Computational Approaches to Characterizing Online Health Communities

Shaodian Zhang

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2016

ABSTRACT

Computational Approaches to Characterizing Online Health Communities

Shaodian Zhang

Online health communities (OHCs) have been increasingly popular among patients with chronic or life-threatening illnesses for the exchange of social support. Contemporary research of OHCs relies on methods and tools to handle analytics of massive user-generated content at scale to complement traditional qualitative analysis. In this thesis, we aim at advancing the area of research by providing computational tools and methods which facilitate automated content analysis, and by presenting applications of these tools to investigating member characteristics and behaviors.

We first provide a framework of conceptualization to systematically describe problems, challenges, and existing solutions for OHCs from a social support standpoint, to bridge the knowledge gap between health psychology and informatics. With this framework in hand, we define the landscape of online social support, summarize current research progress of OHCs, and identify research questions to investigate for this thesis.

We then build a series of computational tools for analyzing OHC content, relying on techniques of machine learning and natural language processing. Leveraging domain-specific features, our tools are tailored to handle content analysis tasks on OHC text effectively.

Equipped with computational tools, we demonstrate how characteristics of OHC members can be identified at scale in an automated fashion. In particular, we build up multi-dimensional descriptions for patient members, consisting of what topics they focus on, what sentiment they express, and what treatments they discuss and adopt. Patterns of how these member characteristics change through time are also investigated longitudinally. Finally, relying on computational analytics, members' behaviors of engagement such as debate and dropping-out are identified and characterized.

Studies presented in this thesis discover static and longitudinal patterns of member characteristics and engagement, which are potential research hypotheses to be explored by health psychologists and clinical researchers. The thesis also contributes to the informatics community by making computational tools, lexicons, and annotated corpora available to facilitate future research.

Table of Contents

List of Figures

List of Tables

Acknowledgments

Foremost, I would like to thank my advisor Dr. Noemie Elhadad. For the past five years, Noemie has been as good as a Ph.D. advisor could realistically be. Before I entered DBMI, it was Noemie's eloquent plea that broke my dream of ending up in silicon valley as an engineer and made me decide to embrace the beauty of health data technology to

which I have so much sense of belonging today. In the past five years, it was my advisor who reshaped me from a technician who cared only about implementation to an authentic researcher who knows how to tell stories that make impact (a.k.a. accepted by journals). It was Noemie who granted me almost absolute freedom of arranging my Ph.D. life and study, and who sometimes canceled meetings at the last minute so that I can happily goof off on weekdays. Most importantly, throughout the five years it was my advisor who relentlessly corrected grammatical errors in my papers, especially singular vs plural issues that I always messed up since they did not exist in my native language or in my intuition. In retrospect, it was an exceptionally enjoyable but productive experience working with Noemie.

I am also fortunate to have worked or interacted with a superb committee and a number of outstanding colleagues. I would like to thank Dr. George Hripcsak, Dr. Suzanne Bakken, Dr. Jason Owen, and Dr. Mark Dredze for their insightful comments and suggestions. Without their contributions this dissertation would not have been possible. Dr. Jason Owen and Dr. Erin Bantum have been influential to my research from the perspective of health psychology. I really enjoyed all the collaborations with them. Dr. Suzanne Bakken has made indispensable contributions to frameworking my dissertation research, and her constructive criticism has helped me significantly improved my work. Sharon Lipsky Gorman, Frank Chen, Drashko Nikikj, Edouard Grave, Rimma Pivovarov, Adler Perotte, David Albers, Tiang Kang have all helped me or collaborated with me in exciting research projects. I would like to recognize their contributions to this dissertation.

Finally, I would like to thank my family and friends for their support. Especially, I would like to thank my wife who have given me a Ph.D. life with love and care.

Part I

Introduction, Framework, and Datasets

Chapter 1

Introduction and Specific Aims

1.1 Background and significance

1.1.1 Popularity of online health community

The Internet has revolutionized the way people seek and exchange health-related information. Pew Research reported that one third of American adults have gone online to research a medical condition, and 80% of Internet users have looked online for health-relevant information, indicating the Internet's increasing impact on health information consumption and health management [Fox and Duggan, 2013]. Traditionally, patients with chronic diseases like diabetes or life-threatening illnesses like cancer obtain information about their conditions primarily from their health care providers; but they are relying more and more on information from the Internet nowadays [Castleton *et al.*, 2011].

Aside from being an increasingly important source of information for patients, the Internet, particularly newly emerging web 2.0 applications such as blogs, forums, and social networks, are revolutionizing how patients exchange social support with care providers, family members, friends, and peer patients. As early as mid-1990s, researchers have created computer-mediated support groups for patients with cancer [Weinberg and Schmale, 1996]. The past two decades, in particular, have witnessed the flourishing of online mailing lists, blogs, forums for health purposes [Davison, 2000]. For example, Breastcancer.org, which was originally a platform for disseminating breast cancer knowledge, has been hosting a massive

discussion board for breast cancer patients and survivors with more than 150,000 registered members who have published more than 1 million posts of discussions [Wang *et al.*, ; Zhang *et al.*, 2014]. More recently, heterogeneous social network services have also become popular among patients. For example, Bender and colleagues reported that there were 620 Facebook groups for breast cancer in 2011, with rapidly increasing popularity and user activity [Bender *et al.*, 2011]. Facebook has also been used in several studies to improve patient communication, such as for weight loss surgery [Das and Faxvaag, 2014], physical activity intervention [Valle *et al.*, 2013], and breast cancer [Bender *et al.*, 2011]. PatientsLikeMe, an expanding platform integrating social networking with tailored health management and social support, is also becoming increasingly popular. In particular, its flagship amyotrophic lateral sclerosis (ALS) community has gathered the largest online population of ALS patients in the world [1]. We also notice that patients rely on Twitter to post health-related messages, which are investigated by epidemiologists and health researchers in studies for different purposes [Aramaki *et al.*, 2011; Hawn, 2009; Lamb *et al.*, 2013].

1.1.2 Significance of online health community research

Online health communities for different patient populations have been the subjects of research for years, for varied purposes such as creating social support interventions [Owen *et al.*, 2004a; Salzer *et al.*, 2010; Hø ybye, 2005], understanding patient behaviors [Wang *et al.*, 2015; Mamykina *et al.*, 2015; Hartzler and Pratt, 2011], assisting community facilitators [Huh *et al.*, 2013], finding critical disease- or medication-specific information [Portier *et al.*, 2013; Tuarob *et al.*, 2014], etc. Previous research suggested that one of patients' primary motivations of using online health communities (OHC) is to to exchange information, practical tips, and stories about their conditions and to get emotional support from their **peers** [Eysenbach *et al.*, 2004; Ziebland *et al.*, 2004; Das and Faxvaag, 2014; Magnezi *et al.*, 2014; Zhang *et al.*, 2014]. It was reported that patients cannot always access the information they need from health professionals [Hartzler and Pratt, 2011], and that sometimes information obtained from care providers can be patchy [Rozmovits and Ziebland, 2004].

[1] http://www.patientslikeme.com/conditions/9-als

On the contrary, peer patients are able to appreciate each other's conditions better than professionals, family members and friends. They are also better at providing necessary emotional support and practical advice of daily health management [Cohen *et al.*, 2000; Bender *et al.*, 2013]. For patients, using online communities to exchange peer support has no temporal [Sharf, 1997] or geographical [Nápoles-Springer *et al.*, 2007] restrictions; these communities are also more accessible for people with disabilities and psychological issues [Setoyama *et al.*, 2011].

One important question worth exploring about OHC is whether participation make positive impact on patients' psychological, social, or physical health, which was investigated in a number of previous studies. Some of them indeed found that participation in OHC produced positive social-support outcomes for patients [Gustafson and Hawkins, 2001; Børøsund *et al.*, 2014; Ruland *et al.*, 2013; Stanton *et al.*, 2013; Lieberman *et al.*, 2003; Winzelberg *et al.*, 2003], but some failed to discover a benefit [Owen *et al.*, 2005; Salzer *et al.*, 2010; Høybye *et al.*, 2010; Lepore *et al.*, 2014]. Traditionally, peer support for patients, especially the issue of its impact on health, has belonged to the realm of health psychology. Existing interventions through OHCs have been carried out in tight experimental setup with full control of the research settings, where researchers can access necessary subjects' information to answer research questions and identify outcomes [Zhang *et al.*, 2016a]. In these interventions, researchers usually follow principles from clinical research design, sample participants from patient populations, conduct randomized experiments, and carry out statistical analyses to examine effects.

1.1.3 Need for computational methods

In the big data era, particularly when studying OHCs that are open to the public (e.g. breast cancer forum or Facebook), researchers have opportunities to access much larger patient populations. Sometimes, subject of a OHC research study can be the entire user population with certain conditions from a massive online community (e.g. Facebook), which is unimaginable in traditional experimental studies [Wang *et al.*, ; Zhang *et al.*, 2014]. As such, contemporary OHC research requires novel informatics methods and tools to handle analytics of the massive data in a more effective way, to complement traditional manual

analysis. Specifically, as public OHCs are getting increasingly popular and are producing vast amount of peer-to-peer interaction content, this is an exciting time with previously unseen potential for advancement of OHC research to rely on sophisticated data-driven computational methods.

In the general domain, computational approaches including techniques from machine learning, natural language processing, data mining, and knowledge discovery, have been applied to analyses of various types of Internet content, including web pages [Liu, 2007], Wikipedia [Gabrilovich and Markovitch, 2007; Milne and Witten, 2008], social media [Liu *et al.*, 2011; Russell, 2013], etc. In the most recent decade, these techniques have been gradually transplanted to OHCs in studying their content, characteristics, user behaviors, and impact of participation. Some of the studies have shown much promise in identifying patterns at scale [Qiu *et al.*, 2011a; Wang *et al.*, ; Zhao *et al.*, 2012; Portier *et al.*, 2013; Zhang *et al.*, 2014; Wang *et al.*, 2015], which can be difficult with only traditional qualitative or manual approaches. However, OHCs are different from other types of online communities in many ways, making method transplantations challenging. For example, creators of communities for general purposes usually focus on activeness and popularity, while OHC creators emphasize critically on the quality of information and how OHCs actually impact members' physical and psychological health [Bouma *et al.*, 2015]. For another example, content in OHCs is usually highly domain- and community-specific with heavy usages of medical sub-language, creating additional challenges to content analysis [Elhadad *et al.*, 2014].

To overcome these difficulties, efforts are being made to develop computational solutions to facilitate OHC research [Qiu *et al.*, 2011a; Wang *et al.*, ; Zhao *et al.*, 2012; Portier *et al.*, 2013; Zhang *et al.*, 2014; Wang *et al.*, 2015]. However, until today there is still a gap between health researchers' and health psychologists' needs for more powerful tools to analyze massive OHC content, and existing progress in the informatics community to create methods tailored to OHC research. A theoretical framework bridging the knowledge from the two sides is also needed, so that computational efforts could be made toward solving psycho-social problems precisely and meaningfully.

In this thesis, we aim at advancing the area of research by providing computational

tools and methods which enable OHC research at scale. We first conceptualize a theoretical framework to systematically describe problems, challenges, and existing solutions of OHCs from a social support standpoint, to bridge the knowledge gap between health psychology and informatics communities. We then build up a series of computational tools for the analyses of OHC content, and demonstrate how these computational approaches can be leveraged to study characteristics of members, user behaviors, and possibly social support impact of OHCs. Specifically, we use computational methods to model individual members from different perspectives, and to investigate longitudinally factors that contribute to certain user behaviors such as dropping-out and debate. This thesis also contributes to the OHC research community by making computational tools, lexical resources, and annotated corpora available to facilitate future research.

1.2 Specific aims and research questions

Our work introduced in this thesis was approved by the Columbia University IRB office. In general, this thesis aims at providing computational tools, based on machine learning and natural language processing, tailored to analyzing online health community content, and at demonstrating how these tools can be leveraged to study characteristics and behaviors of members. The specific aims along with research questions described as follows are investigated in the thesis. Part II, Part III, and Part IV of this thesis will be focusing on these three specific aims, respectively.

1.2.1 Specific aim 1

Specific aim 1: Create computational resources and tools to automate the basic dimensions of large-scale analysis of online health communities, including lexicon creation, named entity recognition, topic classification, sentiment analysis, treatment attribution identification, and debate detection.

Our first specific aim is to provide computational tools for multiple content analysis tasks on online health community data, based on natural language processing (NLP) and machine learning. The methods analyze the content at semantic and pragmatic levels based

on well-established theories and methods of NLP and machine learning, but are heavily tailored to OHC text by leveraging knowledge bases and lexical resources. For example, we create a toolkit to collect domain-specific lexicons from online health community text in an unsupervised manner to support downstream applications. We also propose a supervised learning pipeline for pragmatic analyses, based on which we create tools for multiple tasks such as topic classification, sentiment analysis, and debate detection. In addition to providing such computational tools and evaluating their effectiveness, we also ask following research questions in this thesis:

> **Research question 1**: What is the impact of domain knowledge on both supervised and unsupervised approaches to content analysis in online health communities?
>
> **Research question 2**: What is the impact of feature representation (e.g., word embeddings vs bag of words) on the accuracy of supervised and unsupervised approaches to content analysis? How do syntactic and semantic features impact tools' performance?
>
> **Research question 3**: To which extent are the tools and approaches devised portable from one type of community to another (either communication style or disease)?

1.2.2 Specific aim 2

Specific aim 2: Use computational tools to model individual online health community members, including discovering their trajectories of topics of discussions, patterns in sentiment expressions, and treatment usages.

The second specific aim of this thesis is to rely on the tools created in specific aim 1 to automatically identify several main variables of interest with respect to OHC members, and to establish multi-dimensional characterizations for these members. This includes identifying topics of discussions of user posts, sentiment expressed by members, and treatments used and discussed by members. This specific aim also includes studying the changes of these variables longitudinally, and investigating correlations between these variables toward a multi-variant user modeling. Specifically, we ask following research questions:

Research question 1: What are the most prevalent topics of discussions in online health communities? What are the most prevalent topics stratified by users' self-reported disease profiles?

Research question 2: What do sentiment users express in discussions mostly? Are there any trajectories of sentiment changes from a longitudinal standpoint?

Research question 3: What are the attributions of mentions of treatments in OHC posts? Can we identify evidence of actual usage of drugs from these mentions?

1.2.3 Specific aim 3

Specific aim 3: Use computational methods to study member engagement with respect to community interactions, including detecting and characterizing debates among members, and studying how different factors contribute to user's decision of dropping-out.

