UNDERSTANDING PATIENT EXPERIENCE FROM ONLINE MEDIUM

BY

HYUN DUK CHO

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Associate Professor Roxana Girju, Chair
Professor Chengxiang Zhai
Associate Professor Hari Sundaram
Assistant Professor Michael Paul, University of Colorado at Boulder

ABSTRACT

Improving patient experience at hospitals leads to better health outcomes. To improve this, we must first understand and interpret patients' written feedback. Patient-generated texts such as patient reviews found on RateMD, or online health forums found on WebMD are venues where patients post about their experiences. Due to the massive amounts of patient-generated texts that exist online, an automated approach to identifying the topics from patient experience taxonomy is the only realistic option to analyze these texts. However, not only is there a lack of annotated taxonomy on these media, but also word usage is colloquial, making it challenging to apply standardized NLP technique to identify the topics that are present in the patient-generated texts. Furthermore, patients may describe multiple topics in the patient-generated texts which drastically increases the complexity of the task.

In this thesis, we address the challenges in comprehensively and automatically understanding the patient experience from patient-generated texts. We first built a set of rich semantic features to represent the corpus which helps capture meanings that may not typically be captured by the bag-of-words (BOW) model. Unlike the BOW model, semantic feature representation captures the context and in-depth meaning behind each word in the corpus. To the best of our knowledge, no existing work in understanding patient experience from patient-generated texts delves into which semantic features help capture the characteristics of the corpus. Furthermore, patients generally talk about multiple topics when they write in patient-generated texts, and these are frequently interdependent of each other. There are two types of topic interdependencies, those that are semantically similar, and those that are not. We built a constraint-based deep neural network classifier to capture the two types of topic interdependencies and empirically show the classification performance improvement over the baseline approaches.

Past research has also indicated that patient experiences differ depending on patient segments [1–4]. The segments can be based on demographics, for instance, by race, gender, or geographical location. Similarly, the segments can be based on health status, for example, whether or not the patient is taking medication, whether or not the patient has a particular disease, or whether or not the patient is readmitted to the hospital. To better understand patient experiences, we built an automated approach to identify patient segments with a focus on whether the person has stopped taking the medication or not. The technique used to identify the patient segment is general enough that we envision the approach to be applicable to other types of patient segments.

With a comprehensive understanding of patient experiences, we envision an application system where clinicians can directly read the most relevant patient-generated texts that pertain to their interest. The system can capture topics from patient experience taxonomy that is of interest to each clinician or designated expert, and we believe the system is one of many approaches that can ultimately help improve the patient experience.

To my parents, for their love and support.

ACKNOWLEDGMENTS

First, I must thank my advisor, Professor Roxana Girju, for her patience, empathy, guidance, and inspiration during my journey as a Ph.D. student. She was always around to encourage me to continue when I encountered various roadblocks in research or to show tough love when I needed a bit more push. She was always very understanding of my circumstances and did everything to ensure that I can work around hardships I have encountered. Furthermore, she taught me how to be a successful researcher. Without her, I do not think I would have been able to write this dissertation, and I owe her all my gratitude.

I would also like to thank committee members, Professor Chengxiang Zhai, Professor Hari Sundaram, and Professor Michael Paul. They provided invaluable and insightful feedback on the dissertation. Prof. Zhai was an inspiration and was able to provide a new angle to research problems. He taught me how to look at a research problem, not as an independent project, but to view it from high up. Prof. Sundaram did not hesitate to help out during an unfortunate event, and I would like to express my gratitude for his generosity. Prof. Paul has in-depth knowledge of patient-generated texts, and I gained invaluable insights into the current state of the art in the domain.

This thesis would not have been possible without Prof. Marina Girju, our collaborator at DePaul University as well. She was a lifesaver in numerous circumstances when the situation looked bleak, but was always around to help overcome these events. Furthermore, she also provided an outsiders' view of how one should understand the patient experience.

I would also like to thank both Kannan Achan and Sushant Kumar, both of whom were/are my managers at WalmartLabs. Both of them continuously encouraged me to finish up the Ph.D. degree and were very accommodating when I needed time off to finish up on the dissertation. I would also like to thank Dr. Cinda Heeren for her encouragement and for showing me the value of mentoring students.

Two administrative staff members were continuously helpful throughout my Ph.D. career. I have been in constant contact with these two staff members and would like to show appreciation. Both Kara MacGregor and Vivek Kudaligama would always go above and beyond to ensure that paperwork went through smoothly.

During the duration of my Ph.D. career, I have encountered friends and professional connections for their help, support, and socialization. I would first like to acknowledge Casey Hanson, Pei-Chen Peng, Farzaneh Khajouei, Laura Sloofman, Charles Blatti, and Thyago Duque who were all members of the Bioinformatics group under Professor Sinha, for

accepting me as an unofficial, social lab member. I would also like to thank Majid Kazemian, Kai Zhao, Hongning Wang, and Parikshit Sondhi, who were senior Ph.D. members when I joined and showed me the ropes to being a successful Ph.D. student. I'm thankful for having Fred Douglas, Harry Rocha, Qian Cheng, Thomas Zhang, Yichuan Cao, Connie Park, Mark Pszczolkowski, and Codruta Girlea who provided support and enjoyable conversations throughout the Ph.D. career. I also have to thank Helen Catherine for going above and beyond in breaking a piece of unfortunate news, without which, I do not think I would even have been able to schedule the final defense. Other friends and professional connections were simply inspirational as well. Abhay Kashyap, Lia McLean, and Ryan Panos all showed that it is still possible to finish up a Ph.D. while working. My close friends Sean Massung and Chase Geigle both provided invaluable feedback on my dissertation and on countless hours of listening to my hardships. Ryan Joseph Sherwood was a very responsive and thorough proofreader. Suhan Kumar was a phenomenal mentee while he was interning at WalmartLabs, and also selflessly provided invaluable feedback on the dissertation.

As the famous saying goes, whatever doesn't kill you makes you stronger, and two people demonstrated that sentiment: Prof. Andrew Gallan taught me that sometimes being forced to start from scratch may lead to stronger research work, and Prof. Bruce Schatz taught me the importance of research citations, even though they are occasionally not warranted.

Throughout my Ph.D. career, I was fortunate enough to have mentored numerous undergraduate students. I would first like to thank Promoting Undergraduate Research in Engineering (PURE) program at UIUC for providing invaluable opportunities for graduate students to mentor undergraduate students. I was fortunate enough to have mentored fourteen undergraduates from this program: Rafael Drummond, RJ Kunde, Mariko Wakabayashi, Shravan Gupta, James Lee, Dong Min Shin, Thomas Olson, Ronit Chakraborty, Dawith Ha, Richard Lee, James Chen, Saubhagya Rathi, Rahul Majumdar, and Kevin Lim. While I was officially their mentor, I have also learned a lot from them and owe them my gratitude. I was also able to mentor two other undergraduate students under Prof. Roxana Girju. Tony Gao and Marina Shah were very agreeable, and we were able to conduct exciting research together.

I must also thank my partner, Tom Shepard, for his support, patience, encouragement, and understanding. In the last six months before defending this thesis, we both spent weekdays and weekends working late at night, him on his film, and myself on the thesis. He did not complain even when things were getting intense, and always had very nice things to say.

This thesis would not have been possible without the support of my parents. Despite being thousands of miles away, I was still able to feel their love.

TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION

1.1 PATIENT EXPERIENCE AND HEALTH OUTCOME

Improving patient experience at hospitals leads to better health outcomes [5–7]. Patient experience and health outcomes were once commonly thought to be separate issues, and practitioners believed that improving one aspect of the hospital experience did not influence the other. Perhaps because of these reasons, many physicians and researchers did not investigate ways to improve the patient experience.

According to a meta-study consisting 21 academic articles, however, there are correlations between positive patient experience and desirable health outcomes [8]. The study focused, in particular, on the link between the health outcomes and effective communication between the doctors and the patients. Examples of effective communications are providing a more affirmative statement ('You will be better in the next few days,' over 'I am not sure if you will be better in the next few days'), or providing better information which leads to better health outcomes. Furthermore, an agreement between the patient and the physician about the nature of the problem, a by-product of effective communication, bodes well for a successful treatment outcome. Ultimately, the researchers found that out of all the articles analyzed, 16 of these showed that effective communication between physicians and patients improved health outcomes, 4 showed adverse outcomes, and one showed no difference.

Many later studies replicated the relationship between patient experience and health outcomes. Street et al. [7, 9] suggests good communication between patient and clinicians lead to better health outcomes not only because of the therapeutic nature of the conversation, but also because of improved patient understanding, trust, and clinician-patient agreement. Furthermore, this allows patients to talk more openly about their pain and other health concerns. Another meta study [6] of 55 research articles found positive associations between patient experience, patient safety, and clinical effectiveness. They suggest that patients with a better hospital experience are more likely to be vigilant in tracking their health.

In all of these studies, positive patient experience is often linked with better health outcomes and vice versa. We believe that for patients to have positive hospital experiences, we must first understand their experience. Patient-generated texts are one means through which patients describe their experience. Patient-generated texts refer to those that patients author after a medical event such as taking medication, visiting a doctor's office, or changing their health status (such as, catching a cold, or stop taking medication). Generally, patient-generated texts can be found online on websites such as RateMD (`www.ratemd.com`) or WebMD (www.webmd.com) in the form of patient reviews or online health forums. Patients may ask or respond to questions or share their experience on a wide range of topics in these venues.

Patient-generated texts, in particular, were previously utilized in gathering information from patients on such topics as those regarding prescription drugs, epidemiology or sickness [10–19]. However, there are limited number of works that seek to comprehensively and automatically understand patient experience in these venues. In this thesis, we propose a **comprehensive** and **automated** approach to understanding patient experience from patient generated texts.

1.2 IMPORTANCE OF COMPREHENSIVELY UNDERSTANDING PATIENT EXPERIENCE FROM PATIENT GENERATED TEXTS

We gain a comprehensive understanding of the patient experience by analyzing a wide range of topics on which patients post regarding their experience. Comprehensive understanding of the patient experience refers to representing **what** the patients post, and **who** authors these type of posts. Characterizing **what** topics patients write about is relevant because it allows us to identify different topics present from patient experience taxonomy. In the literature, the practice and science of characterizing concepts, including the principles that underlie such classification is called taxonomy. Without a taxonomy to identify the topics that patients mention, we cannot characterize patient-generated texts, since there is no systematic way to represent what they have written. Furthermore, the taxonomy should cover most, if not all, of the topics patients write about in patient-generated texts. A typical comprehensive patient experience taxonomy covers dozens of topics [20–23]. Only then, can we understand what patients write regarding their experiences in patient-generated texts.

The second instance, understanding **who** posts these types of texts, pertains to identifying patient segments. We define patient segments as a group of patients who share similar traits such as their health status, or medications that they are taking. In an in-depth analysis of patient satisfaction data from Veterans Health Administration (VHA) [24], researchers found that age, health status, and race consistently had a statistically significant effect on satisfaction scores. Of these, health status is determined by whether the patient was readmitted to the hospital, had an adverse drug reaction, or had a particular illness [1–4]. Researchers should also consider patients' segments which allow us to deepen our understanding of who expresses what topics in patients' experiences.

However, comprehensiveness is not enough to properly understand the patient experience from patient-generated texts. There are dozens of websites such as RateMD, and WebMD where patients can write about their experiences. Each of these has millions of patient-generated texts, so manually labeling them is infeasible. To properly understand this massive amount of texts, we need an automated way to annotate these texts comprehensively.

1.3 CHALLENGES IN AUTOMATED EXTRACTION OF COMPREHENSIVE PATIENT EXPERIENCE

Past works in understanding patient experience have focused on only one of the two aspects: they were either comprehensive but not automated, or automated but not comprehensive. Lopez et al [20], Doyle et al [21], Jung et al [22], and Emmert et al [23] for instance, proposed a comprehensive patient experience taxonomy covering dozens of topics. Their research, however, was not automatic and required manual annotations. On the opposite end of the spectrum, Doing-Harris et al [25] proposed an automated approach to identifying different topics in patient experience. However, they only covered seven topics in their analysis, and hence, it is not comprehensive. Unsupervised methods, such as those used by Ranard et al [26] and Wallace et al [27] are highly automatic and require minimal training data. However, these methods are not comprehensive because unsupervised methods, by definition, do not have a structured taxonomy that the topics fall into. We believe that part of the reason why researchers do not approach the problem both automatically and comprehensively is because it is a challenging task. Identifying patient experiences **automatically** and **comprehensively** from patient generated texts is difficult for four reasons.

First, annotating patient-generated texts with comprehensive list of topics related to patient experience is a challenging task. With recent development in machine-learning literature, popular supervised learning approaches such as Logistic Regression, Support Vector Machines and Deep Neural Network frameworks have gained traction in automatically classifying a target dataset. However, to utilize these methods, we must first obtain a labeled dataset. In the absence of one, as was the case in our problem domain, researchers must then manually annotate the dataset with the corresponding labels. The annotated dataset should have a relatively high inter-annotator agreement, a key metric in ensuring the quality of the annotation. However, as the number of labels increases, the more difficult it becomes to achieve high agreement. In a comprehensive patient experience taxonomy, there is a large number of topics. Hence a high-quality annotation process becomes complicated.

Second, capturing both the shallow and syntactic (e.g., those represented by unigram, or dependency parses), and the deep semantic representation of patient-generated texts is a difficult task. Patient-generated texts encode a wide variety of information useful for hospital decision making [28]. However, processing and analyzing these texts brings a series of challenges [25, 29–32] stemming both from how patients write (form, style, grammar, language, etc.) as well as from what they write (the content itself). Unlike other free-form text, in their vast majority, patient comments are short, relatively concise and written with no context. While previous research does well in analyzing long pieces of text, it generally disregards short and vague comments [20, 33]. Furthermore, patient comments abound in shortenings and abbreviations defined by the patients themselves rather than determined consistently by the medical field. What further complicates the task is that their language is informal, with many stylistics (i.e., 'lol', 'wtf') and figurative terms (sarcasm, metaphors, etc.). All these issues pose a significant challenge for text-processing tools.

Furthermore, automatically identifying topics in patient experience from patient-generated texts is a difficult task. When a patient writes about their experience, he or she has a specific list of topics that he or she would like to express. The patient may decide to talk about a single topic or multiple topics. In automatically understanding patient experience, the model should consider whether patients generally talk about a single topic, or if they talk about multiple topics. Furthermore, if a patient talks about multiple topics, some of the topics may naturally co-occur. We define the tendency for two topics to co-occur in a given corpus as topic inter-dependency. However, capturing topic inter-dependency is a difficult task. There are exponentially many combinations on how labels can correlate, making understanding how the topics co-occur an intractable task.

Fourth, inferring patient segments is a difficult task. While some websites allow patients to declare their segments, for instance, gender, geographical locations, or their health status in their user profile, many others opt not to do so in fear of potential privacy issues. However, the lack of publicly available patient profiles poses difficulty in inferring their segments, dampening the possibility of comprehensively understanding patient experience.

1.4 CONTRIBUTIONS

To automatically and comprehensively understand patient experience, we are faced with four separate challenges: high-quality data annotation, feature representation, topic representation, and patient segment representation. In this thesis, we present our contributions to tackle the four separate challenges. With a comprehensive and automated approach to understanding patient experience in place, we explore how this can be applied in a real-world application domain, which is our fifth contribution. We make the following contributions:

Contribution 1.1. We utilized a state of the art patient experience taxonomy to classify topics from the taxonomy on patient-generated texts.

In collaboration with researchers at DePaul University, we built a classifier to an automatically annotate topics from patient experience taxonomy on patient-generated texts. In order to realize this goal, annotated dataset with topics from patient experience is necessary. The researchers at DePaul University first proposed a new patient experience taxonomy motivated by Healthcare Providers and Systems (HCAHPS), a survey which is a patient experience questionnaire already being used by hospitals [34], Lopez et al [20, 35], and Doyle et al [21].

Next, our collaborators at DePaul University annotated patient-generated text based on their proposed taxonomy. The annotators from DePaul University went through a multi-step process to annotate comprehensive and high-quality patient experience taxonomy annotations in patient reviews. We calculated the inter-annotator agreement and using this agreement as a guidance, the collaborators refined the definition of each topic in patient experience taxonomy on each iteration of the annotation process.

With the new annotation in place, we were able to train our models on the new corpus. We discuss the annotation process in more detail, because it is an integral part of the dataset that we use in this thesis in Chapter 3.

Contribution 1.2. We utilized a comprehensive patient experience taxonomy built by the collaborators at DePaul university to develop a classifier which was trained on rich semantic feature representation of the corpus that we identified for patient-generated texts.

Researchers, in understanding patient experience from patient generated texts, typically represent the corpus with a bag-of-words (BOW) model [25–27]. The BOW model treats each word in a document as a set of words, disregarding grammar, and word order. Furthermore, the BOW model ignores context of the words, and is incapable of capturing which words may be semantically similar. As an example: two sentences, '*I want to express my gratitude to my doctor*,' and '*I want to express my appreciation to my doctor*' both capture similar meaning, but a BOW-based classifier will not recognize the resemblance. By capturing the semantic representation of 'gratitude' and 'appreciation,' the classifier can capture the similarity.

To the best of our knowledge, our work is the first in utilizing semantic representation of patient generated texts to classify topics in patient experience taxonomy. We utilized and adapted semantic tools such as WordNet [36], and latent embedding space [37–39] with the aim of capturing complex word usage present in the RateMD corpus. Our classifier was trained on the annotated RateMD dataset that our collaborators at DePaul University annotated with topics from patient experience taxonomy. The in-depth discussion of this contribution is in Chapter 4.

Contribution 1.3. We trained a comprehensive patient experience taxonomy classifier by incorporating label inter-dependencies.

Existing works in understanding patient-generated texts focus on how topics are discussed in the corpus as a whole, for example, by analyzing which unigram features are prominent for a given topic [11, 25]. Topic analysis is conducted on a wide range of areas such as those in understanding patient experience [13–15, 20, 25], health status [16–19, 40], and drug usage [10–12]. However, such an analysis of topics does not capture how the two topics may be related to each other. Consider the following example:

"His staff are VERY rude on the phone and in person. You have to wait at least a month to get in to see him"

In traditional approaches, three topics, 'Professionalism,' 'Friendliness,' and 'Appointment Access' are captured in this snippet. However, these approaches do not capture, in their model representation, whether 'Friendliness' and 'Appointment Access,' for example, are interdependent or not. They assume that each of the topic is independent of each other, i.e., they presume topics follow binary relevance model [41]. We explore whether binary relevance assumption holds or not, which, to the best of our knowledge, is the first to do so in classifying topics in patient experience taxonomy. We built a constraint-based deep neural network model based on two different assumptions to capture inter-dependencies. The first assumption captures how topics that are semantically similar to each other tend to be interdependent with each other. As an example, if a patient expresses that the doctor is friendly, then the patient is likely to think the doctor is professional as well, The second assumption measures how the two semantically different groups may be interdependent with each other. From the above snippet, it is possible that professionalism and appointment access are interdependent, despite the fact they are not semantically similar. We discuss how we capture topic inter-dependencies in Chapter 5.

Contribution 1.4. We trained an automated patient segment classifier.

To comprehensively represent topics from patient experience taxonomy, we also need to identify who posts about the experience. Patients' experiences differ depending on the patient segment that the patient is in. Because the number of patient segments is potentially infinite (as an example, there are thousands of different health ailments, or medications that they may be taking) we instead focus on a single type of patient segment to demonstrate that it is possible to capture patient segments. We, in particular, focus on whether a patient has stopped taking medication or not. This is an important segment where the patients may have a vastly different experience depending on their drug usage [1–4]. We implemented rich semantic features to represent different patient segments which we believe can be applied to extracting other patient segments as well. To the best of our knowledge, none of the previous works specifically focused on automatically segmenting patients to better understand the patient experience. We discuss this contribution in Chapter 6.

Contribution 1.5. We propose an automated recommender system which identifies patient-generated texts that contains the topics that physicians may be interested in.

With a comprehensive and automated patient experience classifier that can annotate patient-generated texts, what are some of its applications? One application is in better facilitating communication between patients and hospital staff members. In patient review websites, patients express their gratitude or frustrations regarding their hospital visits. In each of the reviews, there are a diverse set of topics mirroring those in patient experience taxonomy. Reading through these reviews, addressing concerns that patients have, and passing on gratitude that they had will help improve patient experience in the future.

However, reading through the reviews can be a daunting task. There are over a dozen patient review websites such as RateMD or WebMD, and at least as many online health forums which makes it unrealistic for all of the staff members to sit down and read through all the reviews and posts that are written about the hospital. Hospital staff members are known to be extremely busy, and it would be helpful to receive updates on what patients talked about on the topics that they are interested in. For example, doctors themselves may not be interested in how the financing works, whereas the administrative staff would be interested in learning more about what patients say about these topics. On the other hand, if the doctor sees that patients complain a lot about friendliness, he or she can then focus his or her effort in being more cordial. What we see here is that different staff members are interested in different topics. Hence, a method that can route posts that are of interest to a given staff member will help save time, especially if they are interested in reading patient reviews. Our proposed recommender system learns which topics the corresponding staff member, or expert is interested in. It also assigns one expert per patient-generated text, which ensures each staff members are shown a different set of posts. We discuss our contribution on how to build a recommender system that can utilize topics in patient experience in Chapter 7.

1.5 THESIS ORGANIZATION

The rest of this thesis is organized as follows:

- Chapter 2 provides an overview of relevant techniques and prior works in understanding patient experience from patient-generated texts. We explore related works with a focus on patient experience, patient segments, and applications in patient-generated texts.

- Chapter 3 details the dataset that we used to conduct our experiment on patient-generated texts.

- Chapter 4 shows the justifications and descriptions of each semantic features of the patient-generated texts, and is our first and second contribution of this thesis. We show how the representation improves our understanding of patients' experience by empirical experiments we have conducted. Furthermore, we explore semantic insights we gained on each topic.

- Chapter 5 details how the topic-dependencies can be modeled and corresponds to our third contribution of this thesis. We explore two different hypotheses in modeling topic-dependencies, showing that the two assumptions enable a richer topic representation in patient-generated texts.

- Chapter 6 describes how patient segments can be identified from patient-generated texts. This chapter corresponds to our fourth contribution. We explore and evaluate the semantic representation of the two patient segments with which we experimented.

- Chapter 7 explores a patient-generated text routing system for clinicians which is our fifth contribution of this thesis. We describe how we ensure the routed texts are of interest to clinicians, and to help them save time with one expert per post constraint. We then empirically show the performance results.

- Chapter 8 summarizes our findings and explores future research directions to gain a comprehensive understanding of patient experience.

CHAPTER 2: RELATED WORKS

We discuss past works as to how researchers collected and analyzed patient experience feedback.

2.1 COLLECTING PATIENT EXPERIENCE FEEDBACK

Identifying the relationship between patient experience and health outcome is possible because of the thorough data collection process. There have traditionally been two different approaches to gathering this information, one by utilizing the HCAHPS survey, and another by analyzing surveys sent by hospitals. HCAHPS survey is a formalized questionnaire that asks patients to answer multiple-choice questions in order to gauge their experience at a given hospital. Private hospital surveys, on the other hand, often act as addendums to existing HCAHPS surveys to solicit aspects of patient experience that the HCAHPS do not cover. In the past decade, online media – in particular, customer review or patient review websites, have arisen as an alternative method of collecting patient experience. We describe each different approach in this section.

2.1.1 HCAHPS

According to Manary et al. [42], when designed and administered appropriately, patient-experience surveys provide invaluable insight into the quality of the care. Furthermore, to eliminate confounders and alternative explanations, researchers should take into consideration the interactions and aspects of the quality of care to properly analyze outcome measurements as opposed to treating patient satisfaction as an overall rating that care providers should maximize.

The United States government has instituted a standardized survey, called HCAHPS, to measure patients' outcomes after examining past studies which indicated the importance of collecting patient experience. The HCAHPS survey was a culmination of different studies; it showed that patient satisfaction had decreased over time, indicating that we should be measuring patient health dating back many decades. HCAHPS has since seen broad adoption across the United States.

HCAHPS has been widely successful in that it replaced the old survey that measured patient experience [34]. Furthermore, 30% of incentive funds, estimated at $850 million, is based on how patients rate their hospital experience on the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) patient satisfaction survey.

The survey asks discharged patients 27 standardized questions about their hospital stays. Of these, 18 core questions pertain to critical aspects of patients' hospital experiences. The questions cover a wide range of topics, in numerous verticals such as the hospital environment, communication between the staff members, and responsiveness of the staff members, and is designed to cover essential aspects of a patient's stay at a healthcare institute.

Furthermore, there have been many longitudinal studies on how demographics influence patient responses. HCAHPS are most frequently utilized to understand overall patient satisfaction for a given disease or treatment (for instance, cancer, or a surgical procedure) [43,44]. These are not limited to identifying patients' well-being; some studies focused on the relationship between nurses' perception of work and patients' satisfaction [45]. The survey collects demographic information to determine the outcome.

The limitation, however, is that these often do not collect free response feedback from patients. We are not able to further drill down on why patients responded in a particular direction. Furthermore, these surveys are mandated only to hospitals subject to the Inpatient Prospective Payment System (IPPS) [46], and there is currently no system in place to measure the satisfaction of outpatients, nor of primary care patients.

2.1.2 Proprietary Hospital Surveys

Due to the limitations of the original HCAHPS, some institutes and researchers propose utilizing their patient experience surveys. These are used to identify unknown factors that may influence patient satisfaction, particularly factors that are not covered by HCAHPS. Some of these include parking facilities, opening hours, cleanliness or privacy settings [47–49]. Researchers utilize exit surveys to draw these conclusions. The surveys are mailed to patients' homes, where they are asked to answer and return responses to the hospital.

Hospital surveys have a high response rate [50, 51], often as high as 70%. Of these, face-to-face surveys often had a higher response rate (76%) than those collected by mail (66%). These are fairly high response rates, high enough to indicate general trends. More concerning, however, is the indication that those who do not respond to surveys are more likely to have had a negative experience at the hospital [51]. It is beneficial to understand why patients are dissatisfied to improve hospital experience. Furthermore, these surveys often have specific questions that physicians have in mind and ask targeted questions towards patients [47–49, 52]. Patients, in return, respond back to these questions, but they may not mention the reason behind their response. Hospital surveys, while useful in learning specific aspects of patient experience, may have limitations when it comes to freely expressing patient experience.

2.1.3 Online Patient Experience Analysis

A third approach is to utilize an online medium to learn more about patient experience. Unlike HCAHPS or proprietary hospital surveys, online media is not limited to targeted questions. Instead, the medium provides a place for patients to express whatever concerns or problems they encountered during their hospital visits, allowing patients to describe topics that proprietary hospital surveys and HCAHPS have not covered. Indeed, due to the flexibility they provide, online media have actively been utilized to conduct a wide array of research outside of patient experience. Examples include identifying adverse drug events, epidemics, and analyzing public health policies. We provide a brief overview of the works done online with regard to health, and delve deeper into the works specific to patient experience.

Health and Online Social Media

Online patient reviews or online forums allow patients to talk about their experiences and can be viewed as a supplement to both HCAHPS and hospital surveys [26]. Indeed, some researchers indicate that online reviews provide a rich source of data that may be used to track the quality of care [53], though special care is required to be done to correct for cohort size. These sources are useful because the dataset is generally publicly available and patients are not motivated to be overly positive or negative. Furthermore, culling patient discourses online is useful because these provide insight into concerns that patients may not have mentioned to their primary doctors. As an extreme example, some patients lie to avoid negative consequences, to achieve secondary gain (e.g., to obtain medication or disability payments), out of embarrassment or shame, or to present themselves in a better light (e.g., as dutiful and compliant) [54]. Similarly, they may decide to withhold information due in part to embarrassment or privacy concerns [55, 56]. With respect to this, researchers have found that paying attention to patients' online comments may help uncover topics that the clinicians or patients may have missed.

To learn more about the topics discussed online, researchers have investigated multiple ways to tackle this problem – for instance, by topic modeling. Ranard et al [26] used Latent Dirichlet Allocation (LDA) to discover topics frequently discussed on an online review website (Yelp). They then mapped these topics to those in the HCAHPS survey and found that the review website had numerous topics that were analogous to those found in the HCAHPS survey. They further claim that they discovered topics discussed in the online domain that were not covered by the survey. The compassion of the staff was one of these topics; HCAHPS do not explicitly ask if the hospital staff were empathetic. Another study analyzed what causes patient satisfaction and sentiment [27]; this survey segments topics into three broad categories (interpersonal, technical and system) motivated by Lopez et al [20] and run a variant of LDA in the dataset. They uncover the words indicating patient sentiment most frequently associated with these categories. Zhang et al [57] on the other hand, investigated the intent behind posting online, finding that patients frequently post to either find out how to manage their symptoms, what the cause of the symptoms might be, or if a given medication can cause side effects. These works, which are essential contributions in helping us understand patient-generated contents online, did not explicitly build patient experience taxonomy, unlike our work.

Research was not limited to identifying patient experience in online social media. Understanding drug experiences expressed in an online medium is another popular research area. In understanding drug experience, researchers have focused on adverse drug effects or drug-addicted patients' experiences. Identifying adverse drug events from online media [10–12] is a heavily researched area. The goal is for computer algorithms to detect adverse drug events even before health researchers identify this as a potential problem. If an unusual pattern is identified for the drug of interest, this may merit further investigation to see if the medication does, indeed, have adverse effects that patients have reported. Another line of works summarize drug usage from social media by modeling how different demographics use drugs differently [14, 58]. Some other works focus on experiences of drug addiction [13–15, 59]. This line of work focuses on patients who are addicted to drugs, or how these users utilize social media to seek help or support. While not direct patient experiences, these sources are similar in that they are a means by which patients share their experience with regards to drugs.

Online social media is also used as a signal to identify potential epidemic [16–19] while other lines of research focus more on pharmacovigilance [60, 61]. These works utilize users' tendencies to post on social media about discomfort and aggregates these based on geographical location. In evaluating the validity of these sources, researchers frequently utilize historical data from the Center for Disease Control.

Patient Review Websites

Online patient review websites are places where patients can describe their hospital experiences, and where these writings are available to be viewed by anyone else who comes to the website. Patient review sites such as RateMD consist of numeric ratings along with free text responses. The website allows outsiders to glean from one particular interaction with the healthcare institute. Many works utilizing patient reviews focused primarily on understanding review ratings themselves [62–65] because they not only provide numerical evaluation of the doctors, but free response texts as well. Some of these works focus on rating behavior characteristics [62–64, 66], while other researchers seem more interested in how different patient demographics rate clinicians [65]. Interestingly, patients generally rated pediatricians, general surgery, and subspecialty surgery the highest, while generalists and subspecialists had the lowest overall ratings [40]. Perhaps it is not surprising that a lot of computational models focused on sentiments [27] or predicted physician ratings on online review websites [67, 68].

A different line of work strives to better understand the patient experience from these review websites, arguing that these are a useful supplement to existing HCAHPS and hospital surveys. According to one study, 57% of reviews mention questions that are covered in HCAHPS domain [69]. Topics that were not covered in HCAHPS include financing, including unexpected out of pocket costs and stressful interactions with billing departments; system-centered care; and perceptions of safety. Because of these characteristics, incorporating aspects of HCAHPS and other features that are distinct to patient reviews help us better understand patient experience. To better understand patient experience, we utilize patient reviews to understand the experiences of patients.

2.2 UNDERSTANDING PATIENT EXPERIENCE FROM PATIENT GENERATED TEXTS

Computational research in health care traditionally focused on analyzing medical chart entries and notes made by physicians or discharge summaries written by medical staff with the goal of discovering relationships among medical concepts, patient diagnosis and treatment [70–72]. More recently, thanks to technological advancements in data acquisition as well as increased attention to patients, a growing body of patient data has fueled a new wave of academic research, this time with the goal of investigating patient comments to improve patient satisfaction and experience [73, 74]. This research, based on a new type of data – patient feedback, in the form of solicited and unsolicited free-text collected through patient surveys – has been particularly interesting yet challenging. So far, the research aimed to discover patients' complaints through their negative comments [29, 75, 76]. Various models have been proposed for such analyses, notably content analysis [77, 78] and thematic evaluation [30, 75, 79, 80]. With our research, we broaden the scope of the analysis so that we learn from all types of comments, positive as well as negative.

Directly related to our research are the works of Licong et al [31], Maramba et al [32], Elmessiry et al. [81], Doing-Harris et al. [25], Doyle at al. [21], Lopez, et al. [20], Greaves et al. [82], and Brody et al. [83]. They present different modeling approaches and all show that linguistic analysis of textual comments provided by patients and healthcare practitioners bring new insight to understanding what matters with regard to patient satisfaction and experience. For example, Lopez, et al. [20] provided a detailed analysis of patient comments identifying topics such as *overall excellence, negative sentiment*, and *professionalism* as well as specific factors like *interpersonal manner, technical competence*, and *system issues.* Doyle, et al. [21] refined their topic categories to better search for a meta-analysis of patient experience. We expand on this previous research with our own empirical observations and propose a revised topic schema (with definitions and examples) which we then use to better understand patient comments and automatically identify relevant topics.

From a computational perspective, most previous research relied on shallow text representations in their empirical investigations. For example, based on a term frequency–inverse document frequency (TF-IDF) analysis of a large number of consumer health questions, Licong et al [31] discovered a set of relevant topics. Similarly, Maramba et al. [32] analyzed free-text responses from a post-consultation postal survey. They used a patient's overall satisfaction rating (Dissatisfied, Satisfied) to label relevant vocabulary terms ranked based only on simple word counts; e.g. , 'surgery', 'excellent', 'service', 'good', 'helpful' (for satisfied patients), and 'doctor', 'feel', 'appointment', 'rude' (for dissatisfied patients). In our research, we go beyond simple TF-IDF metrics and representations to better identify a larger number of semantic labels indicative of patient comments supplemented by comment codes and sentiment, and thus, bring in new insights.

Elmessiry et al., 2016 [81] focus only on patients' complaints about their doctors relying on a standard bag of words representation tested on six machine learning classifiers. Similarly, Greaves et al. [82] used a Naïve Bayes multinomial classifier to classify 6,412 free-text online comments about hospitals but they also limit their analysis to only three topics – *overall recommendation, cleanliness*, and *treatment with dignity.* In this thesis, we expand previous research and seek to understand and learn from all patients' comments, positive, negative, mixed or neutral. Further, our approach goes beyond bag of words processing to account for verbal as well as non-verbal context (e.g., sequences of words in a comment as well as metadata, such as comment codes), and thus, dives into the underlying language structure and reduces search noise. Moreover, we go beyond two target labels and recognize that patient complaints involving physicians' practices are very broad and complex on their own as they can refer to many issues, from treatment concerns and decisions, to room environment, physician behavior, and competency questions.

More recent research has provided new insight into the problem through novel methods of analysis, though still focusing on complaints. For example, Doing-Harris et al. [25] employ supervised and unsupervised methods to identify the most common topics (i.e., vocabulary-based and Naïve Bayes classifiers), as well as to discover new, unexpected topics (i.e., topic modeling) discussed in negative patient comments. They augment their model with sentiment features by training a positive-negative classifier. With our approach and dataset, we expand the breadth of the analysis, build better models, and gain new insight – e.g. we analyze all patient comments, and we augment the model with a much richer set of gold standard meta-information (sentiment, ratings), all without additional training.

Researchers have also utilized Deep Neural Network (DNN) framework to aid in tasks in online health domains [84,85]. These methods modify existing networks such as Convoluted Neural Network [86], or sequential models such as Recurrent Neural Network, Long Short Term Memory [87] or their variant in their approaches. Motivated by promising results seen from utilizing these frameworks, we also employ DNNs, but additionally applied constraints to these networks in line with constrained convoluted neural networks [88]. Motivated by promising results of the constraint-based neural network, we also add new soft-constraints to our domain. Unlike the previous works in DNN framework, we utilized a unique hierarchical structure that is present in patient experience taxonomy in Chapter 5. To the best of our knowledge, such approaches have not yet been attempted in capturing patient experience taxonomy.

2.3 APPLICATIONS THAT CAN HELP IMPROVE PATIENT EXPERIENCE

Recommending or finding experts [89, 90] is an ongoing research field that helps question askers find the best domain expert. The goal here is to return a ranked list of best answerers based on the similarity between the query and the answerer's posting behavior. Guo et al. [89] proposed a probabilistic generative model for the QA community and employed a user-question-answer model. They evaluated their method on Yahoo! Answers. Another line of work [90] proposed combining both the user's profile and the questions that they have answered. They utilized a variant of LDA called segmented topic models [91] to recommend posts to users. Similarly, CQARank [92] made use of the Topic Expertise Model to learn about users topical interests, and inferred their model using Gaussian Mixture Model. Such research is especially helpful for CQA websites like Quora[1] where users can ask specified users to answer in exchange for on-website credits. The main goal of these works, however,

[1] http://www.quora.com/

is different from ours. While they focus on recommending experts to question askers (which helps question askers), our aim is to recommend forum posts to experts (which in turn helps experts save time).

Some research has investigated the expertise of such users. These are known as expertise retrieval [93] systems, and are widely used. Some popular examples are ArnetMiner[2] and Microsoft Academic Search[3]. These can be either profile-based [94], where the works leverage existing user profiles to further build expertise of a given user, or document-based [95], which ranks candidates based on a combination of the documents relevance score and the degree to which the person is associated with that document. Our approach builds users' expertise by analyzing documents to which they have responded previously, and hence, is an example of document-based user modeling.

A more closely related line of research to our work is that of question recommender systems. These can be divided into two main areas: 1) those that utilize only the forum text, and 2) those that also leverage user's behavior. In the first line of research, there are works that recommend questions based on users' queries [96], forum threads [97] or posts [98]. Other works utilize answerer behavior to recommend users to posts [99–102]. These methods are based on collaborative filtering [100], probabilistic models [99, 101, 103], or classification models [102]. Similar to our work, these works recommend which questions users should respond to. The difference, however, is that we focus on designated experts while they address only general question answerers, i.e., non-designated experts. To the best of our knowledge, our work is the first to tackle the problem of recommending questions to designated experts in online forum environments. We have noticed that, when it comes to designated experts, only one expert is likely to respond to a given post. On the other-hand, multiple users may respond to the question on CQA websites that do not have designated experts.

Our work on routing texts to experts has multiple objective functions to capture the characteristics of the forum posts and designated experts. Other works have phrased this as social collaborative filtering [104, 105] which captures various social aspects. We further extend this framework by adding constraints, or regularization [106] terms based on the our intuitions and data observation.

[2] http://www.arnetminer.org/

[3] http://academic.research.microsoft.com/

CHAPTER 3: BUILDING AND ANNOTATING PATIENT EXPERIENCE TAXONOMY

In this chapter, we describe how our collaborators at DePaul University built and annotated patient experience taxonomy on patient-generated texts. The work is based on a grant sponsored by DePaul University. We reproduce the work here since we utilize this dataset in understanding patient experience, and the annotation process provides insight into why topics exist in the proposed patient experience taxonomy.

We first discuss the process that our annotators at DePaul University followed to build a comprehensive patient experience taxonomy based on both online patient reviews and existing research works [20, 21, 107]. The annotators then annotated the corpus with the topics from patient experience taxonomy. The challenge of this step of the process was in ensuring that the definition of each of the topics was not ambiguous. To ensure that the annotators agreed on the definition of each topic in the corpus, they went through a multi-step annotation process. In each step, we calculated inter-annotator agreement to gauge and guide annotators in their annotation process. Afterwards, we conducted an analysis of the topic distribution of the new patient experience taxonomy in patient-generated texts. This work is based on our journal article that is currently under submission [108].

3.1 INTRODUCTION

Any machine learning algorithms require annotated datasets to make predictions. These datasets typically consist of input features and desired output labels. The goal of a machine learning algorithm is then to learn to predict output labels from the input features [109,110]. More precisely, a typical machine learning learns $y = f(\mathbf{x})$, where $\mathbf{x}$ is input features, and y is the desired output labels, or prediction labels. The function $f(\cdot)$ learns to map input features to output labels. Without labeled data, it is not possible for an automated system to learn to classify new input. Predicting topics in patient-experience taxonomy from patient-generated texts is no different. To automatically identify which topics are being discussed in patient-generated texts, we first need labeled data.

From a cursory glance, it may seem that obtaining the annotated dataset, especially from patient-generated texts may be a trivial task. There are millions of patient-generated texts available online, whether they are patient review websites such as RateMD, or online health forums such as Healthboards. A researcher can merely implement a web scraper to gather these texts. Indeed, obtaining input features $\mathbf{x}$ from these media is not an issue at all. The unfortunate reality, however, is that obtaining topics, or labels y from patient-generated texts is more challenging. To mitigate the lack of annotated labels in online texts, some of the past research works reframed their problems into predicting meta-data such as hashtags from microblogs [111–113]. However, it is not feasible to annotate patient-generated texts with topics from patient-experience taxonomy by utilizing meta-data. Patients typically do not express their hospital experiences by utilizing hashtags, or other topic-related meta-data. Even if we were able to obtain topics related to patient experience by mining the related hashtags, there are still issues with data annotations. Only a subset of users (typically younger users) actively utilize hashtags which skews the distribution of the dataset. The classifier will not perform well when it is tasked with predicting posts by users who do not typically use hashtags [114].

Because of the limitations in utilizing meta-data to label patient-generated texts with topics from patient-experience taxonomy, we instead opted to obtain manually annotated patient-generated texts, in particular, online patient reviews. Our collaborators at DePaul University first propose a new patient experience taxonomy in which we conduct our analysis throughout this thesis. The taxonomy was built in multiple phases, where our collaborators first conducted an in-depth analysis of potential topics. We then went through multiple stages of agreement analysis to ensure that annotators agreed on label annotations. We describe the annotation process and annotation agreements in this chapter.

