer to the transcripts after de-identification.

Recording Setup

After removing PHI from the transcripts of the selected test visits, the readers read the transcripts for audio recording. To recreate audios that are faithful to the actual test visits, the readers sought to match disfluencies, generic words, and non-grammatical utterances present in the human-generated transcript. Additional information on the reading protocol is in Appendix B.

Multiple recording systems were chosen by considering the accessibility to the average physician (or medical providers in general) in terms of complexity and cost. Five different microphones were selected and placed approximately one foot away from the readers. Selected sources of microphones were 1) external high-end microphone, 2) omni-directional lavalier microphone, 3) cell phones, 4) voice recorder, and 5) laptop. The list of recording devices

Type	Model	Settings	Number of Recorded Visits
External High-end Microphone	Blue Yeti	Omni-directional pickup, 50% Gain, WAV format, 24-bit, 44100Hz	36
Omni-Directional Lavalier	Aputure A.lav	Gain increased to match high-end microphone in software, clipped on the reader playing the 'physician' role, WAV format, 24-bit, 44100Hz	36
Cell Phone	Samsung Galaxy S7	"Smart Recorder - High-quality voice recorder" app; WAV format, 24-bit, 44100Hz	32
	iPhone X	"Voice Memo" app, M4A format, lossless recording quality	4
Voice Recorder	Sony ICD-UX533	WAV- LPCM format, 16-bi, 44100Hz	36
Laptop	Macbook Pro 2016	"Audacity" app, WAV format, 24-bit, 44100Hz	36

Table 5.1: Microphone and recording equipment information.

and the number of visits recorded by each device is in Table 5.1. Two different cell phones were used for different visits but are treated as the same source for simplicity. Audios were all recorded with a sampling rate of 44.1 kHz.

5.1.2 Automatic Speech Recognition (ASR) engines

Table 5.2 shows different types of speech recognition engines that were used for comparison. All ASR engines are cloud-based APIs that are easily accessible to the general public.

Amazon Transcribe from AWS (Amazon web services), Google Cloud Speech-to-text, and Microsoft Azure Speech-to-text were selected for the experiment. For Google Cloud Speech-to-text, there are multiple models available. Three models were chosen, where one of them (Video) is a premium model with an extra cost. Each service had additional functionalities available, such as speaker diarization, however, those options are ignored in the experiment since not all systems provide the same type of information.

Short Name	ASR Engine	Model	Version
Amazon	Amazon Transcribe	Default	1.9.86
Google_Default	Google Cloud Speech-to-text	Default	v1p1beta1
Google_Phonecall	Google Cloud Speech-to-text	Phone call	v1p1beta1
Google_Video	Google Cloud Speech-to-text	Video (premium)	v1p1beta1
Microsoft	Microsoft Azure Speech-to-text	Default	1.5.1

Table 5.2: Automatic Speech Recognition engines used for the study. The version for Amazon Transcribe is the version number of Boto3 (AWS SDK for Python).

Our team also tried AutoTranscribe dictation service from Nuance Dragon Medical Practice Edition, but the results are not included since the service is not optimized for multiple speakers' dialog and did not yield results comparable to the other ASR systems.

Alignment of ASR-generated Transcripts

Aligning the ASR-generated text to the human-generated transcript is a required step to measure the performance. It also helps standardize the ASR outputs since every ASR engine output has a different structure and information. Alignment is done in the standard manner using the backtrace path information while obtaining the edit distance [164] at the word level between the ASR-generated and the human-generated text, with the boundary token inserted at the end of each talk-turn in the human-generated transcript. After locating the talk-turn boundaries in the ASR output, additional talk-turn level information, such as speaker information and the topic label, can be utilized.

5.1.3 Classification model

Gated recurrent units with hierarchical structure (Hier-GRU) from Chapter 4 were used to extract the topic information from the transcripts of talk-turns. The Hier-GRU was selected since it is the best performing model for the same dataset from the experiments in Chapter 4. The Hier-GRU model has two levels of bi-directional gated recurrent units (GRUs), one

at the word level to get a talk-turn representation, and the other at the talk-turn level for incorporating the interactions between the talk-turns and to predict the topic label. A talk-turn is composed of a sequence of words, and 100-dimensional pre-trained GloVE [122] vectors are used to initialize the word embeddings. See Figure 4.3 for illustration. I used the same hyperparameter settings as in Chapter 4.

5.2 Evaluation Methods

5.2.1 Word Error Rate (WER)

For each recorder and ASR type, the average word error rate is calculated between the transcript after removing PHI and the output from the ASR. Word error rate (WER) is a commonly used metric for measuring the performance of a speech recognition system, and is equivalent to the edit distance for words. It is defined as $\frac{Sub+Del+Ins}{Sub+Del+Cor}$, where Sub, Del, Ins, and Cor are the number of substitutions, deletions, insertions, and correct words, respectively. The denominator is the number of words in the original transcript read by the readers. WER is computed for each ASR-generated transcript, and then averaged over the cases that share the same recording modality and the ASR type. Punctuation symbols (e.g., a period or a comma) and non-verbal tokens are removed, but none of the stopwords are removed when calculating WER.

5.2.2 Classification Performance metrics

In order to measure the quality of current speech-to-text systems for extracting topic information, I calculate the accuracy of the predicted topic labels at the talk-turn level and the F1 score at the session level. Specifically, I measure the talk-turn level accuracy and the

session level F1 score for both the human-annotated transcript and the ASR-generated transcript, and compute the difference between the two scores. This helps quantify the amount of reduction in performance by using ASR for transcription.

5.3 Results

For the 36 test visits of the human-generated transcripts, the talk-turn level accuracy was 56.2%, and the visit level F1 score was 74.53%, both micro-averaged. The numbers are different from those in Chapter 4 since the pre-processing pipeline is slightly different, such as removing PHI, and it is the result based on only 36 test visits, rather than full cross-validation.

The scatter plot in Figure 5.1 shows the overall results. The x-axis is the averaged WER calculated between the ASR outputs and the human-transcribed transcripts (after removing PHI), and the y-axis is the classification performance drop by using ASR for transcription, by subtracting the scores from those when using the human-generated transcripts. The scores are based on 36 visits, except for the case where audios were recorded with Cell Phone and transcribed using Microsoft Azure ASR (purple star in the plot). Four for these audio files, the ASR system did not return any output and the result is based on 32 visits.

The values of WER range approximately from 0.18 to 0.34, and the differences in the performance measures range from 0.00 to 0.05 in both talk-turn and visit level. Also, more performance degradation is observed with higher WER.

In general, the results from the same ASR are more similar than the results from the same recording device, as the data points with the same colors in Figure 5.1 are more grouped together than those with the same shapes. The *Google_video* ASR engine, which is a premium model offered by Google Cloud, had the lowest WER overall, as well as the smallest

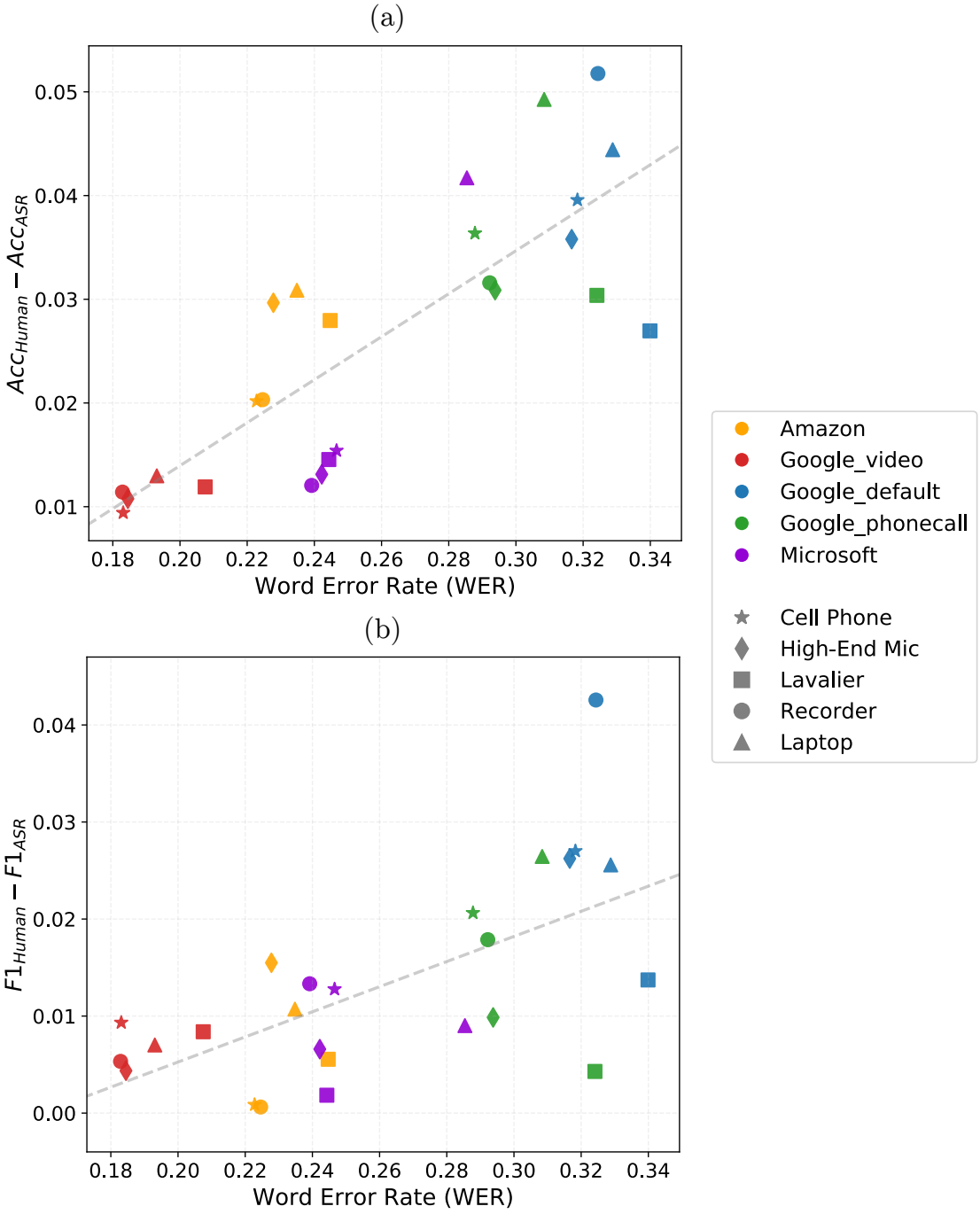


Figure 5.1: The results plotted as word error rate vs. difference in classification performance for using ASR. The colors in the scatter plot represent different ASR engines, and the shapes denote different recording devices. (a) Performance difference in talk-turn level accuracy (b) Performance difference in visit-level F1 score.

performance drop in downstream classification accuracy at the talk-turn level.

5.4 Discussion

The overall pattern in Figure 5.1 is linear to some extent. The data points with lower WER tend to have less performance drop, and it is the opposite for the points with higher WER. Also, we can observe that the variance of performance scores becomes higher as WER increases. However, although there is a linear trend between WER and the difference in performance, the amount of difference in WER does not necessarily transfer to the difference in the classification performance. The reduction in talk-turn level topic classification accuracy is only 5–15% of the difference in WER. Thus, while the drop in topic classification accuracy is correlated with WER, it is also much lower than WER.

There is a considerable amount of variation in classification performance across ASR engines. In terms of WER, *Google_video* (red) typically had about 0.2 WER, whereas *Google_default* (blue) and *Google_phonecall* (green) showed much higher WER. On the other hand, it is hard to find trends when using certain recording devices. This suggests that choosing the right ASR system is more important than selecting a better recording device, and that cell phone recordings may be able to produce a transcript with a similar quality as more expensive equipment would give. It is possible for doctors to make use of their easily accessible recording devices, such as a laptop or a cell phone.

Figure 5.2 provides an overview of the predicted results of a selected test visit. The conversation in the visit flows from left to right, and the colors represent the talk-turns that are predicted or labeled as different topics. The top plot in Figure 5.2 is the original topic sequence labeled by a human labeler, and the bottom three plots are the predicted sequences by Hier-GRU using different types of transcripts. It is easy to observe that the bottom three

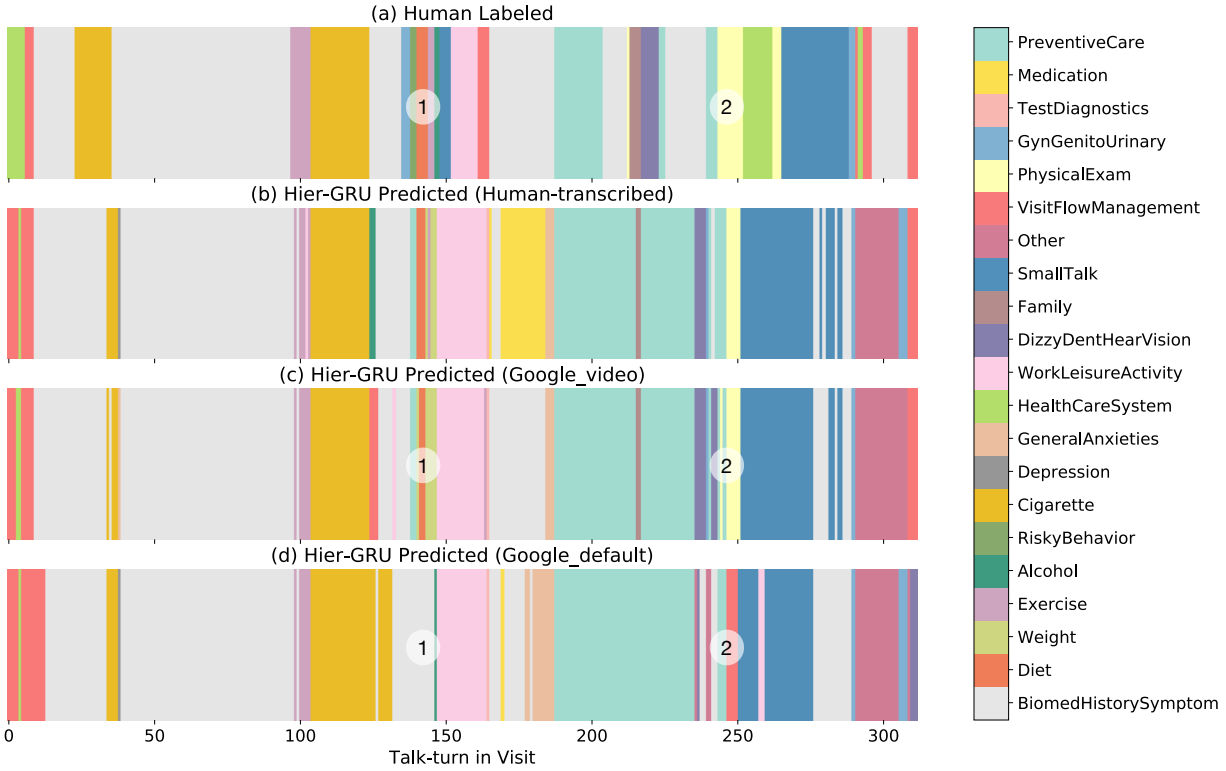


Figure 5.2: The sequences of topic labels for an example test visit. Each topic label is color-coded with the colors shown on the right side. (a) Topic sequences labeled by human labelers. (b) Hier-GRU predicted topic sequences with the human-generated transcript. (c) Hier-GRU predicted topic sequences with the transcript generated by *Google_video* ASR. (d) Hier-GRU predicted topic sequences with the transcript generated by *Google_default* ASR. WER for this specific visit was 0.1959 and 0.3239, respectively for *Google_video* and *Google_default*. The numbered circles are for the examples in Figure 5.3.

plots are similar to each other. The sequence of predicted topic labels generated from the *Google_video* transcript (Figure 5.2 (c)) is very similar to that from human-transcripts (Figure 5.2 (b)). Figure 5.2 (d), which is the predicted topic sequence with *Google_default* ASR output, has a few more areas that are different from (b). For example, talk-turns between 140 and 145, 210 and 260, and the talk-turns after 300.