The final aim of this thesis is to detect certain types of dynamics of community interactions and user behaviors, by leveraging computational approaches developed and member characteristics discovered in previous specific aims. Particularly, we are interested in how different member characteristics impacst users' behavior and decision making, including whether to participate or to withdraw, and if and how debates are triggered among members. Two main research questions are asked as follows:

Research question 1: Is it possible to detect the presence of debate in the discussion threads of an online health community? What are the topics that are more likely to trigger debates among community members?

Research question 2: How do factors such as interactions among community members (e.g. initializing discussion v.s. responding) and users' sentiment influence their own decisions regarding withdrawing participation?

Chapter 2

Synthesizing Current Online Health Community Research

In this chapter we describe how we synthesize previous research of online health communities from a social support standpoint. We propose a framework, which describes the landscape of social support and where online health community is situated, and summarizes research questions investigated. The framework will also be used to organize research questions investigated in this thesis.

2.1 A framework to conceptualize OHC research

The framework has two meta-layers illustrated in Figure 2.1 and is derived as follows [Zhang *et al.*, 2016a]. The upper layer (Conceptualization in Figure 1) synthesizes existing social support theories from [Friedman and Silver, 2007a; Wills, 1991] and identifies three major aspects of social support pertaining to the definition of online health communities.

The first sub-layer in conceptualization lists types of social support, which can be informational, emotional, or instrumental [Friedman and Silver, 2007a]. The second sub-layer represents sources of social support, which can be from lay persons in one's social network and from professional caregivers [Dennis, 2003]. The third sub-layer, setting of support, represents whether social support is exchanged online or offline, and types of online venues [Friedman and Silver, 2007a; Wills, 1991; Sharf, 1997]. It is noteworthy that the proposed

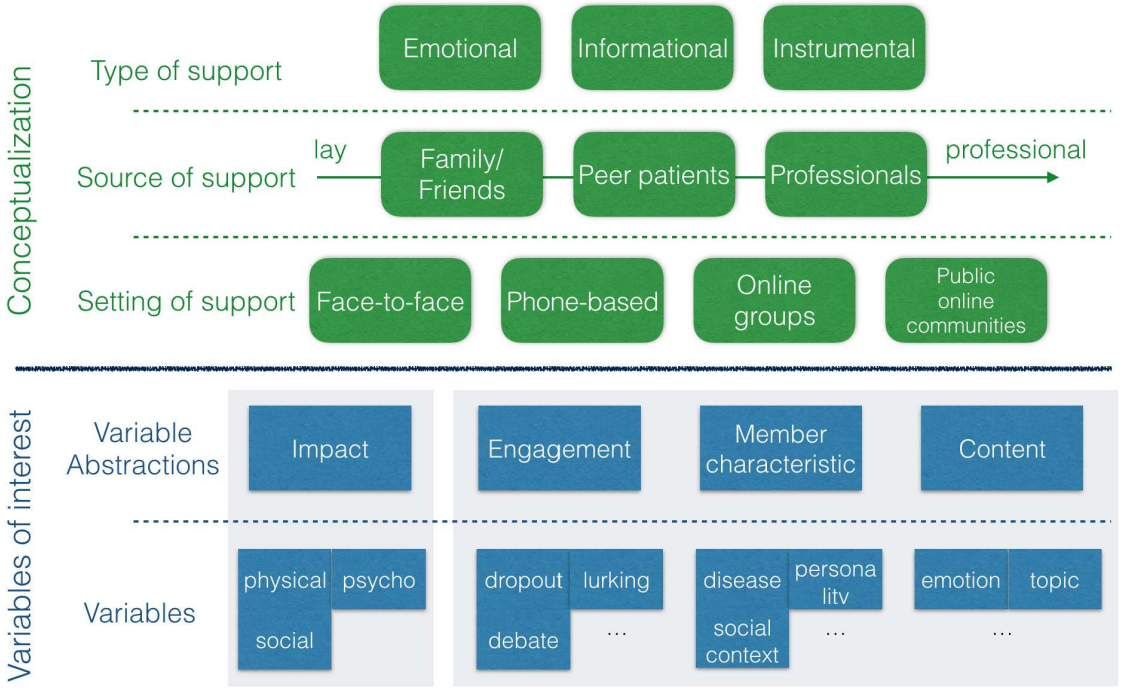


Figure 2.1: A framework for studying online health communities. Two meta-layers, conceptualization and variables of interest, represent how online health community fits in the landscape of social support, and what variables of interests are studied by previous research, respectively.

framework is not able to cover every aspect of social support, and that the three dimensions could have complex interactions in real world interventions. Based on this framework, we identify where online health community fits in the landscape of social support, which we define as **the online groups for patients exchanging peer support, primarily informational and emotional support**.

We use the lower meta-layer in our framework, variables of interest, to synthesize current research of online health communities. In general, research questions could be classified into two categories: impact of participation, and characterizing online health communities. Characterization of communities can further be decomposed into several sub-questions, member characteristics, content, and member engagement. Variables in this meta-layer were obtained through a review of the literature.

The literature search was carried out with the following query on PubMed, and focused on OHC or online social support with emphasis on cancer communities: ("community" OR "communities" OR "network" OR "support" OR "peer-to-peer" OR "forum") AND ("online" OR "internet" OR "on-line") AND "cancer" (constraint: in title). There was no time constraint to the search. The search was executed in July 2015 and returned 140 publications. Out of the 140 publications, 24 were excluded as irrelevant to our focus of study with respect to online peer support. We further expanded the set of publications by including 44 more publications which were either not indexed in PubMed or ones in the reference of our original pool of publications that yet did not match our search query. The literature analysis to identify the variables of interest studied within the peer-support framework was thus carried out over a pool of 116+44 =160 papers.

After collecting the publication pool, we manually coded each publication by finding its primary variable of interest with regards to online health communities. Variables discovered in the annotation were then refined and synthesized. This process was carried out iteratively, in which we refined both the framework and the publication annotations until the framework in Figure 2.1 was obtained. A complete list of all publications along with their associated annotations is given at http://people.dbmi.columbia.edu/noemie/ohc/literature.html. Some of the studies have more than one code according to our framework.

2.2 Defining OHCs from the standpoint of social support

In this section, we review how the concept of online health community emerges in the larger context of social support following the first meta-layer of our framework. For patients, one of the most important purposes of participating in an online health community is to seek and exchange social support with others [Sharf, 1997; Shaw *et al.*, 2000; Lieberman *et al.*, 2003]. Therefore, sorting out basic building blocks of social support that are relevant to OHCs should be helpful for informatics researchers. We describe the connection between social support and online health community here as a guide for organizing the research and potential impact of informatics research on OHCs.

2.2.1 Type of support

The first building block of social support is type of support, which usually contains three specific types: informational support, emotional support, and instrumental support, which are exchanges of information, nurturance, and tangible assistance, respectively [Friedman and Silver, 2007a]. Examples below are snippets of posts from an online health community, the discussion boards of breastcancer.org, which showcase exchanges of the three types of support, respectively.

> **Informational support**: I had a bilateral with radical on the right and prophylactic on the left. I think all you can do is gentle exercise to strengthen your back (yoga).
>
> **Emotional support**:I'm sorry you are going through this You want to talk, anytime! We are all there with you. Good luck!
>
> **Instrumental support**:Can someone help file my insurance claim?

In online settings, informational and emotional support are usually exchanged more frequently than instrumental assistance [Meier *et al.*, 2007a; Wang *et al.*,], via posting textual or multimedia content in computer mediated forums, bulletin boards, or chatrooms.

2.2.2 Source of support

The second building block of social support is the source of support. According to a social classification given by Dennis [Dennis, 2003], social support, which is obtained through one's social relationships, can be from embedded social members like family and friends, as well as from professionally created networks like social support groups. Dennis mentioned that although family members and friends are crucial sources of support for patients, researchers suggested that in distressing times, members in such social networks may not be able to fully appreciate the stressful experience. Instead, peers who share similar problems can be a better choice when one needs emotional support such as empathy and encouragement from others. From the perspective of informational support, Hartzler and Pratt found that patients could provide valuable information based on personal experience, which is not likely to be provided by health professionals or other lay members [Hartzler and Pratt, 2011]. It was suggested that the spirit of pursuing peer support is to find similar others, and the desire to communicate with people who share similar problems is the fundamental motivation of participating in an online health community [Campbell *et al.*, 2004; Shaw *et al.*, 2000; Gorlick *et al.*, 2014].

2.2.3 Setting of support

Traditional face-to-face peer support groups have several limitations [Weinberg and Schmale, 1996]: first, many patients are physically weak and not able to walk or drive to the site for group discussion; second, some patients have full-time jobs, hindering them from participating regularly; finally, for patients living in less populated areas, especially ones with rare diseases, participants may have difficulty finding others with the same conditions.

The Internet has the potential of revolutionizing the way patients exchange peer support, since patients are much more likely to find similar others online than in a restricted geographical area in which traditional offline peer support happens. The fact gives rise to the third variable: the setting in which support is delivered, represented in the third sub-layer in our framework. In the most recent decade we have witnessed a lot of investments from the research community into designing Internet-based peer support groups [Owen *et al.*, 2005; Salzer *et al.*, 2010; Lieberman *et al.*, 2003; Winzelberg *et al.*, 2003;

Gustafson and Hawkins, 2001] and such studies have shown promise in improving psychological wellbeing of patients and in facilitating health management. Aside from online support groups, which are usually created and tightly controlled by researchers, public online health communities such as Breast Cancer Forum [Wang *et al.*, ; Elhadad *et al.*, 2014; Zhang *et al.*, 2014], the CSN network [Portier *et al.*, 2013; Qiu *et al.*, 2011a], and Facebook groups [Bender *et al.*, 2011] are also becoming popular.

2.3 Research questions for the analyses of OHCs

In the second part of our framework, we identify two main categories of research questions of online health communities, one regarding impact of participation, the other regarding characterizing communities, members, and their behaviors. We suggest that most informatics research to date has focused on characterization of OHCs, leaving the potential for utilizing informatics techniques to study the impact of the participation.

2.3.1 Impact of participation

The first research question that can be asked regarding OHCs is whether participation of OHCs makes positive impact, and if so, what kind of benefit can be observed. A wide range of studies have aimed at answering this question by both experimental and observational approaches, but it is noteworthy that most of them are based on non-public online support groups created by health psychologists, while no interventional studies have been carried out on public OHCs. Table 2 lists literature with experimental study designs for online peer support groups for cancer specifically. The Design column in the table lists the different types of study designs used. They mostly are randomized control trials, with a few pre-post studies.

Among the ten randomized controlled trials, 4 rejected the null hypotheses. However, in two of the RCT studies with positive outcome identified, [Gustafson and Hawkins, 2001] and [Børøsund *et al.*, 2014], the intervention packages included multi-purpose web-based health management tools other than peer support. As such, results from these studies cannot be interpreted directly as an evidence that peer support was leading to the benefits.

literature	**subject (# sample)**	**design**	**outcome**
Gustafson et al. 2001	bc (246)	RCT	+ social support
Lieberman et al. 2003	bc (67)	pre-post	+ reduced depression
Winzelberg et al. 2003	bc (72)	pre-post	+ reduced depression
Owen et al. 2005	bc (62)	RCT	- quality of life, - psycho wellbeing, - physical wellbeing
Lieberman et al. 2005	bc (114)	pre-post	+ psycho wellbeing
Salzer et al. 2010	bc (78)	RCT	- psycho distress, - quality of life
Hoybye et al. 2010	cancer (58)	RCT	- mood adjustment, - self-rated health
Ruland et al. 2013	bc and pc (325)	RCT	+ less symptom distress
Osei et al. 2013	pc (40)	RCT	- quality of life
Hwang et al. 2013	cc (306)	RCT	- CRC screening, - fecal occult blood test
Stanton et al. 2013	bc (88)	RCT	+ less depressive symptoms
Borosund et al. 2014	bc (167)	RCT	+ reduced depression
Lepore et al. 2014	bc (184)	RCT	- mental health outcome

Table 2.1: Experimental studies of online peer support groups for cancer. +' indicates an identified impact, and -' means no outcome observed using the measurement. bc: breast cancer; pc: prostate cancer;cc: colorectal cancer; pre-post: pre-post study design with no control group; RCT: randomized control trial

Observational studies also contributed to understanding impact of group participation by suggesting participation's impact on enhancing patient-provider understanding [Sharf, 1997], members' self-empowerment [Høybye *et al.*, 2005; van Uden-Kraan, 2008] and producing better outcomes in terms of stress, depression, and coping [Beaudoin and Tao, 2008]. To date, although online peer support groups are getting more and more popular, sound evidence to support the effectiveness of such interventions is still in development. One of the primary reasons is that in most of the previous experimental studies, the sample size was not sufficiently large, leading to the possibility that confounding factors moderated the outcome more than the independent variable of interest (community participation) did. Factors like health status and offline support reception [Kim and Shin, 2013], self-efficacy of the users [Namkoong *et al.*, 2010], language use in communication [Lewallen *et al.*, 2014], and coping ability and style [Batenburg and Das, 2014] were identified as moderators or predictors of effectiveness, which cannot be completely controlled in an experimental study with only hundreds of participants.

As another increasingly popular source of online peer support, large, asynchronous online health communities such as breast cancer forum or Facebook groups overcome the issue of sample scarcity by attracting large populations of targeted patients. More recently, informatics approaches, particularly automatic content analysis based on computational or statistical methods, have been proposed to study outcome of this type of communities. For example, based on automatic classification of messages, Wang and colleagues 2012 found that emotional support is positively correlated and informational support is negatively correlated with sustained participation [Wang *et al.*,]. These studies of online communities may be less biased in samples, but have limitations in effectiveness of automated methods and the inability to build up causal relationship between usage and outcome because all of their study designs are completely retrospective and essentially observational.

There may have also been disadvantages associated with participating in online health communities. Owen and colleagues found that compared to face-to-face groups, it is more difficult to build up commitment to and cohesion within online groups [Owen *et al.*, 2008]. Furthermore, it is more difficult for members to interpret others' tone and emotion in the absence of physical and non-verbal cues, which might lead to conflicts that quickly escalate

[Friedman and Currall, 2003].

2.3.2 Characterizing OHCs and their members

Given the difficulties of studying the social support impact of online groups and the realization of complexity of online communities, researchers are increasingly interested in characterizing online health communities and their members, where most informatics research lies in. There are a lot of variables to consider regarding communities, their facilitators/moderators, users, and interactions. It is noteworthy that not all variables are included in our framework. For example, purpose of group when it was originally created [Bender *et al.*, 2011], creators' participation in the group [Kraut and Fiore, 2014], and type of group [Gorlick *et al.*, 2014] may be vital to the community development as well.