3.2 ANALYZING FREE-TEXT COMMENTS

3.2.1 Data Description

To understand patient experience from patient-generated texts, we started out with a widely known patient review website, RateMD (`http://www.ratemds.com`), with $58,110$ patient reviews across $19,636$ unique US doctors, as was also the case in Wallace et al [115]. The website lists a vast number of doctors in the United States in many different disciplines such as dentistry, cardiology or family health. Upon interacting with a doctor, the patient can write a review detailing his or her experience. They are first asked to rate service quality in four specific areas (helpful, knowledge, staff, and punctual), and a global rating. The ratings range from 1 (very poor) to 5 (excellent). They then write free-response answers detailing their experience at a hospital.

Consistent with existing research on patient reviews, the majority of the reviews in RateMD dataset was positive [64]. The sentiment distribution can be seen in Table 3.1. We note that approximately two-thirds of the reviews received ‘very good,’ or ‘good’ ratings. Consistent with existing literature, the review website does not appear to be a venue

Meta-data dimension	**Review Rating**					**Total**
	Very Good (5)	**Good (4)**	**Neutral (3)**	**Poor (2)**	**Very Poor (1)**	
Staff	52.70%	15.17%	10.33%	6.88%	14.91%	44,409
Helpful	60.89%	5.71%	4.45%	6.48%	22.47%	52,218
Punctual	46.16%	19.30%	12.40%	7.57%	14.57%	52,067
Knowledge	63.38%	6.29%	6.52%	6.99%	16.82%	52,218
Overall rating	68.15%		6.42%	25.42%		52,240

Table 3.1: Distribution of Patient Comments by Meta-data dimensions.

Sentiment	**Mean**	**Std. Dev.**	**Min**	**Max**	**Total**
Positive	54.88	44.56	1	261	24490 (46.9%)
Negative	82.71	64.57	1	512	6074 (11.63%)
Neutral	70.46	62.81	1	231	122 (0.02%)
Mixed	73.84	57.53	1	273	21554 (41.3%)
Overall	65.98	53.91	1	512	52240 (100%)

Table 3.2: Distribution of Patient Comments by length [words].

where disgruntled patients unreasonably criticize on physicians [116]. Furthermore, because a significant portion of the corpus is positive, analyzing only the negative reviews will not provide us with a comprehensive picture of what patients write.

A more interesting analysis focuses on the length of the reviews, seen in Table 3.2. An average review is 65.98 words long. However, negative reviews (82.71 words) were quite a bit longer than the positive ones (54.88). The above statistics indicate that reviewers have more to say when they have had a negative experience than a positive one.

3.2.2 Data Annotation and Topic Classification Schema

Our collaborators built the patient experience taxonomy and annotated the topics on patient-generated texts. To provide the readers with how the taxonomy was conceived and the corpus annotated, we provide an overview of the annotation process. In this section, we use the word ‘we’ if a particular portion was done by us, whereas we use ‘collaborators,’ ‘annotators,’ ‘they,’ or ‘researchers’ if the collaborators from DePaul University performed a particular section.

To build a patient experience taxonomy, our collaborators first started by manually annotating a sample of 300 random reviews from the RateMD dataset. Two annotators with a background in linguistics, natural language processing, and health care started annotating the comments with topics that they felt best describe the reviews. Each annotator, consistent with many other annotation tasks, worked independently. They then met to compare the topics that they found in the reviews. In parallel, our collaborators conducted thorough research in both computer science and healthcare domains related to patient experience. They found three works to be of interest, those conducted by Doyle et al. [21] and Lopez, et al. [20], and Doing-Harris et al. [25]. These works all describe and propose different patient experience taxonomies, and were all, to different degrees, motivated by the HCAHPS survey. The researchers compared the patient experience topics that exist in the literature and those that the annotators found from RateMD, finding that a significant portion of the topics overlap. They then combined the topics discovered by the annotators and those from existing research works to create a preliminary annotation schema.

To validate the proposed schema, our collaborators conducted another round of annotation study. Two annotators who have prior training in annotating datasets were given a batch of 300 comments to annotate individually and were asked to annotate as many topics as they could for each review. We then calculated their inter-annotation agreement, and the researchers analyzed and discussed any disagreements. The process repeated through four more batches, each with 300 comments, and each time, they refined the annotation guidelines. Furthermore, the inter-annotator agreement improved on each iteration, as can be seen in Figures 3.2, 3.3, 3.4, and 3.5. After they have annotated 1,500 comments, the inter-annotator agreement reached an overall κ agreement of 0.777 with an average agreement percentage of 77.9%. In the existing literature, an inter-annotator agreement of above 0.7 is considered acceptable [117]. This allowed researchers to finalize the annotation schema of 27 topics on a two-layer taxonomy. The researchers also proposed to group the 27 topics into those that are semantically similar to each other. The groupings are referred to as meta-classes and cover eight different general topics (*Interpersonal Manner, Communication, Technical Competence/Skills, Scheduling, Medical Plan, Practice Environment, Other, Overall Patient Experience*).

Our collaborators then annotated 2,090 additional RateMD reviews to create a gold standard based on the finalized annotation schema which we refer to as patient experience taxonomy. The frequencies of each topic are shown in Table 3.3 and Table 3.4.

Because annotators were allowed to annotate more than one topic per review, we also conducted how many topics reviews have on average. On average, patients described 3.32 topics per review. Furthermore, patients may discuss up to 11 topics, but only 10% of the reviews had a single topic which implies that a multi-label multi-class classifier should be employed to model this dataset. We then investigated which topics co-occur frequently, shown in Figure 3.1. We note that frequently occurring topics such as *professionalism*, and *temperament* co-occur with virtually all the other topics, indicating that the two topics

Topic	Topic Definition	Example of Patient Comment	Frequency
Interpersonal Manner			
Helpful *(helpful)*	Useful, giving or ready to give help	A great physician who was very helpful and helping me get back to work.	962
Temperament/ Manner *(temp)*	The medical staffs nature, especially as it permanently affects their behavior, e.g. nice, kind, awful, polite, sarcastic, courteous, frustrated, angry...	Of all the doctors and specialists my family visits (and there are A LOT), Dr. Klomp is my favorite. He's knowledgeable and friendly.	1882
Trust *(trust)*	Belief in the reliability, truth, honesty of medical staff.	Dr. Kramer & staff will promise the world to get your money. He told me after surgery that he & the surgery team decided that I had a young appearance; so he did not do extensive procedure as he & I had agreed upon & for which I paid. His arrogance was unequaled when I voiced my opinion about the unsatisfactory results.	213
Put-at-ease *(ease)*	To cause someone to relax, lessen discomfort, feel welcome.	Dr Kronholm's manner put me at ease instantly.	243
Good with kids *(kids)*	Medical staff does well caring for children	Is it not standard that when you see a new Dentist, they introduce themselves to you; This one barely acknowledges me. Staff and dentist not good with children. Very rough and abrasive. Recommended root canal on a 5 year olds baby tooth.	116
Communication			
Communication Skills *(comm)*	Sharing and exchanging information on patients and treatment	he didnt look me in the eyes, came in for 5 minutes told me i was fine and left	616
Explanation *(explain)*	revealing relevant facts, answers questions, clarifies information in easy to understand terms	Brilliant man. Gave me more info than what I was in for.	472
Follow up *(follow)*	Medical staff communicates with patient on treatment, medication previously given/prescribed.	He told me after surgery that he & the surgery team decided that I had a young appearance so he did not do extensive procedure as he and I had agreed upon & for which I paid.	180
Empathy *(empathy)*	Shows concern about patients feelings and needs and those of their families, shows compassion, respect, encouragement.	I believe Dr. Kronholm is a very good concienous Dr. who cares about his patients and not just the $ sign. He is very informative so you know everything he is doing. Great bed side manner. I can't remember when I was so impressed with a DR.	822
Technical Competence/Skills			
Professionalism *(prof)*	Hospital staff shows competence, knowledge, experience, efficiency, detail in care	The staff is rude and short with me and when i called for help in 2011 during a flare up becuase i had no one else to turn to, they promply turned me away and said he is not even sure he wants to see you anymore; He is very rude and so are the staff.	1955
Clinical-skills *(clinical)*	Any discrete and observable act in the process of patient care	Had he started helping me in 2008 instead of thinking i was just a dumb 21 year old i wouldnt be this bad.	714
Perceived success treatment *(perc treat)*	description of the perceived outcome of treatment	He told me after surgery that he & the surgery team decided that I had a young appearance so he did not do extensive procedure as he & I had agreed upon & for which I paid.	539
Pain *(pain)*	physical suffering or discomfort caused by illness or injury	I was diagnosed with lupus and RA in 2008, but my stats were low, even tho i was ina lot of pain and i was very sick.	215
Medication *(meds)*	All Medicinal cures administered to treat patients	He did not put me on any medications that he thought would help me.	197
Tests quality *(tests)*	The standard of health tests administered to patients	Brilliant man. Gave me more info than what I was in for. Told me I was vit D deficient, needed osteoporosis test(which me internist had overlooked)	46
Perceived bias *(bias)*	See, feel or suffer inclination or prejudice for or against someone	I was diagnosed with lupus and RA in 2008, but my stats were low, even tho i was ina lot of pain and i was very sick.	28

Table 3.3: Taxonomy of Patient Comments for topics Interpersonal Manner, Communication and Technical Competence/Skills

Topic	Topic Definition	Example of Patient Comment	Frequency
Scheduling			
Appt. Access *(appt)*	Easiness to set up medical appointments	One of his nurses was extremely difficult to work with. Almost make it unbearable to call his office.	150
Response/ Wait Time/ Punctuality *(response)*	The time it takes medical staff to start caring for the patient.	Only issue is that he does get behind on his appointments, because of the extra time spent with patients.	420
Time Spent *(spent)*	The amount of time medical staff dedicates to patients and their care	She got in and out of my appointments as fast as she could she never asked if I had questions or concerns.	500
Referrals *(referral)*	Staffs communication with patients after a treatment is put in place, after discharge	He also waited 4 years before he decided to send me elsewhere when he could not cure my headaches. She passed me off to the on call Doctor that night and told her I was getting petocin without ever discussing it with me	132
Medical Plan			
Insurance Coverage *(coverage)*	Any discussion in relation to insurance coverage, e.g. quality of insurance, what covers	This doctor's office says they accept Aetna, but they tricked my wife into signing a form claiming that we would pay any amount not covered by insurance despite what we agreed over the phone.	80
Cost of Care *(cost)*	Any discussion in relation to the cost of care, medical bills, insurance costs	He was willing to do most of our appointments on the phone, because of my money situation... and he did not charge me.	251
Practice Environment			
Privacy *(privacy)*	State/condition of being free from being observed or disturbed by other people or public attention.	Dr Kronholm's manner put me at ease instantly. I believe this is vital when exposing one's most personal physical issues.	9
Office Environment *(patient room and waiting room)*	Setting, status, ambiance of patient rooms, waiting room	He has a collage on his ceiling above the exam table of various attractive male celebrities and cut-outs from magazines of words that say things like: do you love me.	89
Location *(location)*	comment related to the location of the medical office, e.g. far, close, convenient, etc.	The only reason I continue to come here is because I don't want to drive to Colorado Springs. But next time I think I will.	17
Overall Patient Experience			
Overall Patient Experience *(ov exp)*	Comments describing patients liking or disliking of their overall experience with their doctor. It includes likelihood to recommend.	I would recommend Dr. Kerr or his staff to anyone looking to get a cosmetic treatment. If you are considering facial plastic surgery by Dr. Jonathan D. Kramer, run, don't walk away from his office!	884
Other			
Other *(other)*	General comments with no specific topic.	Dr Cordova no longer has a practice in Homer Alaska. He sold his business to Dr Nelson	287

Table 3.4: Taxonomy of Patient Comments for topics Scheduling, Medical Plan, Practice Environment, Overall Patient Experience and Other

are what patients generally discuss when they write patient reviews. Some of the less frequent topics co-occurred frequently as well, for instance, {*response time* and *appointment access*}, {*perceived treatment* and *pain*}, and {*cost* and *coverage*}. Many of the frequently co-occurring topics belong to the same meta-class, indicating that when patients write reviews, they frequently talk about multiple semantically similar topics, i.e., they often talk about various topics that belong to a single meta-classes. Of course, there are outliers, for instance, {*explain* and *time spent*} which belong to different meta-classes, indicating that it is not sufficient to merely capture the topic co-occurrences based on whether the two topics

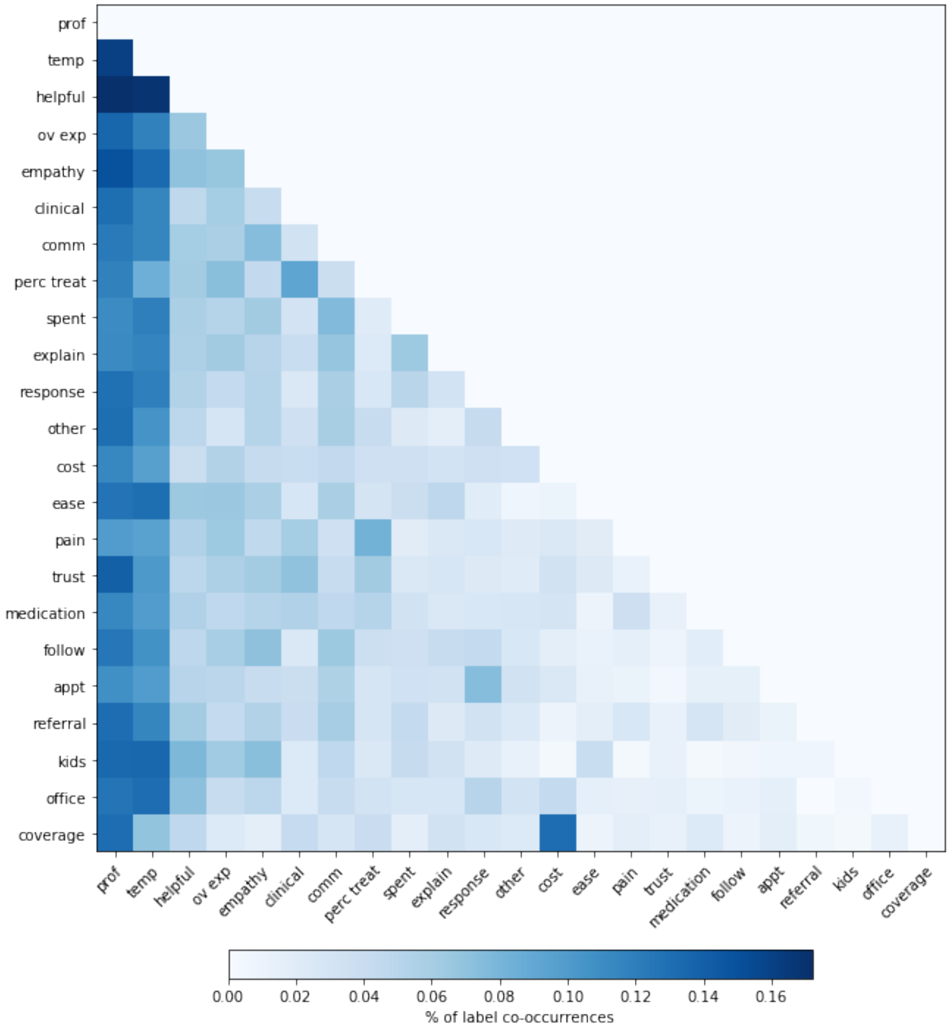


Figure 3.1: Co-occurrence heatmap

fall under the same meta-class or not.

3.3 CREDIBILITY OF PATIENT-GENERATED TEXTS

The primary concern regarding patient-generated texts is their credibility. Mostly anonymous people author patient-generated texts, and there are concerns that the medium may not be an excellent source to learn about patients' experience. As an example, researchers may be concerned that only disgruntled patients may post on patient review websites. Furthermore, it is possible that patient-generated texts may not represent all segments of patients. Younger patients may be more likely to utilize patient-generated texts, than say, older patients.

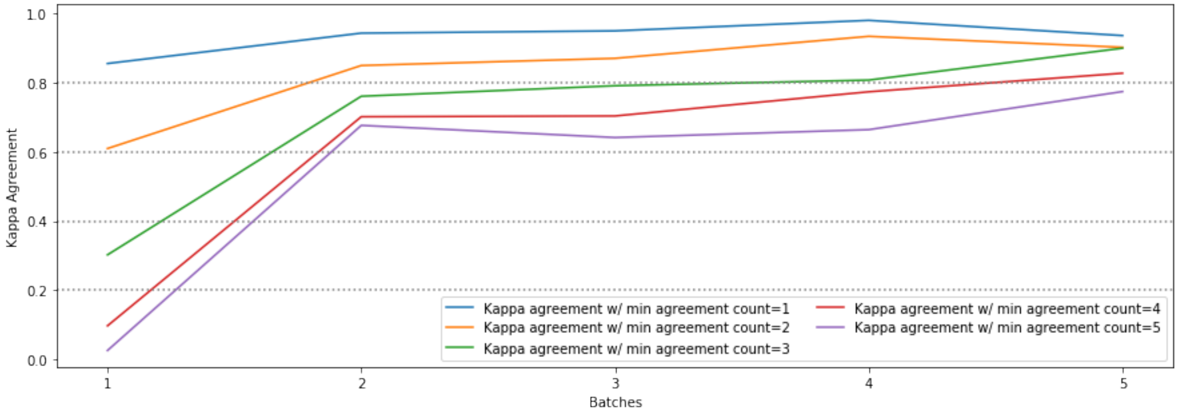


Figure 3.2: κ Inter-Annotator Agreement for Different Target Classes

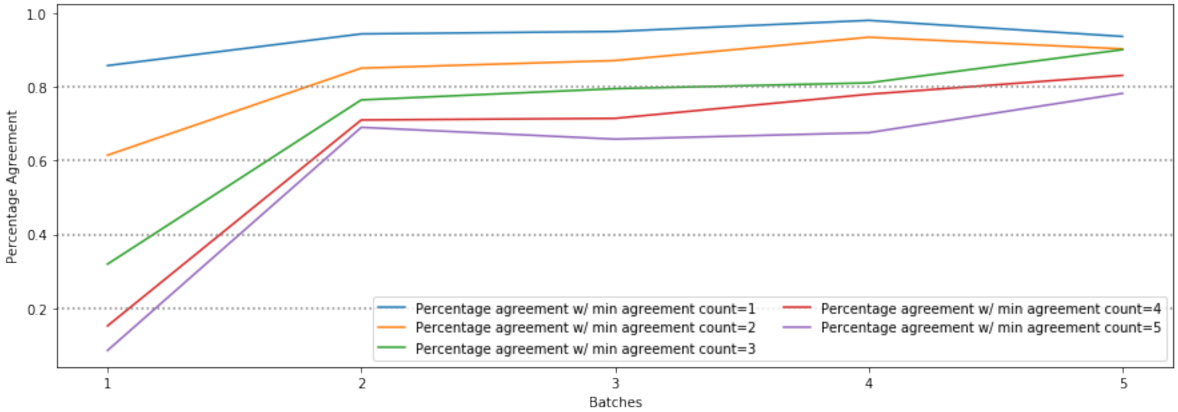


Figure 3.3: Percentage Inter-Annotator Agreement for Different Target Classes

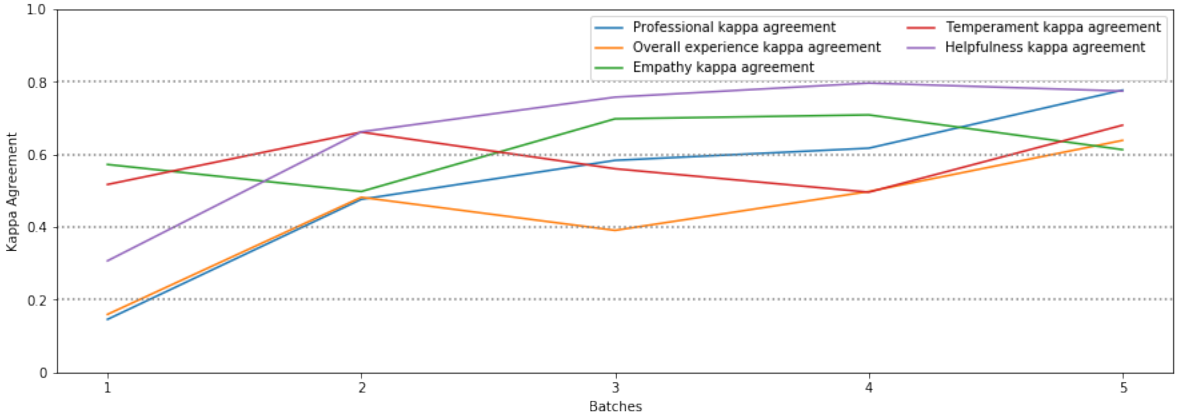


Figure 3.4: κ Inter-Annotator Agreement on *temperament, helpfulness, overall experience, professionalism* and *empathy*

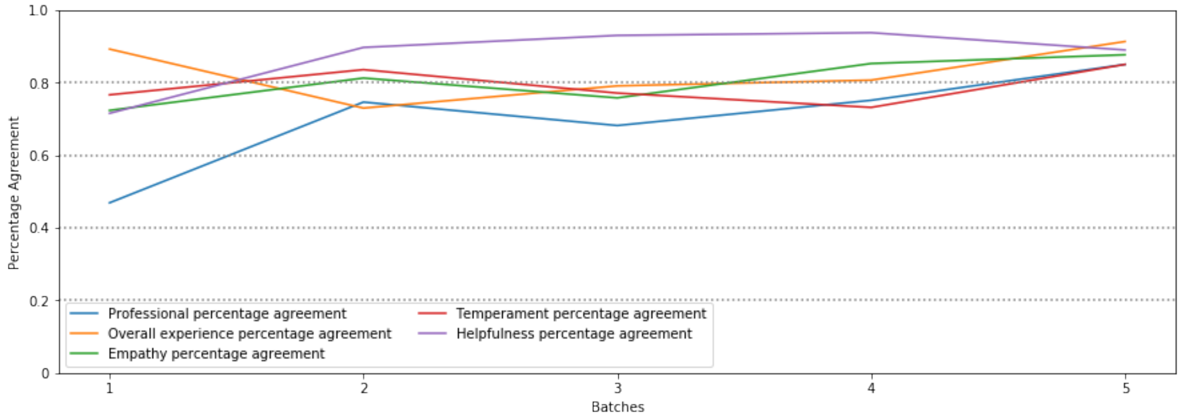


Figure 3.5: Percentage Inter-Annotator Agreement on *temperament, helpfulness, overall experience, professionalism* and *empathy*

There are several lines of works which investigated concerns associated with the quality of patient-generated texts. The first concern is whether the dataset itself is trustworthy, in particular, whether they may be a target of malicious actors writing overwhelmingly negative reviews. Numerous research works, including ours, indicate that online patient reviews are not overwhelmingly negative [23, 26, 40, 64, 118]. Indeed, these past works indicate that anywhere between $60 - 90\%$ of the reviews are positive. The findings are consistent with past works that were performed on proprietary survey [119], where the authors found about 70% of the patient feedback are also positive. However, it is essential to note that the ratio of positive to negative reviews does not prove the presence or absence of malicious reviewers that aim to impact a given hospital's reputation. Indeed, it is possible that there are malicious actors who are spreading false or misleading information on patient review websites, but the ratio of positive to negative reviews is still similar to those seen in proprietary surveys. Furthermore, because patient-generated texts are anonymous, it is impossible to know whether the reviewer is indeed a past patient, or is only acting as one [120, 121]. With recent developments in false and malicious online reviews detection algorithms [122, 123], however, we can mitigate the issues of malicious actors.

Another concern is whether the quality of the patient reviews is good enough to capture patients' hospital experiences. The quality can be captured first by determining whether or not patient-generated texts contain the topics that researchers and clinicians may be interested in. According to a past research [26, 69], patient reviews seem to capture most of the topics that researchers and clinicians are interested in, one claiming that 57% of the reviews are topics seen in HCAHPS survey. Furthermore, patients and physicians generally agree (74.3%) on what constitutes a good physician [62]. The study implies that, when patients write patient reviews, other physicians will agree that the ratings that patients assigned to the doctor are justified. However, there are physicians that question whether patient reviews correctly capture the actual quality of care that patients have received [53, 124]. According to one study [53], there was no significant difference in ratings between hospitals that serve a large number of patients, or those that serve relatively few numbers of patients. While the size of the hospital is not a proxy of the quality of care, the larger hospitals are likely to have more facilities, hence, more likely that they can process large segments of patients.

Furthermore, patient reviews sometimes suffer from data sparsity issues [64, 65]. The median number of reviews each doctor received was in a single digit, indicating that aggregating reviews from different doctors to conduct meta-analysis is a more realistic approach to utilizing this medium. Because of the relatively few reviews an average doctor receives, it may be difficult to conclude what a typical patient's experience is like with said doctor. It is possible, however, as more people start utilizing patient reviews in the future, that the median number of reviews for each doctor may increase.

The three limitations that the patient reviews currently pose – potential for malicious actors, quality of the reviews themselves, and data sparsity issues in terms of individual doctors – can all be overcome, either by deploying malicious online review algorithms when patients post reviews or by designing a procedure that is more conducive to patients writing higher quality patient reviews. It is because the limitations of the dataset can be overcome that we believe it is worth investigating patient experience as expressed in patient reviews.

3.4 DISCUSSION

In this chapter, we describe in detail how our collaborators proposed patient experience taxonomy and annotated the RateMD dataset. This started with analyzing the dataset characteristics where we found that vast majority of the reviews were positive. Based on existing literature and in-depth analysis of the dataset, our collaborators came up with taxonomy which consists of 27 different classes under 8 meta-classes. From taxonomy description, the annotators labeled topics in multiple batches. At the end of each batch, our collaborators conducted annotation studies to help improve agreement. This resulted in a satisfactory κ-agreement. With these labeled patient experience comments, we are able to conduct a more thorough research which we describe in the succeeding chapters.

CHAPTER 4: RICH SEMANTIC REPRESENTATION OF PATIENT EXPERIENCE TAXONOMY

To better understand patient experience as expressed in patient-generated texts, we first investigate what patients write about on each of the topics in the patient experience taxonomy we proposed in the previous chapter. In this chapter, we explore rich semantic representation to capture topics in patient experience taxonomy from patient-generated texts. We then analyze association between semantic features and topics.

This work is based on our journal article that is currently under submission [108].

4.1 INTRODUCTION

The HCAHPS survey is an invaluable tool for understanding patient experience. There have been numerous research works based on the patient responses of the survey – for instance, understanding patients' satisfaction after cancer treatment, or surgical procedure [43, 44]. The vast impact of the survey with regard to understanding patients' perception of care is due in part to its wide use by hospitals. For any hospital under the Inpatient Prospective Payment System (IPPS) [46], the government mandates HCAHPS surveys be collected. Researchers are then able to utilize the surveys to conduct studies on symptoms and diseases from hospitals that are under IPPS.

The survey, however, is not without its faults. Because the survey only has multiple choice questionnaires, we can only collect **what** the patient feels, but we cannot collect the **why**. As an example, consider an actual HCAHPS survey question:

During this hospital stay, how often did nurses treat you with courtesy and respect?

In a typical HCAHPS survey, a patient may indicate that the nurses seldom treated him or her with respect. The reason may be that the nurses did not correctly answer the patient's question, or it may be that the nurses were busy with other patients and did not bother to talk to the unsatisfied patient. However, a typical HCAHPS survey makes it impossible to know why the patient felt that they were treated disrespectfully. Only by collecting free-response texts can we understand why a patient answered in a particular way to a given question.

One possible solution is to utilize proprietary surveys that hospitals send out to their patients. The questionnaires can be better tailored to the needs of these hospitals and may even include free-response questions. While this is an improvement over multiple-choice-only

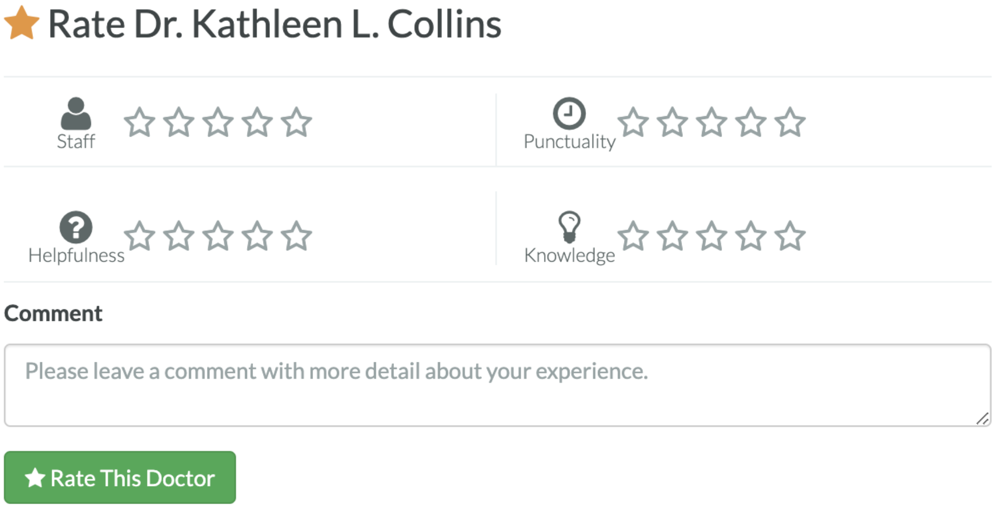

★ Rate Dr. Kathleen L. Collins
Staff ☆☆☆☆☆
Punctuality ☆☆☆☆☆
Helpfulness ☆☆☆☆☆
Knowledge ☆☆☆☆☆
Comment
Please leave a comment with more detail about your experience.
★ Rate This Doctor

Figure 4.1: Example prompt patients are asked to answer at a patient review website. Notice that they are allowed to write about any topics regarding their hospital visit.

questions, the proprietary survey still suffers from two big problems. First, the surveys are, not surprisingly, private. Hospitals are not required to share the survey publicly, nor should they necessarily share them, due to privacy concerns. Second, the proprietary surveys, similar to HCAHPS questionnaires, are top-down. Hospital staff members or researchers come up with questions that they would like to ask, rather than patients coming up with issues that they have had. It is possible that the questionnaire creators miss out on some crucial issues that patients experience.

To remedy both the top-down approach of asking questionnaires to patients and to answer *why* patients respond to particular issues, we propose supplementing existing surveys such as the HCAHPS surveys with patient-generated texts – in particular, patient reviews. Patients can rate their hospital experience, along with several specific topics such as the quality of the staff, punctuality, or helpfulness. More importantly, however, they are free to describe their experience at a given hospital, as seen in an example prompt in Figure 4.1. Such a bottom-up approach allows patients more flexibility in describing their visits and may prompt the discussion of topics that may not be covered in both HCAHPS nor proprietary hospital surveys.

However, the flexibility that patients have in writing patient-generated texts is also its most significant disadvantage. Because topic-specific questions are not asked, it is difficult, especially for an automated system, to identify the topics that are discussed in the text. Even if an automated system can identify topics, each of the topics needs to be interpretable since the purpose of utilizing patient-generated texts to understand topics in patient experience taxonomy was to understand why patients write about a particular topic.

One can be tempted to utilize an existing bag-of-words classifier to automatically identify topics in patient-generated texts, similar to those seen in Doing-Harris, et al [25]. However, such systems may not be able to capture subtleties present in natural language. For example, consider the following three sentences:

I want to express my ***gratitude*** *to my doctor.*
I want to express my ***appreciation*** *to my doctor.*
I want to express my ***frustration*** *to my doctor.*

From the perspective of a unigram-based classifier, the three sentences differ only by a single word. It may not recognize that the first two sentences are similar while the last sentence means almost the complete opposite.

In this chapter, we address the various issues that arise from utilizing patient-generated texts. To identify the topics present in patient-generated texts, we first implemented a classifier based on the bag-of-words model. We then experiment to determine how richer semantic representation and enrichment aid the classification task, finding that, indeed, a simple bag-of-words model is not sufficient to capture each topic representation. Because our model learns which features are strongly associated with each topic, we can analyze and offer insight into each of the topics that our classifier was trained on. To help determine why patients talk about a particular topic, we analyze features in each topic. The analysis provides new insights into how and why patients write about topics in the taxonomy; for example, patients are more concerned about how well their doctors listen to them when they talk about the communication skills of their clinician.

4.2 MODELS

For our model, we start with clustering and classification, the main approaches used to analyze large text datasets in natural language processing, each bringing own benefits and limitations [25, 81, 83]. Clustering, an unsupervised learning method, uses statistical concepts to split datasets into subsets with similar features and does not rely on any previously labeled data. Its output is a set of clusters, with comments in each cluster described by their similarities. On the other hand, classification is a supervised learning method that assigns predefined tags to instances in a dataset. To do this, it relies on a training set to find similarities based on cohesive attributes and then to categorize any new data. Classification will make use of labels for the new categories found based on information gain.

In previous literature, while both methods have been widely used, due to their individual limitations, they offered mixed results. In this study, we test both methods and define a set of experiments to improve on their previous standard applications.

4.2.1 Clustering Models

Clustering, often used as a first step in understanding the topic distributions of a given dataset, allows analysts to get preliminary valuable insights into new possible topics. Because it works fast and at low cost, this method is especially useful when analyzing large text datasets. In our study, the entire collection of comments is large, therefore, in the first phase of our research plan, we also start with clustering. If successful, we proceed by comparing the clustering output to the labels established in previous research as well as our proposed 23 labels. If there is a large overlap, this approach proves very efficient – since it is cheaper to run the clustering algorithm to find the relevant clusters, and then conduct analysis directly on these clusters. However, if unsuccessful, e.g., the clusters are extremely noisy, we learn that it is better to classify the patient data into predefined relevant topics. We are also mindful that clustering uses a statistical approach that is entirely reliant on the text itself, that is on how the patients themselves write their comments with no input from medical personnel. Therefore, even though we may discover new patient topics through clustering, these topics may not be relevant to medical practitioners. This limitation would further indicate that a classification approach is better.

For our analysis, we use K-Means [125] and LDA [126], the two most used clustering methods. As in previous research [25], we use unigrams to represent the data and run the models with different numbers of clusters: 10, 15, 20, 25, and 30. K-Means, via the Expectation - Maximization algorithm, calculates the k centroids for each cluster. Each datum is then assigned a cluster-based distance to a given centroid. LDA (Latent Dirichlet Allocation) is a generative model which learns latent topics from a given corpus, assuming that a word may belong to different topics depending on the document that it appears in.

4.2.2 The Classification Model

As a second approach in identifying the topics of discussion in patients' comments, we develop a classification model. We start with the pre-annotated corpus of 3,590 patient comments labeled with our proposed taxonomy classes and use it for training and testing the classifier, the annotation labels being used as gold standard. Training the classifiers to associate patient comments with their correct topics is extremely important and should rely

Text Rep. (features)	**What it does**	**Why it is beneficial**	**Comment Example**
Unigram	a comment is represented as a bag of words (i.e., unigrams)	Unigrams, as basic units of text representation, capture the most representative words for the target class	”I thought his staff was great” is represented as the bag of words {I, thought, his, staff, was, great}
Dependency parse *(DP)*	captures grammatical relationships between two words in a sentence	allows us to capture multi-word phrases, negations, long dependencies (that cannot be captured otherwise)	”The nurse who helped me during my last visit in May was nice”, (captures long dependency ’nice’ – ’nurse’ encoding temperament)
Word Embedding *(W2V)*	learns semantically related words in the vocabulary during training and enriches every content word in test with similar content words/collocations as identified in training	especially useful since it expands short test comments with semantically similar words that appeared in training	”Nurses always offered me and my family water.” word2vec adds: {nurse practitioner, staff}, {patient, family}, {tea, drink, plate}, {tea, cup, bowl}
Dependency Parse Embedding *(Dep2V)*	for every content word in a review, it adds the top 5 syntactically related words learned during training from the entire vocabulary (content word types)	especially useful since it expands short comments with syntactically similar words which help capture grammatical structures	”Nurses always offered me and my family water.” dep2vec treats dependencies as context - i.e., the context vector for the word ‘water’ is: {nmod:poss(water, my), compound(water family), conj(me, water)}
Document Embedding *(D2V)*	an extension of word2vec for learning document embeddings, i.e., a distributed bag of words over all reviews in training	especially useful if some words appear infrequently; utilizing document’s latent subspace helps mitigate word sparsity issues	doc2vec infers a given comment text into the corresponding n-dimensional embedding feature space
WordNet *(WN)*	a freely available lexical database of English capturing semantic relations (i.e., IS-A) between lexical entries	brings semantic knowledge into the classification, being able to map concepts into semantic classes (i.e., super-concepts); generalizes over lexical tokens, especially when limited training.	in training, words like ’doctor’ and ’nurse’ occur frequently with labels like *empathy.* However, if a test comment contains ’anesthesiologist’ (e.g.,: words related to empathy), WordNet helps the classifier label the comment accordingly (i.e., all are a ’healthcare professional’)
Aspect Ratings *(AR)*	meta-data identifying our aspects of patient experience: helpfulness, staff, punctuality, knowledge	indicate patients’ sentiment of each aspect	The meta-data aspects ratings for ”I thought his staff was great” were all positive

Table 4.1: Text representations (features), description, rationale and examples.

on a good text representation of the comments. We propose to employ a combination of seven approaches starting with unigrams and bigrams representations further augmented with dependency parse representation, word embedding (Word2Vec, Doc2Vec), WordNet [127] and aspect ratings (meta-dimensions). Each of these representations (also called features), provide different ways in understanding patients’ comments, as detailed in Table 4.1. In our analysis, we run a total of 128 scenarios combining these features and then determining which combinations give the best training, and, thus, classification. We report the best 20 scenarios. We are mindful of the risk of overfitting, since the bigger the combination of features used, the bigger the feature vector and thus, the bigger the risk of overfitting. As is standard in classification models [128], we compare the F-1 scores across all text representation scenarios, we show these results and discuss them in Section 4.3, Evaluation) below.

For classification, we build a binary classifier for each of the 23 topics in our proposed taxonomy, e.g. one classifier for *helpful*, one for *temp*, etc. We run the four most widely used machine learning models for classification, i.e. Naïve Bayes, Decision Tree, Perceptron and Support Vector Machines, and choose the best performing classifier[1].

To account for these specifics when modelling unbalanced data distributions, we build 23 classification models, each with four machine learning models, two classification schemas, and a combination of seven feature representations. Our formal classification model is therefore described as follows.

$$h_j(\vec{x}) = \text{argmax}_{i,j} p(y_{i,j}|\vec{x}), \quad \forall j \in \{1 \dots L\}, \forall i \in \{0, 1\}, \tag{4.1}$$

where: j denotes the j's topic being classified out of the L total number of topics; $\vec{x}$ is the set of features representing the textual comment; and $\vec{y}$ is the set of topics that identify it. $y_{i,j}$ identifies whether the comment is labeled with topic j (i.e., $y_{1,j}$) or not (i.e., $y_{0,j}$); $p(y_{i,j}|\vec{x})$ is the prediction output for the label j (i.e., $p(y_{1,j}|\vec{x})$) or the rest (i.e., $p(y_{0,j}|\vec{x})$) for its given feature $\vec{x}$.

We test the performance of all classification models across the two schemas and across the seven text representations using precision, recall and F-1 measures evaluated at micro and macro levels. The Micro method, also reported by Doing-Harris et al [25], sums up the individual true positives (TP), false positives (FP), and false negatives (FN) of all classifiers $i = 1 \dots n$:

$$P_{micro} = \frac{\sum_i \text{TP}_i}{\sum_i \text{TP}_i + \sum_i \text{FP}_i}, \; R_{micro} = \frac{\sum_i \text{TP}_i}{\sum_i \text{TP}_i + \sum_i \text{FN}_i}, \text{and } F\text{-}1_{micro} = 2 \cdot \frac{P_{micro} \cdot R_{micro}}{P_{micro} + R_{micro}} \tag{4.2}$$

The Macro method averages the precision and recall over all classifiers $i = 1 \dots n$:

$$P_{macro} = \frac{1}{N} \sum_{i=1}^{n} P_i, \; R_{macro} = \frac{1}{N} \sum_{i=1}^{n} R_i, \text{ and } F\text{-}1_{macro} = 2 \cdot \frac{P_{macro} \cdot R_{macro}}{P_{macro} + R_{macro}} \tag{4.3}$$

where P_i and R_i are the precision and recall of binary classifier i, respectively.

The Weighted method calculates weighted averages of the precision and recall over all classes $i = 1 \dots n$. Weight of class i is based on the relative frequency of the label:

$$P_{macro} = \sum_{i=1}^{n} p(i) \cdot P_i, \; R_{macro} = \sum_{i=1}^{n} p(i) \cdot R_i, \text{ and } F\text{-}1_{macro} = 2 \cdot \frac{P_{macro} \cdot R_{macro}}{P_{macro} + R_{macro}} \tag{4.4}$$

[1] In our case, SVM was the best performing classifier

where P_i and R_i are the precision and recall of class i, respectively, and $p(i) = \frac{cnt(i)}{\sum_{j=1}^{n} cnt(j)}$. $cnt(i)$ refers to number of times class i appeared in the corpus.

Macro F-1 is used to determine the overall system performance across all the classifiers (and thus, across all labels), but does not account for the number of text instances within each label. Micro-F1 mediates this issue, therefore, we report it and analyze it when comparing system performance. Ideally, a good classifier should perform well on micro, macro and weighted F-1 scores.