Two excerpts from the same visit are presented with accompanying text from the transcript in Figure 5.3 for more detail. The two areas discussed in Figure 5.3 are marked as numbered white circles (1) and (2) in Figure 5.2. In talk-turns from 140 to 145, the doctor and the patient talked about eating habits and working out, as seen in the "Human Generated"

Index	Spkr	Human Generated		Google_video		Google_default	
140	MD	Diet	Okay alright Social history Last year we felt your **diet** could use some improvement ...	Preventive Care	All right social history last year We felt your **diet** could use some improvement ...	Biomed History	social history last year We fell you'd I could use some improvement ...
141	PT	Diet	I actually have been eating more vegetables and fruits	Weight	Actually have been eating more vegetables and fruits	Biomed History	Actually have been eating more vegetables and fruits
142	MD	Diet	Wonderful	Diet	wonderful	Biomed History	wonderful
143	PT	Diet	um which I've always loved I just you know you get distracted and it's easier to grab this ...	Diet	which I've always loved I just you know you get distracted and it's easier to grab this ...	Biomed History	which I've always loved I just you know you get distracted and it's easier to grab this ...
144	MD	Exercise	Okay great You doing any **weights** at the **gym**	Weight	Okay great you doing any **weights** at the **gym**	Biomed History	-
145	PT	Exercise	No no **weights** You did recommend that	Weight	No no **weights** You did recommend that	Biomed History	-
...							
245	MD	Physical Exam	Yes Unfortunately you have to	Physical Exam	Yes unfortunately you have to	Preventive Care	fortunately you have to
246	MD	Physical Exam	Now sit up	Preventive Care	Now sit up	VisitFlow Mgmt	Now sit up **part**
247	PT	Physical Exam	Pardon me	Physical Exam	Pardon me	VisitFlow Mgmt	**of** me
248	MD	Physical Exam	**sit up** Let me see your hands patient is **watched**	Physical Exam	**sit up** Let me see your hands patient is **watched**	VisitFlow Mgmt	Let me see your hands patient is **washed**
249	PT	Physical Exam	walking	Physical Exam	Walking	VisitFlow Mgmt	walking

Figure 5.3: Topic labels and texts generated by human, *Google_video*, and *Google_default* for the same part of the audio. Approximate locations of the two examples are shown as (1) and (2) in Figure 5.2.

column. However, *Google_default* ASR, from the text shown in the rightmost column, missed a lot of keywords such as "diet" in talk-turn 140, and "weights," and "gym" in talk-turns 144–145. In fact, no text was recognized for the talk-turns 144–145. Not being able to transcribe the important words resulted in predicting the whole segment as *BiomedHistory*, which is the most common topic in the dataset (see Table 4.1). Similar results happened in the second excerpt in talk-turns 246–249, due to missing "sit up" in talk-turn 248.

Google_video was able to detect most of the important words, but the misunderstanding of the term "weight" in talk-turns 144–145 produced errors. These errors are due to the misclassification of the classfier, and not necesarily caused by the ASR, as a similar sequence of predicted labels is observed in Figure 5.2 (b) in the same area. Also, *Exercise* and *Weight* are shown to be confused with each other, as discussed in Figure 4.9.

Overall, ASR-generated transcripts create word errors, and the errors affect the topic classification results. However, the difference in classification performance is minimal compared to the WER difference.

5.5 Conclusions

In this chapter, I showed that the current ASR systems provided by major software companies could potentially be used to detect topics that are discussed during primary care visits. An actual deployment of such system could alleviate physician burnout problem and allow physicians to spend more time on direct patient care. Advanced technology for better speaker diarization and resolving concerns in patients' privacy, such as automatic deidentification, are major remaining obstacles that need more attention and effort.

The primary contributions of this chapter include the following:

- I demonstrated a feasibility test for obtaining the structured topic information by adding ASR to doctor-patient conversations, as an extension to Chapter 4.
- I tested with audio files from different recording devices and with different ASRs to capture variability in performance depending on the recording environment or the ASR engine.
- I evaluated the results systematically by measuring the model performance with accuracy and F1 score, and measuring the ASR performance with WER.
- I showed that the drop in topic classification accuracy tends to be much lower than the performance drop in transcribing the audio (measured by WER).

Chapter 6

Predicting Emotion Trajectories from Transcripts of Patient-Physician Dialog

In 2016, there were an estimated 883.7 million medical office visits where patients and medical doctors discussed symptoms and engaged in treatment planning [2]. The quality of the interactions and responsiveness to patient needs and concerns is foundational to patient-centered care [48, 92, 93]. Patient-physician interactions involve more than just discussions of symptoms and biomedical facts. Patients can be in intense emotional states when visiting their doctors. How doctors respond to patient emotions [94, 145] is related to important patient outcomes [143, 144]. Consequently, an important aspect of patient-centered care is the effective navigation of emotions during medical appointments.

Patients may desire that physicians devote more time in medical appointments to their emotional concerns [97]. At present, patient experiences in care are primarily evaluated using surveys [161], which place a burden on the patient, are subject to social desirability,

and are limited in scope. Direct observation and feedback is the gold standard for supporting providers, but it is time-consuming, expensive, and generally not feasible in most clinical settings [77].

Recent advances in machine learning and natural language processing (NLP) have shown enough capability for predicting emotional content in conversations [127, 154], which could be applied to the clinical conversation. However, the majority of prior work has focused on human-to-machine interactions than human conversations. Also, the studies focus more on the performance of the models than the evaluation methods and its applications.

In this chapter, I propose a neural-network-based model for large-scale medical conversation transcripts to predict emotional valence, and I describe the details of the two evaluation methods that are suitable for such data. The evaluation methods are also applied to the human labels to measure the agreement between the human raters, which provides an approximate upper bound on what is achievable by machine learning methods.

6.1 Related Work

6.1.1 Machine Learning Methods for Emotion Recognition

Sentiment analysis has long been one of the major interests in machine learning and natural language processing community [5, 118]. Most of the studies in sentiment analysis and opinion mining have focused on mining online reviews (e.g., Amazon reviews [103]) and social network posts (e.g., Twitter [4]). Various approaches have been adopted ranging from lexicon-based [134, 76] to modern deep neural network models [36, 82]. However, for such data, each document or sentence is treated independent to each other, and thus the methods do not have sequential or contextual structure between sentences.

Recognizing emotion in conversation is a relatively new research area that has gained attention due to the abundance of conversational data and to build more human-like, emphatic agents in conversational AI [127]. Utilizing contextual or sequential information in conversational data is crucial in predicting the emotional valence attached to short utterances that may contain very little information. In the past few years, models based on neural networks (mostly recurrent neural networks investigated in this context) have been actively investigated [100, 59, 66].

However, the sizes of the datasets used in the studies of predicting emotions in conversations are relatively small compared to other text datasets used for sentiment analysis, ranging from 6k to 14k utterances. Also, the average length of conversations is around 10 to 60 utterances depending on the dataset, such as the IEMOCAP [22], AVEC [131], and MELD [126] dataset.[1] Applying the models mentioned above to much longer dialogs, such as conversations in doctor visits, requires more efficient ways of memory management, or modifications in the model structure to only see the surrounding utterances.

In addition, the datasets used in prior work on emotion classification in conversations are either not from human to human dialogs, or do not have any medical context in the vocabulary, and thus, these models may not generalize well to detecting emotions for doctor-patient conversations in primary care visits.

6.1.2 Evaluation Methods for Emotion Recognition in Dialogs

For dialog emotion classification datasets with categorical labels, such as IEMOCAP [22] and MELD [126], the main evaluation metrics used are accuracy, F1, or recall scores [100, 59, 66].

For datasets that have scores on an ordered scale for emotional valence, particularly for the AVEC [131] dataset (or SEMAINE [105]), evaluation methods have relied on correlation

[1] A summary of different datasets is well described in [127].

coefficient and mean absolute error. As an example, Majumder et al. [100] measured the correlation coefficient and the mean absolute error as two performance metrics for the modified AVEC dataset. However, there is no thorough analysis for correlation coefficient results, and the whole sequence of a dialog session is used to calculate correlation coefficient, rather than separately calculating it for each speaker. Also, the dataset is based on human-agent interactions, and the annotation process was carried out in a manner that is significantly different to the dataset used in this chapter.

6.1.3 Text Analysis in the Medical Domain

There has been growing attention to quantitative text analysis in the field of medicine. Electronic health records (EHRs) have been used for various information extraction tasks [25, 98], such as identifying post-operative complications [112] as well as patients that may need further treatment or care [169, 155].

Determining the emotional state of an individual based on their speech or writings is another primary interest in medicine [142, 161]. For example, Pestian et al. [123] developed automated systems for classifying emotions in suicide notes. Similarly, Georgiou et al. [58] classified emotions using lexical features in married couples' discussions. Other work has focused on classifying emotional valence in online text-based psychotherapy for depression [74].

For medical dialogs, recurrent neural networks have been used to chart symptoms [128] and classify the topic of the conversation at each talk-turn [119]. However, although empathetic patient-physician communication is fundamental to the practice of medicine, there has not been any attempt to evaluate the quality of the interactions in primary care visits by automatically capturing the emotional content of the visit.

6.2 Dataset

The dialog transcript dataset used in this chapter is a combination of transcripts from two different studies. The first one is the Mental Health Discussion study by Ming et al. [149] and the other is the Assessment of Doctor-Elderly Patient Transactions (ADEPT) by Teresi et al. [156]. Both datasets are transcripts of doctor-patient visits, where each talk-turn (the utterances spoken by a speaker at each turn), is labeled (or "coded") to one of the topic labels based on the Multi-Dimensional Interaction Analysis coding system (MDIA) coding system [27].

The utterances in this dataset have also been labeled according to their emotional valence labels. Fourteen labelers[2] were asked to assign an emotional valence label to each utterance. In total, there are 353 visits with 210k utterances. An utterance is defined as a discrete sentence unit within a talk-turn, that is separated by periods, exclamation marks, and question marks. The emotional valence takes integer values that range from -3 (very negative) to +3 (very positive), with neutral at 0. For cases where labelers used different boundaries for splitting a talk-turn into multiple utterances, the most fine-grained boundaries are used and the labels are assigned to each of them. Not all labelers labeled all the sequences for each visit—some of them skipped labeling the utterances that they are not confident of. Figure 6.1 shows an example of the labeled data. In this example, some of the talk-turns are split into multiple utterances, and each utterance has multiple emotion valence labels assigned by 6 labelers.

Out of 14 labelers, I removed the labels from the four labelers whose distribution of assigned labels are significantly different from the rest.

Figure 6.2 (a) shows the emotion label distributions from all 10 labelers used for my analysis. 79.21 percent of all the labels were labeled as neutral (0). Figure 6.2 (b) shows the results

[2]All of the labelers were affiliated with the University of Utah.

	Utterance	Avg	Emotional Valence by Human Labelers					
			4	5	9	10	13	14
0	MD: Why hello there.	1.3	2	1	1	1	2	1
1	How are you doing?	1.0	2	1	1	1	0	1
2	PT: Doing good.	1.5	2	1	2	1	2	1
3	MD: Are you?	0.3	1	0	0	0	0	1
4	PT: Doing good, doing good.	1.2	1	1	1	1	2	1
5	Just, you know, got a little more pressure on me.	-0.8	-1	-1	-1	-1	-1	0
6	MD: Yeah.	0.0	0	0	0	0	0	0
7	PT: My stepfather passed June 5th.	-2.0	-2	-1	-2	-1	-3	-3
8	MD: Oh, I'm sorry to hear that.	-1.3	-1	-1	-1	-1	-2	-2
9	PT: So, I'm running, like twice a week I'm going to my mothers out there by Metro, airport, so.	-0.7	-1	0	-1	0	-2	0
10	She's hanging in there, she was married 40 years.	-1.0	-1	-1	-1	-1	-2	0
11	MD: Okay.	-0.2	0	0	0	0	-1	0
	…							
48	MD: So you 're trying to help your mom out.	-0.3	-1	0	0	0	-1	0
49	PT: Right, right.	-0.3	-1	0	0	0	-1	0
50	MD: Yeah.	-0.2	0	0	0	0	-1	0
51	PT: Yeah.	-0.2	0	0	0	0	-1	0
52	MD: Well it seems like your, your blood pressure looks pretty good, yeah?	1.3	1	1	1	2	2	1
53	PT: I'm hanging in there, still roller skating, still swimming, and still walking.	1.3	1	1	1	2	2	1
54	MD: Are you?	1.0	1	0	0	2	2	1
55	You 're pretty active.	1.7	1	1	1	3	2	2
56	PT: Trying to be.	0.8	1	1	0	1	1	1
57	MD: That's darn good.	2.2	2	2	2	2	3	2
58	PT: Trying to be.	1.2	1	1	1	2	1	1
	…							
68	MD: When you check your blood sugar, how often do you check it?	0.0	0	0	0	0	0	0
69	PT: Twice a day, in the morning and in the evening.	0.0	0	0	0	0	0	0
70	MD: Okay.	0.0	0	0	0	0	0	0
71	In the morning what kind of range do you get with the numbers?	0.0	0	0	0	0	0	0
72	PT: It ranges between 90 and 110.	0.0	0	0	0	0	0	0
73	MD: Oh that's good.	1.3	2	1	1	1	2	1
74	That's quite good.	2.0	2	1	2	2	3	2

Figure 6.1: An example of a section of dialog from a particular visit. The six columns on the right side show the emotional valence labels assigned by 6 labelers. The averaged value is shown in the third column.