2.3.2.1 Member characteristics

Member characteristics include members' personal profiles such as health status and personality. In reality, member characteristics could be far more complex than that in the proposed framework. For example, gender of user plays a significant role in online interaction [Klemm *et al.*, 1998], leading to completely different themes of interaction in communities dominated by males and females [Owen *et al.*, 2004b]. Age is another demographic variable that makes a difference [Hoffman *et al.*, 2009b; Zhang *et al.*, 2014].

Disease. The first major member characteristic to consider when studying online peer support is the targeted disease of patients. OHC research, in general, has focused on communities for different diseases with different emphasis, such as diabetes [Ravert *et al.*, 2003], weight loss control [Das and Faxvaag, 2014], depression [Houston *et al.*, 2014], and so on. Davidson and colleagues compared social support groups for 20 categories of diseases from life-threatening ones like cancer and AIDS to chronic ones like diabetes. They found that support seeking was highest for diseases viewed as stigmatizing such as AIDS and breast cancer, and was lowest for less embarrassing but equally devastating disorders such as heart disease [Davison, 2000]. Within the scope of cancer, differences were identified between breast cancer communities and prostate cancer communities [Owen *et al.*, 2004b]. Besides the effect of gender of users, the fact that breast cancer has higher survival rates and more

treatment options is also shaping how and what users discuss: the breast cancer community, in general, witnesses more emotional support but less informational support, than prostate cancer communities. Moreover, results from analyzing a data from the National Health Interview Survey provide evidence that cancer survivors made greater use of community-based support groups than healthy participants or those with other chronic health conditions [Owen *et al.*, 2007].

Personality. The relationship between individual personalities and health has been investigated scientifically for many years. Health psychologists have found that a health event like a heart attack is more likely to develop in persons who are chronically irritated or hostile, and have established models of linkages between personality and health [Friedman and Silver, 2007b]. It is also reported that optimistic users are more likely to positively react to and ultimately benefit from cancer related experiences [Urcuyo *et al.*, 2005]. Batenburg and Das mentioned that in an online peer-to-peer support group, benefit of participation depends critically on users' coping styles: actively dealing with emotions and thoughts was positively related to psychological wellbeing [Batenburg and Das, 2014].

2.3.2.2 Content

In most current online health communities, members communicate via posts that are mostly textual but also contain a rich set of images and links to external resources. Content of the messages deliver information and sentiment, exerting influence on users' perceptions of social support from the group, and even decide users' intention of sustained participation. For example, people adjusted their behavior in response to whether the messages they receive are informational or emotional [Wang *et al.*, ; Vlahovic *et al.*, 2014]. Such differences in message content can affect perceived empathy of members [Nambisan, 2011]. Conversely, content of the messages can also influence whether informational or emotional support is elicited [Wang *et al.*, 2015]. Content analysis also reveals how individuals in communities make sense of community environments collectively [Mamykina *et al.*, 2015]. Recently, natural language processing techniques have been used to analyze OHC content in recent years [Zhang *et al.*, 2014; Vlahovic *et al.*, 2014; Zhao *et al.*, 2012], with the caveat that these techniques are still facing various open research questions [Park *et al.*, 2015]. Two major

dimensions of content are identified as they appear as frequent topics of previous works: topic and emotion [Portier *et al.*, 2013].

Topics. When the Internet first became an option for peer-to-peer communication, Sharf observed that in an online breast cancer group, topics regarding basic classifications or definitions of tumors and diagnosis are most prevalent [Sharf, 1997], indicating that Internet support was primarily a complementary source of information in early years. A variety of themes such as relationship/family issues became popular in online peer discussions later on [Lewallen *et al.*, 2014; Owen *et al.*, 2004b], but disease specific topics like treatment, diagnosis, and interpretation of lab test results are still most prevalent [Civan and Pratt, 2007; Meier *et al.*, 2007b; Cappiello *et al.*, 2007]. Specific topics of discussions were identified as well. For example, based on content analysis, Meier and colleagues found that the most common topics in 10 cancer mailing lists were about treatment information and how to communicate with healthcare providers [Meier *et al.*, 2007b]. Owen and colleagues proposed a topic schema which includes seven categories: outcome of cancer treatment, disease status and processes associated with the cancer, healthcare facilities and personnel, medical test and procedures, cancer treatment, physical symptoms and side effects, and description of cancer in the body [Owen *et al.*, 2004b]. Based on such schema, prevalence of different topics can be quantified to facilitate content analysis of cancer support groups.

Emotions. Members of communities express different emotions depending on the context. Type and amount of expression of emotion and perception can be crucial to attaining optimal benefits for cancer patients [Kim *et al.*, 2012a]. Based on an Internet support group, Owen and colleagues built a relationship between linguistic indicators of emotions and self-report of emotional suppression, observing a significant interaction between emotional suppression and use of cognitive words on mood disturbance [Owen *et al.*, 2006]. Liess and colleagues manually coded content from face-to-face and online cancer support groups according to a categorization of emotion including positive, primary negative, defensive/hostile, constraint, and neutral affects [Liess *et al.*, 2008].

Researchers have realized that human annotation can be costly and inefficient in content analysis. To solve this problem, Pennebaker and colleagues created the linguistic resource of LIWC (Linguistic Inquiry and Word Count), grouping words into psychologically meaning-

ful categories [Pennebaker *et al.*, 2001]. The dictionaries for emotion words in LIWC have been widely used by researchers in automating emotion analysis of text [Liess *et al.*, 2008; Kramer *et al.*, 2004].

Sentiment analysis, also referred to as opinion mining, is a type of technique determining the overall contextual polarity of content to some topic. Sentiment analysis is sometimes regarded as a simplification of emotion analysis that only considers the general polarity of mood [Bo Pang and Lillian Lee, 2006]. Automatic sentiment classification methods based on machine learning [Bo Pang and Lillian Lee, 2006] have been exploited to investigate sentiments of forum posts published by patient users. For instance, studies found that thread originators change their sentiment in a positive direction through reviewing others' replies and self-replying [Qiu *et al.*, 2011b], and such changes largely result from postings of influential users [Zhao *et al.*, 2012]. In a recent study it was also found that sustained participation in peer support communities would make the users express more positive sentiment in their posts [Zhang *et al.*, 2014].

2.3.2.3 Engagement

Here, we refer to the study of behaviors of community participants, such as posting activity, lurking, and dropping out of the community, as well as behaviors of creators and moderators of the community. We discuss two important behaviors of users influencing the activeness of a community, lurking and dropout.

Lurking. Lurking refers to the behaviour of observing but not participating in Internet culture. A rule of 1% indicates that in online communities or social networks, more than 90% of users lurk and only 10% contribute content, the vast majority of which are authored by the 1% super users. Mierlo suggested that the 1% rule also holds true for online health communities [van Mierlo, 2014] by finding more than half of the users lurking.

Researchers show great interest in identifying who and why lurks. Surveys collected from lurkers indicate that the primary reasons for lurking are "reading is enough", "have nothing to offer", "topic not relevant to myself", "want to talk to similar others", etc. [Gorlick *et al.*, 2014; Nonnecke *et al.*, 2006]. Lurkers tend to be older [van Uden-Kraan, 2008], have shorter history of illness [Setoyama *et al.*, 2011] and less depressed [Kim *et al.*, 2012a]. Specific to

cancer, patients with lower stage cancer are more likely to lurk [Mo and Coulson, 2010]. In terms of how lurking affects benefits of participation, most of the studies suggested that lurkers received less benefit, with some exceptions such as having a higher level of perceived functional well-being [Han *et al.*, 2014] and the same level of self-empowerment [Mo and Coulson, 2010], Given the mixed results, better modeling is needed to further understand and analyse the reasoning behind lurking.

Dropout. Dropout, also referred to as attrition, is the phenomenon of quitting participation in the group. In a broader scope, dropout means discontinuation of participating in eHealth applications and the related phenomenon of participants dropping out of eHealth trials. Eysenbach proposed the "the law of attrition" to summarize the phenomenon that the majority of participants, sometimes over 90%, quit Internet-based trials or applications [Eysenbach, 2005]. Dropout of active members can be disastrous to any social networks, drastically lowering the community's activity and cohesion. Studying dropout of peer-to-peer support groups, especially those of public communities, can be difficult for researchers. Unlike lurkers, users who drop out of a community would not even come back and read the content, which makes it impossible to collect any feedback from these users. The only way to study these members is based on retrospective data. For example, Wang and colleagues did a survival analysis on breast cancer forum, showing that users who received emotional support are more likely to keep participating while users who received informational support are more likely to drop out [Wang *et al.*,].

2.4 This thesis's focus within the framework

Although the landscape of online social support is broad, theoretically this thesis is only interested in public online health communities, which are venues for patients to exchange primarily peer informational and emotional support in an asynchronous way (see the colored elements of the first meta layer in Figure 2.2). In practice, we focus on building computational approaches to advance research in the second meta layer of the framework - to characterize online health community members and to study how users behave and interact. Specifically, relying on computational tools, we present how to identify topics of

discussions, sentiment expressions, treatment catalogues, evidences of dropping-out, and debates in dialogs in an automated fashion from public online health communities at scale. In addition, we study how these variables correlate with each other longitudinally and how they contribute to certain user behaviors. For example, we are interested in what contributed to users' decisions of dropping-out, and what topics trigger online debates. Methodologically, these studies will be based on automated methods including machine learning (unsupervised and supervised), natural language processing, information extraction, and longitudinal analysis. Figure 2.2 highlights the elements of interests that will be covered, with an additional layer representing techniques used in the thesis. It is noteworthy that although we only investigate limited number of characteristics, computational methods proposed are generalizable to studying other variables of interest in the framework with no or minimal domain adaptation or task-specific setup.

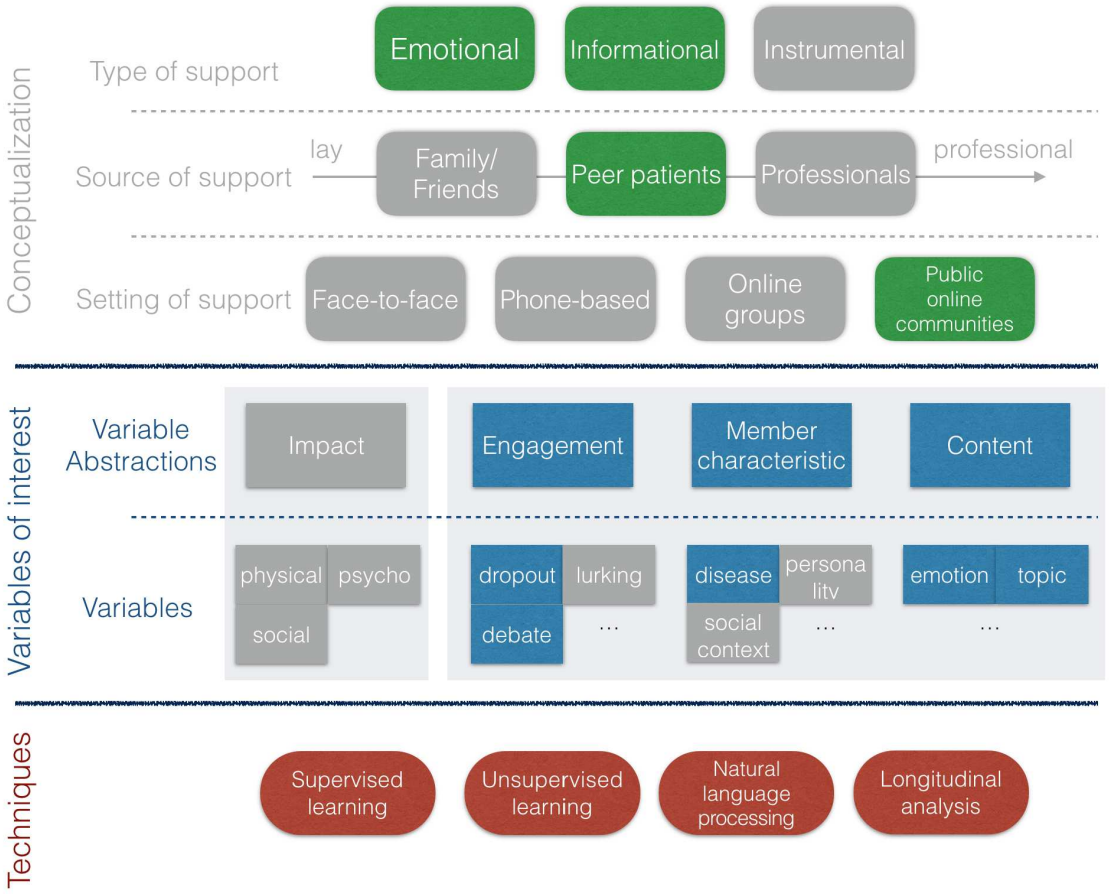


Figure 2.2: Variables of interest discussed in this thesis. Colored elements are the foci of remaining chapters. Compared with our original framework, here we have an additional layer (techniques) which lists major computational approaches we rely on in the studies of this thesis.

Chapter 3

Sources of Data

In this chapter, we introduce all the datasets we rely on for studies in this thesis. A brief summary of all datasets is given in Table 3.1. Most datasets we use are from popular public online health communities, in which discussions are organized in threads of posts. However, we also rely on datasets extracted from other sources, for the sake of investigating adaptability of our methods in specific tasks. In the following chapters, abbreviated names will be used according to the definitions in the table.

3.1 BC: Breast cancer forum

The first dataset we rely on is from the discussion board of the breastcancer.org, which is one of the most active and popular online cancer communities for breast cancer patients

Abbreviation	**Source**	**Content**
BC	breastcancer.org	breast cancer discussions
ASD	autismweb.com and autism-pdd.net	autism discussions
BCC	multiple sources	breast cancer discussions
I2B2	MIMIC	clinical notes
GENIA	GENIA corpus	biomedical literature

Table 3.1: Datasets used in studies of this thesis

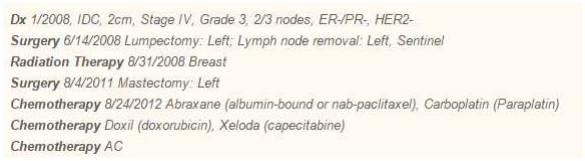


Figure 3.1: A sample of user signature in the BC dataset

Number of sub-forums	78
Number of threads	121,474
Number of posts	3,283,016
Number of authors	58,177
Number of authors with signatures	7,211

Table 3.2: Descriptive statistics of the BC dataset

and survivors. This forum has been the subject of many previous research such as [Nguyen and Rosé, 2011; Wang *et al.*, ; Han *et al.*, 2014], and has kept steady grow in activeness in recent years.