4.2.3 The Classification Features

In multi-class classification models, having a quality text representation (features) is most important in understanding the text and capturing the topics discussed. Because there are numerous features that we experiment with, we identify these features into three different groups. The first group, Surface Features, correspond to the type of features that we can derive from a single review. Unigram, dependency parse, and aspect rating features all fall under this group. The next group utilizes either an existing concept dictionary or word similarity that is learned from the corpus to enrich a given unigram further. Features that fall under this category does not significantly increase the feature size. Instead, they find top k features that are similar to a given unigram. We call this Word Enrichment Features, and the three feature types that fall under the group are WordNet, Word2Vec, and Dep2Vec. The third type of features calculates the latent embedding space features for each review which we denote as Latent Embedding Features. Document embedding features and Dependency document embedding features correspond to this feature group.

Surface Features

Unigram Features. Unigrams, the basic text representation, tokenize all words in the patients' comments and feed them into the classifier. To further improve the classifier, we add stopword removal, to remove words that are not indicative of the meaning of the text, as well as stemming, to generalize over inflections of words. We used the two most recognized stemmers, the Porter Stemmer and the Snowball Stemmer.

Dependency Parser Features. Used widely in many natural language processing applications, dependency parsers capture grammatical relationships between words in a sentence, i.e. multi-word phrases, conjunctions, or negations. Using the Stanford Parser [129], we experiment with two type of dependency parsers: 1) a list of full dependency parsed outputs identifying the dependency relations as well as their arguments in the given review, and 2) a list of dependency relations only, without their arguments in text. Consider the sentence *Nurses always offered me and my family water.*. Here, the first dependency parser creates *nsubj(offered, Nurses), advmod(offered, always), root(ROOT, offered), cc(me, and), nmod:poss(water, my), compound(water, family), conj(me, water)*, while the second parser generates *nsubj, advmod, root, cc, nmod:poss, compound, conj.* As the example shows, the second parser (that of dependency relations) is more general, and we expect it will help the classifier better generalize over the training instances.

Aspect Rating Features. As described in Subsection 3.1, in our dataset, each text review comes with meta-data that has five affiliated ratings. The first is an overall rating which shows how the patient felt about the experience (equivalent to overall sentiment). The next four provide specific information about the doctor's office visit – i.e., staff, punctuality, helpfulness and knowledge – indicating how did the patient feel about them.

We assume that if a patient provides a rating score for an aspect, they indicate that they were interested in addressing that aspect. With this in mind, for each aspect, we split the the rating scores into the following intervals: positive (4 or 5 stars), neutral (3 stars) or negative (1 or 2 stars).

Word Enrichment Features

WordNet Features. WordNet, a freely-available lexical database for English [36], has been used successfully for word sense disambiguation [130] in a large range of applications on product reviews [131]. Through WordNet, the words identified in a patient comment are grouped into sets of synonyms (i.e., synsets) and augments with short definitions, usage examples, and relationships to other synsets – i.e., hypernyms, hyponyms, or meronyms. In our application, WordNet is especially useful in helping our classification model learn the definition of a word, even with limited training data from short patient comments. Our corpus also contains a lot of domain-specific medical terminology (e.g. medicine names, names of medical personnel, etc.) so we have decided to encode some representative hand-picked synsets as part of this feature set.

We tokenized our dataset and the Brown corpus (often used as a reference corpus) and mapped the nouns that were found in WordNet into synsets. Using the probabilities of each synset, we calculated their log ratios both in our dataset and in the Brown corpus. A higher ratio means the given synset is more representative in our dataset, whereas a lower ratio means the given synset appears less frequently than what we would expect. From the top 200 synsets, we picked nouns that best represent our dataset, since we did not perform disambiguation of synsets at this step, and removed emotion-related synsets. For example, *doctor of the church* and *annoyance* were ranked high in our dataset, but were subsequently removed.

The representative synsets were $S = \{$*treatment, food, drug, health professional, physical condition, ill health, pain, body part, hospital, emergency, hospital room, medical instrument, disease, medical procedure*$\}$.

Each word w in a comment W, was mapped to this synset if the word was a hyponym of one of the words in the synsets S. We denote the trail of synsets in between w and $s \in S$ as $t(w, s)$. All synsets in $t(w, s)$ were added to the list of the WordNet features. By including trails of the synset between the word w and synset s, the feature set includes only the words that may be related to healthcare, yet still maintaining enough discriminating power between the two different words which may map to the same hypernyms in S. For example, consider the sentence: *I got to the hospital with a serious flu.* Direct hypernyms of 'flu' are 'respiratory disease,' and 'contagious disease', which have as hypernyms 'disease' and 'communicable disease', respectively. The hypernym of 'communicable disease' is 'disease'. We add all the hypernyms between 'flu' and 'disease' until we ding the first common hypernym 'disease'. The 'hospital' synset is also of interest, so this is added to the feature set as well.

Word Embeddings based Features (Word2Vec). Word embeddings, like Word2vec [132], are vector space-based models [133] in which low dimensional vectors are learned from unlabeled data through a neural network. They have been very successful in representing short corpora such as micro-blogs [134] or patient reviews. In essence, such word embeddings learn the surrounding window context of a given word. If the contexts for w_1 and w_2 are similar, then the words are considered semantically similar. The more similar the words, the more likely they are to refer to the same topic. For example, 'ridiculous' may also mean 'absurd,' and one patient may use the former, whereas another may use the latter. Such a feature is especially important when w_1 occurs relatively frequently, but w_2 does not. In such cases, the classifier would use the word representation of w_1 instead of w_2, thus resolving the inherent sparsity issues.

$$\frac{1}{T} \sum_{t=k}^{T-k} \log p(w_t | w_{t-k}, \ldots, w_{t+k}) \tag{4.5}$$

where k is the size of the window, and t is the t-th index of a given word; T is the length of a sentence. Thus, the prediction task is done via a multi-class classifier as shown by

$$p(w_t | w_{t-k}, \ldots, w_{t+k}) = \frac{e^{y}_{w_t}}{\sum_i e^{y_i}} \tag{4.6}$$

where y represents the learned weights.

In some sense, this is very similar to what we have done with WordNet. Our WordNet enrichment approach maps a given word to a higher order hypernym, thereby generalizing what the word represents. Similarly, word embeddings map all the words that appear in the data into k latent dimensions. There are some notable differences, however: for Wordnet, we focus only on a small subset of synsets and their hyponyms, whereas word embeddings learn the latent representation from the data irrespective of which topics the word may belong to. Please note, however, that our WordNet and word embeddings features do not necessarily capture the same information – this is because semantic information like the one captured by hypernyms and hyponyms identify language abstractions that usually do not occur in free text.

In encoding these features, we first trained a word2vec model on our corpus. Then, for each word $w \in W$, we computed scores s for similar words w', more precisely, $(w', s) = sim(w)$. The similarity score s is based on the closeness of the embedded space. We then encoded the top k similar words as part of the features for the classifier. [2] This allows the classifier to utilize a similar word w' instead of the word w when there are not enough training examples for w.

Dependency Embeddings Features (Dep2Vec). Dependency parsed outputs may suffer from sparsity issues: each word may have a dependency relationship with many other words in a sentence, thus drastically increasing the feature vector size. Thus, instead of directly using dependency parsed outputs, we could train dependency parsed embeddings [135]. Similar to Word2Vec, which relies on a given word w_i, and its context $w_{i-s}, w_{i-s+1}, \ldots w_{i-1}, w_{i+1}, \ldots w_{i+s-1}$ for a window size of s, dependency parsed embeddings (Dep2Vec) are trained on the dependency parsed output as contexts for a given anchor word w. Unlike Word2Vec, the contexts consist of dependency parsed relationships that contain the word w. Consider again the sentence *My son likes the doctor*, with the corresponding dependency parsed outputs *nmod:poss(son, My), nsubj(likes, son), root(ROOT, likes), det(doctor, the), dobj(likes, doctor).* The context for the word 'son,' is *nmod:poss(, My), nsubj(likes,).* Similarly, contexts for the word 'likes' are *nsubj(, son), root(ROOT,), dobj(, doctor).*

Dep2Vec is known to extract functionally similar words, as opposed to semantically similar words that Word2Vec extracts [135]. We hypothesize that, by expanding similar function words, the classification performance improves. Here we encoded the top k similar words (similar to word2vec similarity features).

[2] For all of our experiments, we set $k = 5$ words.

Latent Embedding Features

Document Embeddings Features (Doc2Vec). Doc2Vec [136] captures a latent representation of each document in the corpus (in our case, a document is a patient review and the corpus is the set of reviews obtained after agreement). The model first learns the word embeddings, while treating a given document as an unknown word. It then tries to infer this missing document based on the words that represent the document. Unlike Word2Vec, however, we use here the latent representation of a given document, and add it to the feature set. In training the Doc2Vec model, we used the Skip-gram implementation with a latent vector size of 200. Past literature has shown that after a certain point, neither a bigger window size nor a larger vector would lead to better performance. Another research [137] indicated that choosing a relatively big window size helps when training data is limited. In particular, the bigger the window, the better the model is at capturing similarity between the two words as it uses more contextual information [138].

Dependency Document Embedding based Features (DepD2Vec). We would also like to experiement with another type of features like document embeddings based on dependency parsed outputs. Here, we decided to represent documents as a collection of dependency parsed outputs. More precisely, in a Doc2Vec setting, a collection of words is used to infer the latent dimensions of a given document. Similarly, in Dependency document embeddings setting, we treat dependency parsed outputs as the context, and infer the current document. We call dependency parse based document embeddings model, DepDoc2Vec. The latent embedding space is used as the classification feature. This approach is similar to [139] who also use dependency parses as an embedding space. Similar to Doc2Vec features, we set the latent vector size to 200.

Feature sizes

Overall, we had 8,437 unigram features, 12 aspect rating features, and 14 additional WordNet features. Doc2Vec and DepDoc2Vec both added 200 additional features. Neither Word2Vec nor Dep2Vec have added any additional features as it was finding similar words from the corpus (top 5 most similar words) and using this as its feature set. There were 40 unique dependency parsed relationships that we used with 80,190 dependency parse features. The dependency parsed features, somewhat not surprisingly, are very sparse. We mitigate overfitting by setting tokens that appear less than 30 times as <UNK>.

4.3 RESULTS AND EVALUATION

4.3.1 Clustering Results

We ran two clustering models, K-Means and LDA, on the entire patient experience dataset. With a unigram text representation of patient comments, we investigated 10, 15, 20, 25 and 30 preset clusters. For each cluster scenario, we evaluated the cluster quality with Adjusted Mutual Information (AMI) [140], a standard metric (with the SK-learn implementation from scikit-learn.org) recommended when the gold standard clusters (as set manually) are unbalanced [141]. The metric employs mutual information in calculating the similarity score with the gold standard (with values in [0; 1]; 0 = no agreement; 1 = full agreement) accounting for chance agreement.

Across all the clustering scenarios, we found that the biggest AMI scores were 0.038 and 0.0002 for K-means and LDA, respectively. These numbers indicated a very low performance of the clustering analysis. Further investigation of the clustered comments in the best performance scenario, of 20 clusters, showed the clusters were very noisy, difficult to interpret and label, and not clearly delimited. This outcome is not surprising, given the limitations that come with unsupervised clustering, and given the complexities of the data to be clustered, e.g. short comments, ambiguous language, inconsistent abbreviations, etc. The outcome is in line with findings in previous research as well, where clustering models also proved unsuccessful particularly when analyzing patient comments [25].

4.3.2 Classification Results

We started our classification modelling with a sample of 3,590 patient comments sampled at random from the entire 50k RateMD collection, and annotated with 23 taxonomy classes. On this data, we ran a series of classification models with an 80/20 training/test split and 10-fold cross-validation. In building the highest performing classification model, we devised two major experiments. First, we developed a classification model based on several text and contextual data representations. Second, with semantic knowledge from our proposed meta-class hierarchical taxonomy, we augmented the initial model and created a semantically-informed multi-class classification model of patient experience.

Method	Macro			Micro			Weighted		
	P	R	F-1	P	R	F-1	P	R	F-1
U	0.541	0.378	0.434	0.621	0.513	0.562	0.604	0.496	0.538
U+PS	0.577	0.412	0.470	0.621	0.513	0.562	0.613	0.513	0.552
U+SS	0.576	0.413	0.470	0.619	0.513	0.561	0.612	0.513	0.552
U+SR	0.558	0.373	0.436	0.638	0.492	0.556	0.621	0.492	0.541
X=U+PS+SR	0.594	0.413	0.473	0.635	0.511	0.566	0.626	0.511	0.555

Table 4.2: Baseline Classification Models with Pre-processed Text. U: SVM, unigram with one-vs-rest schema; PS: Porter Stemmer, SS: Snowball Stemmer, SR: Stopword Removal

Experiment 4.1: Multi-class Classification Model

In the first step, we train and test baseline binary classifiers built with four machine learning models with a unigram text representation run with both a one-vs-one schema and then again a one-vs-rest schema. We find that the Support Vector Machines (SVM) model with one-vs-rest schema achieved the highest performance and show its Macro and Micro measures in Table 4.2. We consider this as our baseline classification model and note it as U.

In the second step, we augment U with pre-processing methods, such as stemming and stopword removal. This increases text understanding through generalizations over inflections of words, and through removing words that are not indicative of the meaning of text. As in standard literature, we use the Porter Stemmer and Snowball Stemmer, noted as PS, SS, respectively. Stopword removal is noted as SR. As shown in Table 4.2, the classification performance improves and the best model now becomes X=U+PS+SR.

In the third step, we add a series of eight contextual features to text understanding: *Word2Vec* (W2V), *Doc2Vec* (D2V), *Dep2Vec* (Dep2V), *DepD2Vec* (DepD2V), *Dependency Parser* for dependency parsed output features (DP) and relationship only features from dependency parsed output (DPR), *WordNet* (WN) and *Aspect Ratings* (AR). As shown in Table 4.3, Latent Embedding Features (D2V, DepD2V) and Word Enrichment Features (W2V, Dep2V, WN) generally show improvement over the baseline. This shows that finding functionally (Dep2V), and semantically (W2V) similar words further enriches a given review. WN, despite belonging to word enrichment feature group, had mixed performance, however. In some metrics we see performance gains (macro measures), but in others, there were no discernable performance differences. Part of the reason is WN is general-purpose concept dictionary, so it is not able to capture the contextual information as well as other Word Enrichment Features

Mapping reviews into latent embedding subspace (D2V, DepD2V) acts as summarizing

Method	Macro			Micro			Weighted		
	P	R	F-1	P	R	F-1	P	R	F-1
X	0.594	0.413	0.473	0.635	0.511	0.566	0.626	0.511	0.555
X+W2V	0.603	0.427	0.488	0.645	0.521	0.576	0.636	0.521	0.566
X+D2V	0.601	0.430	0.490	0.639	0.528	0.578	0.632	0.528	0.569
X+Dep2V	0.605	0.416	0.479	0.647	0.519	0.576	0.637	0.519	0.564
X+DepD2V	0.604	0.428	0.487	0.628	0.522	0.570	0.623	0.522	0.561
X+DP	0.600	0.406	0.466	0.651	0.500	0.565	0.616	0.510	0.550
X+DPR	0.582	0.415	0.475	0.655	0.503	0.569	0.621	0.516	0.557
X+WN	0.585	0.415	0.472	0.630	0.511	0.565	0.621	0.511	0.554
X+AR	0.600	0.415	0.476	0.631	0.512	0.565	0.624	0.512	0.555

Table 4.3: Classification Models with Contextual Features. X:U+PS+SR, DP: Full dependency parse features, DPR:Depency Parser by using only the relations as the feature, W2V: Word2Vec similarities features, D2V: Doc2Vec similarities features, WN: WordNet features, AR: aspect rating features, DepD2V: Dependency parser based embeddings features, Dep2V: Dependency parser based similarity features

Method	Macro			Micro			Weighted		
	P	R	F-1	P	R	F-1	P	R	F-1
X+W2V+Dep2V	0.607	0.426	0.489	0.671	0.515	0.582	0.640	0.524	0.568
X+Dep2V+DepD2V	0.611	0.429	0.491	0.662	0.523	0.585	0.654	0.523	0.572
Y=X+W2V+D2V	0.602	0.436	0.495	0.644	0.530	0.581	0.636	0.530	0.572
Y+WN	0.610	**0.443**	**0.504**	0.648	**0.534**	0.585	0.640	**0.533**	0.576
Y+WN+AR	0.606	0.440	0.499	0.643	0.530	0.581	0.637	0.530	0.573
Y+WN+AR+DPR	0.593	0.441	0.496	0.654	0.532	0.587	0.645	0.532	0.576
Y+Dep2V+WN+AR	**0.614**	0.434	0.496	**0.671**	0.522	0.587	0.642	0.532	0.575
Z=Y+DepD2V+Dep2V	0.609	0.438	0.500	0.661	0.529	0.587	0.654	0.529	0.576
Z+WN	0.613	0.436	0.496	0.660	0.527	0.586	0.652	0.527	0.575
Z+WN+AR+DPR	0.592	0.440	0.500	0.657	0.524	0.583	0.648	0.524	0.572
Z+WN+AR	0.609	0.437	0.494	0.663	0.531	**0.590**	**0.654**	0.531	**0.578**

Table 4.4: Classification Models with Combinations of Contextual Features. Y: X+W2V+D2V, Z:X+W2V+D2V+Dep2V+DepD2V.

each of the reviews. Latent embedding features are less prone to noise which helped improve the performance. On the other hand, shallow features (Aspect rating, dependency parses) did not show improvement. DP feature, in particular, results in very large feature space, so the classifier had a difficult time learning each of the features and actually resulted in worse performance. Aspect rating features had mixed performance, and there were no real gain or loss in performance in the classification task.

In the fourth step, we experiment with latices of contextual features to determine their contributions to the overall classification performance and identify the specific best performing combinations. We list the results in Table 4.4.

We see that both embeddings methods, unigram based and dependency based, consistently show improvements over the baseline model. Furthermore, a combination of W2V similarity model and D2V consistently shows improvement over that of D2V similarity model and DepD2V. This seems to show that, for this task, the semantics of the words in a comment is still more important than the dependency structure of that comment. Further, the combination of the two embedding models (DepD2V and D2V) and two similarity models (W2V

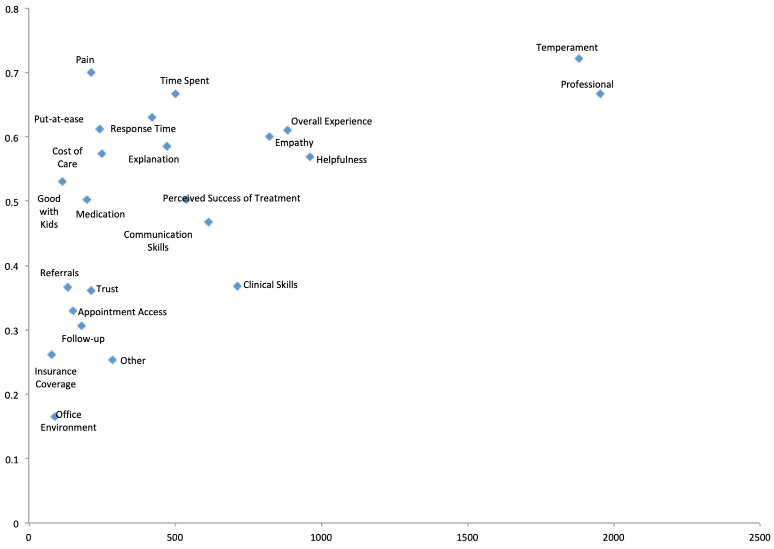


Figure 4.2: Relationship between F-1 value and the frequencies for each labels

and Dep2V) improves performance over those that use only one embedding model or one similarity model. In the table, as shown by micro and weighted F-1 scores, the combination (W2V, D2V, Dep2V, DepD2V, AR, WN) generated the best performing model. This is the multi-class classification model we will use.

In the fifth step, we evaluate the multi-class classification model on all the 23 classes in our taxonomy of patient feedback and present the results in Table 4.5 and Figure 4.2. We find several interesting insights. Overall, the system performed best on *Professionalism* and *Temperament* and least on *Office Environment, Other, Insurance Coverage, Follow, Appointment Access, Trust, Referral*. This is because, with higher occurrences, the classification system had more examples to learn from, hence, had better classification accuracy; the opposite explains the results for the lowest frequency classes. At the same time, there are several notable exceptions. For example, *Pain, Put-at-ease, Good with Kids, Cost of Care* have high accuracy even with low frequencies of occurrence. Comments discussing these classes use many different terms, and have long text representations to learn from. At the other extreme, when discussing *Clinical Skills*, patients seem to use a short list of general words instead of talking about the specific skills of their medical personnel.

Experiment 4.2: Semantically-Informed Multi-Class Classification Model

The patient feedback taxonomy we developed proposed a list of 23 classes that we then grouped semantically into eight multi-classes, e.g. *Interpersonal Manner* = {*helpful, temperament, trust, put-at-ease, good-with-kids*}. We wish to test whether by adding this semantic

Class	Precision	Recall	F-1	Frequency
Helpfulness	0.645	0.522	0.569	962
Temperament	0.741	0.705	0.721	1,882
Trust	0.612	0.272	0.361	213
Put-at-ease	0.816	0.527	0.612	243
Good with Kids	0.663	0.381	0.531	116
Communication Skills	0.587	0.435	0.467	616
Explanation	0.685	0.538	0.585	472
Follow-up	0.518	0.221	0.306	180
Empathy	0.632	0.520	0.600	822
Professionalism	0.662	0.674	0.666	1,955
Clinical Skills	0.455	0.300	0.368	714
Perceived Success Treatment	0.573	0.419	0.503	539
Pain	0.779	0.639	0.702	215
Medication	0.647	0.372	0.502	197
Appointment Access	0.523	0.292	0.330	150
Response Time	0.719	0.556	0.630	420
Time Spent	0.730	0.614	0.667	500
Referrals	0.705	0.284	0.366	132
Insurance Coverage	0.443	0.155	0.261	80
Cost of Care	0.768	0.479	0.574	251
Office Environment	0.483	0.111	0.165	89
Overall Patient Experience	0.697	0.547	0.611	884
Other	0.339	0.160	0.254	287

Table 4.5: Per topic performance for Unigram + WN + Word2Vec + Doc2Vec + Dep2Vec + DepD2Vec + Aspect Rating model.

knowledge from this two-level hierarchy we improve our understanding of patient comments and, thus, increase the classification performance of our multi-class classifier. We start by taking both our baseline (X) and the highest performing models (Z+WN+AR) and evaluate them on classifying our meta-classes. Consistent with our previous analyses above, we find that the (Z+WN+AR) model improves the baseline classification (Table 4.6). Further, if we compare these results with those in Table 4.4 (Z+WN+AR model), we also confirm that the multi-class classifier with semantic knowledge (over multi-classes) is significantly better as shown in all the Macro and Micro performance measures, e.g. F-1 is 0.75 at the multi-class level as compared to 0.59 at the class level. This suggests that semantic knowledge is valuable for a high performing multi-class classification system and is further shown in the high performance measures for each of the nine multi-classes (Table 4.7), where all but two multi-classes (by far the most infrequent) show F-1 between 0.553 and 0.865.

Method	Macro			Micro			Weighted		
	P	R	F-1	P	R	F-1	P	R	F-1
X	0.617	0.522	0.554	0.774	0.713	0.742	0.742	0.708	0.717
Z+WN+AR	0.691	0.549	0.593	0.785	0.718	0.750	0.759	0.718	0.733

Table 4.6: Performance of the Meta-class Classification System. Best baseline model (X) and best performing multi-class classification model (Z+WN+AR)

Taxonomy Meta-Classes	Precision	Recall	F-1
Interpersonal Manner	0.796	0.849	0.821
Communication	0.790	0.624	0.697
Technical Competence/Skills	0.814	0.923	0.865
Scheduling	0.785	0.651	0.711
Medical Plan	0.822	0.422	0.553
Practice Environment	0.483	0.111	0.165
Other	0.429	0.105	0.168
Overall Experience	0.746	0.538	0.625

Table 4.7: Classification results for the Meta-classes of Patient Experience. Z+WN+AR multi-class classification model

As another way to test whether the semantic knowledge improves classification, we also compare the classification of meta-classes with the classification across classes, but grouped in groups of 2, 3, 5, and all the way to 23. Our hypothesis is that the semantically-informed meta-class classification, as opposed to any ad-hoc class grouping, will have higher performance. This is because by grouping classes based on their semantic similarities, rather than with any ad-hoc strategy, our meta-classes will be more different (meaningfully different) from each other, and this means less ambiguity in training our classifier. Indeed as shown in Table 4.8 grouping classes by their frequencies, as was done in previous literature, is worse off than grouping them semantically. In fact, our semantic meta-class classifier performs better than even a 5-way classification (grouping the top most frequent 5 classes), the F-1 measure of the Z+WN+AR system in Table 4.7 is 0.75 as compared to an F-1 of 0.70 for the model with the 5 most frequent classes.

4.4 ANALYSIS

4.4.1 Analysis of Results

With our final multi-topic classification model, we perform several analyses on the patient experience dataset. First, we compare our model and classification results with those in

Num		Macro			Micro			Weighted		
Classes	Method	Precision	Recall	F-1	Precision	Recall	F-1	Precision	Recall	F-1
2	X	0.787	0.837	0.811	0.787	0.837	0.811	0.787	0.837	0.811
3	X	0.772	0.727	0.723	0.741	0.757	0.739	0.739	0.757	0.746
5	X	0.680	0.630	0.652	0.696	0.664	0.680	0.693	0.664	0.677
7	X	0.605	0.525	0.559	0.647	0.581	0.612	0.639	0.581	0.606
9	X	0.609	0.517	0.556	0.664	0.562	0.609	0.654	0.562	0.599
11	X	0.603	0.498	0.542	0.633	0.548	0.587	0.626	0.548	0.581
23	X	0.594	0.413	0.473	0.635	0.511	0.566	0.626	0.511	0.555
2	Z+WN+AR	0.795	0.869	0.830	0.795	0.869	0.830	0.795	0.869	0.830
3	Z+WN+AR	0.746	0.747	0.745	0.761	0.776	0.768	0.759	0.784	0.769
5	Z+WN+AR	0.708	0.647	0.674	0.720	0.684	0.701	0.718	0.684	0.698
7	Z+WN+AR	0.649	0.553	0.594	0.679	0.606	0.640	0.672	0.612	0.637
9	Z+WN+AR	0.651	0.540	0.586	0.677	0.589	0.630	0.671	0.589	0.623
11	Z+WN+AR	0.659	0.531	0.586	0.675	0.574	0.620	0.670	0.574	0.614
23	Z+WN+AR	0.609	0.437	0.494	0.663	0.531	0.590	0.654	0.531	0.578

Table 4.8: X: unigram model (stopwords removed and words stemmed) vs best performing models as number of classes are increased. U: Unigram features D2V: Doc2Vec embedding features, W2V: Word2Vec word similarity features, DepD2V: Dependency Doc2Vec features, Dep2V: Dep2Vec features, WN: WordNet features, AR: Aspect Rating features, Z:U+W2V+D2V+DepD2V+Dep2V

the previous literature. Then, we present a number of findings and further discuss their implications on academic research. We also explore their significance on managing patient experience in the US healthcare marketplace.

Comparing with the previous research of Doing-Harris et al., 2016 [25], our multi-class classifier with enriched features performs better on 5 out of 7 common classes/topics, in spite of classifying threee times as many classes (23 compared with 7) to classify[3]. Our model, based on SVM with the one-vs-rest schema and a combination of seven text features, uses 8,437 unigrams, 12 sentiments, 14 WordNet, 200 Doc2Vec, 200 DepDoc2Vec and 80,190 Dependency features. We attempted to mitigate the possibility of overfitting by labeling tokens that appear less than 30 times as <UNK>.

Label	**Our System**			**Doing-Harris et al. [25]**		
	P	**R**	**F-1**	**P**	**R**	**F-1**
Explanation	0.685	0.538	0.585	0.90	0.64	0.74
Appointment Wait/Response	0.719	0.556	0.630	0.54	0.60	0.57
Access/Time spent	0.730	0.614	0.667	0.44	0.64	0.52
Practice Environment	0.483	0.111	0.165	0.24	0.57	0.34
Friendliness (Temperament)	0.741	0.705	0.721	0.32	0.57	0.40
Empathy	0.632	0.520	0.600	0.18	0.14	0.16
Overall	0.697	0.547	0.611	0.85	0.27	0.41

Table 4.9: Performance Comparison of Our Multi-class Classification System VERSUS Doing-Harris et al. [25].

First, as Table 4.9 shows, some topics are better classified than others, e.g. Friendliness/Temperament, Access/Time spent. This suggests that the task requires a wide range

[3] While both are patient feedback studies, they used different models (Doing-Harris et al., 2016: Naïve-Bayes; Our classifier: SVM) and datasets and thus, are not directly comparable.

of text representation levels, from shallow (i.e., lexical) to very complex (i.e., discourse and pragmatics, and even word knowledge required for textual reasoning). Indeed, a manual evaluation of the results shows that people discuss the well classified topics by using adjectives easily captured by our N-grams and WordNet features (Table 4.10 and Table 4.11): e.g., empathy is all about the empathetic traits of the medical personnel seen indicated by adjectives like *compassionate, considerate.* Many of the top unigrams, not surprisingly, are indicative of the topics of interest. For example, money related words, those related to professionalism, and those that pertain to top unigram words for cost, professionalism, and time spent and response time, respectively. Interestingly, top unigram features for 'communication skills' had numerous 'listen' related terms, whereas those of 'explains' had terms which required the doctor to speak, or to explain. Consistent with literature that indicates that patient satisfaction increases the more the doctor listens [142], patients perceived doctor listens well to be good communicators. Because the topics are, at least on surface, sentiment agnostic, many of the topics contained both positive and negative polarity words amongst top unigram features. For instance, both 'professional,' and 'unprofessional' were amongst top unigram features for the topic 'professional,' 'helpful,' and 'unfriendly' for top helpfulness topic, and for topic ease, there were 'comfortable,' and 'uncomfortable.' Sentiment agnostic topics were: *professional, temperament, helpfulness, overall experience, perceived treatment, ease.* Similarly, there were other topics which tend to predominantly have unigrams from single (either positive or negative) sentiments. Empathy or communication skills, for instance, had predominantly positive unigram features. Topics whose top unigrams had predominantly single sentiment were: *empathy, communication skills, time spent, response time, trust, pain.* We later investigate if these had some correlations with aspect ratings as well.

Because our dataset is obtained from an online source, it was not surprising to see terms that one would expect in an online domain. Examples of these were 'meds,' 'appt,' 'ins,' which were shortenings of medication, appointment, and insurances, respectively. Furthermore, some topics captured synonyms of the most predominant terms. For instance, 'Cost' had captured many forms of money, such as $, money, bill, billing, charges, refund, cost, copay, cost, and bucks.

Some of the top unigram features were shared amongst multiple topics. One such example is 'knowledgeable' which appeared with the topics professional and clinical skills, 'time', which appeared with both response time and time spent, and 'help,' which appeared with both helpfulness and empathy. These seem to indicate that some labels may appear together, but we leave this up for later studies on how this information can be used to further improve the classification performance.

Having a clear list of words that summarize a topic is very helpful for hospital and healthcare administrators as it provides a snapshot of how patients perceive the medical care they receive as well as the medical staff providing that care. Negative descriptions like “rude” immediately directs the hospital to institute personnel training to address this attitude problem. Likewise, seeing positive descriptions is an indication to the hospital to continue its staff management direction.

Class	Top Unigram Features
Helpfulness	helpful, help, helped, helping, unfriendly, best, impatient, ache, uncaring, client, date, phenomenal, beyond
Temperament	friendly, rude, abrupt, arrogant, kind, attitude, especially, knowlegeable, polite, best
Trust	trust, honest, admire, intelligent, realistic, insightful, proactive, $, hands, trusted, ago, medi-cal, praise, final, ethical
Put-at-ease	ease, comfortable, uneasy, safe, well-informed, feel, affordable, uncomfortable, calm, confident, input, upbeat, even, fearful, well
Good with Kids	kids, children, son, daughter, adults, babies, boys, daughters, old, child, really, thoughtful, respect
Communication Skills	talk, listen, listens, communicator, listener, listening, easy, clinician, diagnostician, sits, shooter, information, relate, acute
Explanation	explain, explains, explaining, questions, explained, $, level, tricare, everything, logical, details, evaluate, answered, part
Follow-up	follow, followup, called, responds, pre-natal, prenatal, phoned, show, set, call, calls, pre-op, pick, picked, work
Empathy	cares, caring, compassionate, considerate, compassion, warmth, genuinely, help, concern, left, empathetic, concerned
Professionalism	professional, knowledgeable, deal, efficient, unprofessional, birth, submit, encourage, thorough, today, make, bled, attitude
Clinical Skills	diagnosis, knowlegable, team, identify, newest, fine, birth, quite, recognize, diagnose
Perceived Success Treatment	awesome, eliminated, without, treated, waiting, prostate, research, may, ridiculous, honest, results, cracked

Table 4.10: Top Unigram Features Associated with each Labels for 23 labels

Second, as we focus our analysis on the topics that are classified well, like *Temperament, Time spent*, we discover that our model reached this performance grace to some text representations, e.g. WordNet, that are particularly valuable in understanding patient comments. For example, through WN, the model also learns new concepts like 'health professional'. We take this learning and extend the WN analysis to the other topics modeled then extract all

Class	Top Unigram Features
Pain	pain, traumatic, painful, cough, swelling, headaches, enjoyable, affordable, discomfort, office, fatigue, incomplete, visits, friendly, useless
Medication	meds, medications, medication, pills, prescribed, antibiotics, drugs, even, rushed, worked, given, prescription, find
Appointment Access	appointment, appt, appointments, actually, school, get, sessions, day, manner, within, life, always, year, started
Response Time	wait, waited, late, usually, time, us, waiting, empathetic, talk, knowlegable, know, concerned, scheduled, level, timely
Time Spent	rush, rushed, time, spent, spends, judges, take, spend, distracted, taken, talk, takes, pushes
Referrals	refer, referred, specialist, references, transfered, recommendations, another, referals, transferred, assistance, encourage, discourage, suggestions, refers
Insurance Coverage	coverage, ins, company, medicare, companies, help, go, insurance, n't, check, fix, also, extremely, pediatric
Cost of Care	$, charged, money, bill, billing, charges, worth, refund, one, cost, copay, costs, pay, well, bucks
Office Environment	office, modern, filthy, dirty, dingy, room, know, busy, explains, family, clinic, waiting, needs
Overall Patient Experience	recommend, answered, reccommend, highly, recommended, unbelievable, reduction, busy, horrific, job
Other	probably, far, commited, agree, leave, moved, edwards, think, pharmacy, maybe, richards, grandmas, every

Table 4.11: Top Unigram Features Associated with each Labels for 23 labels (Continued from Table 4.10)

WordNet synsets amongst the top 100 most strongly associated features with a given topic (Table 4.12).

In order to investigate into these features, we retrieved top 500 features for each topics (in terms of learned weights), and listed all the WordNet features that appeared in Table 4.12. There are some topics (professional, helpfulness, and kids) that did not have any WordNet features in the top 500 features, hence we do not report them in the table.

With this, we discover that the topic of pain is further associated with 'body part', 'medical procedure' and 'emergency', as patients discuss not only that they are in pain but also mention the location of their pain (e.g. 'head' and 'stomach'), the medical procedure they were administered and the intensity of their pain. Moreover, we learn that patients also seem to complain that the right medication was not administered, or that it happened after a medical procedure. Two such example are: 1) *"The nurse drew blood which did not* *hurt, but did cause awful bruising. Reckless!"* and 2) *"I was so sick, yet I had to call 2x times to ask for antibiotics."* Such associations could be identified only because we allowed hypernyms of given synsets to be included as feature sets. This analysis suggests that some text representations are particularly beneficial in understanding comments in a healthcare field and thus, we show here that, to some extent, WordNet is a useful text representation in classification models of patient experience. Such associations could be identified only because we allowed hypernyms of given synsets to be included as feature sets. Thus, it is clear that some textual representations are particularly beneficial in understanding comments in a healthcare field and thus, we suggest that WordNet is a useful text representation in classification models of patient experience.

Class	Top WordNet Features
Temperament	health_professional
Trust	treatment, hospital_room, drug, physical_condition
Put-at-ease	ill_health, treatment, health_professional
Communication Skills	treatment, medical_procedure, pain
Explanation	physical_condition, pain, health_professional
Follow-up	disease, hospital_room, ill_health, treatment, hospital
Empathy	disease, hospital
Clinical Skills	medical_procedure, emergency, hospital_room
Perceived Success Treatment	ill_health, physical_condition, medical_procedure, body_part, treatment, hospital_room, hospital
Pain	pain, emergency, drug
Medication	drug, treatment, body_part, ill_health
Appointment Access	emergency, pain, hospital
Response Time	emergency, body_part
Time Spent	disease, treatment, physical_condition
Referrals	health_professional, body_part, hospital, emergency
Insurance Coverage	treatment
Cost of Care	pain, treatment, medical_procedure
Office Environment	body_part
Overall Patient Experience	physical_condition
Other	emergency, drug

Table 4.12: Top WordNet Features

We see some obvious, and less obvious findings. Topics which involve treatment (medication, pain, or perceived treatment, clinical skills) all have medication or treatment related WordNet concepts such as ill health, medical procedure, medication, pain, or treatment. On the other hand, many of the interpersonal related topics had health professionals as its top

WordNet features. These topics include temperament, explains, ease and referral. WordNet concepts emergency and body parts often appeared on both treatment and interpersonal related topics, in part because these terms are used when clinicians explain the procedure (and the body part where the surgery is being conducted), and when the procedure is actually being done on the patient.

Third, we discover unexpected synsets in understanding how to classify some topics. Take for example *communication skills* – based on its definition (Table 3.3), we expect to find the synset health professional to be highly associated with the topic, since patients talk to the medical personnel to learn about their symptoms and treatments, as in this example comment: "*Great doctor, he knows what he's doing. I would recommend him to a friend.. The staff was slightly rude on the other hand.*" Rather, to our surprise, we find treatment, medical procedure, and pain as the most strongly associated synsets. A manual inspection of comments shows that patients tend to refer to doctors by their names or using pronouns, which would not be captured by the 'health professional' synset.

Fourth, looking at the lower classification measurements, we see that some topics like clinical skills are more difficult to identify even when using our seven different features. They would require not only more context but also more informative text representations that allow us to perform reasoning. For example, for comments like:

"*After an accident at work that herniated L-4,L5,S-1, I went to Dr. Cabot on a rec. from a friend. He tried facet injections which were extremely painful and not helpful. He seemed angry when I returned because they had in fact, made me worse. He scheduled a CT-Myleogram. Did not use an IV to replace fluids. Sent me home with a migraine headache. I laid in my bed for 12 days with a migraine. He refused to fix the problem and refused to give pain meds. His unwillingness to do a blood patch allowed blood to seep into the spinal canal. This created scar tissue. IT CAN NOT BE REPAIRED. 11 years later I had to have surgery to try and repair. Ended up worse. I am now perm. disabled and in EXTREME pain every minute. This man destroyed my life - literally. Do not let him treat any part of your spine. You will wake up every day for the rest of your life in agony. Most days the pain is so bad i want to die. Could have fixed it easily. You get what you pay for.*"

It is clear that a unigram representation, even if enriched with semantic information (e.g., from WordNet), is not enough to capture complex topics of discussion like *empathy*. As has been shown previously [143], the person providing the comment usually does not say directly and explicitly that the behavior of the other was unempathetic for fear that his comment can be viewed as pathetic, or that he feels sorry for oneself. Instead, complaining about a non-empathetic clinician, is usually associated with a negative situation that is so uncomfortable that the patient feels that it needs to be described. Correctly identifying *empathy* in patient comments, thus requires discourse-level representations that allow textual reasoning.

Fifth, we further investigated into aspect rating features. There are four aspects (knowledge, staff, punctual, helpful) and three sentiments (positive, neutral and negative) for comment. Similar to the previous section, we took top 500 features from each topic and listed aspect related topics in Table 4.13.

Class	Top Aspect Rating Features
Helpfulness	Helpful - Positive, Knowledge - Negative
Explaination	Knowledge - Neutral
Clinical Skills	Knowledge - Negative, Helpful - Positive
Medication	Helpful - Neutral, Knowledge - negative
Response Time	Punctual - Negative
Office Environment	Helpful - Neutral
Overall Patient Experience	Knowledge - Neutral, Helpful - Positive
Other	Helpful - Negative, Helpful - Neutral

Table 4.13: Top Aspect Rating Features. Takes a form of Aspect - Sentiment. Aspect can be one of knowledge, helpful, punctual, or staff

Aspect rating features were often not among top 500 features for a given topic. However, for those that did have aspect ratings, we find some interesting trends. Patients were more likely to rate punctuality negatively if they were talking about response time. Those who provided positive helpfulness and negative knowledge aspect ratings were more likely to be talking about helpfulness in the free-form comment. This is in-line with what we noticed in the Unigram feature analysis, where the top unigram features for this topic contained both positive and negative sentiment terms. The same trend held for the topic 'overall experience' as well. We further noticed negative aspect were more likely to be associated with topic rather than positive aspect. There were only three positive aspect features related to topics, whereas there were a total of ten neutral or negative aspect features.

4.4.2 Comparison with Latent Aspect Rating Analysis Classifier

We further compared whether or not an unsupervised method can still learn to classify each topic in patient experience taxonomy. For this purpose, we picked LARA [144], which can learn topic representation with minimal supervision. LARA learns representative keywords for each topic from an initial set of seed words. Researchers are expected to come up with the initial set of keywords since these guide how the algorithm selects representative keywords for each topic. An advantage of the method is its ability to identify which topics are present in a given text without training data. The advantage of the method allows researchers to rapidly explore new sets of datasets since annotations are not necessary to explore a given corpus. LARA is considered a robust unsupervised approach that can learn to identify topics in a text.