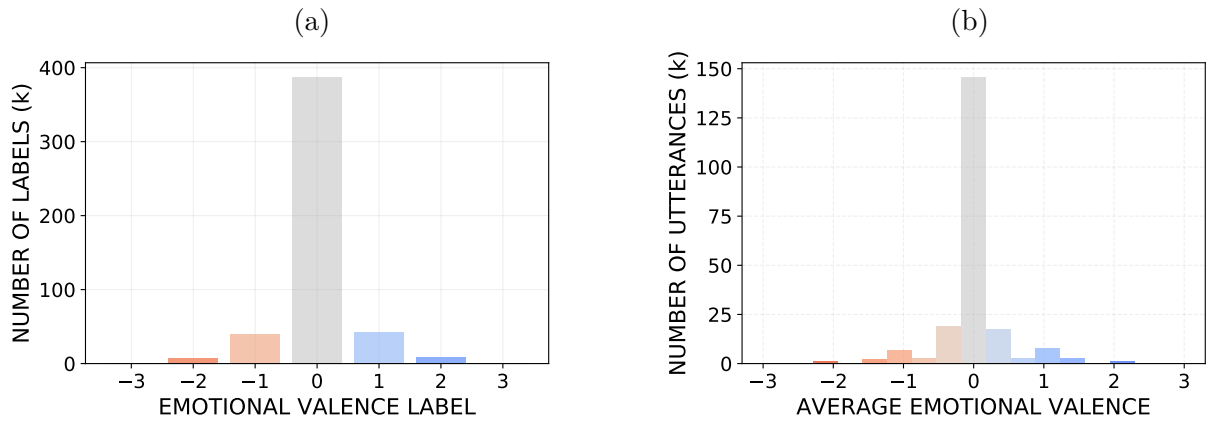


Figure 6.2: (a) Histogram of emotional valence labels from 10 different labelers. Multiple labels may exist for each utterance. The proportion of each label in percentage from left to right is 0.24, 1.63, 8.18, 79.21, 8.61, 1.93, and 0.19. (b) Histogram of the per-utterance emotional valence, where multiple labels assigned to each utterance are averaged to a single value.

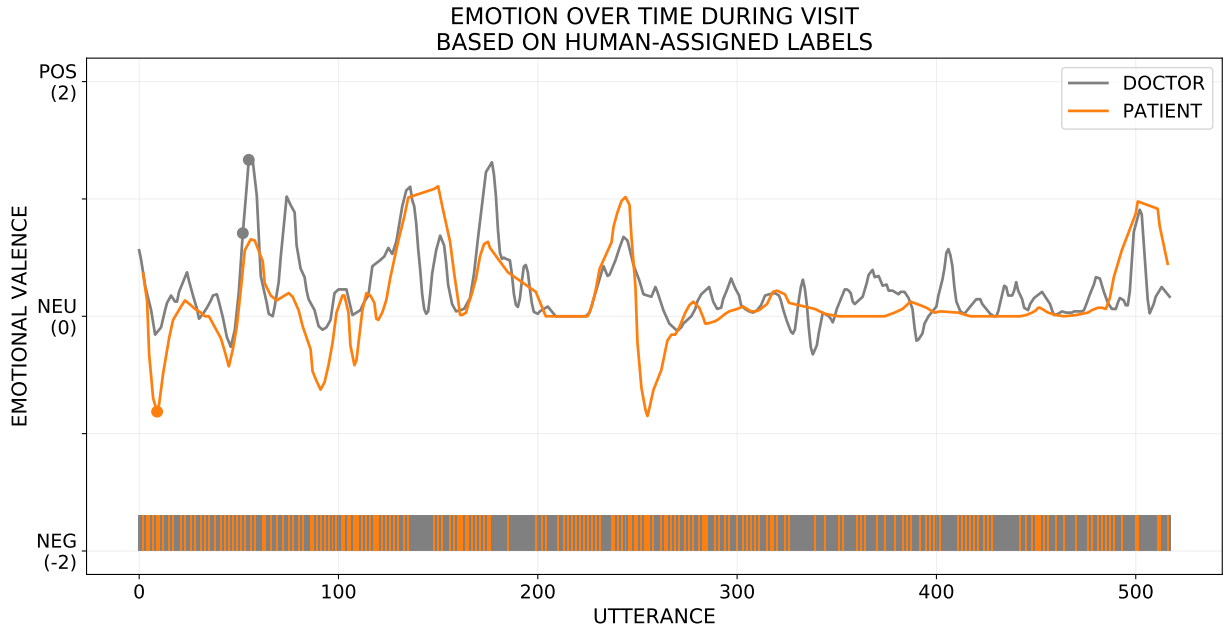


Figure 6.3: Per-utterance mean emotional valence for a doctor (gray) and a patient (orange) during the same visit as in Figure 6.1. The mean emotional scores for each speaker are smoothed with triangular moving average of window size 7 to produce smoother lines. The colors at the bottom shows the parts of utterances spoken by each speaker. For this visit, 63.7% of the utterances (330 out of 518 utterances) are from the medical doctor.

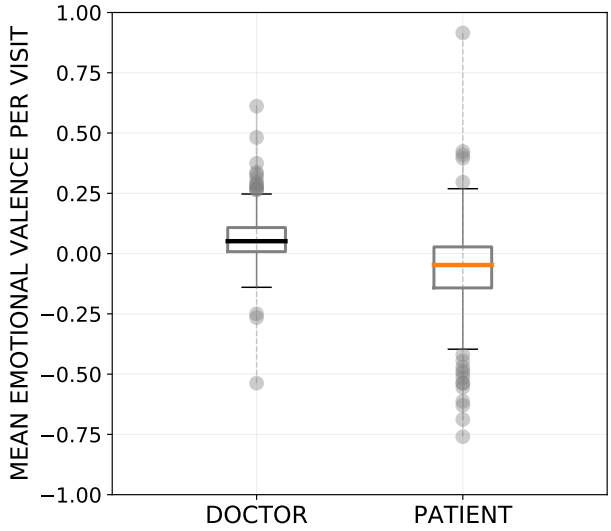


Figure 6.4: Mean emotional valence per visit, for each speaker.

after taking the average of the average of the labeler's valence scores for each utterance. I will use these average scores extensively in this chapter.

The same visit shown in Figure 6.1 can be presented without the text, by showing the trajectory of the mean emotional valence per speaker as in Figure 6.3. The x-axis of the plot is the utterance index, which is the same value as the first column in Figure 6.1. Since the dialog transcripts assume only one speaker speaks at a time, the color bar is plotted at the bottom to show the active parts for each speaker in a visit. To reduce the noise, each sequences of labels are smoothed using moving average with triangular window of size 7 (three utterances before and after each center utterance). The smoothing is done for visualization purposes only. We see that the patient and doctor each transitions through different emotional states during the visit with the patient having more pronounced negative emotion than the doctor.

The utterance indices 7, 52, and 57 in Figure 6.1 are interesting examples to look at. Those are marked as solid circles in Figure 6.3. When the patient talks about his/her recent bereavement (utterance 7), the emotion valence drops to a low (negative value). The doctor listens and tries to empathize with the patient. After that, the doctor tries to brighten the

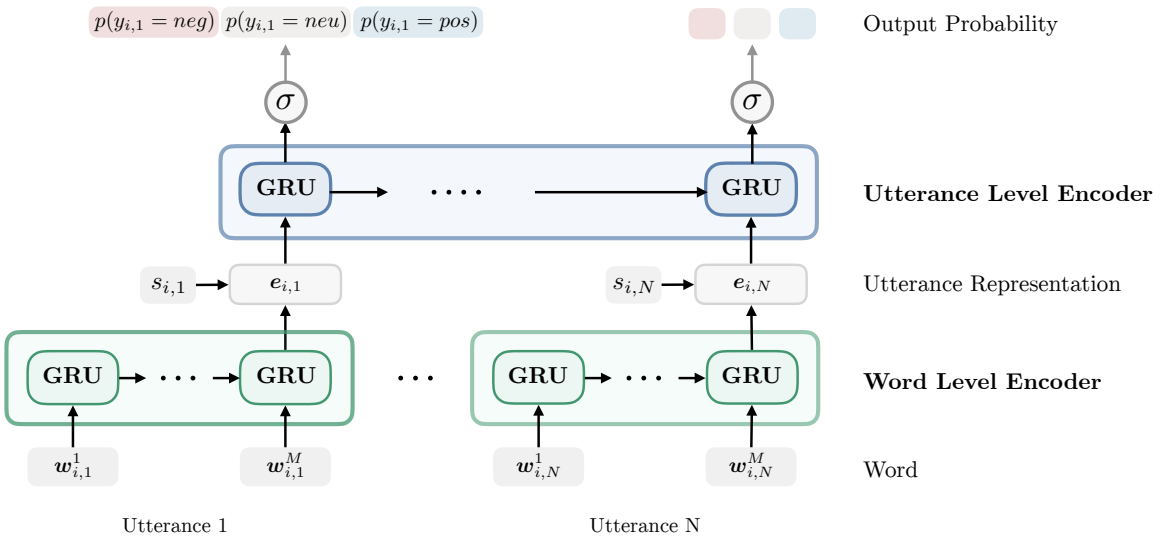


Figure 6.5: Simplified model that illustrates the output probabilities from the softmax function. The subscript i, j represents the jth utterance in visit i. GRUs both at the utterance level and the word level are bi-directional.

mood by talking about the blood pressure (utterance 52), and compliments the patient for being active and exercising often (utterances 54–57). A similar pattern exists later in this visit (around utterance 250–300 in Figure 6.3) as well in other visits, where the patient talks in a negative fashion and the doctor refreshes the conversation by saying positive things.

A similar trend can be seen in Figure 6.4, which shows boxplots of emotional valence per visit, where doctors tend to have more positive emotion than patients overall. For each visit, the emotional valences of the utterances spoken by each speaker are averaged. Each boxplot represents 353 values, one data point for each of the 353 visits.

6.3 Model

The primary type of model I investigate for prediction of emotional valence is similar to the recurrent neural network model (Hier-GRU) used in Chapters 4 and 5.The main difference between the model shown in Figure 4.3 and the model used in this chapter (Figure 6.5) is the

Figure 6.6: A diagram that shows the steps to make an utterance vector $\boldsymbol{e}_{i,j}$ speaker-dependent. A linear projection (FC stands for the fully connected layer) is performed to create a speaker embedding $\boldsymbol{s}_{i,j}$, and the speaker embedding is added to the existing utterance vector after an activation function. The size of the speaker embedding is the same as that of the utterance vector $\boldsymbol{e}_{i,j}$.

speaker embedding. Therefore, the model is referred to as "Hier-GRU-S," with an "S" that stands for "speaker." Unlike the topic labels, even though emotional valence between doctor and patient is often correlated over time (e.g., see Figure 6.3), speakers can have different emotions even when they are communicating at the same time. Also, since the doctors show slightly more positive emotions overall (Figure 6.4), it is reasonable to incorporate the speaker information with each utterance.

One of the simplest ways to incorporate speaker information is to make the utterance vector $\boldsymbol{e}_{i,j}$ speaker-dependent by adding a projection of the speaker embedding, where the projection is followed by an activation function (see Figure 6.6 for an illustration). This type of addition method for having multi-speaker representations has been used successfully in speech recognition [3] and in speech generation [120, 124, 153]. Since the model used here is more of a text classification model where the model size is significantly smaller than those for speech data, the speaker embedding is added to just one part of the model, to the utterance vector $\boldsymbol{e}_{i,j}$. Also, since it is used only once, there is no need to keep a separate low-dimensional embedding, and the size of the embedding is set to the size of the $\boldsymbol{e}_{i,j}$.[3]

Figure 6.5 illustrates the high-level idea of the model. The figure focuses on a single utterance j in visit i. A speaker index for the utterance $s_{i,j}$ (either 0 or 1 to indicate doctor or patient) is embedded and added to the utterance vector (details are simplified in the figure). Although unique speaker IDs were not used, having the information on whether the speaker is a doctor

[3]In prior work, a low dimensional speaker embedding was learned for each speaker and shared across the model by adding a projected version of the embedding to different parts of the neural network model.

or a patient showed an improvement in both of the evaluation methods discussed in the later sections.

The probabilities for each of positive, neutral, and negative are shown at the output after the softmax in Figure 6.5. These probabilities themselves are used for evaluation in addition to the classification decision $y_{ij} \in \{pos, neu, neg\}$, which corresponds to the argmax of the output probabilities.

6.3.1 Model Training

For training, the labels are averaged to a single emotional valence score $\overline{y}_{i,j}$ for each utterance j in visit i. Also, due to the highly skewed label distribution shown in Figure 6.2, labels with extreme values are all treated as just "positive" or "negative," resulting in three categorical labels: negative (-1), neutral (0), and positive(1). The score $\overline{y}_{i,j}$'s that are -0.5 or lower are quantized to negative, and the values that are 0.5 or higher are rounded to positive. All the other values are considered as neutral.

To validate the performance of Hier-GRU-S, an independent logistic regression (LR) classifier is used as a baseline model. The results from the Hier-GRU without speaker information are also reported for comparison.

Text cleaning and preprocessing steps are very similar to the steps described in Section 4.3.1 in Chapter 4. However, stopwords are included for the bag-of-words representation used in the LR classifier since removing stopwords decreased the performance for emotion classification. All unigrams and bigrams are included except for the ones that appeared in 2 or fewer utterances (document frequency of 10^{-5}). The size of the bag-of-words vector for LR was $V = 52278$. For Hier-GRU-S and Hier-GRU, only the unigrams are included without any stopwords removed, with the final vocabulary size of 5489.

Implementation details and hyperparameters settings are described below for each model. All hyperparameters are selected via grid search on a single fold.

Hier-GRU-S

The model was implemented using the AllenNLP toolkit [55], based on PyTorch. The dropout rates for the word embedding layer and the speaker embedding layer were both set to 0.4. The word embedding size was 100, and the hidden size of the GRUs was 128 for both the utterance level and the word level. Both GRUs were bi-directional, and thus the utterance vector $\boldsymbol{e}_{i,j}$ had size 256. The speaker embedding size was set to 256 to match the dimension with the utterance vector. Both GRUs had a dropout rate of 0.5, and had a single layer. I used ReLU for the activation function at the speaker embedding. Other references used softsign or hyperbolic tangent for activation to limit the output scale. However, it was less important in my experiment since there are only two different speakers, and using ReLU gave slight improvement over softsign. The model was trained with cross-entropy loss and was optimized via Adam [83], with 0.001 as a starting rate.