For this thesis, the entire public available content of the discussion board was collected in January 2015. The discussion board is organized in distinct forums, each with threads and posts. In total, 3,283,016 posts organized in 121,474 threads were extracted. We also crawled meta-data of posts and threads, such as timestamps, author names and IDs, and post signatures. It is noteworthy that in this particular forum, user signatures contain users' self-reported diagnosis and treatment information. One example of such user signature is given in Figure 3.1. However, not all users reported such information in their profiles. Detailed statistics of this dataset are given in Table 3.2.

3.2 ASD: Autism forums

The ASD dataset of autism forums were collected from two sources: autismweb.com and autism-pdd.net, which are primarily for autism patients and caregivers. The forum from autism-pdd.net was officially closed in 2015 and could no longer be accessed. We crawled all content that was public available from these two forums in March 2015. These two forums

Number of sub-forums	16
Number of threads	61,817
Number of posts	551,029
Number of authors	10,210

Table 3.3: Descriptive statistics of the ASD dataset

are designed for the same audience and thus have similar functionalities, but the forum from autismweb.com is significantly larger than the one from autism-pdd.net. As such, we merged these two forums into one single dataset, with following information available: sub-forums, threads, posts, and authors. Detailed descriptive statistics of this dataset can be found in Table 3.3.

3.3 BCC: A heterogeneous breast cancer consumer dataset

This dataset consisted of data that was collected from four different sources, which are all generated by breast cancer consumers in interventions or communities for social support purposes [Bantum *et al.*, 2016]: transcripts of online support groups from a distress management intervention for cancer survivors called Health-space.net (n = 174; see [Owen *et al.*, 2014b; Owen *et al.*, 2014a] for a description of the larger study), transcripts of online support groups from the Cancer Support Community (n = 39), transcripts from online support groups from a collaborative study between the Cancer Support Community and the British Columbia Cancer Agency (n =21), individual posts from Breastcancer.org (n = 83), and individual online writings from an online expressive writing study (n = 159; see [Owen *et al.*, 2011] for details on original study).

The purpose of creating this dataset is to apply one of our methods (see Chapter 5) to a more heterogeneous dataset to evaluate its generalizability, given that the BC and BCC datasets are both authored by breast cancer users but are from different types of communities (public v.s. closed). Detailed statistics of this dataset can be found in Table 3.4.

Source of data	# posts or transcripts	# sentences	# authors
Health-space.net	465	60,022	174
Cancer Support Community	30	20,760	60
Breastcancer.org	96	1077	83
Expressing writing	622	14827	159

Table 3.4: Descriptive statistics of the BCC dataset

3.4 I2B2 and GENIA

The I2B2 and GENIA datasets are used to evaluate our unsupervised named entity recognition (NER) system in one of the studies introduced in Chapter 4. The two datasets each has named entities annotated. The i2b2 corpus is a set of clinical notes with Problems, Tests, and Treatments annotated as entities, while GENIA corpus is a collection of biomedical literature consisting of biological entities such as DNA, RNA, and protein. i2b2 and GENIA are mainstream datasets for evaluating NER and were leveraged in two major biomedical named entity recognition shared task events: the i2b2 challenge 2010[1] and the BioNLP/NLPBA 2004[2], respectively.

The I2B2 includes discharge summaries from Partners Health Care, Beth Israel Deaconess Medical Center, and University of Pittsburgh Medical Center (denoted in this paper as Partners, Beth, and Pittsburgh for short). The GENIA corpus is the primary collection of biomedical literature compiled and annotated within the scope of the GENIA project. The corpus was created to support the development and evaluation of information extraction and text mining systems for the domain of molecular biology. The corpus contains Medline abstracts, selected using a PubMed query for the three MeSH terms "human," "blood cells," and "transcription factors." The corpus has been annotated with various levels of linguistic and semantic information. The original GENIA corpus contains 36 classes of entities. A more widely used version of GENIA corpus is the one simplified for the BioNLP/NLPBA shared task, in which entities are grouped into only 5 major classes: protein, DNA, RNA,

[1]https://www.i2b2.org/NLP/Relations/

[2]http://www.nactem.ac.uk/tsujii/GENIA/ERtask/report.html

cell line, cell type. We use these five categories in this paper.

Evaluations and other details of these two datasets are given in [Uzuner *et al.*, 2011b] and [Kim *et al.*, 2003].

Part II

Basic NLP Tools for Online Health Community Research

Introduction

The very first step of studying public online health community at scale is to create computational tools and resources that can support automated content analysis. In this part of thesis, natural language processing (NLP) and machine learning techniques are leveraged to automate various tasks ranging from building lexicons to analyzing sentiment expressions. More broadly, NLP and machine learning techniques have been used in a wide range of applications such as machine translation [Brown *et al.*, 1990], automated question answering [Andrenucci and Sneiders, 2005], and online review opinion mining [Bo Pang and Lillian Lee, 2006]. Recently, statistical NLP based on machine learning has successfully applied in analyzing biomedical text including clinical notes and scientific literature (see [Zhang and Elhadad, 2013] for a review of one important task: named entity recognition), partly thanks to the development of medical ontologies and lexical resources.

The primary advantage of statistical NLP is that it requires no or minimal hand-crafted rules or heuristics which are basic components of traditional rule-based systems [Zhang and Elhadad, 2013]. However, statistical NLP relies critically on the availability of linguistic resources, especially annotated corpora. Recent decades have witnessed breakthroughs in creating biomedical corpora to facilitate statistical NLP tasks, such as the I2B2 and GENIA corpora described in section 3.4. The establishment of these linguistic resources have greatly helped the development of data-driven methods for text mining and content analysis in the biomedical domain [Kim *et al.*, 2004; Uzuner *et al.*, 2011a].

Content of online health communities, however, differs drastically from other genres of biomedical texts that have been traditionally studied. Scientific articles, for instance, are fully technical and exclude personal stories, narrative style, or emotional content. Medical and health news stories, for another example, focus on newsworthy events and provide a mix of narrative and scientific style. In contrast, the language in online health communities is both emotional and technical; their style is often narrative, they are highly interactive (different responders contribute to each thread), and they are peer-reviewed (there is evidence that inaccurate medical statements are rare and quickly corrected by other posters [Esquivel *et al.*, 2006]). This genre of text is also widely different from the ones traditionally considered in the field of information processing outside of the medical domain: the lan-

guage is often quite community specific (participants use many acronyms and abbreviations unknown outside of this medium [Elhadad *et al.*, 2014]); at the same time, the posts are authored in an unstyled and unedited manner, with sometimes informal and ungrammatical expressions. As such, migrating of existing methods and systems trained on non-OHC text usually fails; the lack of linguistic corpora equipped with computational methods to process the corpora have been the main bottleneck for large-scale analysis of OHC content.

In this part of the thesis, we describe how computational tools and linguistic resources (corpora and lexicons) were created to facilitate the basic dimensions of analyses of online health communities. The methods, primarily based on natural language processing and machine learning, are able to handle various content analysis tasks at scale in an automated fashion. Tools described in this part of thesis will also be the foundations of all user modeling and characterization discussed in subsequent parts. With respect to our framework, in this part of the thesis we focus on building up tools based on techniques of supervised machine learning, unsupervised learning, and natural language processing for the basic analysis of OHC content (Figure 9.2).

In order to understand the overwhelming amount of health-related, patient-generated OHC content, two levels of linguistic information need to be learned. We refer them to as understanding the *lexical semantics* and *pragmatics* of the content in this thesis, respectively. To be specific, we define *lexical semantics* as studying the meaning of words and phrases, and hence creating lexicons that can support further semantic analysis of sentences. *Pragmatics*, on the other hand, is defined as understanding meanings of sentences, paragraphs, and discourses of discussions in context. This includes modeling topics of discussions, analyzing sentiment expressions through content, detecting debates and user stances towards certain issues from dialogs, etc. One major difference between *lexical semantics* and *pragmatics*, in this thesis, is whether the task depend on context information including general thematic context of the community and surrounding posts published by other authors. It is also noteworthy that semantic information identified at lexical level will be critical features for pragmatic analyses.

In the next chapter, we describe how semantic representation of words and phrases are created in an unsupervised fashion, and how these representations can be used to build

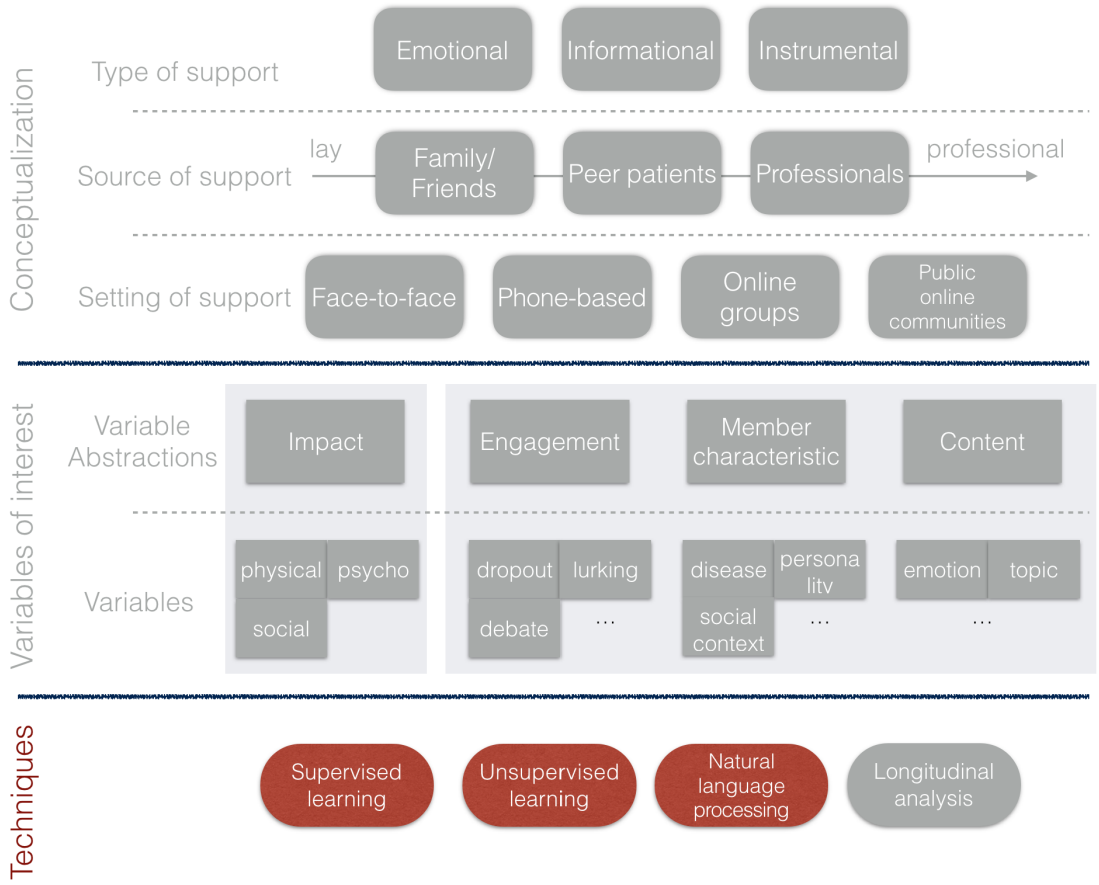


Figure 3.2: Variables of interest discussed in this thesis. Colored elements are the foci of this part of thesis.

lexicons for downstream analyses. In chapter 5, we describe how tools for different pragmatic analyses are built, based on supervised machine learning and lexicon resources we create in the previous step.

Chapter 4

Lexical Semantics of OHC texts: An Unsupervised Approach

4.1 Introduction

One fundamental step to language understanding is to quantify meaning of the basic linguistic units, words and phrases, in a particular domain, which is the focus of lexical semantics in this chapter. In linguistics, a general approach of modeling meaning of words is by looking at context [Martin and Jurafsky, 2000], assuming that words with similar meaning often occur in similar context. In recent years, distributional semantics has been popular, in which meanings of words are represented by vectors of real numbers projected from their context in certain ways. Two highly related tasks of lexical semantics for OHC content, named entity recognition and lexicon creation, will be discussed in this chapter by leveraging principles from distributional semantics.

The first task of lexical semantics is to create lexicons which represent domain knowledge. For many research activities of OHCs, capturing domain knowledge about topics discussed in a community and organizing terms and concepts discussed into lexicon and terminologies is needed for knowledge discovery and information extraction [Overberg *et al.*, 2010; Portier *et al.*, 2013]. Designing automated tools to build these lexicons is a challenging task, however, because the language used in online health communities differs drastically from the genres traditionally considered in the field of information processing and from

the sublanguages already investigated in the biomedical domain [Harris and Harris, 1991; Friedman *et al.*, 2002]. As previously discussed, health community vocabulary is characterized by abbreviations and community-specific jargon, and posts are authored in a style-free and unedited manner, with often informal and ungrammatical language. In addition, the content of the posts is both emotionally charged and dense with factual pieces of information, indicating that specific semantic types of information, like emotions, are more prevalent than in traditional biomedical texts. More importantly, these features can vary community by community, which creates more challenges to build high quality lexicons. As such, a method that is able to capture community-specific characteristics but highly portable to different communities is needed.

Named entity recognition (NER) aims at recognizing all terms from text that belong to certain semantic categories, which is critical to understanding the thematic focus of the content, to extracting salient concepts in discussions, and as features to facilitating downstream content analysis. In NER, terms (either single words or multiple words) of interest are identified and mapped to a pre-defined set of semantic categories. In the clinical and biomedical domain, systems were created including extracting clinical entities from radiology reports [Friedman *et al.*, 1994; Hripcsak *et al.*, 1995; Fiszman *et al.*, 2000], identifying diseases and drug names in discharge summaries [Chapman *et al.*, 2001; Melton and Hripcsak, 2005; Uzuner *et al.*, 2011a], and detecting gene and protein mentions in biomedical paper abstracts [Tanabe and Wilbur, 2002; Settles, 2004; Yeh *et al.*, 2005]. Most of existing NER tools were created not for online health community text, and they are trained specifically on certain types of data (usually scientific literature or clinical notes) which limits their adaptabilities. Given the dramatic difference between these genres of text and OHC text and among content from different communities, an unsupervised NER tool that can be conveniently applied to heterogeneous online health community data is needed.