Because seed words are crucial for the performance of LARA, we conducted two different runs based on the seed words. First, we selected seed words based on the name of each class and its variants. As an example, variants of the class ‘helpful’ are ‘help,’ ‘helps,’ ‘helpful,’ and ‘helpfulness.’ The assumption in this run is that researchers do not have any knowledge of each topic in the taxonomy; hence, selects seed words based on the name of the class. We denote this run as LARA-1. The second run leverages the top unigrams for each topic, learned by our classifier, as shown in Table 4.10 and Table 4.11. The assumption here is that we have an oracle that knows which seed words best represent each topic. The second run represents the best case scenario for LARA on our dataset. We label this run as LARA-2. We compare F-1 scores of LARA-1 and LARA-2 against the supervised method, which we denote as Our Method. The results are shown in Table 4.14

Method	Macro			Micro			Weighted		
	P	R	F-1	P	R	F-1	P	R	F-1
Our Method	0.594	0.413	0.473	0.635	0.511	0.566	0.626	0.511	0.555
$LARA-1$	0.245	0.541	0.257	0.196	0.447	0.273	0.386	0.447	0.332
$LARA-2$	0.303	0.603	0.356	0.301	0.526	0.383	0.442	0.526	0.422

Table 4.14: F-1 score comparison for our method (on unigram features), LARA with seed words based on child-class label variants (LARA-1), and LARA with oracle seed words (LARA-2).

We notice that somewhat unsurprisingly, our proposed classifier, which is a supervised approach, performs better than LARA. A more interesting finding is the performance difference between the case where we have an oracle that knows the most representative seed words, versus the case where we do not. We notice significant performance gains on the oracle set of seed words (LARA-2) over the topic variant seed words (LARA-1). The performance gains of LARA-2 indicate that our proposed classifier and LARA have potentials to complement each other in further understanding patient experience topics in different settings. Take, for example, the researchers wish to explore patient experience topic in a different, unlabeled corpus such as online health forums. The typical supervised approach may perform poorly on a dataset that it is not trained on. The new dataset may follow different feature distributions than those from which the classifier was trained, causing degradation in performance. On the other hand, LARA can use the top unigrams learned from our supervised approach as its seed words, as we have done in LARA-2. LARA-2 can then learn other words that are indicative of each topic in the new corpus, ultimately using these to classify unseen texts in a different type of corpus. Depending on the findings from this new corpus, researchers can further decide to manually annotate training data on the new corpus, which will further improve classification performance. For this thesis, however, we show new directions on how LARA and our approach can complement each other, but leave the actual exploration of this as potential future research work.

4.4.3 Error Analysis

We first conducted error analysis based on whether or not the number of gold labels present in a given review impacted the classification performance. To conduct this analysis, we grouped the RateMD corpus into two separate groups. The first group had at most 3 classes, whereas the second group had at least 4 classes per review. We picked 3 as a demarcation point because the mean number of classes per review was 3.32; hence, we divided up the corpus based on this number. For all the methods, the first group, which had three or fewer classes had an F-1_w, F-1_{micro}, and F-1_{macro} scores of 0.552, 0.560, and 0.481, respectively. For those that had 4 or more classes, F-1_w, F-1_{micro}, and F-1_{macro} scores were 0.603, 0.617, and 0.515, respectively. We were initially surprised at this result, seeing that those with more labels in a given review had better classification performance, but upon careful analysis, we found that more classes mean misclassifying one label had less of an impact on classification performance than the ones with a small number of classes. In other words, if a review text has 8 classes, then misclassifying one or two of these has less of an impact than misclassifying those with 3 classes.

We next analyzed whether the length of the sentence had an impact on the classification performance. Again, we grouped the corpus in two, one which had at most 65 words, and another which had more than that number of words. This number was picked because an average review had 65.98 words. We found that the classifier had more difficulties with longer texts than with shorter ones (F-1_w, F-1_{micro}, and F-1_{macro} scores of 0.594, 0.611, and 0.489, respectively for short texts, versus F-1_w, F-1_{micro}, and F-1_{macro} scores of 0.562, 0.571, and 0.493, respectively, for long texts). Longer texts contain more unique words than shorter ones, which increases confusion for the classifier. We further noticed that the unigram-only model performed quite a bit worse on longer texts than on shorter ones, compared to our proposed method (0.527, 0.539, 0.459 for F-1_w, F-1_{micro}, and F-1_{macro}, respectively, for short texts). On the other hand, the unigram-only model had classification performance comparable to that of the best performing approach (0.582, 0.600, 0.493 for F-1_w, F-1_{micro}, and F-1_{macro}, respectively, for long texts). We see that the gap between the best performing model and the unigram model was around 0.01 for shorter texts, whereas it was 0.04 for longer ones. Our model performed better than the unigram models on longer texts by either 1) reducing the corpus into word embedding subspace, allowing the classifier to train in this subspace properly, or 2) expanding upon similar words, both semantically (Word2Vec) and syntactically (Dep2Vec), which increases the number of times unigram features are trained, and expands upon unigrams that did not appear frequently on training dataset. Furthermore, both DepD2Vec and Doc2Vec map a given corpus into an embedding document space, capturing complex interactions between different words in a corpus.

Consider the following example with which we demonstrate how to compare the peformance of the unigram model to that of our proposed model:

"WONDERFUL OBGYN! He's a terrific practitioner. Kind and gentle, intelligent, great diagnostician, too."

Gold labels for the above snippet are 'Clinical Skills,' 'Helpfulness,' 'Professionalism,' and 'Temperament.' Unigram model predicted only the 'Temperament' class. On the other hand, our model was able to predict three out of the four classes ('Helpfulness,' 'Professionalism,' and 'Temperament'). Part of the reason our model performed better than the unigram-only approach is that our model, unlike the unigram model, can utilize words that the classifier may not have seen a sufficient number of times (for instance, diagnostician and OBGYN).

On the other hand, our combined model tended to under-predict when the label did not appear frequently in the training set, and over-predicted those that appeared often. Classes such as 'Trust' or 'Office Environment,' both of which did not appear often had a very low recall. On the other hand, frequently occurring labels such as 'Professionalism,' or 'Temperament' often were predicted more often than appropriate. Consider the following example:

"I was scared when I went to see Dr. Rajpal and he put me at ease and explained my MRI. I have the utmost confidence in him. He is very caring and personable."

Our model predicted five classes, 'Ease,' 'Empathy,' 'Explanation,' 'Professionalism,' and 'Temperament.' However, the gold labels were 'Ease,' 'Empathy,' 'Explanation,' 'Temperament,' and 'Trust.' One class that is conspicuously missing is the 'Trust' class. Furthermore, even though the 'Professionalism' class was not a gold label, our classifier still predicted the class. This class appeared the most frequently, and the classifier often erred on the side of predicting than not predicting.

Furthermore, our model is not capable of capturing complex text representations that require utilizing external knowledge. In such a scenario, word-based representations are not sufficient because words themselves do not represent a given topic. Instead, it is the combination of word phrases that represent a topic. To accurately infer a topic from a given phrase, the classifier needs outside knowledge.

Consider the following snippet:

"Dr. E. is very professional. He gives real advice, listens without judgment, and is extremely knowledgeable about the latest in medications."

In the above snippet, 'listens without judgment' corresponds to the classes 'Communication Skills' and 'Empathy.' Furthermore, 'knowledgeable about the latest in medications' does not mean that the patient is talking about the drugs that were prescribed. Gold labels for the above snippet are 'Communication Skills,' 'Empathy,' and 'Professionalism.' Our classifier predicted 'Medication' class and failed to predict 'Empathy.' Neither of the predictions (or lack of) is correct. For the classifier to correctly predict 'Empathy,' it would need to know that 'listens without judgment' corresponds to 'Empathy.' Each unigram feature (listens, without, judgment), however, does not capture the class 'Empathy.' It is only by combining the three words into the phrase that it amounts to 'Empathy.' To capture this, however, the classifier needs an external database of what constitutes empathy, which our model does not incorporate.

Our classifier also had a difficult time with labels that were similar to each other, or cases where annotators could have assigned a topic one way or the other. For example, consider the following snippet:

"She is never on time no matter what time of day you book an appointment. For a 5 minute appoinment, plan on spending 2 hours in her office. She is very knowledgeable about what she does, demands perfection as an end result, but has trouble communicating to us regular people not it the industry. She also talks super fast, and has frequent staff turnover."

Gold labels are 'Communication Skills,' 'Professionalism,' 'Response Time,' and 'Time Spent.' Our classifier, on the other hand, predicted 'Communication Skills,' 'Explanation,' 'Response Time,' and 'Time Spent.' The class, 'Explanation' was not predicted by our classifier. On the one hand, this is perfectly reasonable - the phrase 'trouble communicating to us regular people not it the industry' indicates that the doctor was not good at explaining concepts. On the other hand, according to the annotation guideline, the topic 'Explanation' is assigned if it pertains to whether the doctor is explaining some concept or not, rather than the communication skills itself. The above snippet describes more about the doctor's communication skills than how well she is at explaining concepts. However, our classifier was unable to pick up on these subtleties.

Finally, our classifier had difficulties classifying texts with typos – in particular, those that have multiple gold labels and are critical in classifying a review text into one class or the other. Consider the following snippet:

"He was a great Doctor to deal with. Fixed my incisional hernia. Great bedside manner. I would recomend him to anyone. Has a great sense of humor. Has always run ontime. Staff

are always nice and helpful."

In the above snippet, the gold labels are: 'Helpfulness,' 'Overall Experience,' 'Perceived Treatment,' 'Professionalism,' 'Response Time,' and 'Temperament.' Our classifier, on the other hand, predicted 'Helpfulness,' 'Overall Experience,' 'Professionalism,' and 'Temperament.' We notice that the classifier, in particular, failed to predict 'Response Time' even though there is a time-related unigram (ontime). However, the patient did not correctly spell the word in the review snippet. The patient should have written, 'Has always run **on time**,' rather than 'Has always run **ontime**.' However, this typo (*ontime* instead of *on time*) proved to be enough for our classifier to not predict 'Response Time' class.

4.5 DISCUSSION

We started this chapter with a question: can we utilize patient-generated texts to understand patient experience? More precisely, we wanted an automated approach to identifying what topics patients post in patient-generated texts. Furthermore, our next goal was to further investigate why patients write about a given topic by analyzing the word usage of each topic. However, simply utilizing a bag-of-words model is not sufficient to understand patient-generated texts. Each word has a different semantic representation, where some are similar to each other, and others different. Bag-of-words models, however, are agnostic to how similar or different the two words are, so we cannot properly capture the corpus characteristics with this model.

To address the challenges in identifying and analyzing patient experience topics in patient-generated texts, we first proposed eight text features designed to enrich the review corpus. The features can roughly be divided into three different approaches: Surface Features, which include unigram, aspect ratings, and dependency parses; Word Enrichment Features, which include retrieving top k similar words via Word2Vec and Dep2Vec, and WordNet; Latent Embedding Features, which include latent embedding space represented by Doc2Vec and Dep2Vec. Surface Features generally had minimal impact on improving the performance of the algorithm over unigram models. In fact, dependency parses degraded the performance. However, both Word Enrichment Features and Latent Embedding Features improved the classification performance.

With a better approach to understanding and classifying patient experience topics, we analyzed the two immediately interpretable features: unigram and WordNet features. The analysis was conducted on all the topics covered under patient experience taxonomy. We provided some intuitive and counterintuitive analysis by analyzing these two features. To the best of our knowledge, we are the first to conduct such a thorough analysis of the patient experience from patient-generated texts, and the first to answer ‘how’ and ‘why’ patients express each topic.

Because annotation is a laborious process, we were also curious if off-the-shelf clustering algorithms can identify patient experience topic without supervision. Unfortunately, clustering algorithms did not identify these topics. While the finding is discouraging, because the annotation process was based on randomly sampled patient reviews, we believe the classifier will still perform well on unlabeled patient review texts. However, we did not verify the classification performance outside of the annotated dataset.

Finally, in this chapter, we assumed that topics were independent of each other. While such assumption is beneficial when conducting the analysis of each topic, patients do not write reviews and consider each topic in isolation. To better understand patient experience, we also need to characterize how topics co-occur.

CHAPTER 5: TOPIC INTER-DEPENDENCIES IN PATIENT EXPERIENCE TAXONOMY

In this chapter, we first explore whether a single topic or multiple topics should be assigned to each patient-generated text. If there is a single topic per patient-generated text, the problem becomes a single-label multi-class classification task. The goal is to find a single topic that best fits the patient-generated text. On the other hand, if there are multiple topics per patient texts, the problem becomes a multi-label multi-class classification task. We argue that the patient-generated texts are best represented as a multi-label multi-class problem.

We then explore how topics may be dependent of each other. We believe there are two types of topic interdependencies. First, there may exist group of topics that are semantically similar to each other. For example, 'Helpful' may be semantically similar to 'Temperament.' There are eight different semantic groupings in patient experience taxonomy which are referred to as a 'meta-class.' To understand the relationship between topics, we first explore how topics within a given meta-classes may be interdependent of each other.

Second, semantically different topics – i.e., those that belong to different meta-classes may still be interdependent of each other. As an example, a patient may view a doctor who didn't spend time with him/her to be unhelpful. The topics represented in this case are 'Helpful' and 'Time spent.' These two topics are not semantically similar to each other, i.e., they do not belong to the same meta-class.

To understand topic inter-dependencies, we designed a model that explores the two scenarios in mind. By understanding how the topics interact with each other, we gain further understanding of patient-generated texts and their relationship to patient experience taxonomy.

5.1 INTRODUCTION

A patient's experience has a huge impact on his or her health outcome [5–7]. Positive patient experience leads to not only therapeutic experience but also trust and understanding between clinicians and patients. These affect adherence to treatment (such as taking medication), increased access to care, and a better understanding of underlying symptoms.

Traditionally, measuring patient experience was conducted by hospital surveys, most notably HCAHPS [107, 145–147] which is the nation's first standardized patient survey. The survey asks discharged patients 27 standardized questions about their hospital stay. Of these, 18 core questions pertain to critical aspects of patients' hospital experiences which

Question Type	**Example**
Communication with nurses and doctors	During this hospital stay, How often did nurses treat you with courtesy and respect?
Cleanliness and quietness of the hospital environment	During this hospital stay, how often were your room and bathroom kept clean?
Responsiveness of hospital staff	How often did you get help in getting to the bathroom or in using a bedpan as soon as you wanted?
Pain management	During this hospital stay, did you have any pain?
Communication about medicines	During this hospital stay, were you given any medicine that you had not taken before?
Discharge information	During this hospital stay, did you get information in writing about what symptoms or health problems to look out for after you left the hospital?
Overall rating of hospital	Using any number from 0 to 10, where 0 is the worst hospital possible and 10 is the best hospital possible, what number would you use to rate this hospital during your stay?
Recommendation	Would you recommend this hospital to your friends and family?

Table 5.1: HCAHPS Question Types and examples

are shown in Table 5.1[1].

While the HCAHPS survey is an invaluable tool for understanding patient experience, it has its limitations. These surveys are mandated only to hospitals subject to Inpatient Prospective Payment System (IPPS) [46]. The system covers only acute patients. The vast majority of the patients' experiences are not acute; primary doctors are often the first point of contact. There is no such standardized system in place which measures patients' experiences for less acute type patients. Furthermore, survey ratings are not personalized. What is good for one demographic (say, a relatively healthy patient, or a person of a specific

[1] Questions were reprinted from https://www.hcahpsonline.org/globalassets/hcahps/survey-instruments/mail/jan-1-2018-and-forward-discharges/click-here-to-view-or-download-the-updated-english-survey-materials.pdf

Meta Classes	**Child Classes**
Interpersonal Manner	Helpful, Temperament, Trust, Put-at-ease, Good with kids
Communication	Communication Skills, Explanation, Follow-up, Empathy
Technical Competence/Skills	Professionalism, Clinical skills, Perceived success treatment, Pain, Medication, Test quality, Perceived bias
Scheduling	Appointment access, Response/wait time/punctuality, Time spent, Referrals
Medical plan	Insurance coverage, Cost of care
Practice environment	Office environment, Privacy, Location
Overall patient experience	Overall patient experience
Other	Other

Table 5.2: Meta classes and its corresponding child classes.

race) may not be suitable for another demographic [148, 149]. A way to see why patients have rated and talked about a particular topic would mitigate these problems.

Online sources act as an excellent supplement to these surveys [26]. A considerable benefit of review websites such as Yelp or RateMD is that, similar to online product reviews, patients give not only numeric ratings to doctors, but also give a subjective opinion of their hospital experience. There is richness in data that researchers can leverage to understand what patients think about a given topic. To realize this goal, however, we first need to be able to identify the topics of discussion. Unfortunately, patient reviews are unstructured which makes it somewhat difficult to extract topics from what patients have written. A reliable method to identify topics that patients are talking about will be immensely helpful in understanding patient experience.

In the previous chapter, we utilized a topic taxonomy of online patient reviews motivated by past research works in this domain [20, 21, 25]. It consists of 27 topics which roughly mirror those identified in the HCAHPS survey and as well as others mentioned by the previous researchers. The 27 different topics which are referred to as child-classes each have a single parent. The parent class of the child-class is denoted as a meta-class. Similar child-classes are grouped in the same meta-classes, whereas those that are different are grouped into a different meta-class[2]. We reprint the patient experience taxonomy in Table 5.2.

Topics from the patient experience taxonomy are then assigned to patient reviews, allowing

[2] While there are 27 topics in patient experience taxonomy, we built a classifier using only 23 topics since the four topics (Test quality, Perceived bias, Privacy, and Location) appeared very infrequently. We follow this convention and also use 23 topics in conducting experiments

researchers to further investigate different topics. There can be a single topic for a given review instance or multiple topics. Single-label multi-class scheme is an instance where each text is assigned a single topic. However, this does not adequately capture all the topics that are discussed in a corpus. As an example, consider the following patient review: *"The ability to schedule an appointment so quick was great and appreciated. Doctor took time to go over my pain and provide some routes to go with treatment."* In this review snippet, the patients talk about appointment access and explanation skills, covering two topics. A single-label multi-class classification scheme, by definition, can only pick one of the two topics, and hence is not sufficient for the task. Assigning multiple labels to each review instance will better represent the topics that are discussed in reviews.

In this work we claim that the topic classification problem in the patient review corpus should be defined as a multi-label multi-class classification and not as a single-label multi-class one. The multi-label multi-class classification (MLMC) problem is where each review text is assigned multiple topics and is not limited to single labels, as is the case with the single-label multi-class classification setting. Furthermore, in the past works that have considered patient reviews as multi-label multi-class classification problem, researchers have considered the topics to be independent of each other [20, 25]. However, this is an overly simplified assumption and some topics may be closely related to each other. For example, it can be the case that when a patient mentions that the doctor put him or her at ease, he or she also trusts the doctor. Similarly, a patient who has expressed easy appointment access to the hospital may also feel that the staff was professional.

Furthermore, in classifying patient experience taxonomy on patient reviews as a multi-label multi-class classification problem, there is a clear dependency between the meta-classes and child-classes in the classification process. There are two types of dependencies. First, child-classes are grouped into the same meta-class based on how similar they are to each other. Inevitably, this causes the child-classes that share the parents to have strong correlations with each other. If we are aware that a parent class exists for a given patient review (for instance, 'Scheduling' meta-class), we are then confident that at least one child-class (for example, 'Appointment access') of the said meta-class will exist.

Second, outside of the parent-child relationship, child-classes may still correlate with a different meta-class. This occurs because patients may not only talk about similar topics, but different sets of topics as well. For instance, a patient may start by talking about the clinical skills of the doctor, along with the medication that they were administered. They may further mention that the doctor had good communication skills, and explained the purpose of each of the treatments. In the above case, across four child-classes that the patients talked about ('Communication skills,' 'Explanation,' 'Clinical skills,' and 'Medication'), two meta-classes were mentioned ('Communication,' and 'Technical Skills'). By capturing the dependencies between the child-class and the meta-class, we gain deeper representation of the patient experience taxonomy and patient reviews.

In ultimately capturing the label dependencies, we explore our work in a deep neural network framework because of its success in tasks ranging from classification in both images [150,151] and texts [152–154]. The success comes from the ability to automatically infer natural feature representation; hence the reduction in feature engineering. A further benefit of this framework lies in its inherent ability to add constraints to its objective function such as those seen in Constrained Convoluted Neural Network [88]. Because the framework is capable of capturing natural feature representations that may only have been obtained via feature engineering, utilizing this framework forces the proposed constraints to be sufficiently meaningful; otherwise it would already be captured by its input features.

Our contributions are as follows:

1. We show that in hierarchical multi-label multi-class classification tasks, enforcing the structure (parent-child) of the taxonomy better captures label correlations. Meta-classes group similar child-classes in patient experience taxonomy. By enforcing the parent-child relationship, we are able to capture inherent correlations between child-classes that exist within a given meta-class.

2. Rather than capturing correlations between the child-classes which is typically conducted in the multi-label multi-class classification domain, we further demonstrate that we can instead capture correlations between a meta-class and a child-class. Such is beneficial especially in the case where a given child-class does not appear frequently in a given corpus, but its meta-class appears frequently.

3. We propose using Deep Neural Network (DNN) to classify online patient review topics. While the framework has been used in many different areas, to the best of our knowledge, DNN has not been used in classifying patient experience taxonomies.

In Section 5.2, we discuss patient experience and patient reviews, followed by Section 5.3 where we provide justification for picking the multi-label multi-class classification approach to capture patient reviews, and provide the framework that we use to capture class correlations. We then discuss hypotheses and methods that we developed to capture topics in patient reviews in Section 5.4. We evaluate our method in Section 5.5 and conclude in Section 5.6.

5.2 UNDERSTANDING PATIENT EXPERIENCE IN PATIENT REVIEWS

Online patient reviews are places where patients can describe their hospital experience and are publicly available to be viewed by anyone else who comes to the website. An example of such review websites is RateMD which is a leading patient review webpage. To write reviews on this website, patients go through two separate reviewing process to provide their experience at hospital. First, patients are asked to provide meta-information on five different aspects which are helpfulness, punctuality, staff, and knowledge, and their overall experience. Patients can assign anywhere from one star (dissatisfactory) to five stars (satisfactory) in these aspects. Afterwards, patients are asked to describe their experience in free-response format. These free response reviews typically have 65 words across 4 sentences. The vast majority of these (95%) are 10 sentences or fewer.

Within the reviews, patient talk about numerous different topics, where we show commonly mentioned topics in Table 5.2. We refer to these as patient experience taxonomy. The taxonomy is hierarchical in nature, where a parent topic (which we call meta-class) contains multiple child topics (which we refer to as child-class). This is a fairly common setup, where different researchers in patient experience have also proposed hierarchical taxonomy structure [20, 25, 27, 155]. On the other hand, these past works have not utilized the hierarchical structure of the taxonomies, opting to either focus on meta-topics [27, 155], or to focus on child-topics [25].This finding will become evident later on in this paper.

A given review instance can be assigned a single topic from patient experience taxonomy. These are instances of single-label multi-class classification (SLMC), where each review document is assigned a single topic. The benefit of this approach is that assigning a single topic is a lot easier than assigning multiple topics since annotators only need to pick one out of L labels. On the other hand, assigning multiple topics to a patient review is a lot more difficult since there are 2^L possible label combinations. The task of assigning multiple topics to a given review is defined as multi-label multi-class (MLMC) classification problem. In reviews with multiple topics, some of the topics are bound to correlate. Capturing the correlations is, by itself, a different set of problem that needs to be answered in MLMC classification tasks. We explore these issues in the next section.

5.3 SINGLE LABEL AND MULTI-LABEL MULTI-CLASS CLASSIFICATION

The task of classifying multiple texts (i.e., documents) can broadly be defined as a single-label multi-class (SLMC) classification or a multi-label multi-class (MLMC) classification problem. In a single-label multi-class classification setting, the task is to find a single topic that best captures the characteristics of the given dataset. In capturing this single topic, classifier $f_k(\mathbf{x})$ is trained for each topic. Here, $\mathbf{x}$ is the set of features and $s = f_k(\cdot)$ outputs a score s for the k-th classifier which represents the k-th label. The score s represents the fitness of a given label k. Ultimately, the goal of the classifier is to find the best k such that $\hat{y} = argmax_{k \in [1,L]} f_k(\mathbf{x})$.

In the traditional NLP domain, many tasks have taken the form of the single-label multi-class classification problem. Document classification is one instance of these, with the most prominent example being the 20 newsgroup classification task. Furthermore, we can treat parts of speech tagging (POS), semantic role labeling, or named entity recognition as a single-label multi-class classification task. In these use cases, each text unit (i.e., word) is assigned a single label amongst many possible labels (such as verb or noun in the case of POS tagging).

In assigning topics from patient experience taxonomy to patient reviews, we notice that some patient reviews do, indeed, have single topic for a given review snippet. As an example, consider the following snippet:

"Dr. Le is very knowledgeable, especially about the hip problem I have."

In this snippet, the patient primarily talks about how much the doctor knows, so it corresponds to a single topic, 'Clinical Skills.' These types of reviews are suitable for single-label multi-class classification schemes.

However, many of the reviews are not as simple as the above snippet. A more typical review has multiple topics. In such cases, a review may have multiple child topics that share the same meta-class, or may not share the same parent. Consider the following review snippet:

"The ability to schedule an appointment so quick was great and appreciated. Doctor took time to go over my pain and provide some routes to go with treatment."

In this review snippet, the patient primarily talks about 'Appointment Access,' and 'Time Spent.' Both of the topics fall under the 'Scheduling' meta-class, so, at least in the case of a meta-class classification task, a single-label multi-class classification scheme is sufficient to capture the topics discussed in the review.

Other reviews do not even share the same meta-class, as we illustrate in the following snippet:

"Very pleased with the quick availability of an appointment and the results of the appointment. My teenage son was very comfortable with Dr. Le."

In this snippet, the child topics are 'Scheduling Flexibility' and 'Good with Kids.' The snippet has two child-classes and also two meta-classes, 'Scheduling,' and 'Interpersonal Manner.'

We notice from the above example, that while SLMC representation is suitable for some patient reviews, they do not encompass other types of reviews. In assigning topics according to patient experience taxonomies, we found three different types of product reviews: reviews which have only a single child-class, those that have multiple child-classes but single meta-class, and those with multiple child-classes and multiple meta-classes. In framing patient experience taxonomy in patient reviews as a single-label multi-class problem, there is a loss of relevancy, where the prediction does not fully capture the gold label as we have shown in the previous examples. Assigning multiple labels to a given review snippet is able to address the above problem.

In the traditional NLP domain, while single-label multi-class classification tasks have been adapted into a wide array of applications, these are not sufficient for many other use cases, similar to what we have seen in assigning topics to patient reviews. Document tag prediction tasks [156], where the goal is to predict which tag the document is associated with, or product recommendations, are examples where single-label multi-class tasks are not able to capture the given task. The goal of these tasks is to predict more than one tag in the case of the document tag prediction task or to provide multiple product recommendations for the product recommendation tasks.

To address the limitations of single-label multi-class problems, researchers have proposed assigning multiple labels to prediction tasks. These are called multi-label multi-class classification tasks. These tasks assign multiple labels per dataset. More formally, we have

$$F(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_L(\mathbf{x})) \tag{5.1}$$

Each classifier, $f_i(\mathbf{x}), \quad i \in [1, L]$ learns whether a label exists (positive) or not (negative), independent of what the other classifier has learned. The goal is to find label representations such that the distance between $F(\mathbf{X})$ and gold label Y is minimized. From the above formulation, perhaps the most intuitive multi-label multi-class classification approach lies in binary relevance (BR) model [41]. The model assumes that there are no label correlations between labels, and trains L separate classifiers, where L is the number of labels in the dataset.

In the MLMC setting, however, as the number of classes increases, some subset of classes will correlate. Of these, some set of labels are truly related to each other. In such cases, the classifier should take the correlations into account. As an example, a patient who felt that they had easy appointment access are also likely to feel that the staff members were helpful. By learning that appointment access and helpfulness are related, we gain a richer understanding of the dataset. To encode these types of relations, we explore possible ways that are used to capture label correlations in the next section.

5.3.1 Capturing Label Correlations in Multi-Label Multi-Class Classification Task

Given the importance of capturing label correlations, we discuss several lines of work which capture label correlations in the multi-class multi-label classification problem. The lines of work we discuss are label powerset, space reduction, label chains, and explicit relationship encoding. We briefly discuss each of the lines of work in this section.

Label Enumeration Approaches

A naive extension of BR is to encode all possible label correlations. Label powerset enumerates every possible label combinations [157]. More precisely, given set of labels L, label power-set computes correlations for all $l \subseteq L$. While this would capture label correlations, such a powerset will be exponential in size, hence is not desirable. To mitigate the exponential nature of label correlations, other researchers proposed taking k labels to compute correlations [158]. More precisely, suppose there are labels $L = \{l_i\}_{i=1}^{n}$. These approaches then take k subsets of labels $L' \subseteq L$ and treat it as a class, which we denote as $c(L')$. It then feeds $c(L') \subseteq C$ into a single-label multi-class classification framework, where C is the set of k subset. These approaches do not explicitly encode the inherent correlation between the two classes, opting to treat a subset of classes as a single unit instead. Furthermore, there is a loss in expressiveness since the framework can only predict k label subsets. If the gold label for a text is such that $l_g \nsubseteq C$, then there is no possibility that the classifier can predict the true label.

Label enumeration approaches are unlikely candidate for capturing patient experience taxonomy. If we were to opt for the powerset approach, then there will be 2^{23} label correlations that we need to consider, and this is simply not feasible. Furthermore, the number of annotated data in the case of patient experience taxonomies are generally in several thousands, so opting for this approach will lead to overfitting into the dataset. On the other-hand, a k-subset approach leads to losing out on important label correlations and we may end up classifying a limited number, i.e., k number of label sets. However, this limits the usefulnesss of exploring patient experience discussed in patient reviews and some topic combinations may simply not be considered.

Label Space Reduction

Due to both the exponential nature of powersets and limited labels that can be recommended in k-subset approaches, researchers proposed, instead to perform label space reduction in the form of label embeddings [151,159,160]. These approaches embed low dimensional label subspace instead of trying to learn all possible label combinations, leading to mitigation of scalability issue. Label embedding spaces may suffer from loss of information since instead of directly learning the classes, classifiers learn the embedding space. Furthermore, embedding spaces often require a significant amount of data to properly learn reasonable subspace projection, which may not always be available. Similar to what we have argued in the previous section, because in many applications the reviews annotated with patient experience often number in several thousands, label space reduction is also not a feasible candidate for understanding label correlations.

Explicit Relationship Encoding

Rather than trying to learn all possible combinations of labels, or to project labels into a label embedding subspace, another line of work learns the structure of the label correlations. Some of these examples include label chains [161, 162] which treats classes in a multi-label problem as sequences of classes to predict, or by statistical or loss function analysis of label dependencies [86, 154, 163–165]. In label chain approaches, the task is transformed into learning multi-label sequences. The labels are first ordered from the most frequently appearing to the least. Then, the algorithm learns the sequence of labels, similar to parts-of-speech tasks. The intuition is that by learning the sequences, the algorithm learns the relationship between more frequently appearing labels to less frequently appearing labels. Statistical methods, on the other hand, first compute statistical correlations between labels, such as χ^2 or t scores. In computing the multi-class multi-label classification tasks, these statistics are added on top of the prediction layer.

A different line of work opts to directly encode the hypothesized structure of the corpus [88, 166–168]. In this line of work, domain specific constraints are leveraged to aid in multi-label multi-class classification tasks. In one example, they allow only some labels to co-occur together [88], and penalize the objective function when constraints are violated, while another line of work encodes domain knowledge to aid in classification tasks [168]. The benefits of these line of works lie in better capturing the domain characteristics and in requiring less labeled data since the hypothesized structure already encodes what the algorithm, without guidance, would also have had to learn. Furthermore, by encoding the domain characteristics, we are able to test hypotheses directly onto the model, hence allowing us to explore the corpus. Because of the benefits, we opt to encode label relationships explicitly over the other approaches. In particular, we utilize the hierarchical structure of patient experience taxonomy in automatically assigning topics to patient taxonomy.

5.4 METHOD

5.4.1 Hypotheses

In the previous section, we identified three types of patient reviews, where the first type had a single topic, the second type had multiple child classes but a single meta-class, and the third type had multiple child classes and multiple meta-classes. Furthermore, we have argued, in the case of MLMC problem formulations, we need to also capture how labels are correlated with each other.

In the first type of patient reviews, there is no need to worry about label correlations as this is an instance of an SLMC problem. On the other hand, label relationships need to be considered for the second and the third type of reviews. In the second type of patient reviews, all of the child-classes share the same meta-class given a patient review. By explicitly encoding the patient experience hierarchies, for example, that 'Helpful,' or 'Temperament' belong to the same meta-class 'Interpersonal manner,' we can limit the number of possible child-classes that is possible for a given patient review. Algorithm can limit the search space by first identifying the meta-class, and then penalizing the objective function if the model predicts child-classes that do not belong to the identified meta-class. However, limiting the search space is contingent on a well-formed patient experience taxonomy, in particular, when meta-classes form a cohesive collection of child-classes. As an example, a meta-class such as 'Interpersonal manner' is expected to contain child-classes that is related to the said meta-class. If the taxonomy is not well formed – say, some class really should have been a child-class of 'Interpersonal manner' but belongs to a different meta-class, or vice versa – the meta-class representation becomes less cohesive. Linguistically, some of the child-classes will have more similarity to a different meta-class – say, 'Scheduling' – than it is to the meta-class of interest. Capturing the meta-class representation in such cases will prove to be difficult which will lead the meta-class classifier to induce more errors, which will then propagate to child-class classification tasks.

In the third type of patient reviews, those that have different meta-classes and child-classes in a given review, we capture the taxonomy characteristics by learning correlation between a child-class and a meta-class. Let us consider the following review snippet:

"Dr Le is very professional and pleasant. She has a very caring demeanor and listens. I highly recommend Dr Le if you are looking for a healthcare provider."

In this snippet, the corresponding child-classes are 'Professionalism,' 'Temperament,' and 'Communication skills.' The corresponding meta-classes for this snippet are 'Interpersonal manner,' and 'Communication.' An approach that learns correlations between meta-classes will find 'Interpersonal manner' and 'Communication' are correlated from this snippet. In the two meta-classes, there are five and four child-classes in 'Interpersonal manner,' and 'Communication,' respectively, and lumping the child-classes together to encode correlation will lead to a very coarse set of correlations. However, because there are, in patient experience taxonomy, 23 topics, as seen in Table 5.2, this may lead to very large number of class correlations.

A more pressing issue is when a target class does not appear frequently enough to be captured into a correlation relationship. Take, for example, a child class (for instance, 'Trust' which falls under 'Interpersonal manner' meta-class) that does not appear very often in the corpus. In this case, it is infeasible to properly learn the relationship between this class and another class even if it appears frequently (for instance, 'Communication skills'), since the former label ('Trust') appears infrequently. This is because the co-occurrence between the target label and the anchor label ('Communication skills') will not appear enough to properly capture the topic correlations. On the other hand, 'Interpersonal manner' is a frequently occurring meta-class, and capturing correlation between the child-class 'Communication skills,' and the meta-class 'Interpersonal manner' will mitigate this issue.

To capture the relationship between the child class and the meta class, we propose two hypotheses:

- **Child classes are strict subsets of meta-classes.** The proposed two-layer patient experience taxonomy from the previous chapter properly captures the relationship between child and meta-classes. The relationship can be seen in Table 5.2, where the meta-classes, such as 'Interpersonal manner' contains five child-classes, and they are a one-to-many mapping. In a well-formed taxonomy, a model that leverages the hierarchical relationship will better capture patient experience taxonomies in patient reviews. This hypothesis aims to validate the relationship between the meta-class and the child-class.

- **Meta-classes and child-classes correlate outside of their parent-child relationships** In a strict meta-class and child-class relationship, we expect there to be nearly a perfect correlation between a child-class and its parent since there is a hyponym relationship between a child node and its parent node. On the other hand, this hypothesis does not cover the relationship between a meta-class and a child-class that do not have a hypernymic relationship. For example, it is possible that a child-class 'Empathy,' which falls under the 'Communication' meta-class frequently correlates with the meta-class 'Interpersonal Manner' but does not strongly correlate with a 'Scheduling' meta-class. The child-class 'Follow-up,' which also falls under 'Communication' meta-class may correlate more strongly with 'Scheduling' meta-class rather than that of the 'Interpersonal Manner' meta-class, however.

In testing the two hypotheses, we explore optimization frameworks which will allow us to encode different experimental settings. The framework should allow us to incorporate the two hypotheses, and compare with the null hypotheses. The null hypotheses in our case form a simple binary relevance model, where we assume no label correlations. We discuss the optimization framework that we use in this experiment.

5.4.2 Optimization Framework

We utilize a deep neural network (DNN) framework to test out the hypotheses. DNNs have gained popularity recently in part due to their more expressive features than traditional loss function optimization modeling frameworks. Each of the input features in the neural network is embedded into multiple hidden layers which learn latent features that traditional approaches are not able to capture. An additional benefit, which is particularly more pertinent in testing various hypotheses, is its flexibility in encoding different loss functions. Similar to traditional loss function optimization tasks, DNNs allow researchers to modify the optimization objective to better capture modeling hypotheses. In the past, neural network frameworks were combined with existing representations of multi-class multi-label classification such as binary relevance [154], chain classification [162], label embeddings [151, 169] which improved performance. Furthermore, researchers in the past were also able to add various modeling hypotheses on top of DNN frameworks [88, 169]. Techniques that combined neural network with traditional modeling frameworks show improvements over existing using only the traditional modeling approaches. It is because of this flexibility and capability of expressive feature representation that deep neural network frameworks have seen wide applications in both image recognition and text categorization. Since the framework provides both the improvement in feature representation and the flexibility in encoding different research assumptions, we opt to utilize this framework in posing our problem.

5.4.3 Model

We now discuss how we have modeled the two hypotheses proposed in the previous section.

Modeling Child classes are a strict subset of meta-class

We model the first hypothesis in this section. In hierarchical taxonomies, a given meta-class contains at least one child class. More precisely, let us define children of a given meta-class y_p as $Y_{pc} = children(y_p)$. In such cases, we see that $|Y_{pc}| \geq 1 \;\; \forall y_p \in Y_p$, where Y_p is a set of meta classes.

Given the above assumption, we now define the meta-class classifier as $\hat{\mathbf{y}}_p = f_p(\mathbf{x})$ and child-class classifier as $\hat{\mathbf{y}} = f(\mathbf{x})$. For each of the classifiers, we see that $\mathbf{y} \subseteq \mathbf{Y}$, whereas $f_p(\mathbf{x})$ learns to predict meta-classes $\mathbf{y}_p \subseteq \mathbf{Y}_p$ from the training features $\mathbf{x}$. There are anywhere between $0 \leq |f_p(\mathbf{x})| \leq L_p$ and $0 \leq |f(\mathbf{x})| \leq L$, where L_p is the number of meta-class labels, and L is that of child-class labels.

Now let us look at an arbitrary prediction, $\hat{y}_p \in \hat{\mathbf{y}}_p$ from $\hat{\mathbf{y}}_p = f_p(\mathbf{x})$. We expect for each meta-class predictions $\hat{y}_p \in \hat{\mathbf{y}}_p$, $\exists \hat{y}_{pc} \in \hat{\mathbf{y}}_{pc}$ from $\hat{\mathbf{y}}_{pc} = f(\mathbf{x})$ such that $\hat{y}_{pc} \in children(\hat{y}_p)$. If meta-classifier predicted a parent label, then the child-classifier should predict at least one label such that its parent is the same as the predicted parent label.

We formalize this relationship as

$$|f_p(\mathbf{x})| - \sum_{parent(f(\mathbf{x}))) \in Y_p} |parent(f(\mathbf{x}))| \leq 0 \quad \forall f(\mathbf{x}) \in Y_{pc} \tag{5.2}$$

where $parent(\cdot)$ refers to set of all parents of the predicted labels, and both $f_p(\cdot)$ and $f(\cdot)$ are indicator function, i.e., they output either 0 or 1 for a given label, Y_{pc} refers to the set of all children classes for y_p as defined above.

Notice the indicator function is a very stringent constraint. Instead of representing $f(\cdot)$ and $f_p(\cdot)$ as indicator function, we can represent these as probabilities to sets of all possible labels. We call this function as $s(\cdot)$ and $s_p(\cdot)$ for child-class prediction probability and meta-class probability, respectively. For instance, suppose $Y_{pc} = \{a, b, c\}$, and we have a classifier $s(\mathbf{x}) = [0.7, 0.3, 0.4]$. This means we are 70%, 30% and 40% confident in labels a, b, and c, respectively . Notice that these do not sum up to 1 since this is an instance of the multi-class multi-label classification problem.

Using scoring functions $s(\mathbf{x})$ and $s_p(\mathbf{x})$, Equation 5.2 simply becomes

$$s_p(\mathbf{x}) - \sum_{parent(s(\mathbf{x})) \in Y_p} parent(s(\mathbf{x})) \leq 0 \quad \forall s(\mathbf{x}) \in Y_{pc} \tag{5.3}$$

The *parent*$(\cdot)$ simply performs a reduce function to the score from $f(\cdot)$ to its corresponding parent label.

Modeling Meta-class and child-class Relationship

We further hypothesize that meta-classes and child-label may correlate even amongst child and parents which are not a single hop from each other, as per our second hypothesis. To capture this relationship, we first represent the feature network, which is denoted as $F(\mathbf{x})$. The feature network can be represented as an arbitrary deep neural network. For an arbitrary n-hidden layer single layer, this is defined as

$$F(\mathbf{x}) = g(\mathbf{w}_{n-1}F_{n-1}(\mathbf{x})) \tag{5.4}$$

$$F_{n-1}(\mathbf{x}) = g(\mathbf{w}_{n-2}F_{n-2}(\mathbf{x})) \tag{5.5}$$

$$F_1(\mathbf{x}) = g(\mathbf{w}_0\mathbf{x}) \tag{5.6}$$

where F_k, $0 \leq k < n$ is a hidden layer, $\mathbf{w}_i$ is weight on i-th layer, and $g(\cdot)$ is a coordinate-wise non-linearity function, such as ReLu or Sigmoid.