Hier-GRU

This model has the same structure as Hier-GRU-S, except that it does not have the speaker embeddings. The dropout rate at the word embedding layer was set to 0.2, and all other hyperparameters were the same as Hier-GRU-S.

Logistic Regression

The Scikit-learn library [121] implementation of LR was used with the one-vs-rest training option. The inverse regularization strength was set to 3, and the L2 penalty was used.

Subset	Test Labeler	Number of Visits	Number of Utterances
1	9	93	49849
2	13	84	44088
3	10	79	43025
4	4	50	26403
5	14	50	23577
6	5	27	8644
7	8	5	1392
8	2	2	1387
9	7	2	630
10	3	2	630

Table 6.1: The number of visits and utterances for each subset used for evaluation. Each subset consists of utterances that have 3 or more labels including the test labeler (shown in the second column of the table). The rows are sorted by the number of utterances in each subset.

6.4 Evaluation Methods

6.4.1 Comparison with the Human Labels

To provide a baseline for model performance, I evaluate human labeling reliability in the following manner. Since multiple labels may exist for a single utterance, a one vs. rest (OvR) approach is used for each labeler for the human labels to measure their performance. In this approach, the labels from one labeler are treated as values to be tested against the mean of the labels from the other labelers. This way of evaluating quantifies the proportion of labeling errors that humans make relative to each other, or how much humans can agree on each other for rating the emotion labels. Only the visits with three or more labelers (108 visits out of 353) are included in this human OvR evaluation. Also, since each labeler labeled different sets of utterances, different subsets of utterances are selected to calculate the evaluation metric for each labeler. Table 6.1 has the summary information on each subset.

6.4.2 Pearson Correlation Coefficient

The first metric of performance is the Pearson correlation coefficient, which uses the continuous scores from the predicted probabilities. Rather than treating the labels as completely categorical, utilizing the output probabilities can capture more information.

The Pearson correlation coefficient is a statistical measure that quantifies the linear association between two variables. The value ranges from -1 to 1, where 0 corresponds to no correlation. The strongest linear relationship is indicated by -1 or 1 for a negative or a positive relationship, respectively. For evaluating how similar the two emotional valence trajectories are, a higher correlation coefficient value (thus stronger positive relationship) is more desirable. The formula below is used for computing a Pearson correlation coefficient for speaker k (either a doctor or a patient) in visit i.

$$\rho_i^k = \frac{\sum_{j \in \{s_{i,j}=k\}} (a_{i,j} - \overline{a}_i^k)(b_{i,j} - \overline{b}_i^k)}{\sigma_{a_i^k} \sigma_{b_i^k}}, \tag{6.1}$$

where $a_{i,j}$ and $b_{i,j}$ are two different emotional valence scores for utterance j in visit i, and $s_{i,j}$ is the speaker index of the same utterance to indicate whether it was spoken by a doctor or a patient. The numerator term is the covariance between the two, and the overlined symbol stands for the emotional valence mean of the speaker utterances in a visit (e.g., $\overline{a}_i^k = \frac{1}{N_k} \sum_{j \in \{s_{i,j}=k\}} a_{i,j}^k$, where N_k is the number of utterances spoken by speaker k in visit i). $\sigma_{a_i}^k$ and $\sigma_{b_i}^k$ are standard deviations for each of the two sets of scores.

For evaluating the human labels, the correlation coefficients are calculated for each speaker, each visit, and each available labeler in that visit. As described earlier, each labeler is regarded as a "test" labeler, and the correlation coefficient is calculated between the labels from the test labeler and the mean label values from the rest of the labelers who also labeled

that visit. A total of 788 correlation coefficients are calculated for this case.[4]

To calculate the correlation coefficients for the model output, the following score is used as the basis for the model's predictions, for each utterance i in visit j.

$$l_{i,j} = p(y_{i,j} = pos) - p(y_{i,j} = neg). \tag{6.2}$$

This score can capture more fine-grained variations and nuances than the integer labels, particularly because the model is trained with three labels.[5] Then the correlation between the average human valence score $\overline{y}_{i,j}$ and the score $l_{i,j}$ is calculated, for each speaker and for each visit. Only the visits with 3 or more labelers (108 visits) are included to allow a direct comparison with the human OvR evaluation. Therefore, there exist 216 correlation coefficient scores (108 visits $\times$ 2 speakers).

Figure 6.7 shows examples of emotional valence scores during a visit, used for calculating correlation coefficients. In this particular visit, 6 labelers gave scores to each utterance, and the average value per utterance, $\overline{y}_{i,j}$, from all available labelers is shown as the orange line in Figure 6.7 (a). The model predicted score, $l_{i,j}$, is in blue line, and the correlation coefficient is calculated between these two sequences.[6] One labeler (labeler 4) is selected as a test labeler for human evaluation (blue line in Figure 6.7 (b)), and the labels from that labeler are used to compute the correlation coefficient with the labels from the rest of the labelers (orange line in Figure 6.7 (b)). In this figure, we see that both the model's predictions and the single labeler track well with the trajectory of the scores they are being compared to.

For a better visualization, smoothing with triangular window of size 7 is applied to all the sequences in Figure 6.7 (but not used in computing the correlation scores). Also, the scores

[4]Two times (since there are two speaker types) of the sum of the second row in Table 6.1. $2\sum_i NumLabelers_i = 2\sum_c NumVisits_c = 2 \times 394 = 788$, where i and c are the indices for the visit and the labeler, respectively.

[5]This score from the suggested classification model was able to capture positive and negative signals better than using a regression model for the highly skewed dataset used in this chapter.

[6]Although I use the term "sequence," correlation coefficient does not consider the order.

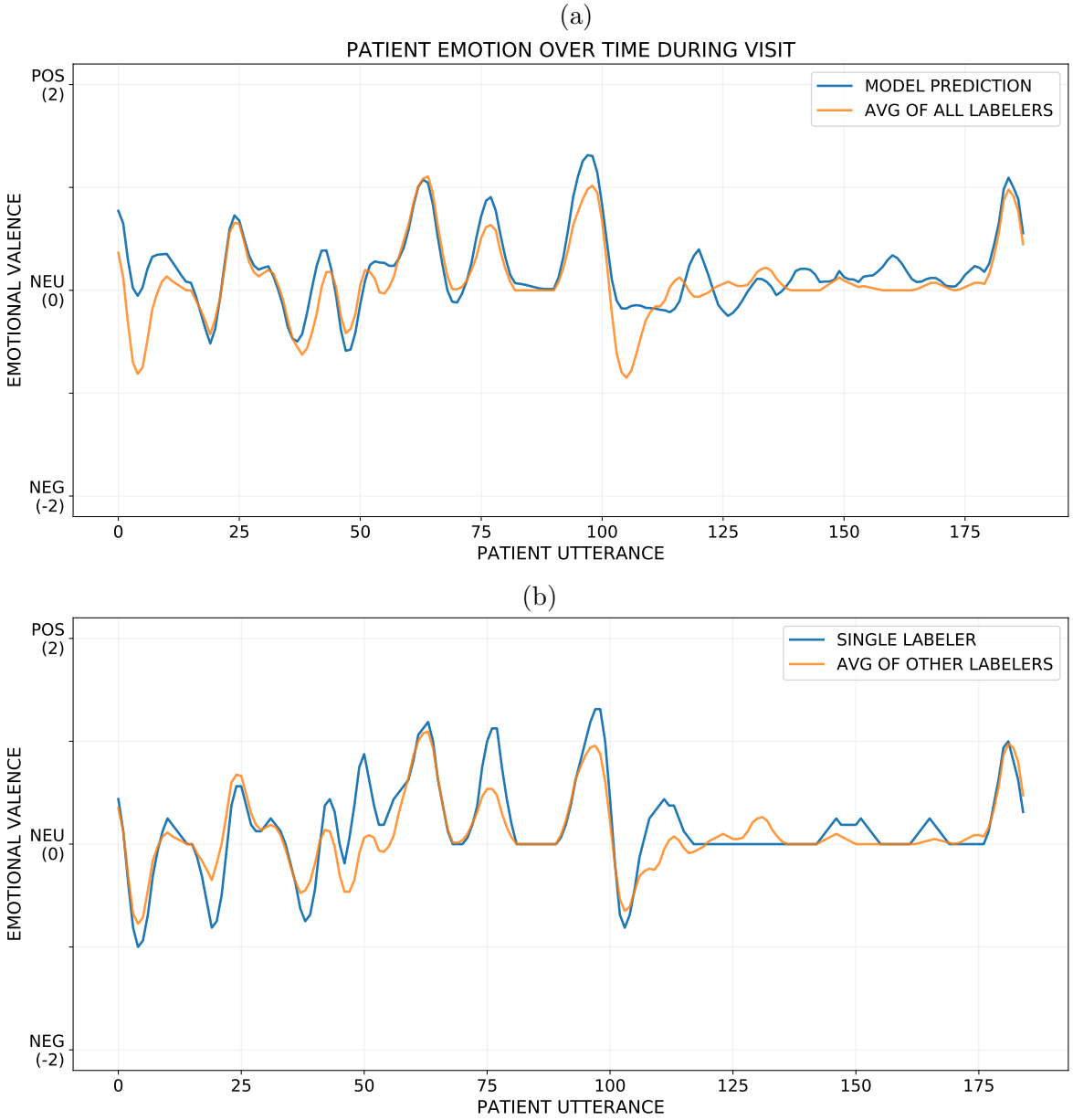


Figure 6.7: Examples of emotional valence trajectories of the patient in a visit (the same visit shown in Figure 6.3) that are used for calculating correlation coefficients. (a) Model predicted score $l_{i,j}$ vs the average label value $\overline{y}_{i,j}$. (b) For human one vs. rest evaluation. In both plots, the orange line is considered as the true or the mean of the majority human label values, and the blue line is the label values to test. The correlation coefficients between the two lines are 0.8168 for (a) and 0.7937 for (b).

of the blue line in Figure 6.7 (a), which is the model predicted score $l_{i,j}$, is scaled to match the variance of the orange line ($\overline{y}_{i,j}$) in the same plot.

6.4.3 R-Precision

R-precision is the second metric that I use as a performance metric. Due to the highly imbalanced data, the base rate for the classification accuracy already reaches around 80% when predicting all the utterances to "neutral." Therefore, it is more important to focus on how many of the positive and negative utterances are correctly classified and how confident the model is for those utterances. R-precision achieves both by calculating the precision at the top R utterances, separately for positive, negative, and neutral.

In information retrieval, R-precision is defined as the precision at the Rth position, where R is the total number of relevant documents for a query that could be retrieved. If there are r relevant documents among the top-R retrieved documents, then the R-precision is calculated as $\frac{r}{R}$.

For this application, an "utterance" is regarded as a "document," and the scores are separately calculated for each categorical or integer label: negative (-1), neutral (0), and positive (1). The utterances are ranked by the predicted probabilities for each emotional valence category to compute the R-precision.

Since the emotional valence label 1 is considered as slightly positive, and -1 is considered as slightly negative, I set the threshold as 0.5 and -0.5 for deciding the number of relevant utterances for positive and negative, respectively. For positive, R is computed by counting all the utterances that have human label values 0.5 or higher. For negative, R is the number of utterances with label values 0.5 or lower. The R for the neutral class is the number of all the other utterances.

In order to make a reasonable comparison between the human and the model, the subsets in Table 6.1 are used to calculate the R-precision for both the human OvR and the model evaluation. For each method (e.g., human OvR or Hier-GRU-S), R-precision scores are calculated for each subset, and the scores are averaged with weights to get a single score. The weights are proportional to the number of utterances in each set and sum to 1.

Additionally, R-precision scores are also compared just between the models. Since the model outputs have labels from all visits, there is no need to evaluate on different subsets. The most straightforward way is calculating R-precision using all the utterance to give a single number. Also, it can be computed for each visit to obtain 353 R-precision scores.

6.5 Results

Similar to Chapter 4, 10-fold cross validation is used to obtain the model prediction results for all 353 visits.

6.5.1 Correlation Coefficient Results

Correlation coefficients are calculated for (a) the human one vs. rest and (b) the three models (Hier-GRU-S, Hier-GRU, and LR). Their distributions are shown as boxplots in Figure 6.8. The highest variance is observed in the human one vs. rest case, however, it also has the highest mean and median values. The distribution of the logistic regression (the rightmost boxplot), which is the non-neural baseline model, is slightly shifted down compared to the other three boxplots, showing the lowest performance in terms of correlation coefficient. The two neural network models, Hier-GRU-S and Hier-GRU both gave similar and comparable results to the human results.

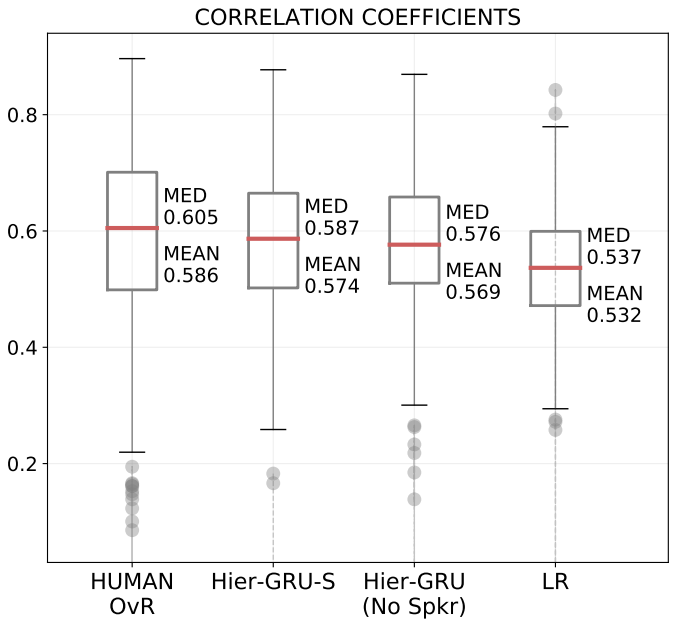


Figure 6.8: The boxplots of correlation coefficients for human and three different models.

Two Sources of Correlation Coefficients	p-value	
Human vs. Hier-GRU-S	0.2935	
Human vs. Hier-GRU	0.1157	
Human vs. LR	0.0000	***
Hier-GRU-S vs. Hier-GRU	0.0034	**
Hier-GRU-S vs. LR	0.0000	***
Hier-GRU vs. LR	0.0000	***

Table 6.2: Results from two-sided T-tests using correlation coefficients from pairs of models, in order to test whether they have identical average values. For the tests with the human scores, independent T-test was used, and all the other tests used dependent T-test.