Intuitively, named entity recognition and lexicon creation are similar tasks sharing the same workflow. They both require identifying salient terms from text, followed by classifying these recognized terms into certain semantic types or lexicons. The major difference lies in the purpose of the task and application scenario of the output. Lexicon creation fo-

cuses on collecting terms that can precisely represent domain knowledge, while named entity recognition is usually used as a step of sentence pre-processing for the sake of dimensionality reduction. As such, in lexicon creation we care more about precision of the terms in representing domain knowledge, while in named entity recognition a balance between precision and coverage must be considered. The two tasks are approached in this thesis using the same unsupervised workflow based on distributional semantics [Zhang and Elhadad, 2013; Elhadad *et al.*, 2014]. The method proposed requires little manual intervention, and can adapt to different types of communities conveniently. Given its high portability, the tool can also be applied to NER and lexicon creation tasks outside of OHC, which is an additional contribution of this part of thesis. In the next section, we present the basic workflow of our unsupervised approach to lexical semantics. In section 4.3, we present a case study of applying our method to one of our OHC datasets, the BC dataset. Section 4.4 and section 4.5 will discuss alternatives and improvements of two specific steps of our method pipeline.

4.2 An unsupervised approach to lexical semantics

The basic setting of the task is that we have a dataset (unannotated) in which terms of interest need to be identified, followed by classifying the identified terms into certain categories. Our method, in a nutshell, is illustrated in figure 4.1.

First, a seed term set is gathered (either from an existing lexicon or from a small manually created one) representative of a given semantic category of interest in NER or in lexicon creation. Second, seed terms and their context, as defined from their occurrences in the online forum, are aggregated into a representative context vector, which reflect the typical context for terms in the category. As such, the representative vector acts as an implementation of the distributional hypothesis, where a word is defined by the context in which it is conveyed. In this step, we could have multiple choices in the distributional semantics, which we discuss in the section 4.5. Third, to identify new terms for the semantic category, candidate terms from the target text are selected and an individual context vector is defined for each. Finally, determining whether a candidate term belongs to a semantic category is achieved by computing the similarity between its individual context vector and

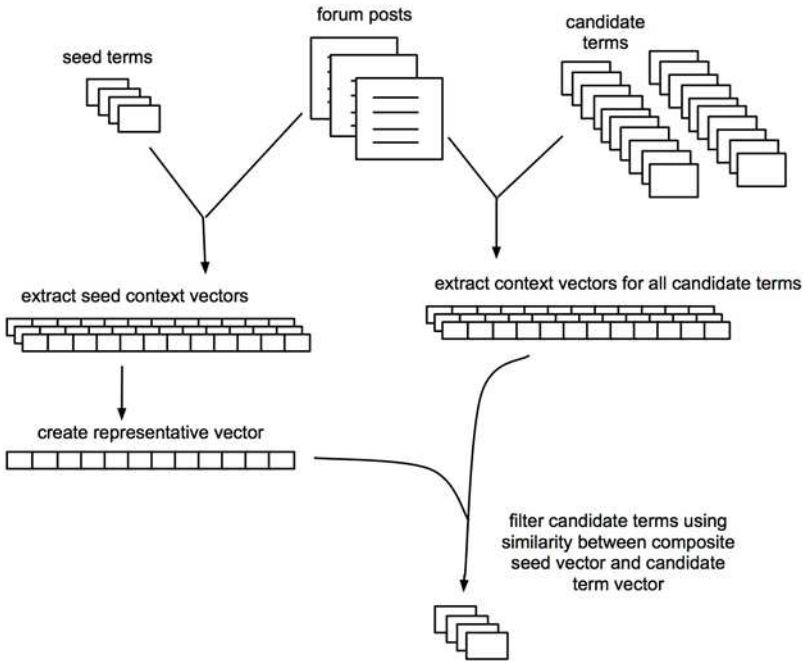


Figure 4.1: Overall pipeline to identify in an online health community the terms representative of a specific semantic category.

the semantic categorys representative vector. If a candidate term is used with words and patterns similar to the ones of the semantic category, it is likely the candidate term belongs.

In the remaining part of this section, we describe the method in detail step by step.

4.2.1 Choosing seeds and candidates

For each semantic category (or entity type), we use an existing lexicon or a manually curated list of terms to gather a set of seed terms that are known to belong to the target category (e.g., medications). Using the unannotated corpora which are usually massive in scale in public OHCs, we also extract a large number of candidate terms that may or may not be members of the target semantic category. The process of selecting candidate terms is where we recognize salient terms from texts, which can be naively approached by collecting terms that match entries in standard terminologies such as UMLS or SNOMED-CT. The next section will present an example of how these terms are collected, and section 4.4 will discuss alternatives to seed and candidate term collection.

4.2.2 Constructing context vectors

Once the sets of seed and candidate terms have been selected, we employ a vector-space distributional similarity method to create context vectors for each term. The context vectors are derived from the vocabulary V found in the dataset with the constraint that a word appears in the corpus. Each element in a terms vector contains a count of the number of times a word in V appeared in a certain context, such as directly preceding our term of interest. Because we use 5 contextual feature types, as described in Table 4.1, each context vector consists of $5|V|$ elements. We chose a set of local, highly specific contextual features to capture similarity in meaning and usage. For instance, in the example given in Figure 4.2, we can capture some of the data contained in exact patterns such as been on X, as well as more general contextual features, such as the presence of the word started somewhere within 3 words of our target. This information helps the method find candidate terms that exhibit similar behavior to our seed terms, and are therefore likely to be in the same semantic category. These context vectors form the underlying representation in our method.

preceding word	A cell for each word in the vocabulary, indicating the number of times it appeared directly before the target term
word at -2	A cell for each word in the vocabulary, indicating the number of times it appeared 2 words before the target term
following word	A cell for each word in the vocabulary, indicating the number of times it appeared directly following the target term
word at +2	A cell for each word in the vocabulary, indicating the number of times it appeared 2 words after the target term
word within 3	A cell for each word in the vocabulary, indicating the number of times it appeared within 3 words before or after the target term

Table 4.1: Feature types used in the vector space model

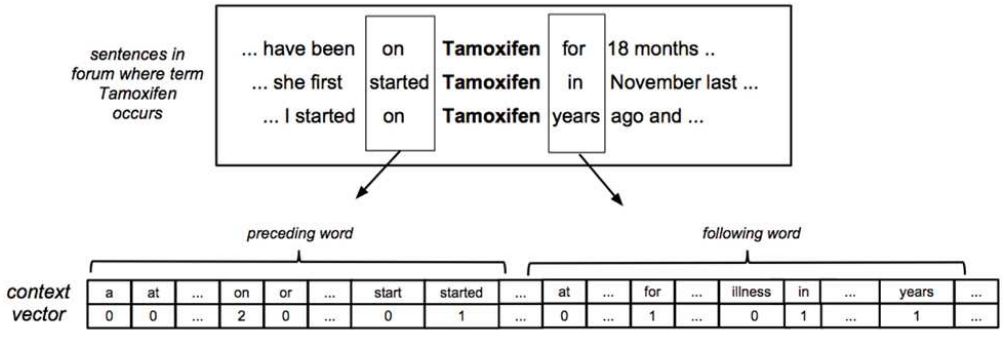


Figure 4.2: Part of the context vector for the seed term Tamoxifen. Each term context vector has a separate set of counts for preceding word, following word, word at -2, word at +2, and word within 3.

In section 4.5, we will discuss alternatives to constructing the context vectors: the choice of semantic representation, and we also present results of a study [Zhang and Elhadad, 2016b] in which we build lexicons for the ASD dataset using different semantic representations.

4.2.3 Creating a representative vector

In order to create a unified representation of a semantic category, the context vectors of the seed terms are merged into a representative vector for the category. Using vector addition, the individual context vectors are added. The vector is normalized by the number of seeds, producing a vector containing the average value for each of the seed vectors. A smoothing step is then performed, in which any values of the representative vector that are specific to only one seed are set to zero. This is intended to remove any contextual information that is unique to a single seed term and does not represent the semantic category as a whole. For example, assuming Arimidex is a seed term for category "medications", and it appears in the sentence I have been on Arimidex (an aromatose inhibitor), we will want to make use of the feature preceding word on, since it is an indicator of a medication term, and will likely be shared by other seeds. However, the word aromatose is specific to the seed Arimidex, and we will discard the associated data unless it is shared by at least one other seed.

To further reduce noise and ensure a high-quality representative vector, a pre-filtering step is employed. The initial representative vector as created above is compared with each of the original seed term vectors using a cosine similarity metric [Manning *et al.*, 2008a]. If the similarity is below a certain threshold, the seed term is considered an outlier, and is removed. The representative vector is then re-created as described above using the filtered group of seed terms.

4.2.4 Calculating similarity

A candidate term t is more likely to belong to a semantic category if its context vector is similar to the representative vector r for the category. Similarity is computed as the cosine metric between the two vectors. If the vector t is composed of (t1, t2, ..., tn) and r is

composed of (r1, r2, ..., rn) then, their cosine is defined as

$$Sim(t,r) = \frac{t \times r}{|||t||| \cdot |||r|||} = \frac{\Sigma_{i=1}^{n} t_i \cdot r_i}{\Sigma_{i=1}^{n} t_i^2 \cdot r_i^2}$$

The values of cosine similarity range from zero, indicating no similarity, to 1, indicating maximal similarity. Thus, our procedure scores each candidate term according to the similarity of its vector to the representative vector for the semantic category. The candidates can then be ranked in descending order of their similarity scores.

4.3 An example study on the BC dataset

In this section, we present an example study which applies the method described above. We aim at building up lexicons representing domain knowledge for the BC dataset, which is from a large breast cancer online forum [Elhadad *et al.*, 2014]. In this study, we focus on three semantic categories of interest: (i) medications, (ii) signs and symptoms, and (iii) emotions and mental states.

4.3.1 Seed and candidate sets

Seed sets are collected separately for each of the three semantic categories as described below.

Medications. To create a set of seed terms denoting names of medications, we use the comprehensive list of medications provided by RxNorm [Nelson *et al.*, 2011]. The list is then ordered by frequency of occurrence in the corpus, and terms appearing with low frequency in our corpus are removed (less than 50 in our experiments), resulting in a seed set of 137 medication terms.

The set of candidate terms for the medication category is defined initially as all out-of-vocabulary words in a standard English dictionary (dictionary from the Aspell program was used in our experiments), following the assumption that medication names are proper names, and thus not part of the standard English vocabulary. We only considered out of vocabulary terms from our corpus, which were frequent enough (50 times at least). This resulted in a set of 1,131 words as potential candidates for medication names.

Signs and Symptoms. We experiment with two medical lexicons for the construction of a set of seed terms denoting signs & symptoms. The first is the Unified Medical Language System (UMLS), where we use a list of all terms assigned to the sign or symptom semantic type [Lindberg, 1993]. The second resource is SIDER, a list of terms denoting side effects extracted from FDA drug labels [Kuhn *et al.*, 2010]. For each of these lists, we filter out all terms that are more than two words long. We then search for occurrences of the remaining terms in our data and extract all single word terms occurring more than 50 times, and all two-word terms occurring more than 20 times. This procedure provides the four seed sets described in Table 4.2. Despite the fact that both UMLS and SIDER seed lists share the most frequent term, there is relatively low overlap between them amongst these high-frequency terms (17 single-word terms, and 21 two-word terms).

Seed Set	**Size**	**Avg. Frequency**	**Most Frequent**	**Coverage**
UMLS single word	84 (45)	1,205 (1,577)	pain	103,695
UMLS two words	136 (63)	134 (228)	hot flashes	37,702
SIDER single word	88 (51)	918 (1,418)	pain	80,780
SIDER two words	92 (38)	166 (335)	hot flashes	31,926

Table 4.2: Number and average frequency of the terms in the four seed sets employed for detecting Signs & Symptoms, before and after (in parenthesis) the filtering procedures, along with the most frequent term in each seed set. The rightmost column specifies the coverage (cumulative frequency of all the terms inside the set) of each unfiltered seed set.

In the case of signs and symptoms, we cannot restrict candidates to out-of-vocabulary terms, as we did for medications, since signs and symptoms are often conveyed using standard-English words and are often multi-words. Instead, we consider any single-word or two-word term as a potential candidate, provided it appears frequently in our data (more than 50 times for single words, and more than 20 times for two-word terms), and consists of well-formed words (does not include numbers or other non-alphabetic characters).

In addition, for two-word terms, we perform another filtering step to reduce the number of candidates and improve quality. This filter is designed to remove multi-word terms that

are very common in the data as a result of the frequency of the component words, rather than the term as a whole. For instance, the two-word term and I appears frequently in our data, but has little meaning as a unit, and its frequency is due to it being composed from two very common words. To filter such cases, we compare the probability of the term as a whole to the expected probability of the component words appearing in adjacent positions by chance, according to their individual probabilities, as shown in Equation 1. The ratio r between these probabilities is compared to a manually specified threshold t (in our experiments, t = 20), and terms with ratios below the threshold are removed from the candidate list. After the selection and filtering procedures, we were left with a candidate list of 10,844 single-word candidates, and 37,015 two-word candidates.

$$Eq.(1) r(word1, word2) = \frac{p(word1\ word2)}{p(word1) \cdot p(word2)} \quad p(x) = \frac{\#\ occurrences of x}{size of data}$$

Emotions. While there exist terminologies for emotions [Pennebaker *et al.*,], we experimented with a very small seed set for emotions. Part of our motivation is to test the robustness of our method to discovering new terms when a limited terminology or none is available. Given the most frequent words in the corpus of posts, we manually selected 10 adjectives as a seed set, which conveyed an emotional state randomly: scared, grateful, sorry, fatigued, guilty, comfortable, nervous, confused, afraid, and happy. Following the filtering step described above to compute the representative vector, there were six emotion seed terms left: scared (frequency of 5,512 occurrences in the corpus), grateful (frequency 1,445), sorry (frequency 20,768), confused (frequency 1,807), afraid (frequency 3465), and happy (frequency 11,338). For the sake of reproducibility, we replicated the experiments with different seed sets chosen randomly and obtained very similar results to the ones given this instance of seed set, and thus only report on these results.

4.3.2 Experimental setup

The output of our method for a given semantic category is a ranked list of terms, which can augment a terminology of known lexical variants for the category (ranking is based on the terms similarity scores to the given semantic category). We asked domain experts (two clinicians and one health psychologist) to review the lists for each of the three categories and

tag each ranked term as a true positive (indeed a term that belongs to the semantic category) or a false positive (a term that does not belong to the semantic category). We report on the Precision at K [Manning *et al.*, 2008a]., a standard evaluation metric for retrieval tasks in which the overall gold standard is unknown in advance with different values of K for the top-K returned results, from K=10 to 50. We also report the cumulative coverage of the true-positive terms retrieved at the different K that is, considering only terms that are not seeds. The coverage is a sanity check that the effort spent on discovering these terms pays off in terms of content that would have been ignored otherwise. For medications, the experts also encountered a number of terms that fell in a gray area. For instance, terms which were general names of treatments, or categories of medications, such as anthracyclines, a class of antibiotics. There were also terms indicating various drug cocktail treatments, as well as names of dietary supplements alternative treatments. Thus, for medication, we report two types of Precisions at K: a strict evaluation, which represents whether the ranked terms were medication names indeed, and one with a less strict definition of medication, which includes medication classes and drug cocktails.