In encoding the parent labels, we take as input $\hat{\mathbf{y}} = f_p(\mathbf{x})$ from a different classifier. We then concatenate the output from the feature network and the parent prediction. We have

$$\mathbf{x}_{fy} = F(\mathbf{x}) \oplus \hat{\mathbf{y}} \tag{5.7}$$

where $\oplus$ is a concatenation symbol. The concatenated feature is then fed into a hidden layer, which feeds to output into the prediction layer. The prediction layer predicts child-classes. This hidden layer is used to embed both the feature representations learned from the features and the parent predictions before the output of the hidden layer is fed into the prediction layer. The benefit of this representation is this allows mutual embedding of features used to predict child classes, represented as $F(\mathbf{x})$, and that of the parent class $\hat{\mathbf{y}}$. Mutual embedding ultimately allows us to encode the correlations between the parent and the child.

To formalize the representation, we have

$$\hat{\mathbf{y}} = F_p(g(F_d(\mathbf{x}_{fy}))) \tag{5.8}$$

$$F_d(\mathbf{x}_{fy}) = \mathbf{w}_d\mathbf{x}_{fy} \tag{5.9}$$

$$F_p(\mathbf{x}) = \mathbf{w}_p\mathbf{x} \tag{5.10}$$

where $F_p(\cdot)$ is a prediction layer, $F_d(\cdot)$ is label and feature embedding layer.

There are some differences to this approach and work that utilized canonical analysis in embedding labels [151]. In the canonical analysis work, researchers do not leverage any hierarchical information. Furthermore, they decouple the feature network and the label network entirely. Their approach learns the feature network and label network independently and only combining the two networks at the prediction layer. The premise behind this is to learn label embedding space and feature embedding space independently. While this is a powerful method, in the case where the dataset is small, label embedding approaches may end up learning noisy label correlations. We instead embed hidden features and parents classes in the same layer, removing the possibility that parent label embeddings are learned independently of the embedded features, hence mitigating the chance of learning noisy label correlations.

Loss Function

We introduced a new constraint which pertains to the parent-child relationship as seen in Equation 5.3. We discuss how we incorporate the constraint into the loss function. We note that there should be some penalty if the parent-child constraint is violated. To do so, we have

$$PC = g(-s_p(\mathbf{x}) + \sum_{parent(s(\mathbf{x}))\in Y_p} parent(s(\mathbf{x}))) \quad \forall s(\mathbf{x}) \in Y_{pc} \tag{5.11}$$

This allows us to penalize the term if and only if the prediction violates the constraints. Similar to the work done in Constrained Convoluted Neural Network [88], we introduce slack variables to allow for convexity. We have,

$$PC = (-s_p(\mathbf{x}) + \sum_{parent(s(\mathbf{x}))\in Y_p} parent(s(\mathbf{x})) - \xi_i)^2 \quad \forall s(\mathbf{x}) \in Y_{pc} \tag{5.12}$$

where $\xi_i \geq 0$ and i refers to i-th label, and ξ_i is slack variable for i-th label.

The optimization function is then given by

$$\min_{W_1 \ldots W_{n+1}} \quad H(\mathbf{y}, \hat{\mathbf{y}}) + \lambda \cdot PC \tag{5.13}$$

where λ determines how much weight to assign to parental constraint. $H(\mathbf{y}, \hat{\mathbf{y}})$ is cross entropy between the prediction and the gold standard. We opt for cross entropy since this is a very widely used metric in the literature, but we also compare this with the squared difference as well.

5.4.4 Comparison with Existing Hierarchical Classification Approaches

Other researchers have also utilized hierarchical taxonomy structures to improve classification performance. In hierarchical taxonomies, class structures follow a directed acyclic graph (DAG), in particular, that of a tree. A typical taxonomy starts from the root of the tree. The root has some number of child classes, which we denote as $c_{(1,ROOT),i}$ where i refers to i-th child, and $ROOT$ denotes that the class is a descendant of the $ROOT$ class. Each of the child class $c_{(1,ROOT),i}$ may have descendents, denoted as $child(c_{1,i}) = \{c_{(2,c_{1,i}),j}\}_j$. In this notation, 2 refers to the level, and j refers to a j-th child of the ancestor class $c_{1,i}$. Notice that we can further define this recursively, more precisely, as $child(c_{k,i}) = \{c_{(k+1,c_{k,i}),j}\}_j$.

There are multiple approaches to encode hierarchical taxonomy structures. The simplest one is by simply ignoring the presence of the hierarchy. These approaches are called flat hierarchy approaches [170]. Notice that flat hierarchies are no different from the non-hierarchical classification algorithm. There have been some studies that indicate in the case where class labels are balanced, flat classification approaches may have an advantage over hierarchical ones [170]. In our case, class labels are imbalanced, so we opt for a hierarchical approach.

There are two different lines of thought in incorporating hierarchy into the classification scheme. The first is to utilize local classifiers [171–175] . These approaches treat each layer as a separate classification problem. Local classifiers first predict ancestor classes and use this information to predict its descendent classes further. Suppose the classified parent node is $c_{k,i}$. In classifying its child node, parent-class label $c_{k,i}$ heavily influences the set of labels its child can be predicted into. The most significant limitation of this approach is misclassifying a parent node, which may propagate errors to the remaining descendent nodes. Furthermore, in local classifiers, both the parent-classes and child-classes are trained independent of each other, and predictions are later corrected via post-processing. The underlying classifiers, because of this reason, do not learn the hierarchical relationship between the child and the parent.

To mitigate the issues with local classifiers, researchers started utilizing global classifiers [176–178]. Rather than classifying each node at a time, global classifiers take into account the class hierarchy as a whole during a single run of the classification algorithm. A major advantage of global classifiers over local classifiers is that misclassifying a parent class does not lead to propagating classification errors to the child classes. Furthermore, unlike local classifiers, global classifiers learn the relationship between child-classes and parent-classes concurrently, which allows us to capture the link between the two types of classes directly. On the other hand, because this is a less intuitive approach, coming up with the proper global model in particular, an objective function that captures the hierarchical structure may prove to be complicated. Coming up with a global model requires researchers to explicitly encode an objective function that characterizes the relationship between child-classes and parent-classes.

Researchers in Deep Neural Network (DNN) communities have also utilized a hierarchical classification scheme, and there is a mix of both local and global classification approaches. The approach utilized by Kowsari et al [179] is an example of local classification approach. In the aforementioned cited work, researchers proposed first classifying parent classes via a typical deep neural network (such as Convoluted Neural Network). They then utilized parent-class classification results to aid in classifying child classes via Recurrent Neural Network (RNN) framework. Unlike typical RNNs, which only takes the feature space as its input, these methods take parent-class predictions as part of its feature space as well. Similarly, Salakhutdinov et al [180], and Roy et al [181] first classified parent-classes, then used these as a prior to further train and predict child-classes. Something we see in local classification approaches is that parent-class classifiers and child-class classifiers are typically trained separately. Because child-classes typically utilize parent-classes in learning to predict labels, parent-class predictions do influence child-class predictions. However, child-class predictions do not influence parent-class predictions.

DNNs have also been utilized for global classification approaches. DNNs typically utilize loss function to guide both the training and prediction tasks. The loss function itself is global; i.e., there is only a single loss function that drives the classification process. Because of its structure, past researchers have extended the loss function to encode the structure of the prediction tasks further. Previous works which utilize similar constraint to ours (parent-child constraints) can be seen in Zhang et al [168]. The researchers, similar to what we have done, propose enforcing parent-child constraints. They propose adding a parent-child constraint to their objective function and minimize the prediction error and constraint violation at the same time. Researchers such as Yan et al [182] similarly incorporated whether or not parent-class predictions and child-class predictions were consistent with each other (the two classifiers both predicted correct parent-child relationships). A general trend we see in global classification approaches that utilize DNN is that their objective functions explicitly measure whether the parent class and child class predictions form a valid relationship or not.

Our approach is a combination of local and global classification approaches. It is local in the sense that the prediction task is still taken as input for the parent class predictions, i.e., our second constraint which feeds in the predictions from a parent classifier. Parent class predictions influence child-class predictions. On the other hand, our first constraint, which enforces parent-child relationships, is a global one. The constraint is encoded as part of the loss function – hence, a global approach. More precisely, we have modified the loss function so that it incorporates the relationship between parent-classes and child-classes (corresponding to global classifier approach) and utilizes the parent-class predictions as an extended set of features to improve child-class predictions. Some of the past works have also combined local and global classification schemes [182, 183] in which researchers utilize parent-class predictions to refine that of child-classes. The difference between these methods and ours lies in the objective function where researchers have different hypotheses on the relationship between parent-classes and child-classes. Fan et al [183] in particular, modeled their DNN on the assumption that parent-classes and child-classes share the same feature space. On the other hand, Yan et al [182] follows a similar assumption to ours. The researchers also enforce parent-child class constraints. However, they have multiple DNN components that classify each child-class, whereas our approach has a single DNN. The difference is that in Yan et al [182] researchers were not interested in capturing label-dependencies as their problem space setting was single-label multi-class classification whereas our goal was to capture multi-label multi-class classification.

5.5 EVALUATION

In this section, we first detail evaluation metrics used to compare different methods of multi-label classification. We then briefly discuss different baselines and our approaches to show the effectiveness of encoding parent information. Using the evaluation metrics and different approaches we then show the evaluation results and end the section with hyper-parameter analysis where we show how sensitive our method is on different values of parameters.

5.5.1 Dataset

In conducting our hypothesis on the relationship between topics in patient experience taxonomy, we utilized RateMD, an online patient review website where patients write about their hospital experience. The doctors range from primary physicians to dentists, and one of the first uses was in analyzing doctor sentiments and latent factors online [155]. We utilized a set of online reviews with 23 different topics across 3,590 online patient reviews as described in Chapter 3.

A unique trait of this dataset is the existence of class hierarchy, where each of the 23 topics belongs to one of eight parent topics. The parent topics are referred to as meta-classes. The relationship between meta-classes and child-classes introduces a unique topic hierarchy that can be exploited in building a classification model. For the rest of the paper, if there is a need to distinguish between one or the other, we refer to the topics as child-classes and their parent classes as meta-classes.

Of the 23 child-classes, the most frequent class appeared 1,882 times, whereas the least frequent one appeared 80 times. On average, there were 3.32 classes assigned per review. Amongst eight meta-classes, the most frequent class appeared 2,715 times, and the least frequent one appeared 89 times. On average, there were 2.65 meta-classes assigned per review.

5.5.2 Experimental Settings

We investigated various shallow features such as unigram, word embeddings, rating information, and dependency parses in Chapter 4. Rather than concoct a new set of features, we focus on how labels interact with other labels. To do so, we need strong baseline features. Motivated by FastText [152], we utilized a word embedding representation of the document. FastText represents averaged out latent embedding space as its document representation. Classification performance on document embedding labels is on-par with unigram representation of a document while, at the same time, requiring significantly fewer features. Since our goal is to focus on the relationship between child-classes and meta-classes in the RateMD dataset rather than to experiment with different features, we opted to utilize the representation motivated by FastText.

To represent embedded document representation, we crawled 50k online reviews from RateMD and trained Doc2Vec by utilizing unigram features. These documents were crawled because we only had 3,590 training examples and it was not sufficient to properly train document embedding space. Each of the documents was then represented as a set of 10 sentences, each with 200 dimensions in latent subspace[3]. We picked 10 sentences because an average review has 4.78 sentences, with standard deviation of 2.92. We found that 95% of review texts have 10 sentences or fewer, hence the justification for picking 10 sentences. For documents with fewer than ten sentences, the remaining vectors were assigned zero vectors. The deep learning framework took the embedding as an input vector, with three hidden layers, each with 1000 dimensions, dropout probability of 0.5, and learning rate of 0.0001. All the neural networks we compared had the same number of hidden layers and the same number of latent embedding space[4]. Furthermore, we predicted a label l if $score(l) > s_l$. During our validation phase, we find the best $s_l \; \forall l \in L$, where L is the set of labels we try

[3] We did not observe a noticeable difference in performance past 200 dimensions

[4] Neither tweaking hidden layers nor the number of dimensions have any measurable difference after the third layer and 1000 dimensions, hence the reason why we picked this set of hyper-parameters

to predict. Threshold s_l was picked between $[0.05, 0.5]$, in 0.05 increments.

In all of our experiments, we ran a series of classification models with an 80/20 training/test split and 5-fold cross-validation.

5.5.3 Evaluation Metrics

Following the methodology used in previous works [151,154,184], we include micro/macro F-1, Hamming Loss, coverage error, ranking loss and label ranking average precision. For the sake of completeness, we also include weighted F-1 on top of micro and macro F-1. We provide a brief overview of each of the evaluation metrics in this section.

Hamming Loss

Hamming loss measures how many times an instance-label pair is misclassified, i.e. a label not belonging to the instance is predicted or a label belonging to the instance is not predicted. More precisely, suppose the true label for i-th data on j-th label is $y_{ij} \in \{0, 1\}$. Given a prediction $\hat{y}_{ij}$, suppose we have $\Delta h_{ij} = |y_{ij} - \hat{y}_{ij}|$. Using this, we have

$$\text{Hamming Loss} = \frac{1}{N}\sum_{i}^{N}\frac{1}{L}\sum_{j}^{L}\Delta h_{ij} \tag{5.14}$$

where N is number of data, L is number of labels. Since this counts the number of miscategorizations, a lower value of Hamming Loss is better than that of a higher one. It is immediately clear that if the expected number of the positive label is significantly smaller than all possible sets of labels, Hamming Loss biases toward not predicting any labels.

Coverage Error

Suppose we can calculate $score(\hat{y}_{ij})$ for all the predictions in i-th data. Using the scores, it is possible to sort all the predictions by their scores in descending order. The rank of the prediction can then be shown as $rank(\hat{y}_{ij})$. Coverage error measures the maximum rank that is needed to predict all positive labels. More precisely, we have

$$\text{Coverage Error} = \frac{1}{N}\sum_{i}^{N} max_{y \in Y_i} rank(y) - 1 \tag{5.15}$$

Similar to Hamming Loss, lower Coverage Error is better, but unlike with the loss, not predicting any labels will result in rather abysmal coverage error.

Ranking Loss

Ranking loss measures the proportion of two label pairs $(\hat{y}_{ij}, \hat{y}_{ij'})$ that are ranked in reverse, or are not properly ranked. If the pairs are ranked in the correct order, the loss is 0, otherwise, the loss is 1. We denote this as $rl(\cdot, \cdot)$. Ranking loss is more formally calculated as

$$\text{Ranking Loss} = \frac{1}{N} \sum_{i}^{N} \frac{\sum_{j}^{L} \sum_{j'}^{L} rl(j, j')}{L \cdot (L-1)} \tag{5.16}$$

Average Precision

Average precision was originally used in information retrieval area which measures the number of correctly identified documents existing in top k retrieved documents. Similarly, we rank each prediction and measure precision. Average precision is formally given as

$$\text{Average Precision} = \frac{1}{N} \sum_{i}^{N} \frac{1}{Y_i} \sum_{y_i \in Y_i} \frac{L_i}{rank(y_i)} \tag{5.17}$$

Micro, Macro, and Weighted F-1

$F-1$ metrics are widely used in evaluating classification performance. We test the performance of all classification models across the two schemas and across the seven text representations using precision, recall and F-1 measures evaluated at micro, macro, and weighted levels. The Micro method sums up the individual true positives (TP), false positives (FP), and false negatives (FN) of all classifiers $i = 1 \ldots n$:

$$P_{micro} = \frac{\sum_i \mathrm{TP}_i}{\sum_i \mathrm{TP}_i + \sum_i \mathrm{FP}_i}, \; R_{micro} = \frac{\sum_i \mathrm{TP}_i}{\sum_i \mathrm{TP}_i + \sum_i \mathrm{FN}_i}, \text{ and } F\text{-}1_{micro} = 2 \cdot \frac{P_{micro} \cdot R_{micro}}{P_{micro} + R_{micro}} \tag{5.18}$$

The Macro method averages the precision and recall over all classifiers $i = 1 \ldots n$:

$$P_{macro} = \frac{1}{N} \sum_{i=1}^{n} P_i, \; R_{macro} = \frac{1}{N} \sum_{i=1}^{n} R_i, \text{ and } F\text{-}1_{macro} = 2 \cdot \frac{P_{macro} \cdot R_{macro}}{P_{macro} + R_{macro}} \tag{5.19}$$

where P_i and R_i are the precision and recall of binary classifier i, respectively.

The Weighted method calculates weighted averages of the precision and recall over all classes $i = 1 \dots n$. Weight of class i is based on the relative frequency of the label:

$$P_{macro} = \sum_{i=1}^{n} p(i) \cdot P_i, \; R_{macro} = \sum_{i=1}^{n} p(i) \cdot R_i, \text{ and } F\text{-}1_{macro} = 2 \cdot \frac{P_{macro} \cdot R_{macro}}{P_{macro} + R_{macro}} \tag{5.20}$$

where P_i and R_i are the precision and recall of class i, respectively, and $p(i) = \frac{cnt(i)}{\sum_{j=1}^{n} cnt(j)}$. $cnt(i)$ refers to number of times class i appeared in the corpus.

Macro F-1 is used to determine the overall system performance across all the classifiers (and thus, across all labels), but does not account for the number of text instances within each label. Micro-F1 mediates this issue. Therefore, we report it and analyze it when comparing system performance. Ideally, a good classifier should perform well on micro, macro and weighted F-1 scores.

5.5.4 Baselines and Our Method

In this section, we test multiple baseline methods against our method. We briefly describe several baseline methods we compared against, along with three approaches that we propose.

Baselines

FastText [152] FastText is a widely used baseline in single-label multi-class text classification tasks. This allows us to compare SLMC with MLMC.

DNN-BCE [153] Acronym for Deep Neural Network-Binary Cross Entropy. Similar to FastText, this is another widely used baseline method in testing multi-label multi-class text classification tasks. This baseline is a binary-relevance model in the context of DNN. We include this baseline to compare classification with and without the hierarchical constraints.

BP-MLL [154] Acronym for Backpropagation for Multi-Label Learning. In both FastText and DNN-BCE, the loss function between the predicted label and the gold label is cross-entropy. BP-MLL uses squared-loss instead, and we include this baseline to test which loss function is better.

C2AE [151] Acronym for Canonical Correlated Auto Encoder. This approach projects labels into label embedding space by jointly learning feature latent space and label embedding space. As argued earlier, in a limited dataset setting, we are unable to learn the label embedding space properly, and we include this baseline to verify the hypothesis. Fur-

Method	HL	CE	RL	AP	F-1$_{micro}$	F-1$_{macro}$	F-1$_w$
FastText	0.132	8.52	0.127	0.650	0.305	0.148	0.257
DNN-BCE	0.130	8.58	0.125	0.672	0.438	0.361	0.488
BP-MLL	**0.128**	8.93	0.133	0.673	0.422	0.336	0.479
C2AE	0.146	12.19	0.230	0.535	0.327	0.220	0.412
HML-PC	0.131	8.55	0.124*	0.672*	0.445*	0.368*	0.494*
HML-PN	0.130	8.42*	0.124*	0.676*	0.540*	0.432*	0.546*
HML-PNC	**0.128**	**8.40**	**0.122**	**0.679**	**0.549**	**0.441**	**0.557**

Table 5.3: Evaluation Results. HL: Hamming Loss, CE: Coverage Error, RL: Rank Loss, AP: Average Precision, F-1$_w$: Weighted F-1. * indicates cases where our method performed better than the best performing baseline for a given metric, bolded text indicates best performing amongst all the algorithms tested.

thermore, C2AE captures correlations in the child-classes space, as opposed to meta-classes space that we opt in our method.

Our Methods

We compare the four baseline methods with three methods that we propose. All three method pertains to how we utilize the parent constraints.

HML-PN Acronym for Hierarchical Multi-Label - Parent Node, this is one of the approaches that we propose, where parent predictions are embedded alongside feature-embeddings. This approach does not encode parent-child constraint.

HML-PC This stands for Hierarchical Multi-Label - Parent Constraints, where we add parent constraints but do not embed parent predictions with feature embeddings.

HML-PNC This is a shorthand for Hierarchical Multi-Label Parent Node & Constraints which adds parent predictions as feature along with parent constraints. HML-PNC is a combination of HML-PN and HML-PC.

5.5.5 Evaluation Results

Performance Comparison

Using the evaluation metric we described in the previous section, we show the classification performance in Table 5.3[5].

[5] While our method was not sensitive to different values of λ, we set $\lambda = 1$ for this experiment

Not surprisingly, FastText, which utilizes a single-label multi-class model performs the worst, especially in F-1 metrics, indicating that our problem is, indeed, best modeled as a multi-label multi-class problem. An average number of topics per review text was 3.32, so a single-label multi-class approach will under-predict labels. Multi-label multi-class classification via label embeddings (C2AE) also did not perform very well. Because of the lack of data to properly learn embeddings, the classification performance of C2AE suffered. Furthermore, we have previously argued that capturing label correlations between two child-labels will lead to very few co-occurrence between the two labels. C2AE ultimately learns the correlations in the child-classes subspace which leads to learning some very sparse labels. We were limited to only thousands of training data which proved to be too sparse to learn label embedding space properly.

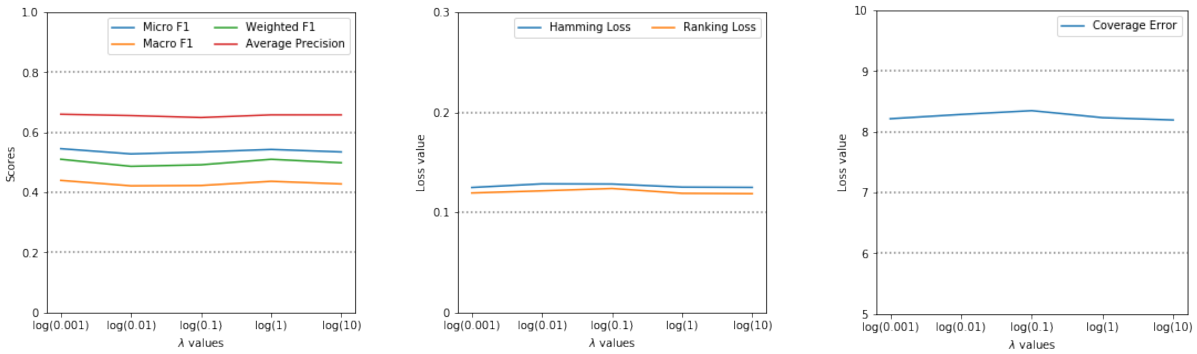
Hamming Loss metric was more interesting, however. We see baseline methods consistently performed better on Hamming Loss than our methods. Upon analysis, baseline approaches performed better, in part because the loss function measures how close the value is to the actual label. Since our data had on average 3.32 labels per review text, predicting a given review instance to negative labels generally reduces the loss.

We notice that, despite significant improvements on F-1 scores, our ranking loss did not differ very much compared to the baseline approach. Ranking loss computes whether irrelevant labels are ranked higher than relevant labels. More precisely, the loss function takes scores $\hat{s} \in Y$, and $\hat{\bar{s}} \in \overline{Y}$, where Y contains scores for relevant (positive) labels, and $\overline{Y}$ contains those for irrelevant (negative) labels. The metric computes the number of times that $\hat{s} < \hat{\bar{s}}$. These numbers are then normalized by $|Y| \cdot |\overline{Y}|$.

The classifier predicts a label l if $score(l) > s_l$. During our validation phase, we find the best s_l $\forall l \in L$, where L is the set of labels we try to predict. Threshold t_l was picked between $[0.05, 0.5]$, in 0.05 increments, hence t_l and $t_{l'}$ may not be the same for two labels $l, l' \in L, l \neq l'$. Because each label can have different threshold t_l at which the label is predicted, it is possible that $\hat{t_l} < \hat{\overline{t_{l'}}}$, but l is still predicted while l' is not predicted for a given review.

Similarly, both average precision and coverage error assume that the threshold t_l is the same for all labels, L. However, we employed dynamic thresholding, which we tuned based on the validation dataset. Thus, it is possible that for two labels, l and l', scores $s_l < s_{l'}$, but the classifier still predicts l but not l'. Dynamic thresholding allows F-1 scores to improve over the baseline even though average precision, coverage error, and ranking loss may not have improved that much[6].

[6] Notice that we ran dynamic thresholding on all the classifiers that we tested.



(a) 1a (b) 1b (c) 1b

Figure 5.1: Parameter sensitivity for different λ values. Notice that our method is not very sensitive to different values of λ

In terms of our method, adding parent predictions to the deep learning framework and the constraints improved performance. Adding parent predictions proved to provide more performance gains than adding constraints. However, adding both parent predictions and constraints (HML-PNC) had the best classification performance, leading us to conclude that the parent-child relationship we annotated on the taxonomy is generally correct, and that there are correlations between child-classes and meta-classes.

Parameter Sensitivity

Our method has a hyper-parameter, in particular, λ to decide how much weight to assign to parent-child constraint. A higher weight indicates that parent-child constraint is more important and vice versa. Notice that additional parent node (HML-PN) does not require λ parameter because it is an additional node to the deep network. Ideally, the algorithm should be less sensitive to the hyperparameter as methods that are sensitive to the parameter are difficult to optimize for best performance. We experimented with parameter values of $\lambda = \{0.001, 0.01, 0.1, 1.0, 10.\}$. We show the results in Figure 5.1. We see that the performance stays consistent over different parameter values indicating that tuning our approach is straightforward.

5.5.6 Analysis

We first measured which loss function works better. The two loss functions that we compared are cross entropy metrics (which is denoted as DNN-BCE) and squared loss (BP-MLL). Consistent with the literature, we noticed that cross entropy performed better than squared loss. Perhaps this is not too surprising since squared loss is more suited to regression problems than classification ones.

We next measured whether encoding child-class correlations is feasible in our dataset. To encode the correlations, the method needs to be resilient to noise, in particular, to prevent the algorithm from making false correlations. Because there are limited numbers of label annotations in our dataset, classes that appear infrequently are more likely to have false class correlations. This occurs because some labels may co-occur together several times by chance, and for classes that appear infrequently, simply co-occurring several times by chance may be enough to trump the true label correlations. To test whether encoding child-class correlations is resistant to making false correlations or not, we compared the performance of C2AE against the cross-entropy baseline (DNN-BCE). C2AE projects child-classes into child-class embedding space, and in doing, they learn how child-classes are correlated with each other. This method, furthermore, uses cross-entropy as its loss function, so a direct comparison is against DNN-BCE. In all of the metics, C2AE performed worse than DNN-BCE baseline. This is not to say that this approach is not good; rather, C2AE is more suited to fitting in bigger datasets where false class correlations are less of an issue.

Upon investigating deeper into the evaluation metrics, we further find that, indeed, infrequent classes caused degradation of performance. This is most evident in comparing macro F-1 between DNN-BCE and C2AE. The metric assigns arithmetic mean of F-1 values for all labels. Frequently appearing classes like 'temperament' and less frequent classes such as 'insurance coverage' are assigned the same weight when computing macro F-1. In this methodology, where frequent labels and infrequent labels have the same weight, there is a 39% drop in macro F-1 scores compared to the baseline (0.361 on DNN-BCE versus 0.220 on C2AE). On the other hand, while weighted F-1 score was lower on C2AE than on DNN-BCE, it represented only 15% decrease in F-1 scores (0.488 versus 0.412). In calculating weighted F-1, higher weights are assigned to frequently occurring classes, whereas infrequently occurring classes have lower weights which is reflected on the F-1 scores we see in C2AE.

Seeing the limitations in C2AE, where the method was sensitive to false corelations, we opted to utilize that between child-classes and meta-classes instead. We again compare the results with the binary cross-entropy baseline method (DNN-BCE) against three of our methods (HML-PC, HML-PN, and HML-PNC). The three methods measure different hypotheses. HML-PC captures our first hypothesis, where we claimed that child classes are a strict subset of meta-classes. Capturing the relationship between the two classes was, indeed, valid, because we were able to see improvement in classification performance. However, the improvements were relatively modest. This suggests that there are relationships outside of that between child-classes and meta-classes. As we see in the evaluation results, HML-PN, which leverages interaction between meta-classes and child-classes, has improved classification performance quite a bit more than HML-PC. Because of this, we surmise that there are different types of label interactions in play outside of child-parent relationships.

HML-PN captures the second hypothesis, where we claimed it is possible to correlate meta-classes and child-classes together. The classification performance improved quite significantly compared to the DNN-BCE baseline method. This indicates that, first, capturing label correlations helps in assigning topics from patient experience to patient reviews. HML-PN is designed in a way that it captures the relationship between meta-classes and child-classes which ultimately helped with the classification peformance. Second, in datasets with a few thousand instances of labeled data such as our patient review corpus, capturing the relationships between meta-classes and child-classes is better than capturing them between child-classes. We have noticed that label embedding method (C2AE) performed significantly worse than even the cross-entropy baseline (DNN-BCE), while HML-PN, which leveraged correlations between meta-classes and child-classes performed quite a bit better. Because some of the child-classes, appear very infrequently, capturing correlation became a very noisy task, hence the reason why we have seen C2AE underperform, whereas HML-PN had performed better.

Our last proposed method, HML-PNC combined both of the proposed methods and it performed better than the two of the methods that we proposed. This suggests that both HML-PN and HML-PC capture different aspects of the interaction between meta-classes and child-classes. Indeed, the two approaches were developed with different goals; HML-PC to capture the relationship between child-classes and meta-classes, and HML-PN to capture how child-classes and meta-classes interact. On the other hand, similar to the performance gains we see in HML-PN, HML-PNC showed only modest improvement over HML-PC. This suggests that interactions between meta-classes and child-classes are stronger than the innate hierarchical structure between the two types of classes, but to properly capture the patient review corpus, we should consider both label interactions and hierarchical structure.

5.6 DISCUSSION

We proposed utilizing parent-child relationships to better understand the patient experience taxonomy applied to the RateMD dataset. The taxonomy is inherently hierarchical where it had meta-classes and child-classes. We implemented two different approaches, one an explicit parent-child constraint and another capturing the relationship between parent and child classes, where the two classes were not in a hypernymic relationship with each other. Both improved various performance metrics over the baseline. The takeaway from these results is that incorporating meta-class allows additional information that is not readily available in the text embeddings themselves. This improved the classification results. We also see that encoding meta-class correlations to predict child-classes improved the performance the most, though enforcing the constraints between meta-class and child-class still improved the performance over the baseline. The combination of the two showed the highest improvement in the performance compared to the baseline. The improvement in performance indicates that the initial taxonomy we proposed is reasonably accurate, and there is an additional parent-child relationship that we can further leverage to enhance our model. In this work, while we were able to embed parent labels as part of the feature-set, we were unable to learn the correlations between input labels. This was because we opted to embed the predicted label into an embedding space, making it difficult to interpret. For future works, we would like to investigate a method to better learn correlations between the two labels. Furthermore, our analysis was limited to only the RateMD data. While this allowed us to re-verify the validity of the proposed taxonomy, it would be of interest to further run the classifier on other datasets.

CHAPTER 6: PATIENT SEGMENT IDENTIFICATION FOR COMPREHENSIVE UNDERSTANDING OF PATIENT EXPERIENCE TAXONOMY

To comprehensively understand patient experience, we first analyzed the semantic representation of each topic in Chapter 4. We then investigated how topics are interdependent of each other in Chapter 5. The methods described in the two chapters described how to automatically identify topics from the comprehensive patient experience taxonomy in patient-generated texts. However, not all patients' experiences are the same. Patient experience is dependent on patients' health statuses, for example, whether the patient was readmitted to the hospital, had an adverse drug reaction, or had a particular illness [1–4]. Identifying the underlying patient segment furthers our understanding of patient experience expressed in patient-generated texts.

We further note, however, that there are nearly infinite variations of patient segments that each person may belong to. Rather than develop models for all possible segments, we sample an example segment and show how the segment can be captured. Of particular interest is whether the patient has stopped taking medication or not. The segment is important because it is connected to a patient's adherence to a given treatment. Drug adherence is a big problem that causes over 300 billion dollars per year in estimated damage. We believe a technique which can identify whether the person is currently taking medication of interest or not can further be extended to enrich other types of patient segments. Segmenting patients allows for a more targeted patient experience analysis, allowing us to improve patients' satisfaction and in turn, leading to a better health outcome.

We answer how we have identified whether the patient has stopped taking the medication or not in this chapter. We first describe how we obtained and annotated training data to identify the segments. Next, we develop both informative and complex sets of features for use in understanding patients' drug usage. The technique to identify different patient segments is based on one of our published works [185].

6.1 INTRODUCTION

Medication non-adherence is a huge issue in the health community where up to about half the patients do not take medication as prescribed [186,187]. Not surprisingly, non-adherence is responsible for an estimated burden of 337 billions dollars per year in direct and indirect health care costs [188]. Analyzing why patients do not adhere to prescribed medications could have a huge impact on cutting health care costs in the long run.

In analyzing medical non-adherence, the first step is to understand why people have stopped taking a given medication, a subset of non-adherence. According to a previous study [189], there are four reasons why patients stop taking medication:

1. Experience of loss by the use of medication (adverse drug reaction)
2. Personal *meanings* associated with taking medication (taking anti-depressants means admitting one is depressed)
3. Different *feelings* evoked by the process of taking medication (emotions, such as disgust or humiliation at the thought of taking the medication)
4. Perceived changes in payoff matrix (efficacy of medication)

However, there are many limitations to this study. For example, the experiments were conducted in a controlled session where researchers asked participants in-depth questions. Due to the formal environment/setting in which the study was conducted, patients may not have expressed themselves as freely as they would have in an informal setting. Moreover, it is very difficult to scale up the research and generalize it to different demographics in such a formal environment.

A possible solution to this problem is to make use of online social media. This has been an active medium for various healthcare tasks such as epidemiologic studies which looked at drug adverse reactions [17], [190], [12]. Social media is also appealing due to its scale – i.e., it is possible to conduct studies on a large number of probems with imput from many people/patients around the world. Furthermore, most of the research seems to indicate that the findings in this medium are consistent with real world findings.

In this work, we tackle the issue of medication use. In particular, we want to analyze why people have stopped taking medication and, as a first step, we focus on identifying whether a person has stopped taking a given medication or not – a binary classification problem given a medication of interest. We note that this is not the only approach to determine how a medication may be used – for example, patients may indicate that they are currently taking the medication, or may simply be asking questions about a given medication. We consider these intents as an 'Others' category and focus primarily on identifying whether a person has stopped taking a given medication or not. We used user messages in health forums as appropriate data and genre because people tend to give more descriptive reasons as to why they may have stopped taking medication compared to micro-blogs despite potentials for bigger amounts of data. We have identified feature sets that were successful in similar tasks, and adapted these to suit our problem at hand. We then analyzed these features to see how different complex features perform with regard to the identification process.

Our contributions are as follows:

1. To the best of our knowledge, we are the first to tackle the problem of whether a person stopped taking medication or not in the online health domain.

2. We adapted existing features that were successful in previous health status tasks (we denote these as informative baselines features), and proposed new complex features for the problem domain. We further analyzed how much each of the different feature types contribute to the classification performance.

3. Our experiments also provide interesting insight not only into the language technology community, but also to health practitioners and the healthcare insurance industry. For example, when patients stop taking anti-depressants, they generally seems to have tried more than one medications, an experience that generates emotions such as disgust and sadness.

6.2 DATASET AND EXPERIMENTATION SETTINGS

In conducting our experiments, we first collected a dataset in which we labeled whether a patient had stopped taking a given medication or not. Unfortunately, we couldn't find any (available) dataset annotated with this kind of information, so we annotated the dataset ourselves. We first crawled a depression forum[1]. The rationale is that depression is one of the most prevalent conditions that affects a wide range of demographics [191]. By manual inspection of a few forum messages, we decided to focus on three popular depression drugs (Zoloft, Paxil and Cymbalta) as our target medication to annotate whether the patient had stopped taking the medication or not. For each medication, we chose an opening post which contained the medication of interest in a given thread and determined whether the person had stopped taking the medication or not. Notice that our task is to classify, for the 'Stop' label, whether the person has stopped taking the medication of interest. The 'Others' label refers to all the other situations, such as they are currently taking the medication, or wanting to learn more about the medication, or citing previous research conducted on the medication. Two annotators, one student in Linguistics and one in Natural Language Processing, labeled 300 forum posts for each medication, for a total of 900 labeled forum posts. These forum posts were randomly selected from the crawled dataset. Next, we calculated the agreements using Cohen's Kappa coefficient, shown here in Table 6.1. The data distribution is shown in Figure 6.1.

[1] www.healthboard.com

Drug	Zoloft	Paxil	Cymbalta
Agreement	0.81	0.75	0.85

Table 6.1: Cohen's Kappa coefficient for Inter-rater agreement for the three different medications considered

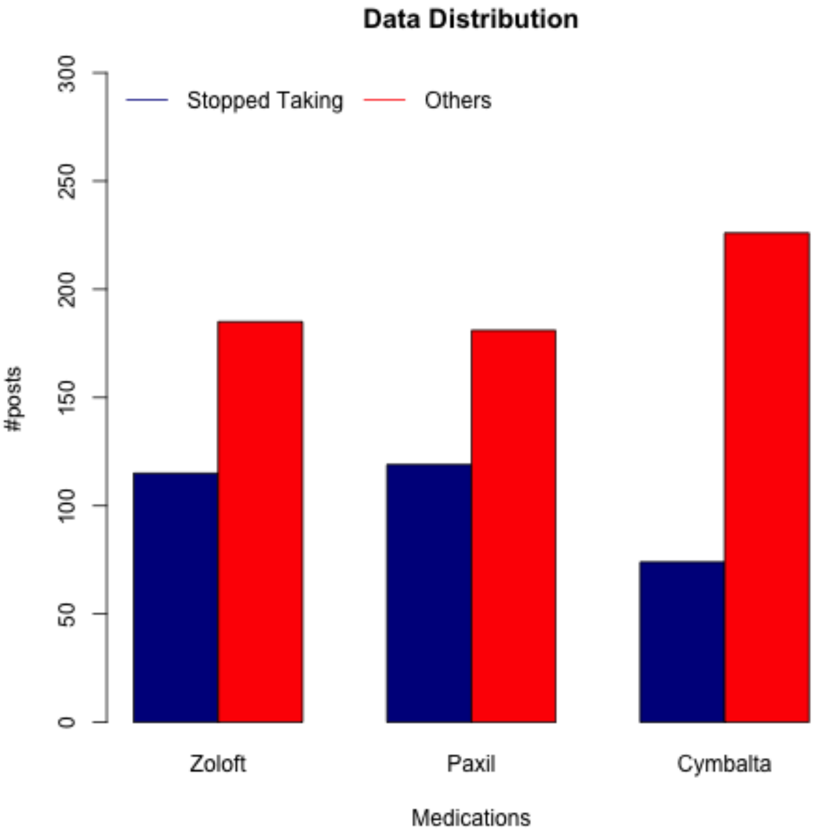


Figure 6.1: Number of labels for different medications

For any forum posts that the labelers disagreed on, there was an adjudication process to determine the final labels. Change of dosage was often a source of confusion, for example *I was on Zoloft 100mg and then double and then i even eat a crocus plant....* Confusions were especially evident when there were multiple mentions of changes of dosage. These were adjudicated after re-reading the forum posts and then agreeing upon the final label. A subset of these errors occurred when patients were weaning off medications. While the intents of these were often to stop taking the medication, unless it is clear from the text that they ultimately stopped taking the medication, these were generally labeled as 'Others'. Another source of confusion is when the forum post was lengthy as there was a lot of information the labelers had to absorb (they had to read the text multiple times).

For all of the experiments, we used Support Vector Machines[2] with 10-fold cross validation. We chose SVM because it is generally considered to be a good classifier on sparse datasets.

6.3 FEATURES

We experiment first with informative baseline features. For this, we identified features from the research literature on projects similar to ours. We then came up with additional features to address common errors the informative baseline features made. Unless otherwise annotated, we considered all our features equally likely, thus having a weight value of one.

6.3.1 Informative Baseline Features: Definitions and Examples

Keyword Pivoted Word Features: In the I2B2 challenge, using a set of good keywords was one of the most popular and effective approach [192, 193]. One of the authors proposed extracting sentences which contain a set of keywords, and ran the K-Nearest Neighbor model on these sentences [193]. Another group proposed using windows of text up to 100 characters instead [194]. Following the spirit of these works, we experimented by curating a list of keywords: *stop, quit, drop, try, wean, switch*. We select a sentence if it has one of the curated keywords and the medication we are interested in, and then extract all the words as our feature set, using TF-IDF to assign weights to words.

Odds Ratio Features: MacLean, et. al., 2015 [59] proposed using odd log ratio for extracting keywords on **post** level as opposed to **sentence** level. This was feasible because the authors were interested in inferring the status of one medication in a given forum post, as opposed to potentially inferring multiple instances of the medication which is what we focus on. Thus, we adapt the odds ratio in such that we only take sentences that either have the medication of interest, or are those that precede, or succeed the drug. The ratio is a measure of strength of association and is given by:

$$OR(t, s) = \frac{f_s(t) \cdot f_{\overline{s}}(\overline{t})}{f_s(\overline{t}) \cdot f_{\overline{s}}(t)} \tag{6.1}$$

where s is the state (such as Stop taking or Others), t is the term we are interested in calculating odds ratio, and $f_s(t)$ is the number of sentences with state s that contain t. $f_s(\overline{t})$ is the number of sentences that does not contain the term t for state s, and $f_{\overline{s}}(t)$

[2] We used Scikit-learn for our experiments. We ran validation on the cost parameter C to set the correct parameter, and set the parameter dual to false so that we can analyze the learned weights that we analyzed later in this section. We did not notice any significant differences in the performance results.