The results are further quantified in Table 6.2 by conducting statistical T-tests with each of the two sets of correlation coefficients. Since the one vs. rest human evaluation generates more coefficient scores than the models, independent T-tests were used when testing with the human OvR correlation coefficient scores. For all the other tests between three models, dependent T-tests for paired samples were used for testing significance.

From the results, the human-labeled emotional valence correlations are not significantly different from the correlations from each of the two neural network models, Hier-GRU-S and Hier-GRU. However, the p-value for the T-test between the human and the logistic regression

is very small, indicating that the null hypothesis of having equal means can be rejected with confidence.

The correlation coefficients from LR are significantly different at the 0.001 level from both Hier-GRU-S and Hier-GRU. Also, although it is not obvious from Figure 6.8, the mean of Hier-GRU-S is different (and higher, from Figure 6.8) from Hier-GRU, and is significant at the 0.01 level.

6.5.2 R-Precision Results

Table 6.3 shows (a) the human OvR precision scores and (b) the model's precision scores relative to all the human labelers.

The scores consist of weighted averages of R-precision scores across 10 different subsets (in Table 6.1). Surprisingly, the human agreement on rating negative and positive utterances was lower than the agreement between the human and the two neural network models' predictions, quantified by the R-precision score.

Since Hier-GRU-S gave the highest weighted mean of R-precisions among the models, I focus next on Hier-GRU-S results by investigating the R-precision scores for each subset. Table 6.4 shows those numbers. The rows in the table are sorted by the number of utterances in each subset, therefore the R-precision scores with larger subsets (subset 1 to 5) have more reliable scores and are weighted with higher weights.

Model	R-Precision (Weighted Avg)		
	Negative	Neutral	Positive
Human OvR	0.437	**0.900**	0.471
Hier-GRU-S	**0.448**	0.897	**0.576**
Hier-GRU (No Speaker)	0.426	0.893	0.573
LR	0.414	0.887	0.528

Table 6.3: Weighted average of R-precision scores.

Subset	Test Labeler	Num Utters	Negative		Neutral		Positive	
			Human OvR	Hier-GRU-S	Human OvR	Hier-GRU-S	Human OvR	Hier-GRU-S
1	9	49849	0.387	0.456	0.881	0.905	0.468	0.579
2	13	44088	0.421	0.446	0.918	0.901	0.477	0.570
3	10	43025	0.471	0.444	0.901	0.899	0.416	0.566
4	4	26403	0.466	0.464	0.916	0.880	0.507	0.570
5	14	23577	0.443	0.435	0.877	0.878	0.481	0.603
6	5	8644	0.641	0.451	0.892	0.896	0.544	0.596
7	8	1392	0.118	0.397	0.929	0.944	0.314	0.570
8	2	1387	0.413	0.385	0.950	0.925	0.526	0.599
9	7	630	0.286	0.333	0.990	0.977	1.000	0.500
10	3	630	0.000	0.333	0.987	0.977	1.000	0.500
Weighted Avg			0.437	0.448	0.900	0.897	0.471	0.576

Table 6.4: R-precision scores for each subset from human OvR and Hier-GRU-S with detailed information.

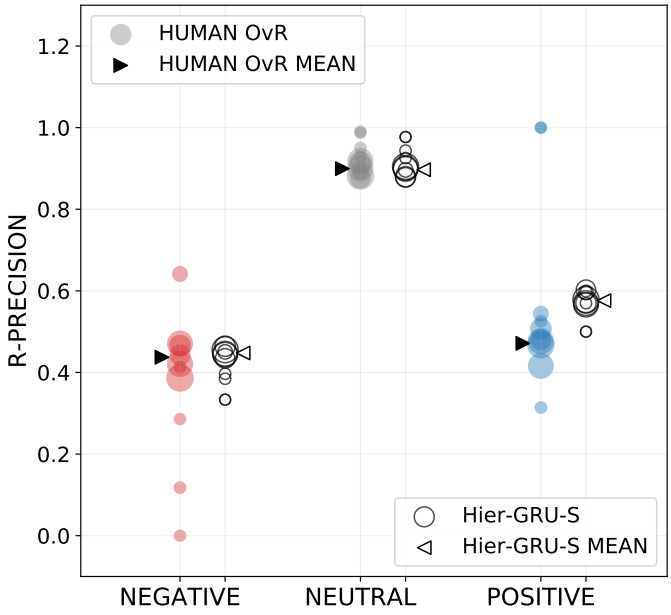


Figure 6.9: R-precision scores for each subset (per-coder subset) from human OvR and Hier-GRU-S. The size of the circles show the size of the subset.

Model	R-Precision		
	Negative ($R = 26036$)	Neutral ($R = 157323$)	Positive ($R = 27378$)
Hier-GRU-S	**0.4557**	**0.8548**	**0.5388**
Hier-GRU	0.4365	0.8511	0.5379
LR	0.4259	0.8458	0.4930

Table 6.5: R-precision scores calculated with all the utterance predictions.

Interestingly, there exist some trends for each subset (or the test labeler). For example, for subset 1, in which the test labeler is 9, human OvR scores are lower than the Hier-GRU-S scores in all three categories. This implies that this labeler tends to disagree with the majority of the other labelers. A more detailed version of the table with precision numbers are in Appendix C for completeness.

The scores are visualized in Figure 6.9, where the solid circles are the R-precision scores from human OvR for each subset, and the empty circles are those from Hier-GRU-S. The size of the circles represents the size of the subset. The larger circles are the ones that are more of interest. The filled triangles (pointing right) are located at the weighted means for human OvR, and the empty triangles (pointing left) show the weighted means for Hier-GRU-S. From the figure, it is easy to see that the negative and positive R-precision scores for human OvR have larger variances than those in Hier-GRU-S. However, the scores that are far away from the weighted mean are from much smaller subsets (since the size of the circles are very small), such as subsets 7, 9, and 10.

Next, I compare the R-precision scores between the three models using the labels from all 210k utterances, which are shown in Table 6.5. Previously, in section 6.5.1, Hier-GRU-S performed significantly better than the other two models when evaluated using the Pearson correlation coefficient. A similar result is observed in Table 6.5, where the Hier-GRU-S has the highest scores for all three categories, and the logistic regression classifier has the lowest scores.

Two sources of R-Precisions	p-value		
	Negative	Neutral	Positive
Hier-GRU-S vs. Hier-GRU	0.0000 ***	0.0000 ***	0.0016 **
Hier-GRU-S vs. LR	0.0002 ***	0.0000 ***	0.0000 ***
Hier-GRU vs. LR	0.0034 **	0.9085	0.0000 ***

Table 6.6: The p-values from the dependent T-tests with paired samples with per-visit R-precision scores. The asterisks represent the level of significance: 0.01 (**), 0.001 (***).

For a more rigorous analysis, dependent T-tests are conducted using per visit R-precision scores. The first column in Table 6.6 shows the two models whose samples are compared. The p-values in Table 6.6 show that most of the pairwise tests were significant at the 0.001 level, except for the two cases that were significant at the 0.01 level. The test results show that the Hier-GRU-S has higher per-visit R-precision scores than the other two models, and the difference is significant. Also, the Hier-GRU model has significantly higher per-visit R-precisions than the logistic regression classifier.

6.5.3 Top Retrieved Utterances

As a qualitative analysis, the top 5 utterances for each label category are presented in Table 6.7. Those utterances are the ones that gave the highest output probabilities in each label category from the model Hier-GRU-S. The Hier-GRU-S model is used for these illustrations since it is the model that gave the best performance. All the retrieved utterances are clear enough to understand that the model has learned emotional valence information. The negative examples in Table 6.7 (a) have very negative expressions and words such as "terrible," "horrible," or "hate." On the other hand, the positive examples shown in Table 6.7 (c) include favorable words and expressions such as "wonderful" or "very good." The utterances that returned the highest neutral probabilities are usually related to physical examinations or prescription. The extended version with top 30 utterances can be found in Appendix D.

(a) Negative

Rank	$p(y_{i,j} = neg)$	Human	Text
1	0.987	-1.5	To wake up to that woman screaming , terrible , terrible .
2	0.978	-2.5	And a lot of times when I 'm cleaning houses myIll have such horrible hot flashes .
3	0.978	-2.5	And they 've been getting worse and worse and worse .
4	0.975	-2.0	The , oh I felt horrible .
5	0.975	-3.0	And I hate that , I just hate that .

(b) Neutral

Rank	$p(y_{i,j} = neu)$	Human	Text
1	0.998	0.0	Push , push , push , push , push , push .
2	0.998	0.0	Cardiac examination finds a regular rhythm with no murmur , gallop , or noted , period .
3	0.998	0.0	Push , push , push , push .
4	0.998	0.0	Okay , Zocor , -num- milligram .
5	0.998	0.0	Push , push , push , push , push , push .

(c) Positive

Rank	$p(y_{i,j} = pos)$	Human	Text
1	0.994	2.0	He 's wonderful .
2	0.994	2.0	Very good , very good .
3	0.993	2.0	That sounds wonderful .
4	0.992	2.2	Were very happy .
5	0.992	2.0	Excellent , excellent .

Table 6.7: Five example utterances with the highest output probabilities generated from Hier-GRU-S. The second column shows the output probability from the model, and the third column shows the averaged value of human-assigned labels for that utterance.

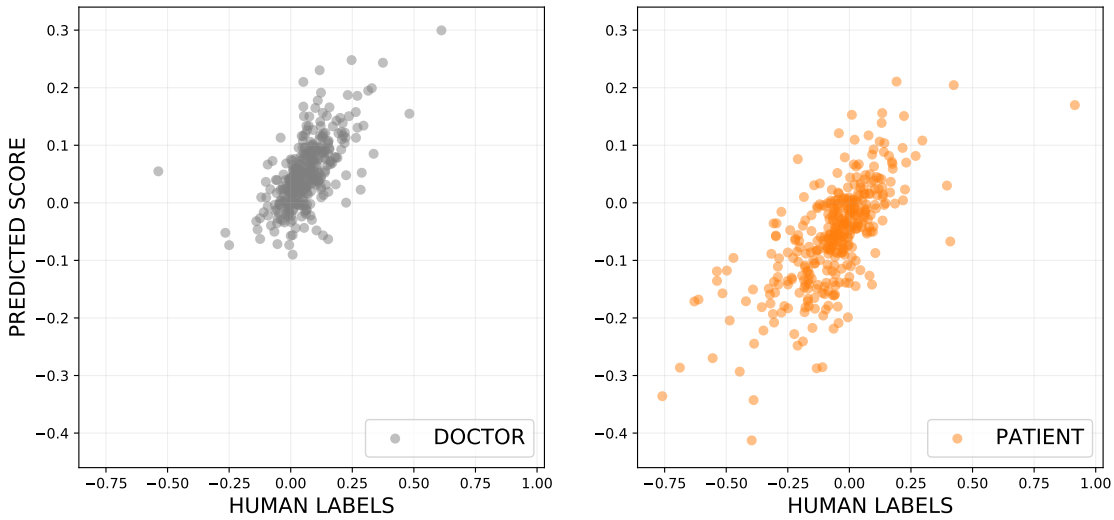


Figure 6.10: Mean emotion score per visit for doctors (left) and patients (right).

6.5.4 Predicting Mean Valence per Visit

Two scatter plots in Figure 6.10 show per-visit emotional valence scores from the human labelers and the Hier-GRU-S model. For each point in a scatter plot, the x value is the mean of all the labels in a visit, and the y value is the mean of the scores $l_{i,j}$ in each visit i. The left plot shows the scores for the utterances spoken by medical doctors, and the right plot shows the same for the utterances by patients.

Both of the plots present that there is a linear relationship between the predicted and the human-rated emotional valence at the visit level, for each speaker. The correlation coefficient between the two mean emotion scores (predicted and human-labeled) for doctor and patient was 0.63 and 0.69, respectively, showing that the model can predict the mean valence per visit. Two plots are shown in the same scale, and it is easy to notice the different areas that the points are located. Overall, patients emotion is more towards negative than that of physicians as it was observed in Figure 6.4.

6.6 Conclusions

Like any other clinical skill, learning to use empathy effectively requires practice and feedback, both of which require an investment of time and effort. However, learning to recognize patient emotions, and consequently identifying and understanding patient experiences and emotions, are necessary first steps.

In this chapter, I established important baselines for automatically predicting the emotions of doctors and patients during primary care visits. Moreover, this work is the first large-scale evaluation of emotional valence recognition in human conversations. Using more than 200k labeled utterances, I trained and evaluated the models by exploiting the two different types of outputs: emotional valence scores from the continuous output probabilities and the predicted categorical labels.

Good performance was observed when comparing the model to a simple baseline (LR) and to human OvR. However, using more recently developed models could possibly yield better results.

In summary, the primary contributions of this chapter include the following:

- I raised the importance of identifying and recognizing emotions in patient-physician conversations.
- I introduced a speaker-dependent neural network model that could be applied to longer dialogs than those in other existing conversation datasets.
- I suggested evaluating the results with Pearson correlation coefficient utilizing the output probability of the models, by treating each speaker's emotional trajectory separately.
- I calculated R-precision scores for each category as an additional evaluation metric.

- I compared the results (from both of the metrics) with the human one vs. rest results using statistical tests.

Bibliography

[1] Ai for social good icml2019 workshop. `https://aiforsocialgood.github.io/icml2019/index.htm`, June 2019. Accessed: 2019-11-14.

[2] National ambulatory medical care survey: 2016 national summary tables. `https://www.cdc.gov/nchs/data/ahcd/namcs_summary/2016_namcs_web_tables.pdf`, Sept. 2019. Accessed: 2019-11-08.

[3] O. Abdel-Hamid and H. Jiang. Fast speaker adaptation of hybrid nn/hmm model for speech recognition based on discriminative learning of speaker code. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7942–7946. IEEE, 2013.

[4] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, and R. Passonneau. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 30–38, 2011.

[5] C. C. Aggarwal. *Machine learning for text*. Springer, 2018.

[6] B. G. Arndt, J. W. Beasley, M. D. Watkinson, J. L. Temte, W.-J. Tuan, C. A. Sinsky, and V. J. Gilchrist. Tethered to the EHR: Primary care physician workload assessment using EHR event log data and Time-Motion observations. *Ann. Fam. Med.*, 15(5):419–426, Sept. 2017.

[7] A. J. Bank and R. M. Gage. Annual impact of scribes on physician productivity and revenue in a cardiology clinic. *ClinicoEconomics and outcomes research: CEOR*, 7:489, 2015.