4.3.3 Results

Medications. In Table 4.3 we list the top ten terms according to the similarity with the representative vector for the medication category, along with their similarity score and frequency in the corpus. For the most part, the system correctly identifies terms indicating medications. There are misspellings (e.g., tamoxifin, benedryl, femera) and abbreviations (e.g., tamox) of medication names. The terms bisphosphonates and hormonals indicate classes of medications.

We can see four classification errors: fatique, carbs, mammos, and lymphedema. The first is a rare misspelling of fatigue in the dataset, with thus little power to be categorized correctly. The term carbs is used in a similar fashion to many medications, since it is an ingested compound and forum users often discuss its effect on their health, much like they discuss medications. In general, we observed that various types of dietary supplements were common in our results for this reason.

Figure 4.3 shows the precision of the classification as we go down the list of retrieved

Retrieved term	Sim. score	Freq.	Retrieved term	Sim. score	Freq.
Tamox	0.888	6,107	bisphosphonates	0.821	549
Hormonals	0.888	1,012	carbs	0.821	326
Tamoxifin	0.880	666	mammos	0.817	704
Benedryl	0.831	402	femera	0.815	452
Fatique	0.827	108	lymphedema	0.815	2,656

Table 4.3: List of top 10 retrieved medication terms not included in seed set, along with their similarity score and their frequency.

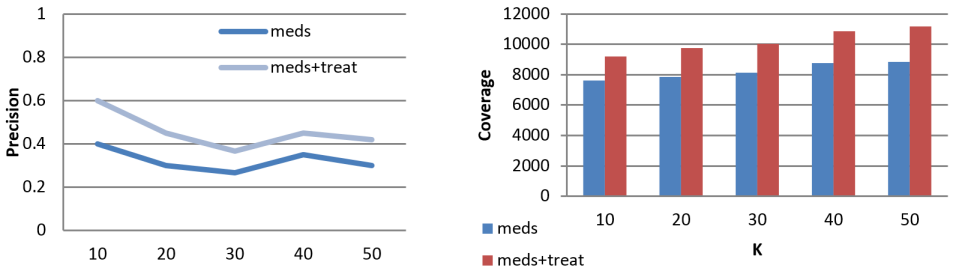


Figure 4.3: Precision (left) and number of instances covered (right) for the top K=10,20,30,40,50 retrieved terms not in the seed set for medication and treatments (meds+treat) and medication names only (meds).

terms (and following our experimental setup where only words outside of RxNorm were assessed for validity). Coverage ranged from 9,188 for K=10 to 11,191 occurrences for K=50 for medications and treatments, and ranged from 7,627 (K=10) to 8,859 occurrences (K=50) for medication names alone.

Signs and Symptoms. Table 4.4 shows the top 10 single-word and two-word terms retrieved as Signs and Symptoms retrieved when using SIDER as seed set. Figure 4.4 shows precision and coverage at K for Signs and Symptoms category using either UMLS or SIDER as seed set.

Retrieved term	**Sim. score**	**Freq.**	**Retrieved term**	**Sim. score**	**Freq.**
itching	0.954	807	joint pain	0.985	2,213
caffeine	0.950	342	mouth sores	0.966	604
chemo	0.950	76,737	body aches	0.959	221
depression	0.950	2,575	acid reflux	0.958	205
discomfort	0.945	1,520	nose bleeds	0.954	131
bleeding	0.942	1,376	hair loss	0.952	1,549
bruising	0.942	336	bone aches	0.949	119
soreness	0.935	476	stomach problems	0.948	101
exhaustion	0.935	248	extreme fatigue	0.947	110
surgery	0.934	35,831	mood swings	0.945	309

Table 4.4: Top 10 single- and two-word terms retrieved as Signs & Symptoms using SIDER as a seed set. Three of the single-word terms identified are not signs or symptoms, but mentions of treatment

As mentioned in the Methods section, we made use of two resources to develop two separate seed sets for this semantic category. In the figure, we see that the different characteristics of the seed set (see Table 4.2), result in differences in performance for our system. The UMLS seed set has better coverage than Sider on single-word terms, for a similar number of words. This means that the single-word terms in the UMLS are more suited to our domain, and this results in higher coverage and precision for the output of our system. For two-word terms the situation is reversed. The SIDER seed set has similar coverage, but is significantly smaller than the UMLS one (see Table 4.2). This means that the seed terms

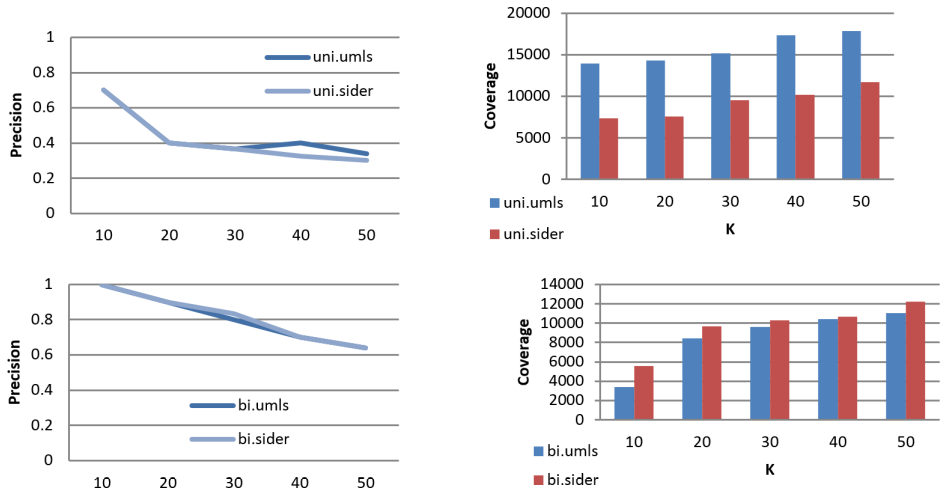


Figure 4.4: Precision (left) and number of instances covered (right) for the top K=10,20,30,40,50 retrieved single-word signs & symptoms (top) and two-word signs & symptoms (bottom), reported for UMLS and Sider as seed set.

are more suited to our domain. For two-word terms, we get better coverage and precision when using SIDER as a seed.

There is another important difference worth noting between single-word terms and two-word ones. In the case of single word terms, the coverage of both the lexicons we employ is quite high. This means it is difficult to find new terms not mentioned in the lexicon, and these are found with lower confidence. This is also the reason for relatively low precision for single-word terms in this semantic category (the precision is measured only for the new terms). For two-word terms, on the other hand, initial coverage of the seed sets is quite low. There are many terms in the data that are strong members of this semantic category, but are not mentioned in the lexicons. This means the system can discover high quality new terms, with higher coverage and better precision.

Emotions. Table 4.5 shows the list of top-10 retrieved emotion terms from the small seed set of six emotion terms. All terms are high-frequency terms in the corpus, except for greatful. Interestingly, the misspelled greatful, despite its low frequency had a high similarity to emotions probably because of its correct spelling grateful was one of the seed term. The precision is much higher with emotions than with the other two semantic categories

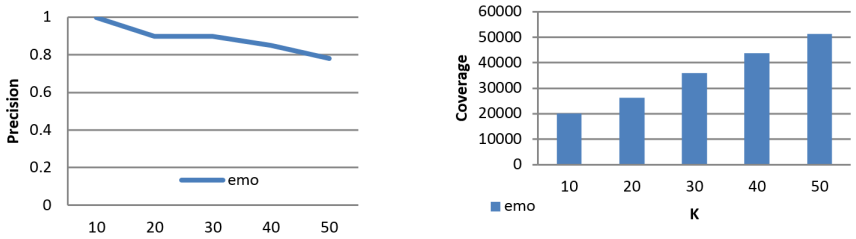


Figure 4.5: Precision (left) and number of instances covered (right) at K=10,20,30,40,50 for the retrieved emotion terms not in the seed set.

medications and signs and symptoms, starting at 100% at K=10 and decreasing to 78% at K=50. For this category, we evaluated up to K=100, with a precision of 64%. Moreover, the coverage of the true-positive emotion terms ranged from 20,076 for K=10 to 51,281 for K=50. This indicates two findings: (i) terms relating emotional states are highly frequent in our corpus, confirming that much emotional support is exchanged amongst the forum members; and (ii) our method is particularly good at discovering new terms when provided with a very small seed set (in this case a set of 6 chosen terms).

Retrieved term	**Sim. score**	**Freq.**	**Retrieved term**	**Sim. score**	**Freq.**
glad	0.878	12,414	thankful	0.741	1,273
relieved	0.847	922	desperate	0.721	252
excited	0.780	1,035	delighted	0.719	152
thrilled	0.769	779	greatful	0.716	80
sad	0.745	2,994	saddened	0.698	175

Table 4.5: List of top 10 retrieved emotion terms not included in seed set, along with their similarity score and their frequency.

4.3.4 Summary of findings

The primary objective of this study was to evaluate the use of lexical semantics in creating lexicons for use in content analysis of online health communities. Existing lexicons, like RxNorm, UMLS, and SIDER are fairly static resources, with potentially low coverage of

the particular sublanguage of online health communities, whose informality often includes unique jargon, misspellings and abbreviations created by community members. Our method aims to fill in these gaps, by generating lexicons to represent the language of members in a given community with respect to different semantic categories.

Our study suggests that using context vectors trained on a small seed set is a viable, robust method to expand existing medical lexicons across a range of potential semantic categories. The method was robust across semantic categories as long as seeds were good representatives of those categories. Furthermore, we showed that the seed set can be very small (e.g., six terms like in our experiments with detecting emotion terms) and still generate viable lexicons with good coverage. Finally, our experiments with UMLS and SIDER suggest that seed set selection should take into account surface characteristics like number of words in phrase. Finally, our studys experimental setup assessed the validity of only terms that were not already covered by existing lexicons. Thus, in the case of a semantic category and a lexicon with good coverage, our method has less opportunity to identify new terms (e.g., RxNorms and medications), but when the existing lexicons are scarce, our method identifies new terms with high accuracy (e.g., emotions).

As online health communities become a standard data source for mining information about patients, the underlying lexicons used to retrieve or assess prevalence of different terms must be representative of the way community members communicate. The lexicons we generated contain variations of known terms, which would be difficult to discover otherwise, as well as terms, which are not covered by existing lexicons.

4.3.5 Impact of seed terms

A common concern when using statistical methods that rely on seed terms is the sensitivity of the method to the choice of seeds. To investigate this issue in our framework, we compared the results of using a variety of different seeds, and examined the effect on the terms retrieved by the system.

First, we compared the output of the system when using the seed set based on UMLS terms to the output when using seeds from SIDER. Despite low overlap between the two seed sets, the output of the system was similar for both. When comparing the top hundred

most highly scored terms, we found an overlap of 91% in the output for two-word terms, and 89% for single word terms. This indicates that the semantic category we are looking for terms indicating signs and symptoms is a well-defined one, with specific usage patterns in the data. A practical implication is that any seed set containing good representatives of the semantic category can be used to successfully retrieve other terms in a fairly robust fashion.

We also experimented to discover if single-word terms could be used as seeds to retrieve multi-word terms in the same semantic category. We used the SIDER single-word seed set to rank the two-word candidates. In this case, however, we found much lower correspondence with the output of the two-word seeds (60% when compared to the UMLS two-word seed group, and 57% compared to the Sider seed). These findings indicate that single-word terms describing side effects are used in a different manner than multi-word expressions, in terms of immediate context, and that it is important to use a seed set of the same type as the candidates that are being ranked (single-word seeds for single word candidates, and multi-word terms as seeds for multi-word candidates).

Finally, on the basis of the success of a small, manually selected seed set for the emotion category, we experimented with using a similar strategy for the medication and signs and symptoms categories. We randomly shuffled the posts in our data and manually selected the first ten terms we saw that belonged to each category i.e., without any reliance on any dictionary. We re-ran our method by filtering these small seed sets and constructing context vectors, and thus the resulting seed sets were at most ten words randomly chosen for each category. Table 4.6 shows the random seed sets in each category. The starred terms were filtered out automatically at the pre-filtering step when creating the representative vector for a given category.

For medication names, using a small set of random seeds was very successful, achieving 66% precision on the top 50 ranked results (74% if names of treatments are included), as compared to 44% and 62% when using RxNorm as basis for the seed. This demonstrates that if the target class is well defined, our method can learn accurate information from only a small number of examples, and a large, manually compiled lexicon is not necessary. For the category of signs and symptoms, the small randomly selected seed sets were also very

effective. For single-word terms, the small seed set achieved 44% precision on the top 50 ranked results, significantly higher than that achieved by using UMLS and SIDER as seed sets, where the accuracy was 38% and 34%, respectively. For two-word terms, the randomly selected seed set achieved similar precision to using UMLS (62% on the top 50), but was not as effective as using SIDER (88%).

Medications	**Signs & Symptoms, Single word**	**Signs & Symptoms, Two words**
tamoxifin	Pain	allergic reactions
herceptin	Leakage	mood swings
taxol	Cyst	* distended abdomen
carboplatin	Nausea	mouth ulcers
taxotere	neuropathy	hot flashes
tylenol	Baldness	high fever
xeloda	Blisters	scar tissue
zofran	Fatigue	* temple pain
percoset	headaches	abdominal pain
avastin	exhaustion	back pain

Table 4.6: Random seed set for Medications, Signs and Symptoms single words, and Signs and Symptoms two words. The terms with asterix were filtered out automatically during the step for construction of the representative vector.

4.4 Improving seed and candidate term selection

As previously mentioned, named entity recognition and lexicon creation share the same workflow, which essentially identifies terms of interest belong to certain semantic types from OHC text in an unsupervised fashion. The two tasks can differ slightly, however, in how the results are interpreted and used. Lexicons, usually used by downstream applications as linguistic resources, require the extracted set of terms to be precise in representing domain semantics. NER, on the other hand, is usually part of pre-processing of text and its output is used as part of the feature set; therefore, for NER, coverage is as critical as precision. In the previous section, we use a quite strict metric to include candidate terms, requiring the candidates to be proper names and to appear more than 50 times in the dataset. For

named entity recognition or lexicon creation which target higher recall, however, the criteria should be re-designed to include more candidate terms.