Informative Baseline Features

Feature type	Example sentence	Example features
Keyword pivoted features	[S]o i was flusterated, and just **stopped** taking my **paxil** for a few days.	TF-IDF weighted words that appears in this sentence
Odds ratio features	I quit the Cymbalta cold turkey and am now going through withdrawal.	quit, cymbalta, cold, turkey
Curated n-gram pattern features	I've been off the paxil for a while.	've/VBP off/IN MEDICATION
Time reatures	I was on Zoloft about 3 years ago	was/VBD on/IN MEDICATION TIME_WORD
Emotion features	Effexor gave me hives, Zoloft gave me horrible nightmares, and Wellbutrin gave me horrible hives after 2 1/2 weeks on it.	sensitivity: -0.51, attention: 0.21, aptitude: -0.52, pleasantness: -0.28

Complex Features

Feature type	Example sentence	Example features
Dependency relations features	I stopped taking the Cymbalta immediately and have now been off it 7 days.	... , (I, PRP) nsubj (stopped, VBD), (taking, VBG) xcomp (stopped, VBD), (Cymbalta, NNP) dobj (taking, VBG), (the, DT) det (Cymbalta, NNP), ...
Frequent sequential pattern features	He put me on effexor.	put on MEDICATION_TOK, me on MEDICATION_TOK
Co-reference resolution features	I've been clinically depressed for the past few months and Effexor hasn't helped me much. True, I've been taking only 75 mg of the **medication**...	Change 'medication' to 'Effexor' and apply all the other features

Table 6.2: Example Features used in our experiment

is for any other states that are not s, how many times the term t appears. Similar to a previous research work [59], we also retain odds ratios which are greater than 2, and set the ratio as the feature weight for the odds ratio word. Furthermore, we do not include words that appeared less than 10 times in our corpus. We compare the results of calculating the odds ratio on a post level versus sentence level in our experiments, and indeed, we see that sentence level odds ratio improves over that of post level.

Curated N-gram Pattern Features: Using carefully curated rules was shown to be effective in detecting whether patients stopped smoking or not in medical notes [195]. This kind of work in particular proposed a set of seven rules to indicate whether a person may be a smoker or not. We adapt this approach and propose an automated rule generation framework which does not involve the stop keywords. In particular, we noticed that when a patient starts a sentence with a verb in past tense, followed by a preposition and a medication, it is likely to be an indication that the patient is not currently taking the medication. This may or may not be followed by a duration/temporal keyword. As an example, in the sentence *I was on Zoloft for 3 months*, we see that the past tense verb (was) is followed by a preposition and a medication name. Similarly, when a sentence starts with a present tense verb, it may be an indication that the patient is currently taking the medication, for example, *I am currently taking Zoloft* (adverbs are important here as well). We further note that patients may mention that they have switched from medication A to medication B. We add these as a feature set as well. More formally, we encode these rules as

[VBD/VBP] [PREP] [MED]

[VBD/VBN] [FROM] [MED] [TO] [MED]

where VBD is past tense verb, VBP is a present tense verb and VBN is verb participle. PREP and MED correspond to preposition and medication, respectively. We also replaced pronouns with the medication name if only one medication was mentioned in the previous sentence (i.e., we resolved the coreference). We further allowed the pattern words to appear few words away from each other, i.e., a PREP may appear 3 words after a VBD or a VBP.

Time Features: Mentions of time expressions has also been a very popular approach in previous research, either as part of a carefully orchestrated set of rules [195] or by utilizing a system that can distinguish expressions that indicate current time, recent past, and distant past [196]. One previous work [195], for example, used time tokens as part of the carefully orchestrated rules to improve the performance of the health status detection. Motivated by these lines of work, we used the rule generation framework from the previous section, and added temporal words at the end. This framework would capture the sentence *I was on*

Dimension	Sensitivity	Attention	Pleasantness	Aptitude
Positive	Anger	Anticipation	Happiness	Trust
Negative	Fear	Surprise	Sadness	Disgust

Table 6.3: Emotions used as features

Method	**Precision**	**Recall**	**F-1**	**Accuracy**
Passage OR	0.540	0.523	0.528	0.680
Sentence OR	0.641	0.562	0.593	0.737

Table 6.4: Performance of odds ratios

Zoloft for 3 months as having a temporal word since the word months appeared after the rule *[VBD/VBP] [PREP] [MED]*. Examples of temporal words are: *currently, now, day, month, week, year.*

Emotion Features: Finally, we experimented with emotion features as well. One of the reasons for which people stop taking medication is the feeling that may be evoked by the act of taking the medication [189]. For instance, if they do not like the idea of taking the medication itself (hatred), or the medication reminds them of some other events which make them feel sad, patients may stop taking the medication. Thus, we used SenticNet [197] which makes use of the idea of hourglass of emotions in modeling the emotions of interest here. The work organizes emotions around four independent, yet concomitant dimensions (sensitivity, attention, pleasantness, aptitude), whose different levels of activation make up the total emotional state of mind. Each of the dimensions may have positive or negative polarity which maps to different set of emotions. As an example, a positive pleasantness may indicate happiness, whereas a negative one may indicate sadness. Table 6.3 shows all 8 emotions we used from SenticNet.. If a sentence contained the medication of interest, we used this emotion ontology and then normalized by the log of the number of words in the sentence. This gave us polarity scores for each of the four dimensions. All of the features described in this section are summarized in Table 6.2.

6.3.2 Informative Baseline Features: Experiments and Analysis

We were first curious about whether it would be beneficial to use words taken from the sentence level or from the passage level. Sentence level odds ratios were calculated using words taken only from those that contain the medication of interest, whereas post level ratios made use of the entire post. Our results are shown in Table 6.4.

We see that sentence level odds ratio outperforms the passage level one. This makes

Sentence OR	stopped, attacks, cold, point, upset, lbs, thinking, saw, made, turkey
Passage OR	baby, become, dad, yrs, mom, failed, per, helping, met, depressants, man

Table 6.5: Top 10 highest ranked odds ratio words

intuitive sense – in a given post, patients are likely to talk about myriads of topics and limiting the odds ratio calculation to the sentence which contains the medication of interest allows us to utilize only the words that are relevant to the medication itself. We show the top 10 highest ranked words for both sentence and passage odds ratios in Table 6.5. In the case of sentence odds ratio, many of the words make intuitive sense. For example, 'cold turkey' is an often used expression when a patient stops taking medication suddenly. We see both 'cold' and 'turkey' in the top 10 highest ranked odds ratios. On the other hand, passage odds ratios do not quite indicate whether a person has stopped taking medication or not. It does, however, show that someone had something to do with medication ('baby', 'dad', 'mom', 'man' and 'depressant'). This can be because in generating our dataset, we chose forum posts which contain the medication of interest.

Next, we were curious to see how much each additional different feature contributed to the performance of the task. Based on the previous results from experimenting with different types of odds ratio, we opted for sentence level odds ratio as the baseline method. Our results are shown in Table 6.6 where we experiment by taking one feature out from all the feature sets that we have described in the previous section.

Other than the odds ratio features which we used as the baseline, curated n-gram pattern features seemed to have the biggest impact on the performance. This indicates that curated n-gram pattern features correctly captured many of the relevant phrases. Keywords played a big factor as well. Along with odds ratio, keyword based TF-IDF features captured the context in which the medication word appeared. Emotions also had some impact in the classification task. This seems to reiterate one of the hypotheses in our introduction where patients stop taking medication because of the feeling that it evokes. Finally, we notice from our evaluations that by combining all of our proposed features we get the best performance.

We next conducted error analysis on false positives and false negatives obtained with the best performing informative baseline and found the following common errors:

Lack of Co-reference Resolution: Users often mention the medication and then indicate they have stopped taking it in the sentences that follow. Consider the following snippet: *Hey guys I took Cymbalta for about three weeks and other than feeling good for the first week I felt like I didn't feel any better. Anyway I stopped it without tapering about 4 days ago and*

Method	Precision	Recall	F-1	Accuracy
Odds Ratio (Baseline)	0.641	0.562	0.593	0.737
Informative Baseline	**0.673**	**0.640**	**0.650**	**0.762**
Informative Feature - o	0.656	0.403	0.491	0.717
Informative Baseline - e	0.639	0.624	0.625	0.746
Informative Baseline - r	0.621	0.589	0.601	0.734
Informative Baseline - k	0.639	0.608	0.620	0.745
Informative Baseline - t	0.633	0.623	0.625	0.745

Table 6.6: Performance of different features. *o*: Odds ratio features, *e*: Emotion features, *r*: Curated n-gram pattern features, *t*: Time features, *k*: keyword-based TF-IDF features

I have been itching all over since then. We see that the pronoun 'it' refers to 'Cymbalta.' However, this has to be solved automatically.

Lack of Dependency Relations: Dependency parsers help us in understanding constraints that may not be straightforward to encode. As an example, the verb 'taken' refers to all the medications in the following sentence: *I have taken Lexapro, Prozac, Zoloft and now Celexa.* Similarly, we have a sentence *Needless to say, I stopped taking the Cymbalta immediately and have now been off it 7 days.* where the word 'stop' is a clausal complement of the verb 'taking.' These are not captured by the informative baseline features, but may be captured by more advanced feature sets.

Unaccounted Frequent Text Patterns: There were various patterns that seemed to appear frequently that were not captured by our basic informative feature sets. As an example, *I was on zoloft and switched over to paxil for about a year.* Here 'was on Zoloft' is a frequently occurring pattern that we did explicitly capture. However, the complex features we introduce in the next section do capture this type of phrases.

6.3.3 Complex Features: Definitions and Examples

Based on the error analysis we have conducted, we propose here three new complex features and use them to further improve the performance of our classifier as well as our understanding of this challenging task.

Dependency Relation Features: Based on the error analysis we have conducted in the previous section, we saw that dependencies between words have a big impact on the classification task. Based on this intuition, we employed the Stanford dependency parser [198] which gives the grammatical relationship between a head word and its modifier word. In particular, this allows us to capture the relationship between a medication word and any other words that may be describing it. Capturing multi-word phrases becomes possible as well, for instance, ‘Stopped taking’ in our previous example which is given a clausal complement label.

Frequent Sequential Pattern Feature: We further noticed that sentences where patients described the times when they stopped taking medications, have similar structures. In order to capture this, we utilized a frequent sequential pattern mining algorithm [199]. The benefit of capturing frequent sequential patterns as opposed to n-grams is that the algorithm can capture words that are farther apart from each other. As an example, *He put me on effexor.* cannot be captured by a simple bigram (put on) since there is a pronoun that appears in between the two words, whereas this can be captured by frequent sequential patterns (as *put on medication*). In running frequent sequential patterns, we tokenized all occurrences of medication as either the current medication of interest (as ‘MEDICATION_TOK’), or other medications that are present in the post, but not something that we are interested in classifying whether the patient has stopped taking or not (as ‘OTHER_MEDICATION_TOK’), and set the weight as the log of the number of occurrences the pattern had in our corpus.

Co-reference Resolution Features: In many sentences, people would mention a medication and indicate if they stopped taking it the sentence after. Despite the usage of an existing off-the-shelf co-reference resolution toolkit [200], it quickly became apparent that the out-of-the-box features were not accurate in deciphering the coreferences. Thus, a heuristics method was used to resolve the problem. Whenever we saw words that may be indicative of medications (such as ’medication,’ ’med,’ and ’anti-depressant’ as well as ’it’) we replaced this with the last medication that we have seen within a two sentence window. We then could apply all the other features to this recovered co-reference as well.

Summaries of the features described in this section are shown in Table 6.2.

6.3.4 Complex Features: Experiments and Analysis

We next experimented with complex features as well. We set the best performing combination of features from informative baseline features as the starting point of this experiment. Our results are shown on Table 6.7.

We can see the complex features we proposed improved both precision and F-1 score. We have further noticed that frequent sequential pattern features had significant boost in improving the performance. Furthermore, the new features we proposed had improved primarily on precision with small reduction on recall. It is also interesting to note that adding co-reference features, by itself, does not improve performance compared to the informative baseline. However, we see an improvement when the co-reference feature is combined with other features.

Method	**Precision**	**Recall**	**F-1**	**Accuracy**
Informative Baseline	0.673	0.640	0.650	0.762
Informative Baseline + d	0.702	0.616	0.652	0.775
Informative Baseline + f	0.768	0.633	0.693	0.809
Informative Baseline + c	0.661	0.634	0.643	0.759
Informative Baseline + d + f	0.756	0.627	0.684	0.802
Informative Baseline + c + f	0.771	0.638	0.695	0.810
Informative Baseline + c + d	0.700	0.606	0.646	0.772
Informative Baseline + d + f + c	**0.786**	**0.636**	**0.701**	**0.817**

Table 6.7: Performance of complex features. *d*: dependency parser features, *f*: frequent sequential pattern features, *c*: co-reference features

There were also cases where the complex features did not perform very well. One common error was when the forum post had sequences of events. As an example, *I got on **Paxil** 2 years ago and switched to Wellbutrin, then I was pregnant and switched to Prozac. I got back on medication, **Paxil**,two months ago.* requires a temporal representation of taking/stop taking the medication, for a given user. A simple 'stop taking' label is difficult because the patient did stop taking the medication at one point, though now the person is taking the medication. While there is no previous research on exactly the problem we tackled here, there is some body of research on generating timelines [201, 202] which we will be able to leverage in future research to mitigate this problem.

Another common error was when the act of stop taking the medication was not explicitly mentioned. From the sentence, *He's been on **Zoloft** but he said that he walked around in a daze all the time and could barely function at work (his job requires that he is alert and ready to respond to problems constantly.) Now he's trying a different med (he takes it sporadically - can't remember the name), and it also makes him tired and like he's in a fog.* We see that the target person has now stopped taking the medication. However, this is inferred from what is written, and one way to classify this is by knowing that by taking a different medication, the person is no longer taking the other medication – i.e., deeper semantic inference. However, there are linguistic devices that will tell us if a person takes multiple pills at the same time, or has switched medication. We will explore such contexts in future work.

6.4 ANALYSIS OF RESULTS

In this section, we analyzed the coefficients from our trained SVM. We have trained SVM using its primal form so that interpreting weights would be easier than if we had used the dual form. The differences in performance were negligible.

Curated n-gram pattern features were a strong indicator, where we found that a lot of patterns which followed *[past tense verb] on medication*, and *[present tense verb] off medication* were associated with 'Stop taking' patterns. Some other patterns that do not involve the prepositions/particles 'on' or 'off' were: 'tried with medication,' 'started with medication,' 'told that medication,' and 'did like medication.' Furthermore, time related features did have a small impact as well, though not as significant as that of the rules. The top 10 important curated n-gram pattern features we discovered are shown in Table 6.8.

Stop Taking	**Others**
was/VBD on/IN MEDICATION TEMPORAL_WORD	have/VBP on/IN MEDICATION TEMPORAL_WORD
have/VBP off/RP MEDICATION TEMPORAL_WORD	've/VBP on/IN MEDICATION TEMPORAL_WORD
was/VBD on/IN MEDICATION	am/VBP on/IN MEDICATION
had/VBD on/IN MEDICATION TEMPORAL_WORD	put/VBD on/IN MEDICATION MEDICATION
have/VBP off/IN MEDICATION TEMPORAL_WORD	have/VBP on/IN MEDICATION
've/VBP off/IN MEDICATION	put/VBD on/IN MEDICATION
got/VBD on/IN MEDICATION MEDICATION	'm/VBP on/IN MEDICATION MEDICATION
am/VBP off/RP MEDICATION TEMPORAL_WORD	started/VBD on/IN MEDICATION TEMPORAL_WORD
've/VBP off/RP MEDICATION TEMPORAL_WORD	'm/VBP on/IN MEDICATION
was/VBD like/IN MEDICATION	got/VBD on/IN MEDICATION TEMPORAL_WORD

Table 6.8: Top 10 important curated n-gram pattern features found by the classifier

As for the Emotion features, 'Stop taking' label had strong inverse correlation among aptitude (0.011), pleasantness (0.00684), attention (0.00526) and sensitivity (0.00291) emotional aspects according to SenticNet. These correspond to disgust, sadness, surprise and fear when patients had stopped taking medications. We note that because of the emotion model we used, there are four different dimensions, and hence, four different emotions for each label. These seem to indicate that when patients stop taking medications, there are, indeed, correlations related to emotions, in particular, that of disgust and sadness.

Frequent sequential patterns dominated most of the highest scoring weights. The algorithm is able to capture many of the frequent patterns that occur. However, they may not make immediate semantic sense for the human analyst. In 'Stop taking' labels, past tense

Stop Taking	**Others**
i MEDICATION_TOK OTHER_MEDICATION_TOK	i taking MEDICATION_TOK
i on OTHER_MEDICATION_TOK	i am MEDICATION_TOK
i MEDICATION_TOK again	been MEDICATION_TOK for
i MEDICATION_TOK off	i of MEDICATION_TOK
i off MEDICATION_TOK	i like MEDICATION_TOK
i and OTHER_MEDICATION_TOK	i just MEDICATION_TOK
off MEDICATION_TOK a	i MEDICATION_TOK mg
i been off	my me MEDICATION_TOK
i on MEDICATION_TOK but	for and MEDICATION_TOK
tried MEDICATION_TOK and	MEDICATION_TOK and if

Table 6.9: 10 most important frequent sequential features the classifier used

verbs ('tried', 'been', and 'was'), and some prepositions/particles ('off') were commonly extracted as part of the patterns, whereas for Others label, present tense verbs ('taking', 'am') and other type of prepositions/particles ('on') were common. For both labels, the current medication of interest (such as Zoloft, Paxil or Cymbalta), annotated as MEDICATION_TOK were usually present in frequent patterns. These findings are consistent with what we have found in curated n-gram pattern features – for example, usefulness of prepositions/particles such as 'off' and 'on,' and phrases that start with 'tried' or 'start.' There were instances where 'MEDICATION_TOK' were not present, but the algorithm still thought it was a frequent pattern. For instance the pattern 'i been off' was considered to be an important feature for the 'Stop taking' label. While this may be a source of ambiguity (in the given frequent pattern, 'i been off,' there is no indication of which medication the author has been on), since a given sentence may be captured by more than one frequent sequential pattern, the fact that some frequent features did not have 'MEDICATION_TOK' is not a problem.

Another interesting observation is that the algorithm was able to find medications that are not the current medication of interest (annotated as 'OTHER_MEDICATION_TOK') to be useful as well, in particular for the Stop_ taking label. For example, we found a very frequent pattern, *I MEDICATION_TOK OTHER_MEDICATION_TOK* (used in the context of, ***I*** *tried* ***Zoloft***, ***Paxil***, *and Cymbalta* if the medication of interest was 'Zoloft.') to be indicative of stop taking medication. This indicates that often times patients mention other medications when they indicate that they have stopped taking a given medication, often times as an enumeration, or that they are now currently taking other medication. The top 10 most important features used by the classifier on both target classes are shown in

Stop Taking	**Others**
longer/RB - neg - no/RB	taking/VBG - dobj - MEDICATION_TOK/NNP
taken/VBN - nsubj - i/PRP	day/NN - det - a/DT
stopped/VBD - xcomp - working/VBG	take/VBP - nsubj - i/PRP
taking/VBG - aux - was/VBD	wondering/VBG - nsubj - i/PRP
had/VBD - neg - never/RB	taking/VBG - dobj - cymbalta/NNP
stopped/VBD - nsubj - i/PRP	have/VBP - nsubj - i/PRP
years/NNS - amod - few/JJ	started/VBD - advmod - just/RB
was/VBD - nsubj - i - P/RP	ago/RB - npadvmod - months/NNS
tried/VBN - nsubj - i/PRP	tell/VB - aux - can/MD
had/VBD - nsubj - i/PRP	'm/VBP - nsubj - i/PRP

Table 6.10: Top 10 most important dependency features found by the classifier

Table 6.9.

We used a dependency parser to capture the relationship between words in a sentence with the hope that it would capture relevant relationships between words. Again, our top 10 most important features are shown in Table 6.10. We discovered some verbs that describe medications, such as tapering/VBG - prep - MEDICATION_TOK/NN, and phrases that indicate (but do not necessarily explicitly mention) stop taking the medications, such as withdrawal/NN - nn - cymbalta/NNP to be associated with the Stop_taking label. On the other hand, dependency relations such as taking/VBG - dobj - MEDICATION_TOK/NNP or prescribed/VBN - dobj - MEDICATION_TOK/NN were associated with 'Others' label. Some of the patterns we found here were also present in frequent sequential patterns. For instance, the algorithm found MEDICATION_TOK/NNP - conj - OTHER_MEDICATION_TOK/NNP to be a useful pattern which is what we have also found via frequent sequential patterns. However, as was shown in our evaluation results, there were enough patterns that the frequent sequential patterns did not capture but were captured by the dependency parser.

We analyzed keyword-based features, and focused our attention on medications that were correlated with 'Stop taking' label or 'Others' label. Other than the medications that we focused on (Paxil, Zoloft and Cymbalta), we found that Lexapro, Prozac, Wellbutrin, Effexor, Celexa, Lamictal, Remeron, Abilify, and Pristiq were associated with 'Stop taking' labels as well. This seems to indicate again that patients often mention multiple medications when they indicate that they have stopped taking a given medication. On the other hand, the medications Seroquel and Buspar had weak negative correlations with the 'Stop taking' label. Seroquel is primarily used as an antipsychotic, and often used in conjunction with other medications to treat depression, and Buspar is often used to treat anxiety. Neither of the medications were primarily used as anti-depressants, which may explain why they had weak negative correlations to the ‘Stop taking’ label. However, we leave these issues for a future in-depth analysis of the complex problem of stop taking medications in social media forums.

6.5 DISCUSSION

In this chapter we looked at different features of various compelxity to learn more about the problem domain. We started out by taking odds ratio of the words that appear in our corpus (i.e., ‘baseline’). We then proposed adding more features motivated by existing approaches that seemed to perform well in similar problem domains and tasks (i.e., ‘informed baseline’). After conducting an error analysis, we went on to propose complex features meant to capture some of the errors that the existing features were unable to capture. These features significangly improved the performance.

For future work, we have several directions. First, we would like to investigate on how to encode the current state of the medication. Some of the common errors that the classifier made were in cases where there were multiple instances of the medication of interest. By constructing these into a timeline, we would be able to capture what the current state of the medication prescription may be. Furthermore, both our keyword-based features and curated n-gram pattern features, while effective at generating relevant phrases, would benefit from a more automated method to increase coverage. We leave this as future work to further increase recall. Another direction, which is an extension of the previous direction in encoding the current status of the medication, is to add a new label, ‘currently taking medication’ as well as constraints to help us classify whether the person has stopped taking the medication or not. We noticed that a person is not likely to take more than one or two medications at a same time, especially if it pertains to the same class of medication (although this is possible). By studying this constraint, we would be able to better classify whether the person has stopped taking the medication or not. Finally, we would like to investigate the **reasons** for which a patient has stopped taking the medication. This would provide useful information to researchers working on healthcare problems or medical doctors on the type of medication the patient should be taking.

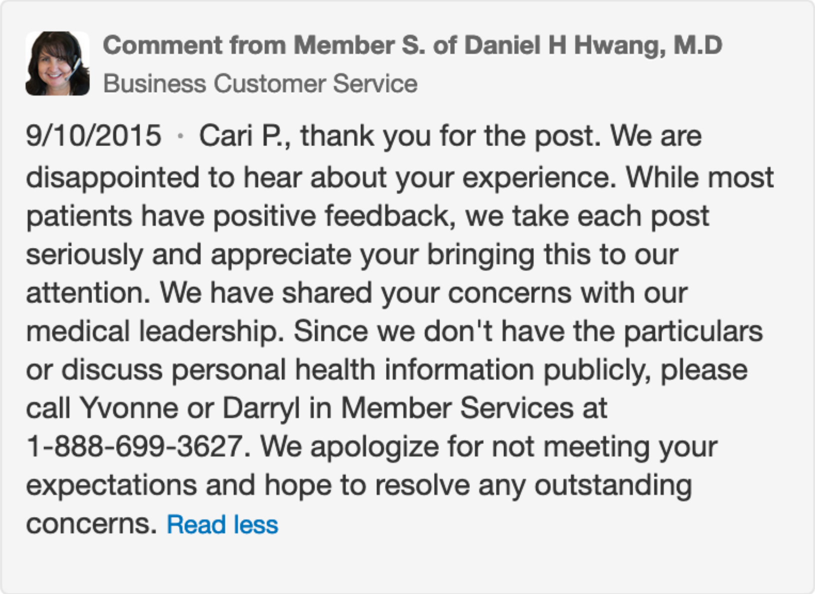
CHAPTER 7: APPLICATIONS OF PATIENT EXPERIENCE TAXONOMY

With a way to identify patient experience taxonomy from patient-generated texts, we explore how we can utilize it to improve the patient experience. An immediate, and perhaps, an obvious application is in the research community, where, similar to work described in Chapter 4, we can analyze each of the topics in more depth. Depending on the type of the community, we further deepen linguistics knowledge, or what patient experience is like for a specific patient segmentation in the case of the clinical research community.

A less obvious application is how we utilize the taxonomy to aid healthcare providers. We find inspiration from online review websites, which have gained popularity in the recent years. There are numerous similar websites where patients can visit and either read or write patient authored reviews. There are domain specific websites such as RateMD(`www.ratemd.com`), WebMD (`www.webmd.com`), Health Grades (`www.healthgrades.com`), and ZocDoc (`www.zocdoc.com`), and general purpose websites such as Yelp (`www.yelp.com`), Google (`www.google.com`) and Facebook (`www.facebook.com`). In each of the review websites, patients can write about their experiences, and understanding what patients write on these websites help healthcare providers identify how they can improve their service. Furthermore, in some review websites, business owners can respond to reviews, as in the example shown in Figure 7.1. These media provide venues for business owners to justify why a particular action was done, which may mitigate misunderstanding between the owner and the customer. Indeed, according to a marketing article [203], responsive business owners are more likely to attract more customers, and it is no different for healthcare provider as well [64, 124, 204].

However, there are too many websites that healthcare providers need to keep track of and respond to. What complicates the issue is that patients may write about numerous different topics, some of which may not be of interest to a given staff member. As an example, administrative staff may be interested in improving financing and scheduling, while nurses may be interested in accurate testing or empathy. A physician may be more interested in his or her communication skills, in ensuring that patients understand their ailments and treatment. Because of the diverse set of topics that the staff members are interested in, they need not be exposed to all the reviews, instead, only to the ones that can help them improve the quality of care. A method to identify which post should be answered by staff members would help save time, not only allowing the providers to reconnect with patients, but also letting them spend more time improving service.

An automated way of routing relevant posts or reviews to the corresponding expert will save time for the staff members, and improve their understanding of patient experience



Comment from Member S. of Daniel H Hwang, M.D
Business Customer Service

9/10/2015 · Cari P., thank you for the post. We are disappointed to hear about your experience. While most patients have positive feedback, we take each post seriously and appreciate your bringing this to our attention. We have shared your concerns with our medical leadership. Since we don't have the particulars or discuss personal health information publicly, please call Yvonne or Darryl in Member Services at 1-888-699-3627. We apologize for not meeting your expectations and hope to resolve any outstanding concerns.

Figure 7.1: Example of a hospital staff responding to patient's review.

directly from their patients, leading to improvement in healthcare service quality. We implemented a system that captures topics of interest for the expert. Furthermore, since we want to minimize the number of patient-generated texts they need to read through, we enforce a soft-constraint which assigns one expert per text. We describe how we route relevant patient-generated texts to designated experts in more detail in this chapter. While the experiment we conducted are not on patient experience taxonomy, we believe the approach can easily be translated to route texts to designated experts. We answer this research question based on one of our published works [205].

7.1 INTRODUCTION

Recommender systems are developed to help many different types of users. Many of the existing systems focus on helping average users [101, 206] while others help experienced users [89, 100, 207]. Our work focuses on a specific type of user called *designated experts*. These users are defined as those who have significant domain knowledge and are formally designated by the forum administrators to aid average users. For instance, medical doctors are designated experts on medical help websites such as MedHelp[1] , while instructors and course staffs are designated experts on Massive Online Open Course (MOOC) sites like

[1] http://www.medhelp.org

Coursera. These users, performing many roles, including answering questions posted by average users, are tasked to answer questions average users post on question-answer forums.

Designated experts should respond to many people at once. In the case of medical forums, medical experts have to answer questions submitted by users. For example, MedHelp, which is one of the leading medical websites, has over 12 million users every month. However, there are less than 300 designated medical experts who respond to patients' questions.

Furthermore, only a small portion of forum posts are answered by designated medical experts. These are people who have medical expertise that are certified by the website that hosts health forums. One work analyzed the role of designated medical experts on online health forums [208] The results show that designated medical experts play a critical role in responding to patients' information needs. In particular, 62.1% of online forum posts may benefit from clinical expertise. However, only 4.7% had responses from medical experts.

MOOCs similarly suffer from imbalance between the number of designated experts and users. Coursera has 5.2 million students and 532 courses[2], which corresponds to an average of approximately $10,000$ students per course. By building an effective recommender systems for designated experts, it would be possible to mitigate such imbalance problems, in particular by minimizing the time experts spend on looking for questions to answer. The system should utilize current approaches used by available recommender systems as well as exploit the unique behavior that designated experts might have.

With this goal in mind, we analyzed the behavior of designated experts. We analyzed 160 thousand threads from the 'Ask a Doctor' forum (which consists of more than 60 subforums related to health) from MedHelp. This forum consists of designated experts (medical doctors) who answer average users' (patients') questions. Other non-designated users can also respond to the patients' questions. We found that less than 1% of threads had more than one expert who responded to a patient's query. This is a unique behavior of designated experts which does not seem to exist on CQA websites. We believe this is because the designated experts are very good at giving good answers to question askers. Other bystander experts may feel it is unnecessary to give an addendum to these answers. This contrasts to the behavior seen on 'Medical Support Communities' from MedHelp. These communities consist of average patients who seek support from fellow users. About 30% of the threads on these communities have more than one user who responded to the patient seeking help.

In this chapter we focus on this unique behavior designated experts have. In particular, we propose to capture it by using matrix factorization with a regularization framework and show that, by capturing how designated experts behave, the retrieval performance of the

[2] http://blog.coursera.org/post/64907189712/a-triple-milestone-107-partners-532-courses-5-2

recommendation system improves.

Our contributions are as follows:

1. To the best of our knowledge, our work is the first to address the problem of recommending forum threads to designated experts. Although some previous research has focused on recommending questions and answers to users in CQA websites [92, 99, 100, 103, 207, 209], we were unable to find other works that address recommending forum threads to designated experts. Other related research has focused on e-mail routing commercial software such as IM an Expert[3]. However, these allow users to provide expertise and take place in a corporate environment, while our work focuses on online sources.

2. We propose a collective matrix factorization framework [105] to incorporate 1) forum posts, 2) user profiles and 3) user similarity.

3. We propose a matrix regularization framework [210] to capture the behavior of designated experts. In particular, we note that only one expert is likely to answer a given forum post. We propose a model that captures such characteristics.

The rest part of this chapter is organized as follows: Our approach is described in detail in Section 7.2. Section 7.3 describes and shows results of our experiments. In particular, we first analyze the impact of each semantic type we introduce in this chapter, then the impact of the parameters. We then summarize and discuss our findings in Section 7.4.

7.2 OUR APPROACH

7.2.1 Base Formulation

Our formulation is defined using low-rank matrix formulation with implicit user feedback. In low-rank matrix factorization formulation, the goal is to approximate the feedback matrix R with two low-rank matrices. The general formulation is given as follows:

$$\mathcal{L}_{base}(U, P) \;=\; ||C \odot (R - U \cdot P^\top)||_F \tag{7.1}$$

where $U \in \mathbb{R}^{n \times k}$ and $P \in \mathbb{R}^{m \times k}$ are low-rank matrices that corresponds to experts and posts, respectively, and $|| \cdot ||_F$ is Frobenius norm. The element-wise product is represented by $\odot$. The matrix C is relative importance of a given entry.

[3] http://research.microsoft.com/en-us/projects/imanexpert/

The feedback matrix R is captured by using binary user feedback, in particular, by determining whether an expert has responded to the given post or not. More formally, if a user i has responded to a j-th thread, then $R_{ij} = 1$, 0 otherwise. Furthermore, not all responses are equal. A user may not have responded to a particular thread because he was not interested in the thread, or because he has not seen it. We utilize a well known weighting metric [211] to mitigate the problem. The method proposes to frame the problem as weighted user feedback. The weight C is defined as follows:

$$C_{ij} = \begin{cases} 1, & \text{if user } i \text{ responded on } j\text{-th thread} \\ 1 - \text{sim}(\vec{U_i}, \vec{P_j}), & \text{otherwise} \end{cases} \tag{7.2}$$

The vectors $\vec{U_i}$ and $\vec{P_j}$ are row vectors of matrices U and P, respectively. Our similarity metric $\text{sim}(\vec{U_i}, \vec{P_j})$ is cosine similarity between the word vector of the i-th user used and that of the j-th thread.

7.2.2 Modeling Post Content

Pure collaborative filtering approaches are not sufficient to handle the semantics. Past literature has suggested incorporating the context of the items [212,213] or user profiles [214] to improve performance. We follow a similar formulation and model document content into the matrix factorization framework. Our formulation is as follows:

$$\mathcal{L}_{posts}(P, W) \;=\; ||D - P \cdot W^{\top}||_F \tag{7.3}$$

where $P \in \mathbb{R}^{m \times k}$ and $W \in \mathbb{R}^{|\vec{w}| \times k}$ are low-rank matrices for posts and words, respectively. The matrix $D \in \mathbb{R}^{m \times |\vec{w}|}$ is modeled from $TF\text{-}IDF$ weighting across the documents.

Modeling post content is an improvement over the base model as it explicitly states the document formulation. However, as will be evident in the experiments section, this approach does not model experts' interests.

7.2.3 Modeling Expert Interest

Similar to what we have done to model post contents, we wish to capture experts' interest as well. The challenge here is the user's profile page is often blank, or near blank. Instead of modeling these profile pages, we leverage the threads that the expert has responded to, and use these as the feature set. Our formulation is given as follows:

$$\mathcal{L}_{users}(U, W) \;=\; ||E - U \cdot W^{\top}||_F \tag{7.4}$$

where $U \in \mathbb{R}^{n \times k}$ and $W \in \mathbb{R}^{|\vec{w}| \times k}$ are low-rank matrices for experts and words, respectively. Similar to the document matrix D representation, the matrix $E \in \mathbb{R}^{n \times |\vec{w}|}$ is modeled using $TF\text{-}IDF$ weighting.

7.2.4 Modeling Expert Similarities

We cannot rely on interaction between two experts in our problem setting since in most cases, only one expert responds to a given thread. More precisely, since only one expert is likely to answer a given forum post, we cannot rely only on collaborative filtering to recommend posts for them to answer. Instead, we measure the similarities between the two experts based on the threads that they have responded to. If two experts are similar, they would be interested in similar forum threads, and vice versa for two experts that are not similar to each other. These are called social regularization framework in the context of social collaborative filtering [215, 216]. Our formulation to capture this intuition is given as

$$\mathcal{L}_{sim}(U, V) \;=\; ||S - U \cdot V^{\top}||_F \tag{7.5}$$

where $U \in \mathbb{R}^{n \times k}$, $V \in \mathbb{R}^{n \times k}$ are expert low-rank matrices and their weights, respectively. $S \in \mathbb{R}^{n \times n}$ is a user similarity matrix.

Similarity matrix S is modeled using cosine similarity [217] between the two experts. For a given user i and i', we have $S_{i,i'} = \cos(\vec{E_i}, \vec{E_{i'}})$, where $S_{i,i'} \in S$, $\vec{E_i}$ is the word vector for the user i, and $\cos(\cdot, \cdot)$ is the cosine similarity between the two vectors.

7.2.5 Modeling Designated Expert Constraints

One response per post.

We have noted only one expert is likely to respond to a given forum post. This formulation is written as follows:

$$\sum_i R_{i,j} = 1, \; \forall j \in \{1, \ldots, m\} \tag{7.6}$$

This equation means some expert would answer the post j. Based on our formulation, $R_{i,j} \sim \vec{U_i} \cdot \vec{P_j}$, we now have, for each posts P_j

$$\sum_i \vec{U_i} \cdot \vec{P_j} = 1,\ \vec{U_i} \cdot \vec{P_j} = \{0, 1\},\ \forall j \in \{1, \ldots, m\} \tag{7.7}$$

as the constraints.

We note this is an NP-complete problem [218] since it is an instance of 0-1 integer linear programming and is not easily solvable. Instead, we relax the binary optimization problem to an optimization problem as the following

$$\sum_i \vec{U_i} \cdot \vec{P_j} = 1,\ \forall j \in \{1, \ldots, m\}, \vec{U_i}, \vec{P_j} \in \mathbb{R}^k \tag{7.8}$$

This can be reformulated, by taking its square, as

$$\mathcal{C}_{1j}(U, P) = \left(\sum_i \vec{U_i} \cdot \vec{P_j} - 1 \right)^2 = 0,\ \forall j \in \{1, \ldots, m\} \tag{7.9}$$

Propensity to answer.

Each experts has a different propensity to answer questions. Some may be prolific at answering questions, while others may not be. To capture this intuition, we introduce a constraint to capture the expected number of posts each expert may answer in the future. We note that $\sum_j R_{i,j}$ is the number of forum threads the expert i has responded to, or expected number of posts expert U_i is thought to have answered. We denote this as $E[U_i]$. Following similar logic as before, we propose the following formulation:

$$\textstyle\sum_j R_{i,j} = E[U_i],\ \forall i \in \{1, \cdots, n\}, \qquad \sum_j \vec{U_i} \cdot \vec{P_j} \approx E[U_i],\ \forall i \in \{1, \cdots, n\} \tag{7.10}$$

Intuitively, this equation forces how often the expert responds to a given post. In the case the real value of $E[u]$ is not known, $E[u]$ can computed by calculating $m \times p(U = u|P)$, where m is the number of posts in forum, and $p(U = u|P)$ is the probability that the expert u would respond to a random post based on historic data. The constraints representation of this formula is given as follows:

$$\mathcal{C}_{2i}(U, P) = \left(\sum_j \vec{U_i} \cdot \vec{P_j} - E[U_i] \right)^2 = 0,\ \forall i \in \{1, \cdots, n\} \tag{7.11}$$

7.2.6 Learning Algorithm

We note that all of our constraints are convex which allows us to leverage KKT condition. We can express the constraints into part of the objective function as the following:

$$\mathcal{C}(U,P) = \sum_{j=1}^{m} \lambda_{c1j}\ \mathcal{C}_{1j}(U,P) + \sum_{i=1}^{n} \lambda_{c2i}\ \mathcal{C}_{2i}(U,P) \tag{7.12}$$

This allows us to encode both the constraints and the objective functions that we introduced in the previous section into an unconstrained optimization problem as the following:

$$\begin{aligned} \underset{U,P,V,W}{\arg\min} \quad & \mathcal{L}_{base}(U,P) + \lambda_p \cdot \mathcal{L}_{posts}(P,W) \\ & + \lambda_u \cdot \mathcal{L}_{users}(U,W) + \lambda_s \cdot \mathcal{L}_{sim}(U,V) \\ & + \lambda_c \cdot \mathcal{C}(U,P) \\ & + \lambda_\beta \cdot (||U||_F + ||P||_F + ||W||_F + ||V||_F) \end{aligned} \tag{7.13}$$

Algorithm 7.1 Thread Recommender System for Experts

```
Input: $R \in \mathbb{R}^{n\times m}$, $D \in \mathbb{R}^{m\times|w|}$, $E \in \mathbb{R}^{n\times|w|}$, $S \in \mathbb{R}^{m\times m}$
Output: $U \in \mathbb{R}^{n\times k}$, $V \in \mathbb{R}^{n\times k}$, $P \in \mathbb{R}^{m\times k}$, $W \in \mathbb{R}^{|w|\times k}$
  $currIdx = 0$
  while Not converged do
    if $currIdx \ \% \ nm = 0$ then
      Precompute $\sum_i \vec{U_i} \cdot \vec{P_j}\ \forall j \in \{1, \cdots, m\}$
      Precompute $\sum_j \vec{U_i} \cdot \vec{P_j}\ \forall i \in \{1, \cdots, n\}$
    end if
    Sample row index $(i, j, l, i')$ from $U$, $P$, $W$, and $V$
    Update $U_{i,*}$, $P_{j,*}$, $W_{l,*}$, $V_{i,i'}$ from $\partial\mathcal{L}(U,P,W,V)$
    $currIdx{+}{+}$
  end while
```

In optimizing Equation 7.13, we applied a *stochastic gradient descent* (SGD) approach. SGDs are widely used in approximating low-rank matrices and have been shown to show good convergence and scalability properties [219]. We further note that computing equation 7.9 and equation 7.11 is rather expensive since it requires summation across rows and columns, respectively. Rather than computing this every iteration of SGD, we precompute these values every *nm* iterations. This proved to be a good compromise between the run time and the performance. Pseudo-code for our algorithm is shown in Algorithm 7.1.

7.3 EVALUATION

7.3.1 Dataset and Experimental Settings

MedHelp (`http://www.medhelp.org/`) is a website where users can satisfy their medical information needs. The website consists of medical news, health diaries, support community forums and an 'Ask a Doctor' forum. Of special interest to us is the 'Ask a Doctor' forum. Average users post question on this forum and designated experts answer them if particular question suits their interest. There are 67 different forum categories, ranging from 'addiction' to 'undiagnosed symptoms.' For the purpose of our system, we only evaluated a forum if it had at least 5 active designated experts. In total, our dataset had 168 experts, 56,194 threads, and 18 forum categories.