[8] N. Banovic, T. Buzali, F. Chevalier, J. Mankoff, and A. K. Dey. Modeling and understanding human routine behavior. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, May 7-12, 2016*, pages 248–260, 2016.

[9] M. C. C. Baranauskas et al. Socially aware computing. In *ICECE 2009 VI International Conference on Engineering and Computer Education*, pages 1–5, 2009.

[10] P. J. Barr, M. D. Dannenberg, C. H. Ganoe, W. Haslett, R. Faill, S. Hassanpour, A. Das, R. Arend, M. C. Masel, S. Piper, H. Reicher, J. Ryan, and G. Elwyn. Sharing annotated audio recordings of clinic visits with Patients-Development of the open

recording automated logging system (ORALS): Study protocol. *JMIR Res. Protoc.*, 6(7):e121, July 2017.

[11] L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state markov chains. *The annals of mathematical statistics*, 37(6):1554–1563, 1966.

[12] M. C. Beach, S. Saha, P. T. Korthuis, V. Sharp, J. Cohn, I. B. Wilson, S. Eggly, L. A. Cooper, D. Roter, A. Sankar, and R. Moore. Patient-provider communication differs for black compared to white HIV-infected patients. *AIDS Behav.*, 15(4):805–811, May 2011.

[13] R. S. Beck, R. Daughtridge, and P. D. Sloane. Physician-patient communication in the primary care office: a systematic review. *J. Am. Board Fam. Pract.*, 15(1):25–38, Jan. 2002.

[14] Y. Bergner, D. Kerr, and D. E. Pritchard. Methodological challenges in the analysis of MOOC data for exploring the relationship between discussion forum views and learning outcomes. In *Proceedings of the EDM Conference*, pages 234–241. International Educational Data Mining Society (IEDMS), 2015.

[15] E. P. Bettinger, L. Fox, S. Loeb, and E. S. Taylor. Virtual classrooms: How online college courses affect student success. *American Economic Review*, 107(9):2855–75, 2017.

[16] S. Bhat, P. Chinprutthiwong, and M. Perry. Seeing the instructor in two video styles: Preferences and patterns. In *Proceedings of the EDM Conference*, pages 305–312. International Educational Data Mining Society (IEDMS), 2015.

[17] S. Bird, E. Klein, and E. Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. "O'Reilly Media, Inc.", June 2009.

[18] M. S. Boroujeni, K. Sharma, Ł. Kidziński, L. Lucignano, and P. Dillenbourg. How to quantify student's regularity? In *Proceedings of the 11th European Conference on Technology Enhanced Learning (EC-TEL 2016)*, pages 277–291, Lyon, France, sep 2016. Springer, Cham.

[19] T. Brijs, D. Karlis, G. Swinnen, K. Vanhoof, G. Wets, and P. Manchanda. A multivariate Poisson mixture model for marketing applications. *Statistica Neerlandica*, 58(3):322–348, 2004.

[20] C. G. Brinton and M. Chiang. MOOC performance prediction via clickstream data and social learning networks. In *Proceedings of the INFOCOM Conference*, pages 2299–2307. IEEE, 2015.

[21] J. Broadbent and W. Poon. Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *The Internet and Higher Education*, 27:1–13, 2015.

[22] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335, 2008.

[23] J. M. Calabrese, J. L. Brunner, and R. S. Ostfeld. Partitioning the aggregation of parasites on hosts into intrinsic and extrinsic components via an extended Poisson-gamma mixture model. *PLoS ONE*, 6(12):1–9, 2011.

[24] D. D. Caperton, D. C. Atkins, and Z. E. Imel. Rating motivational interviewing fidelity from thin slices. *Psychol. Addict. Behav.*, 32(4):434–441, June 2018.

[25] D. S. Carrell, R. E. Schoen, D. A. Leffler, M. Morris, S. Rose, A. Baer, S. D. Crockett, R. A. Gourevitch, K. M. Dean, and A. Mehrotra. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. *Journal of the American Medical Informatics Association*, 24(5):986–991, 2017.

[26] M. Casarrubea, G. Jonsson, F. Faulisi, F. Sorbera, G. D. Giovanni, A. Benigno, G. Crescimanno, and M. Magnusson. T-pattern analysis for the study of temporal structure of animal and human behavior: A comprehensive review. *Journal of Neuroscience Methods*, 239:34 – 46, 2015.

[27] R. Charon, M. G. Greene, and R. D. Adelman. Multi-dimensional interaction analysis: a collaborative approach to the study of medical discourse. *Soc. Sci. Med.*, 39(7):955–965, Oct. 1994.

[28] X. Chen and C. Carroll. Fgs in post-secondary education: A look at their college transcripts (nces 2005-171). us department of education. *National Center for Education Statistics. Washington, DC: US Government Printing Office. Retrieved from http://nces. ed. gov/pubs2005/2005171. pdf*, 2005.

[29] C.-C. Chiu, A. Tripathi, K. Chou, C. Co, N. Jaitly, D. Jaunzeikare, A. Kannan, P. Nguyen, H. Sak, A. Sankar, J. Tansuwan, N. Wan, Y. Wu, and X. Zhang. Speech recognition for medical conversations, 2018.

[30] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, pages 1082–1090, 2011.

[31] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8), 2014*, 2014.

[32] M. C. Chuah and F. Fu. ECG anomaly detection via time series analysis. In *Frontiers of High Performance Computing and Networking ISPA 2007 Workshops, ISPA 2007 International Workshops SSDSN, UPWN, WISH, SGC, ParDMCom, HiPCoMB, and IST-AWSN Niagara Falls, Canada, August 28 - September 1, 2007, Proceedings*, pages 123–135, 2007.

[33] G. Claeskens. Statistical model choice. *Annual Review of Statistics and its Application*, 3:233–256, 2016.

[34] R. Cmejla, J. Rusz, P. Bergl, and J. Vokral. Bayesian changepoint detection for the automatic assessment of fluency and articulatory disorders. *Speech Communication*, 55(1):178–189, 2013.

[35] R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[36] N. Colneriĉ and J. Demsar. Emotion recognition on twitter: Comparative study and training a unison model. *IEEE Transactions on Affective Computing*, 2018.

[37] S. Crossley, L. Paquette, M. Dascalu, D. S. McNamara, and R. S. Baker. Combining click-stream data with NLP tools to better understand MOOC completion. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge*, LAK '16, pages 6–14. ACM, 2016.

[38] J. R. Curtis, S. Sathitratanacheewin, H. Starks, R. Y. Lee, E. K. Kross, L. Downey, J. Sibley, W. Lober, E. T. Loggers, J. A. Fausto, C. Lindvall, and R. A. Engelberg. Using electronic health records for quality measurement and accountability in care of the seriously ill: Opportunities and challenges. *J. Palliat. Med.*, 21(S2):S52–S60, Mar. 2018.

[39] D. Davis, G. Chen, C. Hauff, and G.-J. Houben. Gauging MOOC learners' adherence to the designed learning path. In *Proceedings of EDM Conference*, pages 54–61. International Educational Data Mining Society (IEDMS), 2016.

[40] L. Deleger, K. Molnar, G. Savova, F. Xia, T. Lingren, Q. Li, K. Marsolo, A. Jegga, M. Kaiser, L. Stoutenborough, and I. Solti. Large-scale evaluation of automated clinical note de-identification and its impact on information extraction. *J. Am. Med. Inform. Assoc.*, 20(1):84–94, Jan. 2013.

[41] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1):1–38, 1977.

[42] F. Dernoncourt, J. Y. Lee, O. Uzuner, and P. Szolovits. De-identification of patient notes with recurrent neural networks. *J. Am. Med. Inform. Assoc.*, 24(3):596–606, May 2017.

[43] T. Dvorak and M. Jia. Online work habits and academic performance. *Journal of Learning Analytics*, 3(3):318–330, 2016.

[44] I. A. Eckley, P. Fearnhead, and R. Killick. *Bayesian Time Series Models*, chapter 10 Analysis of changepoint models, pages 205–224. Cambridge University Press, Cambridge, 2011.

[45] J. L. Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[46] G. C. Elvers, D. J. Polzella, and K. Graetz. Procrastination in online courses: Performance and attitudinal differences. *Teaching of Psychology*, 30(2):159–162, 2003.

[47] G. Elwyn, P. J. Barr, and S. W. Grande. Patients recording clinical encounters: a path to empowerment? assessment by mixed methods. *BMJ Open*, 5(8):e008566, Aug. 2015.

[48] R. M. Epstein, K. Fiscella, C. S. Lesser, and K. C. Stange. Why the nation needs a policy push on patient-centered health care. *Health Aff.*, 29(8):1489–1495, Aug. 2010.

[49] D. T. Eton, J. L. Ridgeway, M. Linzer, D. H. Boehm, E. A. Rogers, K. J. Yost, L. J. Finney Rutten, J. L. Sauver, St, S. Poplau, and R. T. Anderson. Healthcare provider relational quality is associated with better self-management and less treatment burden in people with multiple chronic conditions. *Patient Prefer. Adherence*, 11:1635–1646, Sept. 2017.

[50] W. Feng, Y. Liu, J. Wu, K. P. Nephew, T. H. M. Huang, and L. Li. A Poisson mixture model to identify changes in RNA polymerase II binding quantity using high-throughput sequencing technology. *BMC Genomics*, 9(Suppl 2):S23, 2008.

[51] P. K. Foo, R. M. Frankel, T. G. McGuire, A. M. Zaslavsky, J. E. Lafata, and M. Tai-Seale. Patient and physician race and the allocation of time and patient engagement efforts to mental health discussions in primary care. *J. Ambul. Care Manage.*, 40(3):246–256, July 2017.

[52] C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631, 2002.

[53] M. W. Friedberg, P. G. Chen, K. R. Van Busum, F. Aunon, C. Pham, J. Caloyeras, S. Mattke, E. Pitchforth, D. D. Quigley, R. H. Brook, F. J. Crosson, and M. Tutty. Factors affecting physician professional satisfaction and their implications for patient care, health systems, and health policy. *Rand Health Q*, 3(4):1, Dec. 2014.

[54] F. Gannon. Science for society. *EMBO reports*, 7(6):561–561, 2006.

[55] M. Gardner, J. Grus, M. Neumann, O. Tafjord, P. Dasigi, N. F. Liu, M. Peters, M. Schmitz, and L. S. Zettlemoyer. Allennlp: A deep semantic natural language processing platform. 2017.

[56] G. Gaut, M. Steyvers, Z. E. Imel, D. C. Atkins, and P. Smyth. Content coding of psychotherapy transcripts using labeled topic models. *IEEE J Biomed Health Inform*, 21(2):476–487, Mar. 2017.

[57] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian data analysis*. CRC press, 2013.

[58] P. G. Georgiou, M. P. Black, A. C. Lammert, B. R. Baucom, and S. S. Narayanan. thats aggravating, very aggravating: Is it possible to classify behaviors in couple interactions using automatically derived lexical features? In *International Conference on Affective Computing and Intelligent Interaction*, pages 87–96. Springer, 2011.

[59] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[60] Y. Goldberg. *Neural Network Methods in Natural Language Processing*. Morgan & Claypool Publishers, Apr. 2017.

[61] M. Goldstein and S. Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLOS ONE*, 11(4):1–31, 04 2016.

[62] A. Graves, A. Mohamed, and G. E. Hinton. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, pages 6645–6649, 2013.

[63] J. A. Hall, D. L. Roter, and N. R. Katz. Meta-analysis of correlates of provider behavior in medical encounters. *Med. Care*, 26(7):657–675, July 1988.

[64] K. W. Hammond, S. T. Helbig, C. C. Benson, and B. M. Brathwaite-Sketoe. Are electronic medical records trustworthy? observations on copying, pasting and duplication. *AMIA Annu. Symp. Proc.*, pages 269–273, 2003.

[65] M. Hasan, A. Kotov, A. Carcone, M. Dong, S. Naar, and K. B. Hartlieb. A study of the effectiveness of machine learning methods for classification of clinical interview fragments into a large number of categories. *J. Biomed. Inform.*, 62:21–31, Aug. 2016.

[66] D. Hazarika, S. Poria, A. Zadeh, E. Cambria, L.-P. Morency, and R. Zimmermann. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2122–2132, 2018.

[67] D. G. Hewett, B. M. Watson, C. Gallois, M. Ward, and B. A. Leggett. Communication in medical records: Intergroup language and patient care. *J. Lang. Soc. Psychol.*, 28(2):119–138, 2009.

[68] R. G. Hill, Jr, L. M. Sears, and S. W. Melanson. 4000 clicks: a productivity analysis of electronic medical records in a community hospital ED. *Am. J. Emerg. Med.*, 31(11):1591–1594, Nov. 2013.

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

[70] M. Hoerger, R. M. Epstein, P. C. Winters, K. Fiscella, P. R. Duberstein, R. Gramling, P. N. Butow, S. G. Mohile, P. R. Kaesberg, W. Tang, S. Plumb, A. Walczak, A. L. Back, D. Tancredi, A. Venuti, C. Cipri, G. Escalera, C. Ferro, D. Gaudion, B. Hoh, B. Leatherwood, L. Lewis, M. Robinson, P. Sullivan, and R. L. Kravitz. Values and options in cancer care (VOICE): study design and rationale for a patient-centered communication and decision-making intervention for physicians, patients with advanced cancer, and their caregivers. *BMC Cancer*, 13:188, Apr. 2013.

[71] P. Hofgesang and J. P. Patist. Online change detection in individual web user behaviour. In *Proceedings of WWW Conference*, pages 1157–1158. ACM, 2008.

[72] M. Hojat. The interpersonal dynamics in Clinician–Patient relationships. In M. Hojat, editor, *Empathy in Health Professions Education and Patient Care*, pages 129–150. Springer International Publishing, Cham, 2016.

[73] S. L. Hotle. *Applications of clickstream information in estimating online user behavior*. PhD thesis, Georgia Institute of Technology, 2015.

[74] C. Howes, M. Purver, and R. McCabe. Linguistic indicators of severity and progress in online text-based therapy for depression. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 7–16, 2014.

[75] W. C. Hsiao, D. B. Yntema, P. Braun, D. Dunn, and C. Spencer. Measurement and analysis of intraservice work. *JAMA*, 260(16):2361–2370, Oct. 1988.

[76] C. J. Hutto and E. Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Eighth international AAAI conference on weblogs and social media*, 2014.

[77] Z. E. Imel, M. Steyvers, and D. C. Atkins. Computational psychotherapy research: scaling up the evaluation of patient-provider interactions. *Psychotherapy*, 52(1):19–30, Mar. 2015.