We also learned from the example study on BC dataset the importance of seed term set, which is another issue to be further investigated in this section. Particularly, we have found that terms collected from standardized terminologies are effective in representing domain knowledge. In this section, we make this part of our method more general, by defining seed term collection in a more systematic way. Also, since our method pipeline for the lexical semantics is unsupervised, and the only manual work is to create a seed term set for each semantic category, the method is also able to work on other tasks of NER or lexicon creation outside of OHC, which is an additional issue we will explore in this section.

In order to introduce a novel way of candidate term selection, to standardize seed term collection, and to evaluate both coverage and precision of the system (unlike in the previous section, we evaluated only the precision, basically), we applied our methods on a clinical dataset (I2B2) and a biomedical dataset (GENIA) (see section 3.4 for details of the two data sets) which are used as benchmarks in many previous studies. Details of the evaluations can be found in the original paper [Zhang and Elhadad, 2013], and in this thesis we only discuss the two steps of interest, seed term collection and candidate term collection.

Along with this study, we created an unsupervised named entity recognition tool which can be used to identify any types of biomedical entities, and we made the source code and tool available online at `http://people.dbmi.columbia.edu/~shz7004/ner.html`.

4.4.1 Seed term collection based on UMLS semantic type mapping

We generalize the definition of seed term set by mapping them to corresponding UMLS semantic types, semantic groups [McCray *et al.*, 2001], or specific concepts which best represent the semantic meanings of the classes. Two of the entity classes, Medications and Signs&Symptoms, in our previous study on BC dataset, can be represented by following semantic groups in UMLS:

- **Signs&Symptoms**: Sign or Symptom (Semantic type)
- **Medications**: Clinical Drug (Semantic type)

The representation can also be applied to other types of tasks outside of OHC. For

example, for the I2B2 dataset, the three entity classes Problem, Treatment, Test can be represented by following semantic types or semantic groups:

• **Problem**: Disorders (Semantic group)

• **Treatment**: Therapeutic or Preventive Procedure (Semantic type) + Clinical Drug (Semantic type)

• **Test**: Laboratory Procedure (Semantic type) + Laboratory or Test Result (Semantic type) + Diagnostic Procedure (Semantic type)

For GENIA dataset, following semantic representations are assigned to entity classes:

• **protein**: Amino Acid, Peptide, or Protein (Semantic type)

• **DNA**: C0012854 (UMLS Concept)

• **RNA**: C0035668 (UMLS Concept)

• **cell type**: C0007600 (UMLS Concept)

• **cell line**: C0449475 (UMLS Concept)

Notice that the choices of semantic representations might not be absolutely accurate (actually, for some entity types like Problem, there is no clear UMLS semantic type). However, as our method allows noises in the seed term set, it is acceptable to pick the most likely representation based on one's expertise. Once the semantic representation is determined for a class, all the UMLS concepts (and their lexical variants) which belong to the representative semantic types or groups are extracted from the UMLS metathesaurus as part of the seed term set for that target entity class. If the domain representation of a class is defined by individual UMLS concepts, then all *is-a* descendants of those concepts will be included into the seed term set. For example, there is no proper semantic type or semantic group that could be mapped to the entity type "cell type" in the GENIA corpus. Instead, the individual UMLS concept "C0449475: cell type" is a good choice for the representation; thus, we collect all the *is-a* descendants of C0449475 (including all its lexical variants), as seed terms for "cell type." A mixed representation of semantic types/groups and UMLS concepts is also allowed for an entity class.

At the end of this step, we will have a dictionary for each target entity class, which we assume to be the set of seed terms for that class. Semantic representations and number of seed terms collected according to the representations for entity classes in BC, I2B2, and

GENIA are described in Table 4.7.

The generalized definition of seed term set is easier to be customized in different tasks, requiring no manual selection of seed term sets. One can define seed term set just by setting up proper mappings to UMLS semantic types that can represent the semantic category of interest.

Dataset	Class	Domain representation	# Seed terms
BC	Signs	Sign or Symptom (ST)	11,002
	Medications	Clinical Drug (ST)	12,939
i2b2	Problem	Disorders (SG)	398,725
	Treatment	Therapeutic or Preventive Proc. (ST) + Clinical Drug (ST)	153,084
	Test	Laboratory Proc. (ST) + Laboratory or Test Result (ST) + Diagnostic Proc. (ST)	66,015
GENIA	protein	Amino Acid, Peptide, or Protein (ST)	35,351
	DNA	C0012854 (C)	45,671
	RNA	C0035668 (C)	1,029
	cell type	C0007600 (C)	423
	cell line	C0449475 (C)	264,729

Table 4.7: Domain representations for entity classes in BC, i2b2 and GENIA corpora (ST: semantic type; SG: semantic group; C: concept).

4.4.2 Candidate term collection by NP phrase boundary detection

Instead of manually picking candidate terms, using strict threshold of occurrence, or allowing only proper names, we relax the inclusion criteria of candidate term by hypothesizing that all noun phrases (NPs) can be candidate terms, and use an NP chunker to approximate the set of NPs. Although full parsing is needed to find all NPs in a sentence, chunking is more time efficient and its coverage is quite acceptable in most applications. However, it is clear that not all noun phrases in the text can be entities. In order to remove those noun phrases that are clearly not entities of interest, we employ an inverse document frequency (IDF) based technique to filter candidates generated by the NP chunker. The intuition behind this filter is that noun phrases that are most common in the texts, such as "the patient" and "date of birth," are very unlikely to be entities. IDF is a measure of whether a term is common or rare across all documents [Manning *et al.*, 2008b]. Given a corpus D

of documents (sentences in our case) d and a specific term t, IDF is defined as:

$$IDF(t, D) = log(|D|/|d \in D : t \in d|) \tag{4.1}$$

We calculate IDF value for every word in the dataset, and obtain the IDF value for a noun phrase by averaging the IDFs of the words it contains. Then we filter all the candidate NPs whose IDF value is lower than a pre-determined threshold (set to 4 in our experiments). The reason of using such averaged IDF for a noun phrase instead of calculating the IDF value of its own directly is to handle the inherent sparsity of the copora: there are much more possible noun phrases than words in a given dataset.

In order to verify the hypothesis that entities are NPs, we report the coverage of noun phrase chunks on entities (Figure 4.6) in the datasets of I2B2 and GENIA. In all the three corpora of i2b2 as well as GENIA, around 45% of the entities are NP chunks, and nearly 30% of the entities are part of (but not) NP chunks. Only less than 5% of them are completely out of NP chunks without any overlapping words with them. Thus, if we use the collection of NP chunks as an approximation of entity candidate set, around half of entities will be covered. If we allow fuzzy match (i.e., we do not expect the boundaries to be exactly matched with ground truth), only a very small portion of the entities will be missing.

	NP	**Sub-phrase of NP**	**Overlap with NP**	**Out of NP**	**Total**
Pittsburgh	15,254(48.96%)	11,594(37.22%)	2,945(9.45%)	1,361(4.37%)	31,154
Beth	4,657(45.24%)	4,234(41.13%)	963(9.35%)	441(4.28%)	10,295
Partners	3,268(52.51%)	2,072(33.30%)	679(10.91%)	204(3.28%)	6,223
GENIA	18,456(41.83%)	22,797(51.67%)	2,138(4.84%)	730(1.65%)	44,121

Table 4.8: Numbers and percentages of entities that are noun phrases(NPs), sub-phrase of NPs, overlapped with NPs, and out of NPs

To evaluate the effectiveness of the IDF filter followed by the NP chunking, we look into the candidate sets before and after IDF filtering for Pittsburgh dataset. Table 4.10 gives numbers of true positives, false positives, and false negatives of recognizing NPs as entities before and after IDF filtering, regardless of entity classes. Before IDF filtering, the NP chunker finds 72,768 noun phrases from the text, 15,254 of which are entities in gold standard and 57,514 of which are not. The IDF filter removes 17,058 (30%) incorrect

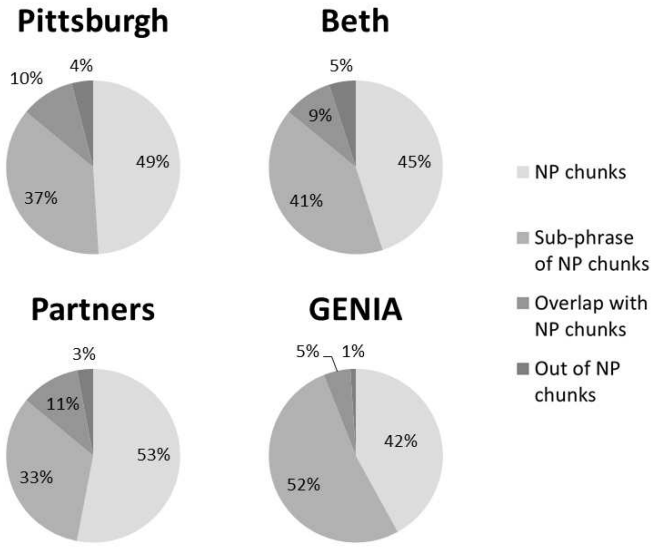


Figure 4.6: Proportions of entities in the corpora that are noun phrases (NPs), sub-phrases of an NP, overlap with an NP, and out of any NP.

	NP	**Sub-phrase of NP**	**Overlap with NP**	**Out of NP**	**Total**
Pittsburgh	15,254(49%)	11,594(37%)	2,945(9%)	1,361(4%)	31,154
Beth	4,657(45%)	4,234(41%)	963(9%)	441(4%)	10,295
Partners	3,268(53%)	2,072(33%)	679(11%)	204(3%)	6,223
GENIA	18,456(42%)	22,797(52%)	2,138(5%)	730(2%)	44,121

Table 4.9: Numbers and percentages of entities that are noun phrases(NPs), sub-phrase of NPs, overlapped with NPs, and out of NPs

candidates successfully, at the expense of only wrongly removing 967 (6%) NPs that should be entities. This supports our hypothesis that phrases that are too common tend not to be entities, and demonstrates the effectiveness of using averaged IDF value to filter candidates.

	True positives	**False positives**	**False negatives**
Before filtering	15254	57514	15900
After filtering	14287	40459	16867
Difference	967	17058	-967

Table 4.10: Effectiveness of IDF filter on Pittsburgh dataset

4.5 Alternatives to distributional representations: word embedding v.s. bag of words

One critical step in our method is to create a vector to represent the context of each term as the distributional semantic representation. In the most straightforward setting, the context vectors are derived from the vocabulary V found in the dataset, and each element in a terms vector contains a count of the number of times a word in V appeared in the term's context of interest. This bag-of-word model assumes numbers of occurrences of words in context as the units of distributional semantic representations. Studies introduced in the previous two sections were using this model, which is also popular in a wide range of NLP applications [Martin and Jurafsky, 2000].

In recent years, a novel type of feature representation has been proposed for NLP tasks such as text classification, parsing, and sentiment, namely, word embeddings [Collobert *et al.*, 2011; Turian *et al.*, 2010]. Similar with bag-of-word model, word embedding uses a vector of real numbers to represent semantics of words. However, word embedding is not obtained directly through counting words from context, but is usually learned via neural networks. Instead of using each word in V as a dimension in the vector, words are represented as vectors which could encode rich contextual information. Intuitively, embedding models encode hidden linguistic information that a word can convey in different context into a vector of certain dimensions. Previous research shows

that the embedding space is more powerful than the one-hot representation (e.g., bag-of-words), and that it makes breakthroughs in many NLP tasks as it conveys richer semantic meanings and is particularly useful in overcoming sparsity [Collobert *et al.*, 2011; Turian *et al.*, 2010]. Word embedding can also be seen as a dimensionality reduction on bag-of-words, in which dimensions are reduced significantly but important information are preserved.

4.5.1 Impact on lexicon expansion

In order to compare the effectiveness of the two models, word embedding and vanilla bag-of-words, in representing semantics of words in online health community text, we carried out a set of experiments to identify entities of treatment on the ASD dataset. Description of the dataset can be found in section 3.2, and details of the data annotation can be found in section 5.5.1. Basically, 4,264 entities representing treatments in 500 posts were manually annotated as the ground truth. In this study, we ignore the attribution labels of entities which will be discussed later, and only focus on how accurately the distributional semantic methods can help capture entities of treatment.

The experiment is a process of expanding lexicons of domain knowledge iteration by iteration using the two assumptions respectively. We are interested in how accurately these two approaches can capture semantically similar new words in the process of lexicon expansion. The method that better captures semantic meaning of word should be able to generate lexicons with higher precision and recall. The experiment went as follows in a semi-supervised bootstrapping fashion, which was a slightly modified version of the algorithm for lexical semantics described previously in section 4.2.

Step 1, we collect a seed term set which consists of 45 common drug names for autism patients from UpToDate (a complete term list see appendix A).

Step 2, for all the words appeared in the ASD dataset, we compute two semantic vectors through using bag-of-words and word embedding, respectively, by using all the data in ASD. For word embedding, we use the word2vec tools Continuous Bag of Words (CBOW) model [Mikolov *et al.*, 2013], and set vector size N = 100, iteration number 20 and all other parameters default. For bag-of-words, we rely on the methods described previously in this

chapter.

Step 3, for each seed term in the seed term set, we find its most similar words by computing the cosine similarities between vector of the seed term and vectors of other words, using the bag-of-words and word embedding vectors, respectively.

Step 4, for each seed term and each type of representation (bag-of-words v.s. word embedding), we find top 5 similar words and put them into the seed term set, and continue with repeating step 2.

As such, two seed term sets were expanded by adding similar terms using bag-of-words and word embedding vectors, respectively, iteration by iteration. Since all terms in the original seed term set are treatment names, words identified by this methods should be treatment names as well, ideally, according to the distributional semantic hypothesis. We ran the algorithm for both seed term sets for 3 iterations, as the seed term sets expanded rapidly. Then we used the expanded seed term sets after each iteration, along with the original seed term set, to do term matchings on the 500 annotated data. We evaluated how accurately (in precision, recall, and F) the two series of expanded term sets identified treatment mentions by comparing the results with manual annotated ground truth. Results for the first 3 iterations are presented in table 4.11.

In general, recalls elevated as seed term sets expanded, while precisions declined. For each iteration, word embedding was able to capture terms more precisely than bag-of-words, although bag-of-word model included more terms every time. In this particular application of identifying treatment entities from OHC text, it seems to suggest that word embedding is able to represent semantics and similarities between meanings of words better than bag of words.