For all of our experiments, we set $k = 100$ as the dimension of our latent features. We calculated all of the performance results by running the algorithm 10 times on each of the parameter settings. Unless otherwise mentioned, for each runs of evaluation, we randomly chose 80% of the posts as training data and the rest as testing data. The performance results we report here are based on taking the mean performance value of all 18 categories based on the best performing parameter combination for the given forum with the exception of the relative weights of the two constraints, λ_{c1j} and λ_{c2i}. Both of these parameters were set to 1 throughout all the forums. Based on the sensitivity analysis that we have run, the relative weights do not cause significant changes in performance. We set $\lambda_\beta = 0.001$, which prevents the algorithm from overfitting. Finally, our algorithm converges if the objective function changes less than 0.001 compared to the previous iteration.

7.3.2 Evaluation Metrics

Performance of each method is measured using Precision at 1, 5, 10 (P@1, P@5, P@10), Mean Average Precision (MAP) and Mean Reciprocal Ranking (MRR). P@1, P@5 and P@10 measures the proportion of recommended items that are ground-truth items amongst the top k retrieved results. This represents how many relevant threads an expert would see if they were presented with 1, 5 and 10 recommended threads. MAP is the arithmetic mean of average precision values over a set of all the retrieved results, and represents a generalization of what experts would see. MRR on the other hand, measures the inverse rank at which the relevant ground-truth appears. This measures how many recommended threads a user would need to browse to find a relevant thread.

7.3.3 Competing Algorithms

In order to evaluate our algorithm, we compare our algorithm against various other methods.

NMF : Non-negative Matrix Factorization [220] is a widely used algorithm in recommender system literature. These capture how users and items interact with each other. This is our first baseline approach.
MF-D : Matrix factorization with post content model. This approach combines $\mathcal{L}_{base}(U, P)$ and $\mathcal{L}_{posts}(P, W)$ to recommend posts to designated experts. Notice that this method does not take expert's interest into the model. This is our second baseline approach.
MF-ED : Matrix factorization with **MF-D** and expert interest model $\mathcal{L}_{users}(U, W)$. This is our third baseline approach.
MF-SED : Matrix factorization with **MF-ED** and expert similarity $\mathcal{L}_{sim}(U, V)$. This captures how similar experts are to each other.
MF-DEC : Matrix factorization with **MF-ED** and constraints $\mathcal{C}(U, P)$.
MF-all : Matrix factorization with **MF-ED**, expert similarity $\mathcal{L}_{sim}(U, V)$ and constraints $\mathcal{C}(U, P)$. This is our proposed method.

7.3.4 Performance Comparison

To evaluate the comparative performance of the algorithms, we ran all competing algorithm on all 18 forum categories. The performance comparison for all our methods is shown in Table 7.1.

Algorithm	$P1$	$P5$	$P10$	MAP	MRR
NMF	0.1959	0.2005	0.1953	0.2263	0.1975
MF-D	0.0736	0.0587	0.0703	0.1914	0.0907
MF-ED	0.3114	0.2708	0.2639	0.2999	0.3066
MF-SED	0.3117	0.2827	0.2750	0.3068	0.3067
MF-DEC	0.3408$*$	0.3139$*$	0.3062$*$	0.3389$*$	0.3200
MF-all	**0.3698**$*$	**0.3281**$*$	**0.3172**$*$	**0.3438**$*$	**0.3390**$*$

Table 7.1: Average performance on 18 different forum categories. * denotes statistical significance over baseline methods at $\alpha = 0.05$.

We notice that NMF does not perform very well. This is because NMF only captures the interaction between users (designated experts) and forum posts. In particular, NMF does not capture the content of the forum itself. We notice a huge performance difference between MF-

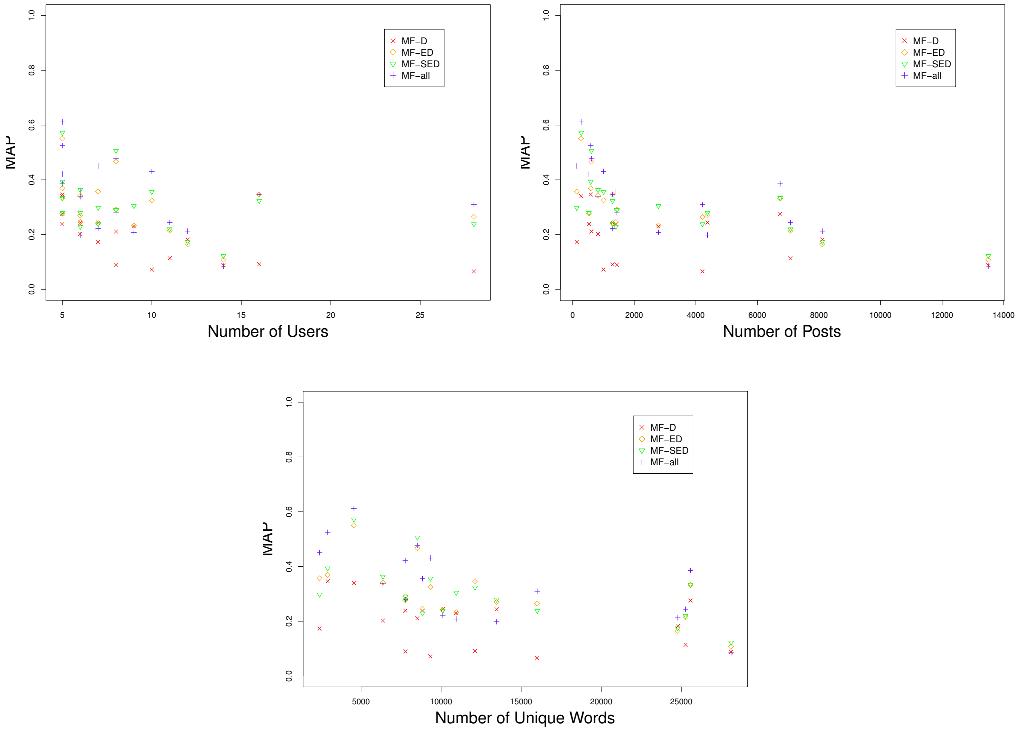


Figure 7.2: Impact of number of users, number of posts and number of words on the performance across all 18 forum categories.

D and MF-ED. This shows that, in recommending forum posts to experts, incorporating both the content of the post and that of the expert is both important. Adding expert similarity gives some performance improvements over MF-ED. In a given forum, there are experts whose interests intersect. Similarity characteristics are captured by adding expert similarity factor, which is consistent with existing literature in social collaboration [104, 106, 221].

A significant boost is seen for MF-DEC over MF-SED and MF-ED. This indicates the constraints are suitable at capturing the macro level behavior in the existence of multiple designated experts. As we have argued in the preceeding sections, only one expert is likely to respond to a given post. The performance gain indicates that, indeed, constraints are well-modeled and improve the performance. Finally, the combined method performs the best which implies that all factors are complementary.

7.3.5 Impact of the Size of the Dataset

The next question we asked was, are there any differences in performance as we increase the size of the data set? In particular, how does the number of posts, unique words and users affect the performance of each of the algorithms? In this experiment, we plot the performance of all 18 forum categories into scatter plots. The x axis indicates number of users, posts or words, and y axis denotes the value for MAP. Our results are shown in Figure 7.2. In almost all forum categories, regardless of the size of the vocabulary, number of posts or users, MF-all consistently outperformed all the other comparison methods. This shows that our results are not biased by subset of forums on which our methods perform very well, or not penalized because of a few forums on which it performs poorly.

7.3.6 Sensitivity Analysis

We also conducted a sensitivity analysis of each parameter. There were four parameters that we tuned: These are λ_p which assigns how much weight to give to post contents, λ_u, which captures expert's interest, λ_s, which assigns relative weights to expert similarities, and λ_c which dictates the importance weight of the constraints. The results that we present are based on averaging the performance across all 18 forums for each of the parameters that we have ran.

We first tuned λ_p from MD-D method. We see here that this method is not very sensitive to the parameter. On the other hand, λ_u from MD-ED seems to more sensitive to the parameter. The two remaining parameters, λ_s and λ_c from MD-SED and MD-DEC, respectively, were not very sensitive to the parameters. This experiment shows that the two constraints which capture experts' behaviors are fairly robust. Furthermore, we notice that encoding expert similarities based on contents they write is robust as well.

We were also curious on how sensitive assigning different weights to the two constraints are. In particular, these weights are λ_{c2i} and λ_{c1j} as shown in Equation 7.12. Tuning each individual λ_{c1j} and λ_{c2i} for all of i and j would be rather expensive. Instead, we give uniform weights to all of λ_{c1j} and λ_{c2i} and vary the parameters. We note from this experiment that both λ_{c1j} and λ_{c2i} are not very sensitive to the parameters. All of our results are in Figure 7.3. Based on the two sensitivity analysis, we learn that our algorithm is robust to varying parameters. In particular, once λ_c is set, differing relative weights of the two types of constraints do not really affect performance.

7.4 DISCUSSION

In this chapter, we introduced the problem of recommending forum posts to designated experts. In particular, we designated an algorithm which captures the experts' behavior from websites such as online health web forums. For most of the questions that average users

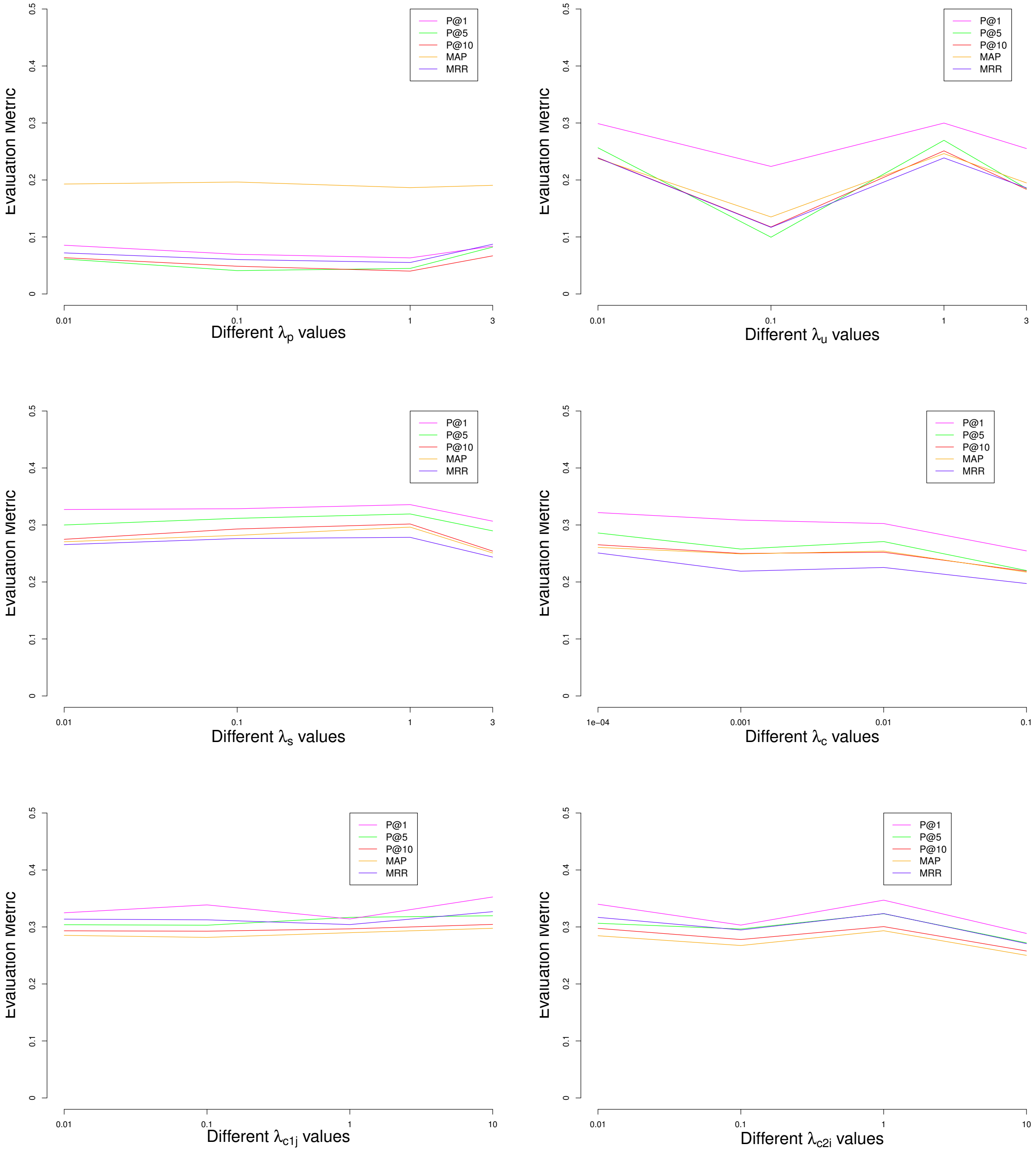


Figure 7.3: Parameter sensitivity on λ_p, λ_u, λ_s, λ_c, λ_{c1j} and λ_{c2i}

post, only one designated expert will answer them. To capture this intuition, we introduced constraints such as only one expert answering a given post. The proposed constraints have improved performance based on 18 different forum categories that we have on which we have run the experiments on. We solved the recommendation problem using a matrix factorization framework. This was chosen because the matix factorization framework allows for an easy extension to add additional constraints.

We believe the proposed solutions to routing health forums to designated experts can also be applied to aid hospital staff members find the most relevant patient reviews. Each of the texts can be assigned topics from patient experience taxonomy. The model can be trained on the assigned topics. Similar to the work conducted in health forums, only one hospital staff members is likely to respond to a patient review.

For future work, one can envision modeling how long an expert takes to answer a question, and incorporating this into the model. Furthermore, we only focused on the behavior of designated experts in our work. We also did not explore whether or not semantics may affect an expert responds to a post. Studying the impact of language semantics may help us further understand designated experts' behavior.

CHAPTER 8: CONCLUSION AND FUTURE WORKS

8.1 RESEARCH SUMMARY

In this thesis, we discussed how we can better understand patient experience from patient-generated texts. Patient experience is important because it affects patient satisfaction and health outcomes [6, 7, 9]. We stated in Chapter 1 that to improve patient experience, we must first understand it.

To realize the goal of learning more about patient experience, we utilized a state-of-the-art patient experience taxonomy which was built and annotated on patient-generated texts by our collaborators in Chapter 3. The taxonomy was based on a combination of an official government's patient experience questionnaire, the HCAHPS survey, along with past research conducted to understand patient experience [20, 21]. One challenge was in ensuring high quality taxonomy annotation. By iterating multiple times, the annotation quality, as measured by inter-annotator agreement improved.

With an annotated corpus, we next investigated how to classify topics from patient experience taxonomy. We explored two different approaches to capture the characteristics of patient experience in patient-generated texts. We first implemented a rich semantic representation of patient-generated texts which improved the performance of an automated and comprehensive patient experience classifier in Chapter 4. From the semantic feature representation of the texts, we learned how patients describe themselves when they talk about a particular topic. For example, patients find clinicians who listen well to be strong communicators. Furthermore, we found that some topics are primarily positive, or negative. For instance, patients generally perceive the topic 'Time spent,' to be negative, because they feel rushed or feel that hospital staff members did not spend adequate time with them.

Topic inter-dependency was important to further understand patient-generated text as it pertains to patient experience; we described how we utilized this in Chapter 5. We developed a model that captured two different topic interactions, one where they were semantically similar, another where they were different. The hypotheses were tested by utilizing rich feature representation afforded by a deep neural network, and by encoding the two hypotheses constraints. Both of the hypotheses helped improve our understanding of the patient experience, though the second hypothesis had stronger signals in improving the classifier performance than the first one.

Because patient experience differs from one patient segment to another, we developed an approach to identify an example segment in Chapter 6. We focused on learning more about the patients' medical history – in particular, whether they are still taking a given medication or not. We further explored linguistic features to further our understanding of this particular patient segment. While this is not the only patient segment that influences patient experience, we believe future researchers can bootstrap this method to learn to identify different patient segments.

In Chapter 7, we proposed how physicians and patients can benefit from extracting topics related to patient experience in patient-generated texts. We argued that while hospital staff members are very busy, they benefit from learning more about their patients' experience. In routing relevant patient-generated texts to hospital staff members, we enforced various constraints. This captured hospital staff members' behaviors as interact with patient-generated texts. The behavioral constraints increased the relevancy of the recommended texts.

8.2 POTENTIAL IMPROVEMENTS

We discussed how to comprehensively and automatically understand patient experience, and proposed its potential application. There are, however, limitations with our models, and improving upon these can further our understanding of patient experience. We discuss limitations and areas that can further improve our research in this section.

8.2.1 Rich Semantic Representation of Patient Experience Taxonomy

We introduced semantic features that can capture the document characteristics of the RateMD dataset in this work. While these features provided us with valuable insights, there are still limitations with our method. First, our proposed set of features did not substantially improve the classification performance. In particular, while our embedding based methods (Word2Vec, Dep2Vec) did improve performance over the baseline, other features such as WordNet or aspect rating had very marginal improvement. Part of this can be due to us not adequately exploring the given feature. We opted for a shallow feature representation of both WordNet and aspect rating. By investigating whether a given shallow set of features improve performance or not, we can, in later successive research works, decide to improve the set of shallow features to better capture the corpus. Our work, in this sense, is exploratory. In the literature there has been work which utilizes WordNet [222, 223] or aspect ratings [144, 224] to represent the corpus. Furthermore, while Word2Vec and Dep2Vec both are powerful embedding methods, researchers have also further extended upon these methods as well. An exciting venture may have been, how we can combine Word2Vec and Dep2Vec, but we did not explore this possibility.

Second, as is the case with many classifiers, our algorithm had low evaluation performance results on classes that appeared infrequently. Such is the critical limitation of supervised approaches. We tried to mitigate this issue by utilizing LARA, which is an unsupervised method. The benefit of the method is that we do not need a large number of labeled data; instead, well-curated seed words are sufficient to identify topics present in texts. To test the upper bound of LARA, we utilized representative keywords learned from our classifiers. However, we were unable to observe improvements over our supervised approach even after utilizing LARA. To ultimately improve the classification performance of infrequent classes, we need more labeled data. However, if we were to annotate the dataset at random, we are more likely to end up annotating those that appear very frequently, for example, 'Temperament' class, and still very few instances of infrequently occurring classes (such as 'cost'). A sampling approach that can sample infrequent labels from the corpus may help. However, we did not explore how we can better label more data in this work.

Third, while we analyzed which features are closely related to each topic, we did not investigate why a particular review may be classified into the said topic. As an example, a patient may have had a bad experience with an unfriendly doctor and detail what exactly happened at the doctor's office. Our proposed approach can identify that the patient has talked about the friendliness of the doctor. However, it is unable to extract, from a human-readable form, what exactly the patient is talking about for the said topic (in the above example, the details of what happened at the doctor's office). This problem can potentially be solved by utilizing existing text summarization approaches. Text summarization techniques can summarize what patients most frequently mention for each topic in a human-readable form, which is an improvement over the feature analysis we conducted. While this is an important avenue to explore, we did not investigate how we can summarize each topic in the patient experience taxonomy.

8.2.2 Topic Inter-dependencies in Patient Experience Taxonomy

First, while the proposed loss function incorporates both the child-class prediction errors and encodes the hierarchical patient experience taxonomy, we have only experimented with squared difference loss to enforce the structural constraints (parent-child relationships). In typical prediction tasks, researchers found that cross-entropy yields the best results. However, we did not explore whether a different type of loss, perhaps one motivated by cross-entropy models would work better or not. Furthermore, we assumed that violation between two different parent-child relationships should be assigned the same weight, regardless of which meta-class or child-class violated the structural constraints. However, it is possible that some parent-child constraint violations are more significant than others. For instance, it is possible that violating the parent-child relationship between 'Interpersonal manner' (a parent class) and 'Temperament,' (a child class) is perfectly acceptable. On the other hand, the same between 'Scheduling,' and 'Time spent' may not prove to be so. In such cases, we should assign higher weights if we violate the latter than the former. In this work, however, we treated both types of parent-child class violations with the same weight.

Second, while our model was driven by two hypotheses, both of which improved the classification performance, the model is unable to learn which two child-classes are closely associated with each other. If two child-classes belong to the same meta-class, then the assumption was that these two should be tightly correlated with each other. Furthermore, we assumed that the label interactions are driven by the meta-class prediction as well. However, we did not explore if the two classes – for example, 'Temperament,' and 'Helpfulness' – are closely correlated or not. Part of the limitation is that directly interpreting label correlations is infeasible in our approach. We utilized a Deep Neural Network (DNN) framework to conduct the study, and while DNNs frequently improve the classification performance, the trained models are generally not interpretable. Our proposed approach also suffers from this drawback.

Third, we proposed a generic hierarchical taxonomy classifier that can be applied to other datasets that are hierarchical. Testing the method described in this work to other datasets has numerous benefits if we can show that hierarchical multi-label (HML) classification approach discussed in this work performs better than non-hierarchical multi-label classification models on other datasets. This may indicate that HML classification approach is flexible and generalizable. On the other hand, if we do not observe any performance gains, then the method is better suited for RateMD dataset but not the other ones. However, we have only tested our method in RateMD dataset, and have not shown whether our approach can also be applicable in other hierarchical problem domains.

8.2.3 Patient Segments Identification for Comprehensive Understanding of Patient Experience Taxonomy

We proposed utilizing online health forums to identify patient segments. However, throughout this thesis, we proposed how we can better understand patient experience from the RateMD dataset. It is perfectly reasonable to question whether the approach we proposed is also applicable on RateMD dataset. While we have not tested our method in the review dataset, we argue that there is still benefit in identifying different patient segments from health forums. The proposed set of features can still be utilized to identify patients' segments on RateMD dataset. While we have not tested whether the features will still be valid in RateMD dataset, they are a good starting point to build a user segments classifier. Recently, many websites have integrated Google ID or Facebook account as well. We can further conceive that a unified online identification can be utilized to classify users' identity in one medium (for instance, in health forums), and use this information to understand patients' experiences better when they author patient reviews. The ID integration can then be utilized to learn patient experiences from different patient segments gathered from online health forums.

Second, we only explored a single type of patient segment. In particular, we proposed an approach to identify whether a patient has stopped taking medication or not. However, there are many different types of patient segments which we did not explore in this work – for example, an automated approach to identify the type of illness that patients have may help us better understand patients' experiences. We argue that techniques developed to identify whether a patient has stopped taking medication or not can still be applicable for other types of patient segments, but we leave this for future work.

Third, while we described and evaluated an approach to identify different patient segments, we did not utilize this information to better understand patients' experiences. Part of this was due to the first limitation we described, i.e., the inability to connect what patients had posted in online reviews to patient segments we identified discussed in this work. Once we can connect patient segments with online reviews that patients have posted, either by Google or Facebook ID integration, or by running the patient segment identification algorithm on the RateMD dataset, we believe we will be able to understand patient experience in different patient segments.

8.2.4 Applications of Patient Experience Taxonomy

We proposed classifiers that can automatically identify topics from patient experience taxonomy in Chapter 4 and Chapter 5. On the other hand, the recommender system we built was trained to recommend posts relevant to a doctor's specialization (such as cancer or heart disease). More precisely, the recommender system was trained on disease taxonomy. While the recommender system is general enough that it can be trained from patient experience taxonomy rather than from disease taxonomy as we have conducted in Chapter 7, we did not demonstrate that learning to recommend posts to hospital staff based on the patient experience taxonomy is, indeed, possible. It would be beneficial, in the future, to show that the recommender system that we built can, indeed, be trained on patient experience taxonomy.

Furthermore, our recommender system was trained on online health forums, whereas the patient experience taxonomy classifier was trained on RateMD patient review datasets. For the recommender system to work, it would either need to be trained on RateMD datasets, or the patient experience taxonomy classifier would need to be trained on online health forums.

Third, while we have conducted a thorough offline test that measured various evaluation metrics (MRR, Precision at k, MAP), we did not conduct a real-world test. A real-world test is beneficial to demonstrate how useful the application is to hospital staff members. For example, a test showing that the vast majority of the hospital staff members found the recommender system to be useful might result in the system's adoption by more hospital staff members.

8.3 FUTURE WORKS

We envision multiple potential future works arising from this thesis. In the thesis, we answered the question of whether the topic interactions can be encoded into the model or not. The answer was a resounding yes where such interactions not only exist but also can be utilized to improve topic identification in the review corpus. However, the work did not investigate **how** the topics may be correlated with each other. It may be the case that two topics, say, 'good with kids,' and 'empathy' are correlated, but the current neural network model was unable to capture such a relationship. The limitation was due to how the deep neural network framework encodes features. With few notable exceptions, it is challenging to analyze features in deep neural networks. Traditional machine learning approaches, such as integer linear programming approach [225] may be one way to explore this venue, but such an approach may suffer, classification performance wise, compared to those developed on deep neural networks. Care must be taken to ensure that the newly proposed method is at least as competitive with the neural networks assuming that they are fed the same feature sets.

We opted for a supervised learning approach where the model was trained based on the patient experience taxonomy. While the supervised approaches capture the set of topics that are of interest to researchers, this requires training data which requires an annotated, or labeled, dataset. Unsupervised approaches, on the other hand, ease this process since they do not need annotated data. However, we did not explore into this avenue, and future works which look into unsupervised methods will allow researchers to identify new patient experience topics more rapidly. At the same time, the newly discovered topics should align with what healthcare researchers and physicians are interested in.

Furthermore, most of our analysis was conducted on a patient review website, with a focus on how patients describe their hospital experience. However, patient review websites are not the only places where patients share their experience, nor are hospital experiences the only type of experience patients have. There are online health forums, micro-blogs such as Twitter or Weibo, where patients may talk about a different set of experiences. Some of these examples are different treatment techniques or experiences of taking particular drugs. Furthermore, comparing different venues of online patient-generated texts – for example, online health forums and online patient reviews – may be of interest. It is possible that patients express themselves differently. For example, micro-blogs are succinct whereas health forums and patient reviews tend to be longer. However, in this thesis, we did not explore how the patients may describe their experience differently, depending on which online medium they write.

Assigning topics from patient experience taxonomy to patient-generated texts opens up new application areas. We proposed utilizing a patient experience taxonomy classifier to aid hospital staff members to retrieve patient-generated texts that may be of interest to them. There are even more application areas in which researchers can utilize patient experience taxonomy classifier, however. Past works investigated sentiments expressed in patient-generated texts [226, 227]. We envision going a step further where each identified topic is assigned a sentiment score. This allows patients to either filter hospitals that meet their criteria for a given topic (for instance, finding hospitals whose staff members are generally friendly), or even allows patient review websites to recommend hospitals based on patients' preset criteria.

It is our belief that some of these applications arising from understanding patient experience will improve the well-being of future patients.

REFERENCES

[1] A. Alexopoulou, S. P. Dourakis, D. Mantzoukis, T. Pitsariotis, A. Kandyli, M. Deutsch, and A. J. Archimandritis, "Adverse drug reactions as a cause of hospital admissions: a 6-month experience in a single center in Greece," *Eur. J. Intern. Med.*, vol. 19, no. 7, pp. 505–510, Nov 2008.

[2] K. M. Hakkarainen, K. Hedna, M. Petzold, and S. Hagg, "Percentage of patients with preventable adverse drug reactions and preventability of adverse drug reactions–a meta-analysis," *PLoS ONE*, vol. 7, no. 3, p. e33236, 2012.

[3] A. J. Forster, H. J. Murff, J. F. Peterson, T. K. Gandhi, and D. W. Bates, "Adverse drug events occurring following hospital discharge," *J Gen Intern Med*, vol. 20, no. 4, pp. 317–323, Apr 2005.

[4] A. J. Leendertse, D. Visser, A. C. Egberts, and P. M. van den Bemt, "The relationship between study characteristics and the prevalence of medication-related hospitalizations: a literature review and novel analysis," *Drug Saf*, vol. 33, no. 3, pp. 233–244, Mar 2010.

[5] J. L. Jackson and K. Kroenke, "Difficult patient encounters in the ambulatory clinic: clinical predictors and outcomes," *Arch. Intern. Med.*, vol. 159, no. 10, pp. 1069–1075, May 1999.

[6] C. Doyle, L. Lennox, and D. Bell, "A systematic review of evidence on the links between patient experience and clinical safety and effectiveness," *BMJ Open*, vol. 3, no. 1, Jan 2013.

[7] 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*, vol. 74, no. 3, pp. 295–301, Mar 2009.

[8] M. A. Stewart, "Effective physician-patient communication and health outcomes: a review," *CMAJ*, vol. 152, no. 9, pp. 1423–1433, May 1995.

[9] R. L. Street, "How clinician-patient communication contributes to health improvement: modeling pathways from talk to outcome," *Patient Educ Couns*, vol. 92, no. 3, pp. 286–291, Sep 2013.

[10] M. J. Paul and M. Dredze, "Drug extraction from the web: Summarizing drug experiences with multi-dimensional topic models," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2013. [Online]. Available: http://aclweb.org/anthology/N13-1017 pp. 168–178.

[11] M. J. Paul and M. Dredze, “Experimenting with drugs (and topic models): Multi-dimensional exploration of recreational drug discussions,” in *Information Retrieval and Knowledge Discovery in Biomedical Text, Papers from the 2012 AAAI Fall Symposium, Arlington, Virginia, USA, November 2-4, 2012*, 2012. [Online]. Available: http://www.aaai.org/ocs/index.php/FSS/FSS12/paper/view/5573

[12] S. Wang, Y. Li, D. Ferguson, and C. Zhai, “Sideeffectptm: An unsupervised topic model to mine adverse drug reactions from health forums,” in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, ser. BCB ’14. New York, NY, USA: ACM, 2014. [Online]. Available: http://doi.acm.org/10.1145/2649387.2649398 pp. 321–330.

[13] R. Eshleman, D. Jha, and R. Singh, “Identifying individuals amenable to drug recovery interventions through computational analysis of addiction content in social media,” in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov 2017, pp. 849–854.

[14] M. J. Paul, M. S. Chisolm, M. W. Johnson, R. G. Vandrey, and M. Dredze, “Assessing the Validity of Online Drug Forums as a Source for Estimating Demographic and Temporal Trends in Drug Use,” *J Addict Med*, Jul 2016.

[15] S. J. Kim, L. A. Marsch, J. T. Hancock, and A. K. Das, “Scaling Up Research on Drug Abuse and Addiction Through Social Media Big Data,” *J. Med. Internet Res.*, vol. 19, no. 10, p. e353, 10 2017.

[16] D. A. Broniatowski, M. J. Paul, and M. Dredze, “National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic,” *PLoS ONE*, vol. 8, no. 12, p. e83672, 2013.

[17] M. Santillana, A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, and J. S. Brownstein, “Combining Search, Social Media, and Traditional Data Sources to Improve Influenza Surveillance,” *PLoS Comput. Biol.*, vol. 11, no. 10, p. e1004513, Oct 2015.

[18] M. J. Paul, M. Dredze, and D. Broniatowski, “Twitter improves influenza forecasting,” *PLoS Curr*, vol. 6, Oct 2014.

[19] A. J. Lazard, E. Scheinfeld, J. M. Bernhardt, G. B. Wilcox, and M. Suran, “Detecting themes of public concern: a text mining analysis of the Centers for Disease Control and Prevention’s Ebola live Twitter chat,” *Am J Infect Control*, vol. 43, no. 10, pp. 1109–1111, Oct 2015.

[20] A. Lopez, A. Detz, N. Ratanawongsa, and U. Sarkar, “What patients say about their doctors online: a qualitative content analysis,” *J Gen Intern Med*, vol. 27, no. 6, pp. 685–692, Jun 2012.

[21] C. Doyle, L. Lennox, and D. Bell, “A systematic review of evidence on the links between patient experience and clinical safety and effectiveness,” *BMJ Open*, vol. 3, no. 1, Jan 2013.

[22] Y. Jung, C. Hur, D. Jung, and M. Kim, "Identifying key hospital service quality factors in online health communities," *J. Med. Internet Res.*, vol. 17, no. 4, p. e90, Apr 2015.

[23] M. Emmert, F. Meier, A. K. Heider, C. Durr, and U. Sander, "What do patients say about their physicians? an analysis of 3000 narrative comments posted on a German physician rating website," *Health Policy*, vol. 118, no. 1, pp. 66–73, Oct 2014.

[24] G. J. Young, M. Meterko, and K. R. Desai, "Patient satisfaction with hospital care: effects of demographic and institutional characteristics," *Med Care*, vol. 38, no. 3, pp. 325–334, Mar 2000.

[25] K. Doing-Harris, D. L. Mowery, C. Daniels, W. W. Chapman, and M. Conway, "Understanding patient satisfaction with received healthcare services: A natural language processing approach," *AMIA Annu Symp Proc*, vol. 2016, pp. 524–533, 2016.

[26] B. L. Ranard, R. M. Werner, T. Antanavicius, H. A. Schwartz, R. J. Smith, Z. F. Meisel, D. A. Asch, L. H. Ungar, and R. M. Merchant, "Yelp Reviews Of Hospital Care Can Supplement And Inform Traditional Surveys Of The Patient Experience Of Care," *Health Aff (Millwood)*, vol. 35, no. 4, pp. 697–705, Apr 2016.

[27] B. C. Wallace, M. J. Paul, U. Sarkar, T. A. Trikalinos, and M. Dredze, "A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews," *J Am Med Inform Assoc*, vol. 21, no. 6, pp. 1098–1103, 2014.

[28] E. N. Demner-Fushman D, "Aspiring to Unintended Consequences of Natural Language Processing: A Review of Recent Developments in Clinical and Consumer-Generated Text Processing," *Yearbook of Medical Informaticsl*, vol. 1, pp. 224–233, 2016.

[29] A. Asprey, J. L. Campbell, J. Newbould, S. Cohn, M. Carter, A. Davey, and M. Roland, "Challenges to the credibility of patient feedback in primary healthcare settings: a qualitative study," *Br J Gen Pract*, vol. 63, no. 608, pp. e200–208, Mar 2013.

[30] S. H. Richards, J. L. Campbell, E. Walshaw, A. Dickens, and M. Greco, "A multi-method analysis of free-text comments from the UK General Medical Council Colleague Questionnaires," *Med Educ*, vol. 43, no. 8, pp. 757–766, Aug 2009.

[31] L. Cui, S. Tao, and G. Q. Zhang, "A Semantic-based Approach for Exploring Consumer Health Questions Using UMLS," *AMIA Annu Symp Proc*, vol. 2014, pp. 432–441, 2014.

[32] I. D. Maramba, A. Davey, M. N. Elliott, M. Roberts, M. Roland, F. Brown, J. Burt, O. Boiko, and J. Campbell, "Web-based textual analysis of free-text patient experience comments from a survey in primary care," *JMIR Med Inform*, vol. 3, no. 2, p. e20, May 2015.

[33] H. Cho and J.-S. Lee, "Data-driven feature word selection for clustering online news comments," in *2016 International Conference on Big Data and Smart Computing (BigComp)*, Jan 2016, pp. 494–497.

[34] E. E. Zusman, "HCAHPS replaces Press Ganey survey as quality measure for patient hospital experience," *Neurosurgery*, vol. 71, no. 2, pp. N21–24, Aug 2012.

[35] A. Detz, A. Lopez, and U. Sarkar, "Long-term doctor-patient relationships: patient perspective from online reviews," *J. Med. Internet Res.*, vol. 15, no. 7, p. e131, Jul 2013.

[36] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995. [Online]. Available: http://doi.acm.org/10.1145/219717.219748

[37] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, 2014, pp. 1188–1196.

[38] O. Levy and Y. Goldberg, "Dependency-based word embeddings," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, 2014, pp. 302–308.

[39] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[40] E. W. Black, L. A. Thompson, H. Saliba, K. Dawson, and N. M. Black, "An analysis of healthcare providers' online ratings," *Inform Prim Care*, vol. 17, no. 4, pp. 249–253, 2009.

[41] E. S. Xioufis, G. Tsoumakas, and I. P. Vlahavas, "An empirical study of lazy multilabel classification algorithms," in *Artificial Intelligence: Theories, Models and Applications, 5th Hellenic Conference on AI, SETN 2008, Syros, Greece, October 2-4, 2008. Proceedings*, 2008. [Online]. Available: https://doi.org/10.1007/978-3-540-87881-0_40 pp. 401–406.

[42] M. P. Manary, W. Boulding, R. Staelin, and S. W. Glickman, "The patient experience and health outcomes," *N. Engl. J. Med.*, vol. 368, no. 3, pp. 201–203, Jan 2013.

[43] T. C. Tsai, E. J. Orav, and A. K. Jha, "Patient satisfaction and quality of surgical care in US hospitals," *Ann. Surg.*, vol. 261, no. 1, pp. 2–8, Jan 2015.

[44] J. D. Shirk, H. J. Tan, J. C. Hu, C. S. Saigal, and M. S. Litwin, "Patient experience and quality of urologic cancer surgery in US hospitals," *Cancer*, vol. 122, no. 16, pp. 2571–2578, 08 2016.

[45] C. Boev, "The relationship between nurses' perception of work environment and patient satisfaction in adult critical care," *J Nurs Scholarsh*, vol. 44, no. 4, pp. 368–375, Dec 2012.

[46] "Acute inpatient pps," https://www.cms.gov/medicare/medicare-fee-for-service-payment/AcuteInpatientPPS/index.htmll, accessed: 2019-01-30.

[47] G. Adhikary, M. S. R. Shawon, M. W. Ali, M. Shamsuzzaman, S. Ahmed, K. A. Shackelford, A. Woldeab, N. Alam, S. S. Lim, A. Levine, E. Gakidou, and M. J. Uddin, "Factors influencing patients' satisfaction at different levels of health facilities in Bangladesh: Results from patient exit interviews," *PLoS ONE*, vol. 13, no. 5, p. e0196643, 2018.

[48] B. N. Dang, R. A. Westbrook, M. C. Rodriguez-Barradas, and T. P. Giordano, "Identifying drivers of overall satisfaction in patients receiving HIV primary care: a cross-sectional study," *PLoS ONE*, vol. 7, no. 8, p. e42980, 2012.

[49] J. L. Jackson, J. Chamberlin, and K. Kroenke, "Predictors of patient satisfaction," *Soc Sci Med*, vol. 52, no. 4, pp. 609–620, Feb 2001.

[50] J. Sitzia and N. Wood, "Response rate in patient satisfaction research: an analysis of 210 published studies," *Int J Qual Health Care*, vol. 10, no. 4, pp. 311–317, Aug 1998.

[51] T. V. Perneger, E. Chamot, and P. A. Bovier, "Nonresponse bias in a survey of patient perceptions of hospital care," *Med Care*, vol. 43, no. 4, pp. 374–380, Apr 2005.

[52] C. P. Lewis and J. N. Newell, "Patients' perspectives of care for type 2 diabetes in Bangladesh -a qualitative study," *BMC Public Health*, vol. 14, p. 737, Jul 2014.

[53] J. Segal, M. Sacopulos, V. Sheets, I. Thurston, K. Brooks, and R. Puccia, "Online doctor reviews: do they track surgeon volume, a proxy for quality of care?" *J. Med. Internet Res.*, vol. 14, no. 2, p. e50, Apr 2012.

[54] J. J. Palmieri and T. A. Stern, "Lies in the doctor-patient relationship," *Prim Care Companion J Clin Psychiatry*, vol. 11, no. 4, pp. 163–168, 2009.

[55] I. T. Agaku, A. O. Adisa, O. A. Ayo-Yusuf, and G. N. Connolly, "Concern about security and privacy, and perceived control over collection and use of health information are related to withholding of health information from healthcare providers," *J Am Med Inform Assoc*, vol. 21, no. 2, pp. 374–378, 2014.

[56] "Alcohol screening and counseling," https://www.cdc.gov/vitalsigns/alcohol-screening-counseling/index.html, accessed: 2010-09-30.

[57] T. Zhang, J. H. D. Cho, and C. Zhai, "Understanding user intents in online health forums," *IEEE J. Biomedical and Health Informatics*, vol. 19, no. 4, pp. 1392–1398, 2015.

[58] J. H. D. Cho, V. Q. Z. Liao, Y. Jiang, and B. R. Schatz, "Aggregating personal health messages for scalable comparative effectiveness research," in *ACM Conference on Bioinformatics, Computational Biology and Biomedical Informatics. ACM-BCB 2013, Washington, DC, USA, September 22-25, 2013*, 2013, p. 907.

[59] D. L. MacLean, S. Gupta, A. Lembke, C. D. Manning, and J. Heer, “Forum77: An analysis of an online health forum dedicated to addiction recovery,” in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015, Vancouver, BC, Canada, March 14 - 18, 2015*, 2015, pp. 1511–1526.

[60] R. W. White, R. Harpaz, N. H. Shah, W. DuMouchel, and E. Horvitz, “Toward enhanced pharmacovigilance using patient-generated data on the internet,” *Clin. Pharmacol. Ther.*, vol. 96, no. 2, pp. 239–246, Aug 2014.

[61] G. N. Noren, “Pharmacovigilance for a revolving world: prospects of patient-generated data on the internet,” *Drug Saf*, vol. 37, no. 10, pp. 761–764, Oct 2014.

[62] F. Rothenfluh and P. J. Schulz, “Physician Rating Websites: What Aspects Are Important to Identify a Good Doctor, and Are Patients Capable of Assessing Them? A Mixed-Methods Approach Including Physicians’ and Health Care Consumers’ Perspectives,” *J. Med. Internet Res.*, vol. 19, no. 5, p. e127, 05 2017.