[78] R. R. Jayasekare, R. Gill, and K. Lee. Modeling discrete stock price changes using a mixture of Poisson distributions. *Journal of the Korean Statistical Society*, 45(3):409–421, 2016.

[79] F. Jelinek. *Statistical methods for speech recognition*. MIT press, 1997.

[80] D. Jurafsky and J. H. Martin. *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2009.

[81] A. M. Kazerouni, S. H. Edwards, and C. A. Shaffer. Quantifying incremental development practices and their relationship to procrastination. In *Proceedings of the 2017 ACM Conference on International Computing Education Research*, ICER '17, pages 191–199. ACM, 2017.

[82] Y. Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751, 2014.

[83] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1532–43, 2014.

[84] C. Kirch and J. Tajduidje Kamgaing. Detection of change points in discrete valued time series. In R. Davis, S. Holan, R. Lund, and N. Ravishanker, editors, *Handbook of Discrete Valued Time Series*, chapter 11, pages 219–244. Chapman and Hall, 2014.

[85] A. Kotov, M. Hasan, A. Carcone, M. Dong, S. Naar-King, and K. BroganHartlieb. Interpretable probabilistic latent variable models for automatic annotation of clinical text. *AMIA Annu. Symp. Proc.*, 2015:785–794, Nov. 2015.

[86] G. D. Kuh, T. M. Cruce, R. Shoup, J. Kinzie, and R. M. Gonyea. Unmasking the effects of student engagement on first-year college grades and persistence. *The Journal of Higher Education*, 79(5):540–563, 2008.

[87] J. Kupiec. Robust part-of-speech tagging using a hidden markov model. *Computer Speech & Language*, 6(3):225–242, 1992.

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

[89] S. Lai, L. Xu, K. Liu, and J. Zhao. Recurrent convolutional neural networks for text classification. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 2267–2273, 2015.

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

[91] C. Learning. Community insights: Emerging benchmarks and student success trends from across the civitas. Technical report, December 2016.

[92] W. Levinson. Patient-centred communication: a sophisticated procedure. *BMJ Qual. Saf.*, 20(10):823–825, Oct. 2011.

[93] W. Levinson, C. S. Lesser, and R. M. Epstein. Developing physician communication skills for patient-centered care. *Health Aff.*, 29(7):1310–1318, July 2010.

[94] W. Levinson, W. B. Stiles, T. S. Inui, and R. Engle. Physician frustration in communicating with patients. *Med. Care*, 31(4):285–295, Apr. 1993.

[95] J. Li and H. Zha. Two-way Poisson mixture models for simultaneous document classification and word clustering. *Computational Statistics & Data Analysis*, 50(1):163–180, 2006.

[96] Z. Liu, R. Brown, C. Lynch, T. Barnes, R. S. Baker, Y. Bergner, and D. S. McNamara. MOOC learner behaviors by country and culture; an exploratory analysis. In *Proceedings of the EDM Conference*, pages 127–134. International Educational Data Mining Society (IEDMS), 2016.

[97] B. Löwe, U. Schulz, K. Gräfe, and S. Wilke. Medical patients' attitudes toward emotional problems and their treatment. what do they really want? *J. Gen. Intern. Med.*, 21(1):39–45, Jan. 2006.

[98] M. Lyman, N. Sager, L. Tick, N. Nhan, F. Borst, and J.-R. Scherrer. The application of natural-language processing to healthcare quality assessment. *Medical Decision Making*, 11(4_suppl):S65–S68, 1991.

[99] M. S. Magnusson. Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior research methods, instruments, & computers*, 32(1):93–110, 2000.

[100] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria. DialogueRNN: An attentive RNN for emotion detection in conversations, 2019.

[101] F. Maqbool, H. Bahadar, and M. Abdollahi. Science for the benefits of all: The way from idea to product. *Journal of Medical Hypotheses and Ideas*, 8(2):74–77, 2014.

[102] E. Mayfield, M. B. Laws, I. B. Wilson, and C. Penstein Rosé. Automating annotation of information-giving for analysis of clinical conversation. *J. Am. Med. Inform. Assoc.*, 21(e1):e122–8, Feb. 2014.

[103] J. McAuley and J. Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172. ACM, 2013.

[104] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[105] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 3(1):5–17, 2011.

[106] G. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*, volume 382. John Wiley & Sons, 2007.

[107] P. D. McNicholas. Model-based clustering. *Journal of Classification*, 33(3):331–373, 2016.

[108] E. G. Mishler. *The Discourse of Medicine: Dialectics of Medical Interviews*. Greenwood Publishing Group, 1984.

[109] A. L. Montgomery, S. Li, K. Srinivasan, and J. C. Liechty. Modeling online browsing and path analysis using clickstream data. *Marketing science*, 23(4):579–595, 2004.

[110] T. B. Moyers, T. Martin, J. K. Manuel, S. M. L. Hendrickson, and W. R. Miller. Assessing competence in the use of motivational interviewing. *J. Subst. Abuse Treat.*, 28(1):19–26, Jan. 2005.

[111] C. Mulryan-Kyne. Teaching large classes at college and university level: Challenges and opportunities. *Teaching in Higher Education*, 15(2):175–185, 2010.

[112] H. J. Murff, F. FitzHenry, M. E. Matheny, N. Gentry, K. L. Kotter, K. Crimin, R. S. Dittus, A. K. Rosen, P. L. Elkin, S. H. Brown, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. *Jama*, 306(8):848–855, 2011.

[113] K. P. Murphy. *Machine learning - a probabilistic perspective*. Adaptive computation and machine learning series. MIT Press, 2012.

[114] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman. Natural language processing: an introduction. *J. Am. Med. Inform. Assoc.*, 18(5):544–551, Sept. 2011.

[115] B. L. L. Ng, W. C. Liu, and J. C. K. Wang. Student motivation and learning in mathematics and science: A cluster analysis. *International Journal of Science and Mathematics Education*, 14(7):1359–1376, 2016.

[116] K. H. R. Ng, K. Hartman, K. Liu, and A. W. H. Khong. Modelling the way: Using action sequence archetypes to differentiate learning pathways from learning outcomes. In *Proceedings of the EDM Conference*, pages 167–174. International Educational Data Mining Society (IEDMS), 2016.

[117] U. D. of Health, H. Services, et al. Guidance regarding methods for de-identification of protected health information in accordance with the health insurance portability and accountability act (hipaa) privacy rule. *US Department of Health and Human Services, Washington, DC) Available at: https://www. hhs. gov/hipaa/for-professionals/privacy/special-topics/de-identification/index. html. Accessed September*, 26:2018, 2012.

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

[119] J. Park, D. Kotzias, P. Kuo, R. L. Logan IV, K. Merced, S. Singh, M. Tanana, E. Karra Taniskidou, J. E. Lafata, D. C. Atkins, M. Tai-Seale, Z. E. Imel, and P. Smyth. Detecting conversation topics in primary care office visits from transcripts of patient-provider interactions. *Journal of the American Medical Informatics Association*, 09 2019. ocz140.

[120] J. Park, K. Zhao, K. Peng, and W. Ping. Multi-speaker end-to-end speech synthesis. *arXiv preprint arXiv:1907.04462*, 2019.

[121] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[122] J. Pennington, R. Socher, and C. D. Manning. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[123] J. Pestian, H. Nasrallah, P. Matykiewicz, A. Bennett, and A. Leenaars. Suicide note classification using natural language processing: A content analysis. *Biomedical informatics insights*, 3:BII–S4706, 2010.

[124] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller. Deep voice 3: 2000-speaker neural text-to-speech. In *International Conference on Learning Representations*, 2018.

[125] P. R. Pintrich. A conceptual framework for assessing motivation and self-regulated learning in college students. *Educational Psychology Review*, 16(4):385–407, 2004.

[126] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.

[127] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *arXiv preprint arXiv:1905.02947*, 2019.

[128] A. Rajkomar, A. Kannan, K. Chen, L. Vardoulakis, K. Chou, C. Cui, and J. Dean. Automatically charting symptoms from Patient-Physician conversations using machine learning. *JAMA Intern. Med.*, Mar. 2019.

[129] K. Roberts and S. M. Harabagiu. A flexible framework for deriving assertions from electronic medical records. *J. Am. Med. Inform. Assoc.*, 18(5):568–573, 2011.

[130] C. Rudin and K. L. Wagstaff. Machine learning for science and society. *Machine Learning*, 95(1):1–9, Apr 2014.

[131] B. Schuller, M. Valster, F. Eyben, R. Cowie, and M. Pantic. Avec 2012: The continuous audio/visual emotion challenge. In *Proceedings of the 14th ACM International Conference on Multimodal Interaction*, ICMI '12, pages 449–456, New York, NY, USA, 2012. ACM.

[132] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[133] A. Shachak and S. Reis. The impact of electronic medical records on patient-doctor communication during consultation: a narrative literature review. *J. Eval. Clin. Pract.*, 15(4):641–649, Aug. 2009.

[134] S. Shaheen, W. El-Hajj, H. Hajj, and S. Elbassuoni. Emotion recognition from text based on automatically generated rules. In *2014 IEEE International Conference on Data Mining Workshop*, pages 383–392. IEEE, 2014.

[135] T. D. Shanafelt, O. Hasan, L. N. Dyrbye, C. Sinsky, D. Satele, J. Sloan, and C. P. West. Changes in burnout and satisfaction with Work-Life balance in physicians and the general US working population between 2011 and 2014. *Mayo Clin. Proc.*, 90(12):1600–1613, Dec. 2015.

[136] C. G. Shultz and H. L. Holmstrom. The use of medical scribes in health care settings: a systematic review and future directions. *The Journal of the American Board of Family Medicine*, 28(3):371–381, 2015.

[137] H. B. Simon. The write stuff: how good writing can enhance patient care and professional growth. *Am. J. Med.*, 126(6):467–471, June 2013.

[138] K. Singh, S. R. Meyer, and J. M. Westfall. Consumer-Facing data, information, and tools: Self-Management of health in the digital age. *Health Aff.*, 38(3):352–358, Mar. 2019.

[139] C. A. Sinsky. On presence: A tale of two visits. `https://catalyst.nejm.org/electronic-health-record-tale-two-visits/`, Dec. 2016. Accessed: 2019-6-27.

[140] A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, J. Grue Simonsen, and J.-Y. Nie. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 553–562. ACM, 2015.

[141] P. Steel and J. Ferrari. Sex, education and procrastination: an epidemiological study of procrastinators' characteristics from a global sample. *European Journal of Personality*, 27(1):51–58, 2013.

[142] M. Stewart, J. B. Brown, and W. W. Weston. Patient-Centred interviewing part III: Five provocative questions. *Can. Fam. Physician*, 35:159–161, Jan. 1989.

[143] M. A. Stewart. Effective physician-patient communication and health outcomes: a review. *CMAJ*, 152(9):1423–1433, May 1995.

[144] R. L. Street, G. Makoul, N. K. Arora, and R. M. Epstein. How does communication heal? pathways linking clinician–patient communication to health outcomes. *Patient Educ. Couns.*, 74(3):295–301, Mar. 2009.

[145] C. D. Stults, J. Elston Lafata, L. Diamond, L. MacLean, A. L. Stone, T. Wunderlich, R. M. Frankel, and M. Tai-Seale. How do primary care physicians respond when patients cry during routine ambulatory visits? *J. Commun. Healthc.*, 7(1):17–24, Mar. 2014.

[146] K. M. Styck. Best practices for supporting upward economic and social mobility for first-generation college students. *The School Psychologist*, 72(2):50–57, 2018.

[147] R. Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[148] M. Tai-Seale, E. Dillon, Y. Yang, R. Nordgren, R. L. Steinberg, T. Nauenberg, T. C. Lee, A. Meehan, J. Li, A. S. Chan, Frosch, and Dominick. Physicians' Well-Being linked to In-Basket messages generated by algorithms in electronic health records. *Health Aff.*, 38(7), 2019.

[149] M. Tai-Seale, L. A. Hatfield, C. J. Wilson, C. D. Stults, T. G. McGuire, L. C. Diamond, R. M. Frankel, L. MacLean, A. Stone, and J. Elston Lafata. Periodic health examinations and missed opportunities among patients likely needing mental health care. *Am. J. Manag. Care*, 22(10):e350–e357, Oct. 2016.

[150] M. Tai-Seale, T. McGuire, C. Colenda, D. Rosen, and M. A. Cook. Two-minute mental health care for elderly patients: inside primary care visits. *J. Am. Geriatr. Soc.*, 55(12):1903–1911, Dec. 2007.

[151] M. Tai-Seale, T. G. McGuire, and W. Zhang. Time allocation in primary care office visits. *Health Serv. Res.*, 42(5):1871–1894, Oct. 2007.

[152] M. Tai-Seale, C. W. Olson, J. Li, A. S. Chan, C. Morikawa, M. Durbin, W. Wang, and H. S. Luft. Electronic health record logs indicate that physicians split time evenly between seeing patients and desktop medicine. *Health Aff.*, 36(4):655–662, Apr. 2017.

[153] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani. Voiceloop: Voice fitting and synthesis via a phonological loop. In *International Conference on Learning Representations*, 2018.

[154] M. Tanana, A. Dembe, C. S. Soma, Z. Imel, D. Atkins, and V. Srikumar. Is sentiment in movies the same as sentiment in psychotherapy? comparisons using a new psychotherapy sentiment database. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 33–41, 2016.

[155] P. L. Teixeira, W.-Q. Wei, R. M. Cronin, H. Mo, J. P. VanHouten, R. J. Carroll, E. LaRose, L. A. Bastarache, S. T. Rosenbloom, T. L. Edwards, et al. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. *Journal of the American Medical Informatics Association*, 24(1):162–171, 2016.

[156] J. A. Teresi, M. Ramírez, K. Ocepek-Welikson, and M. A. Cook. The development and psychometric analyses of adept: an instrument for assessing the interactions between doctors and their elderly patients. *Annals of Behavioral Medicine*, 30(3):225–242, 2005.

[157] S. Thielke, K. Hammond, and S. Helbig. Copying and pasting of examinations within the electronic medical record. *Int. J. Med. Inform.*, 76 Suppl 1:S122–8, June 2007.

[158] K. Tóth, S. Greiff, C. Kalergi, and S. Wüstenberg. Discovering students' complex problem solving strategies in educational assessment. *Proceedings of the 7th International Conference on Educational Data Mining (EDM)*, pages 225–228, 2014.

[159] A. K. Uysal and S. Gunal. The impact of preprocessing on text classification. *Information Processing & Management*, 50(1):104–112, 2014.

[160] M. van Osch, S. van Dulmen, L. van Vliet, and J. Bensing. Specifying the effects of physician's communication on patients' outcomes: A randomised controlled trial. *Patient Educ. Couns.*, 100(8):1482–1489, Aug. 2017.