[63] T. Lagu, N. S. Hannon, M. B. Rothberg, and P. K. Lindenauer, “Patients’ evaluations of health care providers in the era of social networking: an analysis of physician-rating websites,” *J Gen Intern Med*, vol. 25, no. 9, pp. 942–946, Sep 2010.

[64] M. Emmert, U. Sander, and F. Pisch, “Eight questions about physician-rating websites: a systematic review,” *J. Med. Internet Res.*, vol. 15, no. 2, p. e24, Feb 2013.

[65] H. Hao, “The development of online doctor reviews in China: an analysis of the largest online doctor review website in China,” *J. Med. Internet Res.*, vol. 17, no. 6, p. e134, Jun 2015.

[66] K. Okike, T. K. Peter-Bibb, K. C. Xie, and O. N. Okike, “Association Between Physician Online Rating and Quality of Care,” *J. Med. Internet Res.*, vol. 18, no. 12, p. e324, 12 2016.

[67] F. Greaves, D. Ramirez-Cano, C. Millett, A. Darzi, and L. Donaldson, “Use of sentiment analysis for capturing patient experience from free-text comments posted online,” *J. Med. Internet Res.*, vol. 15, no. 11, p. e239, Nov 2013.

[68] R. Dev Sharma, S. Tripathi, S. Sahu, S. Mittal, and A. Anand, “Predicting online doctor ratings from user reviews using convolutional neural networks,” 11 2015.

[69] N. S. Bardach, A. Lyndon, R. Asteria-Penaloza, L. E. Goldman, G. A. Lin, and R. A. Dudley, “From the closest observers of patient care: a thematic analysis of online narrative reviews of hospitals,” *BMJ Qual Saf*, vol. 25, no. 11, pp. 889–897, 11 2016.

[70] R. H. Baud, A. M. Rassinoux, and J. R. Scherrer, “Natural language processing and semantical representation of medical texts,” *Methods Inf Med*, vol. 31, no. 2, pp. 117–125, Jun 1992.

[71] 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, and T. Speroff, "Automated identification of postoperative complications within an electronic medical record using natural language processing," *JAMA*, vol. 306, no. 8, pp. 848–855, Aug 2011.

[72] C. A. Bejan and J. C. Denny, "Learning to identify treatment relations in clinical text," *AMIA Annu Symp Proc*, vol. 2014, pp. 282–288, 2014.

[73] C. Wright, S. H. Richards, J. J. Hill, M. J. Roberts, G. R. Norman, M. Greco, M. R. Taylor, and J. L. Campbell, "Multisource feedback in evaluating the performance of doctors: the example of the UK General Medical Council patient and colleague questionnaires," *Acad Med*, vol. 87, no. 12, pp. 1668–1678, Dec 2012.

[74] C. Graham and P. Woods, "Understanding and Using Health Experiences:Improving patient care," *Oxford University Press*, 2013.

[75] E. Riiskj?r, J. Ammentorp, and P. E. Kofoed, "The value of open-ended questions in surveys on patient experience: number of comments and perceived usefulness from a hospital perspective," *Int J Qual Health Care*, vol. 24, no. 5, pp. 509–516, Oct 2012.

[76] G. C. Liu, M. A. Harris, S. A. Keyton, and R. M. Frankel, "Use of unstructured parent narratives to evaluate medical student competencies in communication and professionalism," *Ambul Pediatr*, vol. 7, no. 3, pp. 207–213, 2007.

[77] G. C. Liu, M. A. Harris, S. A. Keyton, and R. M. Frankel, "What are patients seeking when they turn to the Internet? Qualitative content analysis of questions asked by visitors to an orthopaedics Web site," *Med Internet Res.*, vol. 10, no. 5, Oct 2003.

[78] C. Tang, Paul, W. Black, and C. Young, "What are patients seeking when they turn to the Internet? Qualitative content analysis of questions asked by visitors to an orthopaedics Web site," *Med Internet Res.*, vol. 10, no. 5, Oct 2003.

[79] E. R. Burner, M. D. Menchine, K. Kubicek, M. Robles, and S. Arora, "Perceptions of successful cues to action and opportunities to augment behavioral triggers in diabetes self-management: qualitative analysis of a mobile intervention for low-income Latinos with diabetes," *J. Med. Internet Res.*, vol. 16, no. 1, p. e25, Jan 2014.

[80] L. Ashley, H. Jones, J. Thomas, A. Newsham, A. Downing, E. Morris, J. Brown, G. Velikova, D. Forman, and P. Wright, "Integrating patient reported outcomes with clinical cancer registry data: a feasibility study of the electronic Patient-Reported Outcomes From Cancer Survivors (ePOCS) system," *J. Med. Internet Res.*, vol. 15, no. 10, p. e230, Oct 2013.

[81] A. Elmessiry, W. O. Cooper, T. F. Catron, J. Karrass, Z. Zhang, and M. P. Singh, "Triaging Patient Complaints: Monte Carlo Cross-Validation of Six Machine Learning Classifiers," *JMIR Med Inform*, vol. 5, no. 3, p. e19, Jul 2017.

[82] F. Greaves, C. Millett, and N. P, "Incorporating Anecdotal Reports From Consumers into Their National Reporting System: Lessons for the United States of What to Do or Not to Do?" *Medical Care Research and Review*, vol. 71, pp. 65S–80S, 2014.

[83] S. Brody and N. Elhadad, "Detecting salient aspects in online reviews of health providers," *AMIA Annu Symp Proc*, vol. 2010, pp. 202–206, Nov 2010.

[84] H. Khanpour and C. Caragea, "Fine-grained information identification in health related posts," in *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, 2018. [Online]. Available: https://doi.org/10.1145/3209978.3210132 pp. 1001–1004.

[85] S. Chae, S. Kwon, and D. Lee, "Predicting Infectious Disease Using Deep Learning and Big Data," *Int J Environ Res Public Health*, vol. 15, no. 8, 07 2018.

[86] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, "Deep convolutional ranking for multilabel image annotation," *CoRR*, vol. abs/1312.4894, 2013. [Online]. Available: http://arxiv.org/abs/1312.4894

[87] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *CoRR*, vol. abs/1402.1128, 2014. [Online]. Available: http://arxiv.org/abs/1402.1128

[88] D. Pathak, P. Krähenbühl, and T. Darrell, "Constrained convolutional neural networks for weakly supervised segmentation," in *International Conference on Computer Vision (ICCV)*, 2015.

[89] J. Guo, S. Xu, S. Bao, and Y. Yu, "Tapping on the potential of Q&A community by recommending answer providers," in *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, ser. CIKM '08. New York, NY, USA: ACM, 2008. [Online]. Available: http://doi.acm.org/10.1145/1458082.1458204 pp. 921–930.

[90] F. Riahi, Z. Zolaktaf, M. Shafiei, and E. Milios, "Finding expert users in community question answering," in *Proceedings of the 21st International Conference Companion on World Wide Web*, ser. WWW '12 Companion. New York, NY, USA: ACM, 2012. [Online]. Available: http://doi.acm.org/10.1145/2187980.2188202 pp. 791–798.

[91] L. Du, W. L. Buntine, and H. Jin, "A segmented topic model based on the two-parameter poisson-dirichlet process," *Machine Learning*, vol. 81, no. 1, pp. 5–19, 2010.

[92] L. Yang, M. Qiu, S. Gottipati, F. Zhu, J. Jiang, H. Sun, and Z. Chen, "Cqarank: Jointly model topics and expertise in community question answering," in *Proceedings of the 22Nd ACM International Conference on Conference on Information &#38; Knowledge Management*, ser. CIKM '13. New York, NY, USA: ACM, 2013. [Online]. Available: http://doi.acm.org/10.1145/2505515.2505720 pp. 99–108.

[93] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si, “Expertise retrieval,” *Found. Trends Inf. Retr.*, vol. 6, no. 2&#8211;3, pp. 127–256, Feb. 2012. [Online]. Available: http://dx.doi.org/10.1561/1500000024

[94] K. Balog, L. Azzopardi, and M. de Rijke, “Formal models for expert finding in enterprise corpora,” in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’06. New York, NY, USA: ACM, 2006. [Online]. Available: http://doi.acm.org/10.1145/1148170.1148181 pp. 43–50.

[95] D. Petkova and W. Croft, “Hierarchical language models for expert finding in enterprise corpora,” in *Tools with Artificial Intelligence, 2006. ICTAI ’06. 18th IEEE International Conference on*, Nov 2006, pp. 599–608.

[96] Y. Cao, H. Duan, C.-Y. Lin, Y. Yu, and H.-W. Hon, “Recommending questions using the mdl-based tree cut model,” in *Proceedings of the 17th International Conference on World Wide Web*, ser. WWW ’08. New York, NY, USA: ACM, 2008. [Online]. Available: http://doi.acm.org/10.1145/1367497.1367509 pp. 81–90.

[97] A. Singh, D. P, and D. Raghu, “Retrieving similar discussion forum threads: A structure based approach,” in *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’12. New York, NY, USA: ACM, 2012. [Online]. Available: http://doi.acm.org/10.1145/2348283.2348305 pp. 135–144.

[98] H. Duan and C. Zhai, “Exploiting thread structures to improve smoothing of language models for forum post retrieval,” in *Proceedings of the 33rd European Conference on Advances in Information Retrieval*, ser. ECIR’11. Berlin, Heidelberg: Springer-Verlag, 2011. [Online]. Available: http://dl.acm.org/citation.cfm?id=1996889.1996935 pp. 350–361.

[99] Y. Zhou, G. Cong, B. Cui, C. S. Jensen, and J. Yao, “Routing questions to the right users in online communities,” in *Data Engineering, 2009. ICDE ’09. IEEE 25th International Conference on*, March 2009, pp. 700–711.

[100] S. Chang and A. Pal, “Routing questions for collaborative answering in community question answering,” in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM ’13. New York, NY, USA: ACM, 2013. [Online]. Available: http://doi.acm.org/10.1145/2492517.2492559 pp. 494–501.

[101] M. Liu, Y. Liu, and Q. Yang, “Predicting best answerers for new questions in community question answering,” in *Proceedings of the 11th International Conference on Web-age Information Management*, ser. WAIM’10. Berlin, Heidelberg: Springer-Verlag, 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1884017.1884036 pp. 127–138.

[102] T. C. Zhou, M. R. Lyu, and I. King, "A classification-based approach to question routing in community question answering," in *Proceedings of the 21st International Conference Companion on World Wide Web*, ser. WWW '12 Companion. New York, NY, USA: ACM, 2012. [Online]. Available: http://doi.acm.org/10.1145/2187980.2188201 pp. 783–790.

[103] G. Burel, P. Mulholland, Y. He, and H. Alani, "Predicting answering behaviour in online question answering communities," in *Proceedings of the 26th ACM Conference on Hypertext &#38; Social Media*, ser. HT '15. New York, NY, USA: ACM, 2015. [Online]. Available: http://doi.acm.org/10.1145/2700171.2791041 pp. 201–210.

[104] J. Noel, S. Sanner, K.-N. Tran, P. Christen, L. Xie, E. V. Bonilla, E. Abbasnejad, and N. Della Penna, "New objective functions for social collaborative filtering," in *Proceedings of the 21st International Conference on World Wide Web*, ser. WWW '12. New York, NY, USA: ACM, 2012. [Online]. Available: http://doi.acm.org/10.1145/2187836.2187952 pp. 859–868.

[105] A. P. Singh and G. J. Gordon, "Relational learning via collective matrix factorization," in *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '08. New York, NY, USA: ACM, 2008. [Online]. Available: http://doi.acm.org/10.1145/1401890.1401969 pp. 650–658.

[106] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King, "Recommender systems with social regularization," in *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, ser. WSDM '11. New York, NY, USA: ACM, 2011. [Online]. Available: http://doi.acm.org/10.1145/1935826.1935877 pp. 287–296.

[107] "Hcahps," https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-instruments/hospitalqualityinits/hospitalHCAHPS.html, accessed: 2019-01-30.

[108] J. H. Cho, M. Girju, and R. Girju, "Discovering semantic topics in patient feedback: Making sense of patients' free-text comments," *International Journal on Natural Language Computing (in submission)*, 2019.

[109] S. J. Russell and P. Norvig, *Artificial Intelligence - A Modern Approach (3. internat. ed.).* Pearson Education, 2010. [Online]. Available: http://vig.pearsoned.com/store/product/1,1207,store-12521_isbn-0136042597,00.html

[110] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*, ser. Adaptive computation and machine learning. MIT Press, 2012. [Online]. Available: http://mitpress.mit.edu/books/foundations-machine-learning-0

[111] Y. Gong and Q. Zhang, "Hashtag recommendation using attention-based convolutional neural network," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, ser. IJCAI'16. AAAI Press, 2016. [Online]. Available: http://dl.acm.org/citation.cfm?id=3060832.3061010 pp. 2782–2788.

[112] D. Preoţiuc-Pietro, Y. Liu, D. Hopkins, and L. Ungar, "Beyond binary labels: Political ideology prediction of twitter users," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Vancouver, Canada: Association for Computational Linguistics, July 2017. [Online]. Available: https://www.aclweb.org/anthology/P17-1068 pp. 729–740.

[113] X. Wang, F. Wei, X. Liu, M. Zhou, and M. Zhang, "Topic sentiment analysis in twitter: A graph-based hashtag sentiment classification approach," in *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, ser. CIKM '11. New York, NY, USA: ACM, 2011. [Online]. Available: http://doi.acm.org/10.1145/2063576.2063726 pp. 1031–1040.

[114] F. Kunneman, C. Liebrecht, and A. van den Bosch, "The (un)predictability of emotional hashtags in twitter," in *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM).* Gothenburg, Sweden: Association for Computational Linguistics, Apr. 2014. [Online]. Available: https://www.aclweb.org/anthology/W14-1304 pp. 26–34.

[115] U. S. T. T. Byron Wallace, Michael Paul and M. Dredze, "A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews," *Journal of the American Medical Informatics Association: JAMIA*, vol. 21, no. 6, pp. 39–41, Nov. 2014.

[116] G. G. Gao, J. S. McCullough, R. Agarwal, and A. K. Jha, "A changing landscape of physician quality reporting: analysis of patients' online ratings of their physicians over a 5-year period," *J. Med. Internet Res.*, vol. 14, no. 1, p. e38, Feb 2012.

[117] R. Artstein and M. Poesio, "Inter-coder agreement for Computational Linguistics," *Computational Linguistics*, vol. 34, no. 4, 2008.

[118] A. Lopez, A. Detz, N. Ratanawongsa, and U. Sarkar, "What patients say about their doctors online: a qualitative content analysis," *J Gen Intern Med*, vol. 27, no. 6, pp. 685–692, Jun 2012.

[119] K. Doing-Harris, D. L. Mowery, C. Daniels, W. W. Chapman, and M. Conway, "Understanding patient satisfaction with received healthcare services: A natural language processing approach," *AMIA Annu Symp Proc*, vol. 2016, pp. 524–533, 2016.

[120] J. Segal, "The role of the Internet in doctor performance rating," *Pain Physician*, vol. 12, no. 3, pp. 659–664, 2009.

[121] D. Strech, "Ethical principles for physician rating sites," *J. Med. Internet Res.*, vol. 13, no. 4, p. e113, Dec 2011.

[122] Y. Lin, T. Zhu, H. Wu, J. Zhang, X. Wang, and A. Zhou, "Towards online anti-opinion spam: Spotting fake reviews from the review sequence," in *Proceedings of the 2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '14. Piscataway, NJ, USA: IEEE Press, 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=3191835.3191887 pp. 261–264.

[123] J. K. Rout, A. Dalmia, K. R. Choo, S. Bakshi, and S. K. Jena, “Revisiting semi-supervised learning for online deceptive review detection,” *IEEE Access*, vol. 5, pp. 1319–1327, 2017. [Online]. Available: https://doi.org/10.1109/ACCESS.2017.2655032

[124] A. Pasternak and J. E. Scherger, “Online reviews of physicians: what are your patients posting about you?” *Fam Pract Manag*, vol. 16, no. 3, pp. 9–11, 2009.

[125] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, “An efficient k-means clustering algorithm: Analysis and implementation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, July 2002. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2002.1017616

[126] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003. [Online]. Available: http://dl.acm.org/citation.cfm?id=944919.944937

[127] G. A. Miller, “Wordnet: A lexical database for english,” *Commun. ACM*, vol. 38, no. 11, pp. 39–41, Nov. 1995. [Online]. Available: http://doi.acm.org/10.1145/219717.219748

[128] K. P. Murphy, *Machine learning: a probabilistic perspective*, Cambridge, MA, 2012.

[129] D. Chen and C. D. Manning, “A fast and accurate dependency parser using neural networks,” 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*, 2014. [Online]. Available: http://aclweb.org/anthology/D/D14/D14-1082.pdf pp. 740–750.

[130] D. Yarowsky, “Unsupervised word sense disambiguation rivaling supervised methods,” in *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ser. ACL ’95. Stroudsburg, PA, USA: Association for Computational Linguistics, 1995. [Online]. Available: https://doi.org/10.3115/981658.981684 pp. 189–196.

[131] M. Hu and B. Liu, “Mining and summarizing customer reviews,” in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD ’04. New York, NY, USA: ACM, 2004. [Online]. Available: http://doi.acm.org/10.1145/1014052.1014073 pp. 168–177.

[132] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *CoRR*, vol. abs/1301.3781, 2013. [Online]. Available: http://arxiv.org/abs/1301.3781

[133] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” *CoRR*, vol. abs/1310.4546, 2013. [Online]. Available: http://arxiv.org/abs/1310.4546

[134] Y. Yu, X. Wan, and X. Zhou, “User embedding for scholarly microblog recommendation,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Berlin, Germany: Association for Computational Linguistics, August 2016. [Online]. Available: http://anthology.aclweb.org/P16-2073 pp. 449–453.

[135] O. Levy and Y. Goldberg, “Dependency-based word embeddings,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Baltimore, Maryland: Association for Computational Linguistics, June 2014. [Online]. Available: http://www.aclweb.org/anthology/P14-2050 pp. 302–308.

[136] Q. V. Le and T. Mikolov, “Distributed representations of sentences and documents,” in *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, 2014. [Online]. Available: http://jmlr.org/proceedings/papers/v32/le14.html pp. 1188–1196.

[137] A. Herbelot and M. Baroni, “High-risk learning: acquiring new word vectors from tiny data,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, September 2017. [Online]. Available: https://www.aclweb.org/anthology/D17-1030 pp. 304–309.

[138] O. Levy and Y. Goldberg, “Dependency-based word embeddings,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, 2014. [Online]. Available: http://aclweb.org/anthology/P/P14/P14-2050.pdf pp. 302–308.

[139] A. Komninos and S. Manandhar, “Dependency based embeddings for sentence classification tasks,” in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, California: Association for Computational Linguistics, June 2016. [Online]. Available: http://www.aclweb.org/anthology/N16-1175 pp. 1490–1500.

[140] J. E. Nguyen Xuan Vinh and J. Bailey, “Information theoretic measures for clustering comparison: Is a correction for chance necessary?” *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*, pp. 1073–1080, 2009.

[141] S. Romano, N. X. Vinh, J. Bailey, and K. Verspoor, “Adjusting for chance clustering comparison measures,” *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 4635–4666, Jan. 2016. [Online]. Available: http://dl.acm.org/citation.cfm?id=2946645.3007087

[142] P. Ranjan, A. Kumari, and A. Chakrawarty, “How can Doctors Improve their Communication Skills?” *J Clin Diagn Res*, vol. 9, no. 3, pp. 01–4, Mar 2015.

[143] F. M. Lachmann, *Transforming Narcissism: Reflections on Empathy, Humor, and Expectations*. New York, NY, USA: Taylor & Francis Group, 2007.

[144] H. Wang, Y. Lu, and C. Zhai, "Latent aspect rating analysis on review text data: A rating regression approach," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '10. New York, NY, USA: ACM, 2010. [Online]. Available: http://doi.acm.org/10.1145/1835804.1835903 pp. 783–792.

[145] J. I. Merlino, C. Kestranek, D. Bokar, Z. Sun, S. E. Nissen, and D. L. Longworth, "HCAHPS Survey Results: Impact of Severity of Illness on Hospitals' Performance on HCAHPS Survey Results," *J Patient Exp*, vol. 1, no. 2, pp. 16–21, Nov 2014.

[146] L. Tefera, W. G. Lehrman, and P. Conway, "Measurement of the Patient Experience: Clarifying Facts, Myths, and Approaches," *JAMA*, vol. 315, no. 20, pp. 2167–2168, 2016.

[147] S. J. Mehta, "Patient Satisfaction Reporting and Its Implications for Patient Care," *AMA J Ethics*, vol. 17, no. 7, pp. 616–621, Jul 2015.

[148] M. N. Elliott, W. G. Lehrman, E. Goldstein, K. Hambarsoomian, M. K. Beckett, and L. A. Giordano, "Do hospitals rank differently on HCAHPS for different patient subgroups?" *Med Care Res Rev*, vol. 67, no. 1, pp. 56–73, Feb 2010.

[149] L. S. Morales, M. N. Elliott, R. Weech-Maldonado, K. L. Spritzer, and R. D. Hays, "Differences in CAHPS adult survey reports and ratings by race and ethnicity: an analysis of the National CAHPS benchmarking data 1.0," *Health Serv Res*, vol. 36, no. 3, pp. 595–617, Jul 2001.

[150] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.*, 2012. [Online]. Available: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks pp. 1106–1114.

[151] C. Yeh, W. Wu, W. Ko, and Y. F. Wang, "Learning deep latent space for multi-label classification," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, 2017. [Online]. Available: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14166 pp. 2838–2844.

[152] A. Joulin, E. Grave, P. Bojanowski, and T. Mikolov, "Bag of tricks for efficient text classification," *CoRR*, vol. abs/1607.01759, 2016. [Online]. Available: http://arxiv.org/abs/1607.01759

[153] J. Nam, J. Kim, E. L. Mencía, I. Gurevych, and J. Fürnkranz, "Large-scale multi-label text classification — revisiting neural networks," in *Proceedings of the 2014th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II*, ser. ECMLPKDD'14. Berlin, Heidelberg: Springer-Verlag, 2014. [Online]. Available: https://doi.org/10.1007/978-3-662-44851-9_28 pp. 437–452.

[154] M. Zhang and Z. Zhou, "Multi-label neural networks with applications to functional genomics and text categorization," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 10, pp. 1338–1351, 2006. [Online]. Available: https://doi.org/10.1109/TKDE.2006.162

[155] B. C. Wallace, M. J. Paul, U. Sarkar, T. A. Trikalinos, and M. Dredze, "A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews," *JAMIA*, vol. 21, no. 6, pp. 1098–1103, 2014. [Online]. Available: https://doi.org/10.1136/amiajnl-2014-002711

[156] F. Godin, V. Slavkovikj, W. De Neve, B. Schrauwen, and R. Van de Walle, "Using topic models for twitter hashtag recommendation," in *Proceedings of the 22Nd International Conference on World Wide Web*, ser. WWW '13 Companion. New York, NY, USA: ACM, 2013. [Online]. Available: http://doi.acm.org/10.1145/2487788.2488002 pp. 593–596.

[157] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. P. Vlahavas, "Multi-label classification of music into emotions," in *ISMIR 2008, 9th International Conference on Music Information Retrieval, Drexel University, Philadelphia, PA, USA, September 14-18, 2008*, 2008. [Online]. Available: http://ismir2008.ismir.net/papers/ISMIR2008_275.pdf pp. 325–330.

[158] G. Tsoumakas, I. Katakis, and I. P. Vlahavas, "Random k-labelsets for multilabel classification," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 7, pp. 1079–1089, 2011. [Online]. Available: https://doi.org/10.1109/TKDE.2010.164

[159] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "Cnn-rnn: A unified framework for multi-label image classification," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[160] K. Bhatia, H. Jain, P. Kar, M. Varma, and P. Jain, "Sparse local embeddings for extreme multi-label classification," in *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, pp. 730–738. [Online]. Available: http://papers.nips.cc/paper/5969-sparse-local-embeddings-for-extreme-multi-label-classification.pdf

[161] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine Learning*, vol. 85, no. 3, pp. 333–359, 2011. [Online]. Available: https://doi.org/10.1007/s10994-011-5256-5

[162] P. Yang, X. Sun, W. Li, S. Ma, W. Wu, and H. Wang, "SGM: sequence generation model for multi-label classification," in *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, 2018, pp. 3915–3926.

[163] K. Dembczynski, W. Waegeman, W. Cheng, and E. Hüllermeier, "On label dependence and loss minimization in multi-label classification," *Machine Learning*, vol. 88, no. 1-2, pp. 5–45, 2012. [Online]. Available: https://doi.org/10.1007/s10994-012-5285-8

[164] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-task CNN model for attribute prediction," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1949–1959, 2015. [Online]. Available: https://doi.org/10.1109/TMM.2015.2477680

[165] L. Chekina, D. Gutfreund, A. Kontorovich, L. Rokach, and B. Shapira, "Exploiting label dependencies for improved sample complexity," *Machine Learning*, vol. 91, no. 1, pp. 1–42, 2013. [Online]. Available: https://doi.org/10.1007/s10994-012-5312-9

[166] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 695–711.

[167] Y. Liu, R. Jin, and L. Yang, "Semi-supervised multi-label learning by constrained non-negative matrix factorization," in *Proceedings, The Twenty-First National Conference on Artificial Intelligence and the Eighteenth Innovative Applications of Artificial Intelligence Conference, July 16-20, 2006, Boston, Massachusetts, USA*, 2006. [Online]. Available: http://www.aaai.org/Library/AAAI/2006/aaai06-067.php pp. 421–426.

[168] L. Zhang, S. Shah, and I. Kakadiaris, "Hierarchical multi-label classification using fully associative ensemble learning," *Pattern Recogn.*, vol. 70, no. C, pp. 89–103, Oct. 2017. [Online]. Available: https://doi.org/10.1016/j.patcog.2017.05.007

[169] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ser. ICML'13. JMLR.org, 2013. [Online]. Available: http://dl.acm.org/citation.cfm?id=3042817.3043076 pp. III–1247–III–1255.

[170] R. Babbar, I. Partalas, E. Gaussier, and M.-R. Amini, "On flat versus hierarchical classification in large-scale taxonomies," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'13. USA: Curran Associates Inc., 2013. [Online]. Available: http://dl.acm.org/citation.cfm?id=2999792.2999816 pp. 1824–1832.

[171] T. Fagni and F. Sebastiani, "Selecting negative examples for hierarchical text classification: An experimental comparison," *JASIST*, vol. 61, no. 11, pp. 2256–2265, 2010. [Online]. Available: https://doi.org/10.1002/asi.21411

[172] R. Eisner, B. Poulin, D. Szafron, P. Lu, and R. Greiner, "Improving protein function prediction using the hierarchical structure of the gene ontology," in *Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2005, Embassy Suites Hotel La Jolla, La Jolla, CA, USA, November 14 & 15, 2005*, 2005, pp. 354–363.

[173] M. Ceci and D. Malerba, "Classifying web documents in a hierarchy of categories: a comprehensive study," *J. Intell. Inf. Syst.*, vol. 28, no. 1, pp. 37–78, 2007. [Online]. Available: https://doi.org/10.1007/s10844-006-0003-2

[174] C. DeCoro, Z. Barutçuoglu, and R. Fiebrink, “Bayesian aggregation for hierarchical genre classification,” in *Proceedings of the 8th International Conference on Music Information Retrieval, ISMIR 2007, Vienna, Austria, September 23-27, 2007*, 2007. [Online]. Available: http://ismir2007.ismir.net/proceedings/ISMIR2007_p077_decoro.pdf pp. 77–80.

[175] P. N. Bennett and N. Nguyen, “Refined experts: improving classification in large taxonomies,” in *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009, Boston, MA, USA, July 19-23, 2009*, 2009. [Online]. Available: https://doi.org/10.1145/1571941.1571946 pp. 11–18.

[176] A. Sun and E. Lim, “Hierarchical text classification and evaluation,” in *Proceedings of the 2001 IEEE International Conference on Data Mining, 29 November - 2 December 2001, San Jose, California, USA*, 2001. [Online]. Available: https://doi.org/10.1109/ICDM.2001.989560 pp. 521–528.

[177] P.-Y. Hao, J.-H. Chiang, and Y.-K. Tu, “Hierarchically svm classification based on support vector clustering method and its application to document categorization,” *Expert Syst. Appl.*, vol. 33, no. 3, pp. 627–635, Oct. 2007. [Online]. Available: http://dx.doi.org/10.1016/j.eswa.2006.06.009

[178] C. Vens, J. Struyf, L. Schietgat, S. Dzeroski, and H. Blockeel, “Decision trees for hierarchical multi-label classification,” *Machine Learning*, vol. 73, no. 2, pp. 185–214, 2008. [Online]. Available: https://doi.org/10.1007/s10994-008-5077-3

[179] K. Kowsari, D. E. Brown, M. Heidarysafa, K. J. Meimandi, M. S. Gerber, and L. E. Barnes, “Hdltex: Hierarchical deep learning for text classification,” in *16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017, Cancun, Mexico, December 18-21, 2017*, 2017. [Online]. Available: https://doi.org/10.1109/ICMLA.2017.0-134 pp. 364–371.

[180] R. Salakhutdinov, J. B. Tenenbaum, and A. Torralba, “Learning with hierarchical-deep models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1958–1971, Aug. 2013. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2012.269

[181] D. Roy, P. Panda, and K. Roy, “Tree-cnn: A deep convolutional neural network for lifelong learning,” *CoRR*, vol. abs/1802.05800, 2018. [Online]. Available: http://arxiv.org/abs/1802.05800

[182] Z. Yan, H. Zhang, R. Piramuthu, V. Jagadeesh, D. DeCoste, W. Di, and Y. Yu, “HD-CNN: hierarchical deep convolutional neural networks for large scale visual recognition,” in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, 2015. [Online]. Available: https://doi.org/10.1109/ICCV.2015.314 pp. 2740–2748.

[183] J. Fan, T. Zhao, Z. Kuang, Y. Zheng, J. Zhang, J. Yu, and J. Peng, "HD-MTL: hierarchical deep multi-task learning for large-scale visual recognition," *IEEE Trans. Image Processing*, vol. 26, no. 4, pp. 1923–1938, 2017. [Online]. Available: https://doi.org/10.1109/TIP.2017.2667405

[184] G. Chen, D. Ye, Z. Xing, J. Chen, and E. Cambria, "Ensemble application of convolutional and recurrent neural networks for multi-label text categorization," in *2017 International Joint Conference on Neural Networks, IJCNN 2017, Anchorage, AK, USA, May 14-19, 2017*, 2017. [Online]. Available: https://doi.org/10.1109/IJCNN.2017.7966144 pp. 2377–2383.

[185] J. H. D. Cho, T. Gao, and R. Girju, "Identifying medications that patients stopped taking in online health forums," in *11th IEEE International Conference on Semantic Computing, ICSC 2017, San Diego, CA, USA, January 30 - February 1, 2017*, 2017. [Online]. Available: https://doi.org/10.1109/ICSC.2017.24 pp. 141–148.

[186] W. W. Fleischhacker, M. A. Oehl, and M. Hummer, "Factors influencing compliance in schizophrenia patients," *J Clin Psychiatry*, vol. 64 Suppl 16, pp. 10–13, 2003.

[187] H. Rittmannsberger, T. Pachinger, P. Keppelmuller, and J. Wancata, "Medication adherence among psychotic patients before admission to inpatient treatment," *Psychiatr Serv*, vol. 55, no. 2, pp. 174–179, Feb 2004.

[188] M. R. DiMatteo, "Variations in patients' adherence to medical recommendations: a quantitative review of 50 years of research," *Med Care*, vol. 42, no. 3, pp. 200–209, Mar 2004.

[189] D. Roe, H. Goldblatt, V. Baloush-Klienman, M. Swarbrick, and L. Davidson, "Why and how people decide to stop taking prescribed psychiatric medication: exploring the subjective process of choice," *Psychiatr Rehabil J*, vol. 33, no. 1, pp. 38–46, 2009.

[190] A. Signorini, A. M. Segre, and P. M. Polgreen, "The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic," *PLoS ONE*, vol. 6, no. 5, p. e19467, 2011.

[191] L. Andrade, J. J. Caraveo-Anduaga, P. Berglund, R. V. Bijl, R. De Graaf, W. Vollebergh, E. Dragomirecka, R. Kohn, M. Keller, R. C. Kessler, N. Kawakami, C. Kilic, D. Offord, T. B. Ustun, and H. U. Wittchen, "The epidemiology of major depressive episodes: results from the International Consortium of Psychiatric Epidemiology (ICPE) Surveys," *Int J Methods Psychiatr Res*, vol. 12, no. 1, pp. 3–21, 2003.

[192] P. J. McCormick, N. Elhadad, and P. D. Stetson, "Use of semantic features to classify patient smoking status," *AMIA Annu Symp Proc*, pp. 450–454, 2008.

[193] E. Aramaki, T. Imai, K. Miyo, and K. Ohe, "Patient Status Classification by Using Rule based Sentence Extraction and BM25 kNN-based Classifier," *Processing for Clinical Data*, 2006.

[194] A. M. Cohen, "Five-way smoking status classification using text hot-spot identification and error-correcting output codes," *J Am Med Inform Assoc*, vol. 15, no. 1, pp. 32–35, 2008.

[195] R. Wicentowski and M. R. Sydes, "Using implicit information to identify smoking status in smoke-blind medical discharge summaries," *J Am Med Inform Assoc*, vol. 15, no. 1, pp. 29–31, 2008.

[196] C. Clark, K. Good, L. Jezierny, M. Macpherson, B. Wilson, and U. Chajewska, "Identifying smokers with a medical extraction system," *J Am Med Inform Assoc*, vol. 15, no. 1, pp. 36–39, 2008.

[197] E. Cambria, D. Olsher, and D. Rajagopal, "Senticnet 3: A common and common-sense knowledge base for cognition-driven sentiment analysis," in *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada.*, 2014, pp. 1515–1521.

[198] D. Chen and C. Manning, "A fast and accurate dependency parser using neural networks," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, October 2014, pp. 740–750.

[199] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M. Hsu, "Prefixspan: Mining sequential patterns by prefix-projected growth," in *Proceedings of the 17th International Conference on Data Engineering, April 2-6, 2001, Heidelberg, Germany*, 2001, pp. 215–224.

[200] E. Bengtson and D. Roth, "Understanding the value of features for coreference resolution," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008, pp. 294–303.

[201] J. Li and C. Cardie, "Timeline generation: Tracking individuals on twitter," in *Proceedings of the 23rd International Conference on World Wide Web*, ser. WWW '14. New York, NY, USA: ACM, 2014, pp. 643–652.

[202] Z. Chen and H. Ji, "Language specific issue and feature exploration in chinese event extraction," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, ser. NAACL-Short '09. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009, pp. 209–212.

[203] C. Homburg, M. Grozdanovic, and M. Klarmann, "Responsiveness to customers and competitors: The role of affective and cognitive organizational systems," *Journal of Marketing*, vol. 71, no. 3, pp. 18–38, 2007. [Online]. Available: http://www.jstor.org/stable/30163979

[204] A. Mostaghimi, B. H. Crotty, and B. E. Landon, “The availability and nature of physician information on the internet,” *J Gen Intern Med*, vol. 25, no. 11, pp. 1152–1156, Nov 2010.

[205] J. H. D. Cho, Y. Li, R. Girju, and C. Zhai, “Recommending forum posts to designated experts,” in *2015 IEEE International Conference on Big Data, Big Data 2015, Santa Clara, CA, USA, October 29 - November 1, 2015*, 2015. [Online]. Available: https://doi.org/10.1109/BigData.2015.7363810 pp. 659–666.

[206] J. Jeon, W. B. Croft, J. H. Lee, and S. Park, “A framework to predict the quality of answers with non-textual features,” in *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’06. New York, NY, USA: ACM, 2006. [Online]. Available: http://doi.acm.org/10.1145/1148170.1148212 pp. 228–235.

[207] B. Li and I. King, “Routing questions to appropriate answerers in community question answering services,” in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ser. CIKM ’10. New York, NY, USA: ACM, 2010. [Online]. Available: http://doi.acm.org/10.1145/1871437.1871678 pp. 1585–1588.

[208] J. Huh, D. W. McDonald, A. Hartzler, and W. Pratt, “Patient moderator interaction in online health communities,” *AMIA Annu Symp Proc*, vol. 2013, pp. 627–636, 2013.

[209] J. H. D. Cho, P. Sondhi, C. Zhai, and B. R. Schatz, “Resolving healthcare forum posts via similar thread retrieval,” in *Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, ser. BCB ’14. New York, NY, USA: ACM, 2014. [Online]. Available: http://doi.acm.org/10.1145/2649387.2649399 pp. 33–42.

[210] F. Bach, J. Abernethy, J.-P. Vert, and T. Evgeniou, “A new approach to collaborative filtering: Operator estimation with spectral regularization,” *CoRR*, vol. abs/0802.1430, 2008.

[211] R. Pan, Y. Zhou, B. Cao, N. Liu, R. Lukose, M. Scholz, and Q. Yang, “One-class collaborative filtering,” in *Data Mining, 2008. ICDM ’08. Eighth IEEE International Conference on*, Dec 2008, pp. 502–511.

[212] Y. Shi, M. Larson, and A. Hanjalic, “Mining mood-specific movie similarity with matrix factorization for context-aware recommendation,” in *Proceedings of the Workshop on Context-Aware Movie Recommendation*, ser. CAMRa ’10. New York, NY, USA: ACM, 2010. [Online]. Available: http://doi.acm.org/10.1145/1869652.1869658 pp. 34–40.

[213] L. Baltrunas, B. Ludwig, and F. Ricci, “Matrix factorization techniques for context aware recommendation,” in *Proceedings of the Fifth ACM Conference on Recommender Systems*, ser. RecSys ’11. New York, NY, USA: ACM, 2011. [Online]. Available: http://doi.acm.org/10.1145/2043932.2043988 pp. 301–304.

[214] Y. Li, J. Hu, C. Zhai, and Y. Chen, “Improving one-class collaborative filtering by incorporating rich user information,” in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ser. CIKM ’10. New York, NY, USA: ACM, 2010. [Online]. Available: http://doi.acm.org/10.1145/1871437.1871559 pp. 959–968.

[215] S.-H. Yang, B. Long, A. Smola, N. Sadagopan, Z. Zheng, and H. Zha, “Like like alike: Joint friendship and interest propagation in social networks,” in *Proceedings of the 20th International Conference on World Wide Web*, ser. WWW ’11. New York, NY, USA: ACM, 2011. [Online]. Available: http://doi.acm.org/10.1145/1963405.1963481 pp. 537–546.

[216] P. Cui, F. Wang, S. Liu, M. Ou, S. Yang, and L. Sun, “Who should share what?: Item-level social influence prediction for users and posts ranking,” in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’11. New York, NY, USA: ACM, 2011. [Online]. Available: http://doi.acm.org/10.1145/2009916.2009945 pp. 185–194.

[217] J. Han, *Data Mining: Concepts and Techniques.* San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005.

[218] R. Karp, “Reducibility among combinatorial problems,” in *Complexity of Computer Computations*, R. Miller and J. Thatcher, Eds. Plenum Press, 1972, pp. 85–103.

[219] L. Bottou, “Large-scale machine learning with stochastic gradient descent,” in *Proceedings of COMPSTAT’2010*, Y. Lechevallier and G. Saporta, Eds. Physica-Verlag HD, 2010, pp. 177–186. [Online]. Available: http://dx.doi.org/10.1007/978-3-7908-2604-3_16

[220] D. D. Lee and H. S. Seung, “Algorithms for non-negative matrix factorization,” in *NIPS*, 2000. [Online]. Available: citeseer.ist.psu.edu/lee01algorithms.html pp. 556–562.

[221] Q. Yuan, L. Chen, and S. Zhao, “Factorization vs. regularization: Fusing heterogeneous social relationships in top-n recommendation,” in *Proceedings of the Fifth ACM Conference on Recommender Systems*, ser. RecSys ’11. New York, NY, USA: ACM, 2011. [Online]. Available: http://doi.acm.org/10.1145/2043932.2043975 pp. 245–252.

[222] S. Scott and S. Matwin, “Text classification using WordNet hypernyms,” in *Usage of WordNet in Natural Language Processing Systems*, 1998. [Online]. Available: https://www.aclweb.org/anthology/W98-0706

[223] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, “Short text classification in twitter to improve information filtering,” in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR ’10. New York, NY, USA: ACM, 2010. [Online]. Available: http://doi.acm.org/10.1145/1835449.1835643 pp. 841–842.

[224] 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*, ser. RecSys ’13. New York, NY, USA: ACM, 2013. [Online]. Available: http://doi.acm.org/10.1145/2507157.2507163 pp. 165–172.

[225] D. Roth and W.-t. Yih, “A linear programming formulation for global inference in natural language tasks,” in *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*. Boston, Massachusetts, USA: Association for Computational Linguistics, May 6 - May 7 2004. [Online]. Available: https://www.aclweb.org/anthology/W04-2401 pp. 1–8.

[226] I. Korkontzelos, A. Nikfarjam, M. Shardlow, A. Sarker, S. Ananiadou, and G. H. Gonzalez, “Analysis of the effect of sentiment analysis on extracting adverse drug reactions from tweets and forum posts,” *J. of Biomedical Informatics*, vol. 62, no. C, pp. 148–158, Aug. 2016. [Online]. Available: https://doi.org/10.1016/j.jbi.2016.06.007

[227] Y. Peng, M. Moh, and T. Moh, “Efficient adverse drug event extraction using twitter sentiment analysis,” in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2016, San Francisco, CA, USA, August 18-21, 2016*, 2016. [Online]. Available: https://doi.org/10.1109/ASONAM.2016.7752365 pp. 1011–1018.