[161] M. K. Venetis, J. D. Robinson, K. L. Turkiewicz, and M. Allen. An evidence base for patient-centered cancer care: A meta-analysis of studies of observed communication between cancer specialists and their patients. *Patient Education and Counseling*, 77(3):379 – 383, 2009. Patient-Centered Cancer Communication Research.

[162] W. Ventres, S. Kooienga, N. Vuckovic, R. Marlin, P. Nygren, and V. Stewart. Physicians, patients, and the electronic health record: an ethnographic analysis. *Ann. Fam. Med.*, 4(2):124–131, Mar. 2006.

[163] A. Verghese, N. H. Shah, and R. A. Harrington. What this computer needs is a physician, 2018.

[164] R. A. Wagner and M. J. Fischer. The string-to-string correction problem. *Journal of the ACM (JACM)*, 21(1):168–173, 1974.

[165] K. J. Walker, W. Dunlop, D. Liew, M. P. Staples, M. Johnson, M. Ben-Meir, H. G. Rodda, I. Turner, and D. Phillips. An economic evaluation of the costs of training a medical scribe to work in emergency medicine. *Emerg Med J*, 33(12):865–869, 2016.

[166] B. C. Wallace, M. B. Laws, K. Small, I. B. Wilson, and T. A. Trikalinos. Automatically annotating topics in transcripts of patient-provider interactions via machine learning. *Med. Decis. Making*, 34(4):503–512, May 2014.

[167] G. Wang, X. Zhang, S. Tang, H. Zheng, and B. Y. Zhao. Unsupervised clickstream clustering for user behavior analysis. In *CHI Proceedings*, pages 225–236. ACM, 2016.

[168] H. Wang, D. Zhang, and K. G. Shin. Change-point monitoring for the detection of dos attacks. *IEEE Trans. Dependable Sec. Comput.*, 1(4):193–208, 2004.

[169] S. V. Wang, J. R. Rogers, Y. Jin, D. W. Bates, and M. A. Fischer. Use of electronic healthcare records to identify complex patients with atrial fibrillation for targeted intervention. *Journal of the American Medical Informatics Association*, 24(2):339–344, 2017.

[170] X. Wang, D. Yang, M. Wen, K. R. Koedinger, and C. P. Rose. Investigating how student's cognitive behavior in MOOC discussion forum affect learning gains. In *Proceedings of the EDM Conference*, pages 226–233. International Educational Data Mining Society (IEDMS), 2015.

[171] A. White and M. Danis. Enhancing patient-centered communication and collaboration by using the electronic health record in the examination room. *JAMA*, 309(22):2327–2328, June 2013.

[172] J. W. You. Examining the effect of academic procrastination on achievement using LMS data in e-learning. *Journal of Educational Technology & Society*, 18(3):64, 2015.

[173] J. W. You. Identifying significant indicators using lms data to predict course achievement in online learning. *The Internet and Higher Education*, 29:23–30, 2016.

[174] A. Zhang, B. Culbertson, and P. Paritosh. Characterizing online discussion using coarse discourse sequences. 2017.

[175] B. J. Zimmerman. Self-regulated learning and academic achievement: An overview. *Educational psychologist*, 25(1):3–17, 1990.

Appendices

A List of Tokens

-NAME-

-PATIENT-ID-

-PHONE-NUMBER-

-LOCATION-

-AGE-

-DATE-

-NUM-

Below are the non-verbal tokens that are removed for training and testing.

-POSITIVE-TONE-

-NEGATIVE-TONE-

-GASP-

-GROAN-

-HUMMING-

-SIGH-

-LAUGH-

-CRY-

-BREATH-

-COUGH-

-MUMBLE-

-PAUSE-

-CLAP-

-KNOCK-

-BEEP-

-PHONE-RING-

-NOISE-

-DOOR-SOUND-

B Reading Protocols

1. Before reading, clap once. This will generate a unique signature on the audio waveforms, and allow for syncing of audio in post.

2. Complete the reading as naturally in character as able. Refer to reference information below regarding pacing, reading transcribed punctuation, proper nouns, numbers, and typos.

General Pacing

- If a talk-turn ends with a period, pause for 1 second.
- If a talk turn does not end with punctuation [Interrupt] on next talk-turn.
- If a talk-turn includes repeated punctuation, repeat the effect.

Punctuation

- Period: Pause briefly as if ending a statement.
- Comma: Pause briefly.
- Question mark
  - End statement with slight upward inflection, as if asking a question.
  - If occurring in isolation, pause briefly.
- Exclamation mark: Emphasize sentence.
- Semicolon, colon: not used.
- Hyphen / Dash

- Hyphen: Read as a compound adjective
- Em dash: Pause briefly, as if introducing a list
- En dash: Read as a compound adjective (e.g. nobel-prize-winning) or time span (e.g. 100-400).

Numbers

- Numbers are replaced with the token, `-NUM-`.
- Replace these with any number you believe would most naturally fit into the talk turn.
  - e.g. -NUM- is read as any cardinal number (e.g. one, two, three, four, forty-four) in I took -NUM- pills yesterday.
  - e.g. -NUM- is read as any ordinal number (1st, fourth) in the talk turn, Yeah, Id like my prescription sent to the pharmacy on -NUM- street
- For the dates, use the date of recording.

Names

- Replace these with any name that you deem is appropriate.
- Replace name of feature with generic term (e.g. valley, island, downtown, city, etc.).

Disfluencies

- Include all in reading as naturally as able (e.g. uh, um, uh-huh, mm-hmm, huh).

Sounds

- `-DOORSOUND-, -CLAP-`: Ignore and pause

Stopping

- One reader claps and pause.

C R-Precision Score for Each Subset

Detailed tables with R-precision scores for each model. Each row in tables correspond to the subset shown in Table 6.1.

Test Labeler	Num Visits	Num Utters	Human		Models			
			OvR	R	Hier-GRU-S	Hier-GRU	LR	R
9	93	49849	0.3865	4354	0.4558	0.4291	0.4162	4318
13	84	44088	0.4210	2100	0.4458	0.4213	0.4131	3905
10	79	43025	0.4709	3113	0.4438	0.4247	0.4017	3565
4	50	26403	0.4661	1682	0.4643	0.4432	0.4320	2692
14	50	23577	0.4426	1787	0.4345	0.4164	0.4121	2327
5	27	8644	0.6414	1001	0.4505	0.4182	0.4332	868
8	5	1392	0.1176	34	0.3968	0.3651	0.3651	63
2	2	1387	0.4130	46	0.3846	0.4231	0.4103	78
7	2	630	0.2857	7	0.3333	0.3889	0.3333	18
3	2	630	0.0000	7	0.3333	0.3889	0.3333	18
Weighed Average			0.2739		0.3829	0.3981	0.3792	

Table C.1: R-precision score for each subset, for negative class.

Test Labeler	Num Visits	Num Utters	Human		Models			
			OvR	R	Hier-GRU-S	Hier-GRU	LR	R
9	93	49849	0.8813	40812	0.9050	0.9017	0.8956	40773
13	84	44088	0.9182	38795	0.9011	0.8976	0.8922	35875
10	79	43025	0.9009	36417	0.8988	0.8955	0.8876	34905
4	50	26403	0.9155	22599	0.8796	0.8759	0.8699	20489
14	50	23577	0.8767	19306	0.8783	0.8751	0.8701	18154
5	27	8644	0.8916	6737	0.8964	0.8902	0.8887	6902
8	5	1392	0.9285	1272	0.9437	0.9405	0.9324	1243
2	2	1387	0.9501	1263	0.9254	0.9237	0.9220	1167
7	2	630	0.9904	622	0.9770	0.9803	0.9753	608
3	2	630	0.9871	622	0.9770	0.9803	0.9753	608
Weighed Average			0.9487		0.9414	0.9407	0.9352	

Table C.2: R-precision score for each subset, for neutral class.

Test Labeler	Num Visits	Num Utters	Human		Models			
			OvR	R	Hier-GRU-S	Hier-GRU	LR	R
9	93	49849	0.4679	4683	0.5786	0.5738	0.5324	4758
13	84	44088	0.4767	3193	0.5701	0.5692	0.5304	4308
10	79	43025	0.4160	3495	0.5662	0.5622	0.5067	4555
4	50	26403	0.5071	2122	0.5701	0.5686	0.5087	3222
14	50	23577	0.4807	2484	0.6027	0.6005	0.5646	3096
5	27	8644	0.5442	906	0.5961	0.5973	0.5538	874
8	5	1392	0.3140	86	0.5698	0.5581	0.4767	86
2	2	1387	0.5256	78	0.5986	0.5845	0.6268	142
7	2	630	1.0000	1	0.5000	0.5000	0.5000	4
3	2	630	1.0000	1	0.5000	0.5000	0.5000	4
Weighed Average			0.6177		0.5522	0.5457	0.5237	

Table C.3: R-precision score for each subset, for positive class.

.

D Top 30 Retrieved Utterances

Rank	$p(y_{i,j} = pos)$	Human	Text
1	0.994	2.0	He 's wonderful .
2	0.994	2.0	Very good , very good .
3	0.993	2.0	That sounds wonderful .
4	0.992	2.2	Were very happy .
5	0.992	2.0	Excellent , excellent .
6	0.992	1.5	Thank you !
7	0.991	1.5	Good , good .
8	0.991	2.0	Good , good , good .
9	0.991	1.5	Thanks .
10	0.991	2.0	Great , great .
11	0.990	2.0	Thank you .
12	0.990	0.7	Thank you .
13	0.990	1.5	Thank you , yes , thank you .
14	0.990	2.0	Good , thanks .
15	0.990	1.8	Happy anniversary .
16	0.990	1.3	Wonderful , wonderful .
17	0.990	2.0	Wonderful .
18	0.990	2.0	Wonderful .
19	0.990	0.5	That is wonderful .
20	0.990	1.0	I 'm doing pretty good .
21	0.990	1.3	I 'm good thanks .
22	0.990	2.0	Wonderful .
23	0.989	0.0	Wonderful , good .
24	0.989	2.5	That is great .
25	0.989	1.3	Good , great .
26	0.989	2.5	I am wonderful .
27	0.989	2.0	Excellent , I 'm glad to hear that , very glad to hear that .
28	0.989	2.2	Thank you .
29	0.989	1.3	Good , good .
30	0.989	1.5	Here are your wonderful labs .

Table D.4: Top 30 utterances for positive.

Rank	$p(y_{i,j} = neu)$	Human	Text
1	0.998	0.0	Push , push , push , push , push , push .
2	0.998	0.0	Cardiac examination finds a regular rhythm with no murmur , gallop , or noted , period .
3	0.998	0.0	Push , push , push , push .
4	0.998	0.0	Okay , Zocor , -NUM- milligram .
5	0.998	0.0	Push , push , push , push , push , push .
6	0.998	0.0	Flu vaccine and pneumonia shot
7	0.998	0.0	Okay , take one pill at bedtime only as needed .
8	0.998	0.0	Any surgeries , N ?
9	0.997	0.0	Okay , breathe normally .
10	0.997	0.0	Celebrex , -NUM- milligrams daily or as needed .
11	0.997	0.0	Okay , so I 'll send the , do it get it , um , through mail order ?
12	0.997	0.0	No pap smear necessary .
13	0.997	0.0	Any black or tarry looking stools ?
14	0.997	0.0	Okay , any sexual transmitted diseases ever ?
15	0.997	0.0	Any vitamins or supplements ?
16	0.997	0.0	Okay , deep breath .
17	0.997	0.0	Okay , now , what about , um , flu shot and pneumonia vaccine ?
18	0.997	0.0	refill , yeah , refill .
19	0.997	0.0	Rectal examination finds a normal prostate in a symmetric fashion with no nodule , period .
20	0.997	0.0	vitamins or supplements ?
21	0.997	0.0	Okay , open your eyes .
22	0.997	0.0	Cardiac examination finds a regular rhythm without murmur , gallop or .
23	0.997	0.0	Breathe normally .
24	0.997	0.0	Mammogram , ultrasound and bone density .
25	0.997	0.0	Uh , prescription refill will be given .
26	0.997	0.0	Okay , switch hands .
27	0.997	0.0	CBC , chemistry profile .
28	0.997	0.0	Nasal examination finds some mild left septal deviations with the right side nasal polyp .
29	0.997	0.0	Lasix -NUM- milligrams once a day .
30	0.997	0.0	Does the Zantac relieve these symptoms when

Table D.5: Top 30 utterances for neutral.

Rank	$p(y_{i,j} = neg)$	Human	Text
1	0.987	-1.5	To wake up to that woman screaming , terrible , terrible .
2	0.978	-2.5	And a lot of times when I 'm cleaning houses myIll have such horrible hot flashes .
3	0.978	-2.5	And they 've been getting worse and worse and worse .
4	0.975	-2.0	The , oh I felt horrible .
5	0.975	-3.0	And I hate that , I just hate that .
6	0.975	-0.7	the sheets were wet but I go , Oh my God , this is terrible .
7	0.974	0.0	My mom passed and I know that I 've been feeling really tired .
8	0.965	-1.3	And then all of a sudden I felt like I was going to get sick , so I ran back in the bathroom , I felt clammy .
9	0.961	-0.8	It 's just really weird .
10	0.961	-1.0	I mean my back is sore , my neck is sore .
11	0.958	-2.0	She went snip , snip , snip , snip , snip .
12	0.958	-2.5	I 'm , I 'm buried alive in paperwork and , and just more shit gets thrown at me , sorry .
13	0.957	-0.8	That was terrible .
14	0.956	-0.8	Weird things happen .
15	0.955	-2.5	It just it hurts .
16	0.955	-2.5	And all of a sudden now for the last six , eight months it 's just been horrible .
17	0.954	-2.0	And I 'm tired of telling people I 'm sorry for crying .
18	0.953	-0.7	Cause then you 're real nervous and shaky .
19	0.953	-1.0	It was terrible .
20	0.952	-2.5	I almost passed out .
21	0.952	-2.5	It 's terrible .
22	0.951	-2.0	I 'm so stressed at work
23	0.950	-2.5	It 's getting worse now .
24	0.949	-1.5	It hurts for days .
25	0.948	-1.0	Well like right now my , my neck is sore .
26	0.947	-0.7	And sometimes when you get afraid , it makes it even worse .
27	0.947	-1.0	It hurts .
28	0.947	-1.0	my arthritis is driving me crazy .
29	0.947	-0.5	It 's like , my breasts are tender .
30	0.946	-2.0	I do n't know why my back hurts so much .

Table D.6: Top 30 utterances for negative.