Term set	**# of terms**	**precision**	**recall**	**F**
Seed	45	73.82	6.89	12.60
BOW-1	311	41.79	32.67	36.67
BOW-2	1128	32.06	55.11	40.54
BOW-3	3096	18.76	61.49	28.75
W2V-1	203	57.95	27.59	37.39
W2V-2	709	37.16	54.02	44.03
W2V-3	2114	23.46	61.59	33.98

Table 4.11: Performance measured by precision, recall, and F of keyword matching by using different term sets. BOWs represent seed term sets expanded by using bag of word representations. W2Vs represent term sets expanded by using word embedding vectors. Numbers following BOW and W2V represent numbers of iterations of expanding carried out before obtaining the term sets.

Chapter 5

Pragmatics of OHC Conversations: A Supervised Learning Approach

5.1 Introduction: tasks and methods

Content analysis for online health community requires not just semantic modeling at lexical level, but also understanding meanings of sentences, paragraphs, posts, and threads in dialogs, which we refer to as pragmatic analysis of OHC conversations. Identification of many variables of interest in our framework introduced in section 2.1 depends critically on pragmatic analysis. For example, to identify topic of a post, it would not be sufficient to rely only on semantic representations of words in the post; ordering of words, occurrences of certain domain-specific keywords, as well as thematic context of the conversation where the post locates, all contribute to the identification.

In this chapter, we present how we create different tools for pragmatic analyses for OHC content based on supervised machine learning. Supervised machine learning, in general, is about learning knowledge from (manually) annotated data, which can be applied to unannotated (unseen) data to make predictions. It has been applied to a broad range of fields such as information retrieval, speech recognition, computer vision, robotics, and natural language processing [Michalski *et al.*, 2013]. Compared with unsupervised learning which requires no labelled data and which is used in identifying lexical semantics in the previous chapter, supervised learning usually has the potential to reach much higher performance

[Michalski *et al.*, 2013]; however, it requires annotated training data which usually needs to be coded by human experts. One contribution of this thesis is thus to create high quality annotated datasets to support supervised learning tasks for online health communities. Equipped with these annotated corpora, we construct the pipeline of machine learning to build tools for different pragmatic tasks.

Specifically, we created gold-standard datasets, built and evaluated tools for four example pragmatic tasks which are also vital research questions explored in previous studies and are building blocks of our framework: identification of topic of discussion, sentiment analysis, debate and stance detection, and treatment attribution classification. The first task, identifying topic of discussion, aims at categorizing posts by the topic of information they convey. Sentiment analysis is the task of identifying the overall emotional polarity (positive or negative) authors express through writing. Debate detection, followed by stance identification, is the task of detecting arguements in OHC threads where users have conflicting opinions toward certain issues, as well as stances of participants in debates. The final task, learning treatment attributions, focuses on classifying mentions of treatment names in OHC texts, by whom the mentions should be attributed to (the speakers themselves, or some other ones, etc.). The four variables are representative ones of content, member characteristics, and member engagement in our framework. In this thesis, all these tasks are formulated as **classifications**, and as such can be approached by similar methodological pipeline based on supervised machine learning. Based on the pipeline, we devise four tools solving these problems respectively.

The overall pipeline of the supervised learning for all the tools is given in figure 5.1, consisting of steps from dataset selection to making predictions on unlabelled data. Definitions of all steps are given as follows. Some detailed information of methods (e.g. model selection, feature engineering) may not be covered in this chapter, and could be found in the original papers for topic [Zhang *et al.*, 2016c], sentiment [Zhang *et al.*, 2014], debate [Zhang *et al.*, 2016b], and attribution [Zhang and Elhadad, 2016b].

Dataset choosing. In this step, we pick the dataset for manual annotation, training, and evaluation. While online health communities do share some characteristics (e.g. most of them are organized in threads consisting of sequential posts), communities can differ

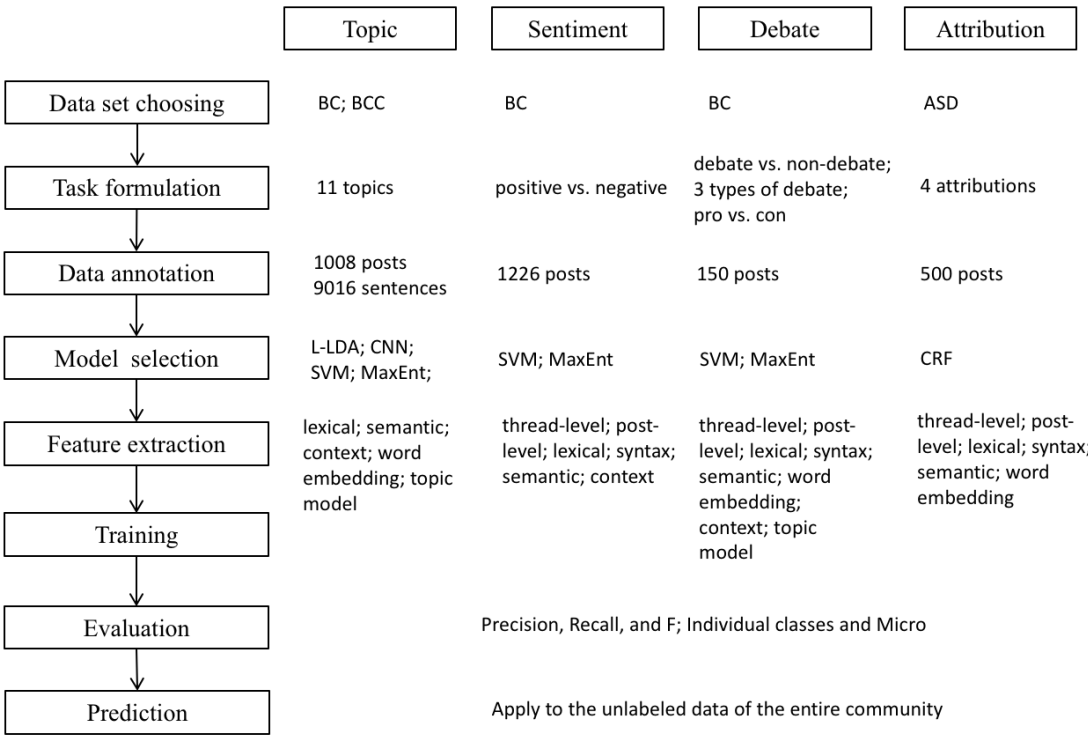


Figure 5.1: The methodological pipeline for four tasks of pragmatics of online health communities. A brief is given for each task at each step.

dramatically in ways members interact and in vocabularies of content, particularly when they target different patient populations, as we reported in Chapter 2. As such, tools trained on annotated content from one community might not be able to applied directly to content of other communities. For three of our tools, topic, sentiment, and debate, we relied on the BC dataset from the most popular breast cancer online forum. Our previous literature survey (see chapter 2) suggested that breast cancer, as a prevalent cancer with relatively high survival rate but long recovery period, and with patients primarily female, has been important subject of social support research and has attracted a lot of attention from psycho-oncologists and OHC creators. We hope that our studies based on the BC dataset can contribute to this active research area by providing new computational solutions. For the fourth tool, learning attributions of mentions, we relied on the ASD dataset because 1) we hope to apply our supervised learning framework on a different community of a different disease rather than breast cancer, and 2) autism communities usually involve both users who are patients (adult ASD patients) and users who are caregivers of patients (parents of autistic children), and thus has more demand of distinguishing the attributions of expressions.

Task formulation. All the four tasks are formulated as classifications in this thesis. For topic identification, we rely on a schema consisting of 11 topics, which will be introduced in the next section. For sentiment, we only consider positive and negative sentiment as most previous studies did. For debate detection and stance identification, we consider different categorization schemas, which we will be presenting in section 5.4.1. For attribution learning, a schema consisting of 5 categories will be introduced in section 5.5.1. It is noteworthy that the first three tasks are carried out at post level, while attribution learning is carried out at token level. This makes attribution learning slightly unique in task formulation, which creates opportunity of leveraging the Markov properties of sequences of tokens in the classification [Rabiner and Juang, 1986].

Data annotation. Figure 5.1 includes an overview of how many posts or sentences we annotated manually for each task as training data. Details of the annotation process will be discussed in corresponding sections. The annotated datasets underwent rigorous quality control, and are one of the main contributions of this thesis.

Model selection. Various machine learning models are leveraged depending on task formulation. Classical discriminative machine learning models such as support vector machine (SVM) [Suykens and Vandewalle, 1999] and logistic regression (MaxEnt) [Della Pietra *et al.*, 1995] were exploited in identifying topic, sentiment, and debate. For topic identification, we also experimented with convolutional neural networks (CNN) [Kim, 2014] and generative graphical models such as labelled latent Dirichlet allocation (L-LDA) [Ramage *et al.*, 2003]. For learning attributions of treatment mentions, conditional random fields (CRF) was used because the classification was conducted at phrase or token level, rather than post level. CRF model, based on the Markov assumption, has been proved to be an ideal choice in token-based sequential learning NLP tasks [Lafferty *et al.*, 2001].

Feature extraction. Several types of features were extracted for each post in each task. With respect to the organization of content in OHCs, features from threads, posts, and authors were extracted. With respect to linguistic hierarchy, lexical, syntactical, semantic, and discourse features were leveraged. Following types of features, in particular, were used in different tools.

Thread-level features refer to features that are identical across all posts in a thread. This includes number of posts in thread, number of authors in thread, average length of posts of a thread, as well as thread meta-information such as creation time and thread originator.

Post level features are those ones 1) from the meta-information of the posts, such as time stamp, author name, and author ID, and 2) basic statistics of the post content, such as number of words, number of sentences, etc.

Lexical features refer to words, lemmas, part-of-speech tags of the content. In our studies, we rely on existing open source tools such as NLTK [Loper and Bird, 2002] or OpenNLP [ope,] to extract these features from raw contents. The feature also includes occurrences of non-semantic tokens, such as question marks, exclamation marks, and mentions of user names.

Syntactical features are ones relying on the parse tree of the sentences. For example, in attribution classification, subject and predicate of sentences are important syntactical features used. Sentences were parsed by the StanfordNLP toolkit [Klein and Manning, 2003] in our experiments.

Semantic features refer to those ones representing domain knowledge, relying on lexicons created based on our unsupervised lexicon creation or named entity recognition methods, or existing lexicons such as WordNet [Miller, 1995] or UMLS [Bodenreider, 2004].
For all studies on the BC dataset, we used the lexicons described in section 4.3 to generate lexicon features. A straightforward keyword match of terms from the lexicons is carried out on the post to extract these features.
Word embedding refers to embedding vectors of words in content, which were introduced in section 4.5.
Topic model is the feature obtained through applying the LDA [Blei *et al.*, 2003] clustering on the raw content to achieve dimensionality reduction. For different tasks, we experimented with different sets of features, as shown in Figure 5.1.

Evaluation. All classifiers were evaluated with 5-fold or 10-fold cross validations using precision, recall, and F score as evaluation metrics, which are defined as follows respectively:

$$Precision = true\ positive/(true\ positive + false\ positive) \tag{5.1}$$

$$Recall = true\ positive/(true\ positive + false\ negative) \tag{5.2}$$

$$F = 2 * Precision * Recall/(Precision + Recall) \tag{5.3}$$

In order to evaluate the overall performance of the system across all classes in addition to accuracy of classification of individual categories, micro average precision, recall and F are also calculated for each task [Yang, 1999].

$$Micro\ precision = \frac{\sum_c true\ positive(c)}{\sum_c true\ positive(c) + \sum_c false\ positive(c)} \tag{5.4}$$

$$Micro\ recall = \frac{\sum_c true\ positive(c)}{\sum_c true\ positive(c) + \sum_c false\ negative(c)} \tag{5.5}$$

$$Micro\ F = \frac{2 * Micro\ precision * Micro\ Recall}{Miro\ Precision + Micro\ Recall} \tag{5.6}$$

In following sections, we present more details for each tool, with particular emphasis on data annotation and our tools' performance on specific tasks, since they vary by tasks and they are the primary contributions of this part of thesis.

5.2 Tool 1: A topic classifier

5.2.1 Data annotation

Data annotation of topics was carried out based on the BC dataset described in section 3.1. To enable reliable and useful annotation of topics, we established a coding schema of discussion topics through a literature review of information needs in online health communities, with an emphasis on breast cancer communities [Meier *et al.*, 2007c; Civan and Pratt, 2007; Blank *et al.*, 2010; Skeels *et al.*, 2010; Bender *et al.*, 2013; Kim *et al.*, 2013]. Our objectives were (i) to devise a coding scheme that is both relevant to describing the information needs of community members as well as applicable to and robust enough for automatic topic classification; and (ii) to design a coding scheme that can be applied to characterizing topics of discussion for either an entire post or its individual sentences. Furthermore, the annotation schema is such that each unit of annotation can be labeled according to one or more topics. For instance, a given post, and even a given sentence can simultaneously convey information about a treatment and the health system. To keep such topical heterogeneity as much as possible, our manual annotation is conducted at sentence level. Topics at post level is obtained through aggregating topics of sentences in post.

The coding scheme was developed using an iterative process to reflect the main topics of discussion of post content. Preliminary coding of 439 sentences (corresponding to 37 posts) provided the initial categories and guidelines for coding. Upon review and discussion, infrequently used categories were collapsed into larger concepts, and the 439 sentences were coded again to verify sufficient agreement between the two initial coders. The 439 sentences and their codes were used as training instances for the later coders, along with the coding guidelines.

Our final topical scheme contains 11 topics, as listed in Table 5.1. It is noteworthy that the topics focus on informational support, rather than emotional dimensions, and range from clinical to daily matters.

Since manual annotation of topics could be labor intensive and time consuming, we are unable to provide manual annotation of topics for all contents in the dataset. Instead, we selected a subset of posts, which contains 1008 posts, from the original dataset described

Topic	**Abbreviation**	**Description**
Alternative	ALTR	alternative and integrative medicine
Daily	DAIL	daily cancer-related experience
Diagnosis	DIAG	diagnoses, measurements, and results of tests
Finding	FIND	health finding, sign, symptom or side effect
Health Systems	HSYS	health systems patients interact with, including nurses, doctors, practices, hospitals, and insurance companies
Miscellaneous	MISC	greetings, uninformative sentence, or any sentence, which does not fit under any other annotation label
Nutrition	NUTR	Nutrition
Personal	PERS	personal information
Resources	RSRC	link, pointer, or quote towards an external information resource
Test	TEST	testing procedures (but not results of tests)
Treatment	TREA	treatments, including procedures, medications and therapeutic devices

Table 5.1: Annotation schema for breast cancer forum text

above. The posts were selected from the different forums, where each forum focuses on specific aspects of breast cancer management, such as diagnosis and treatment options, support through chemotherapy, nutrition, alternative treatments, and daily life. Posts were thus grouped in batches of 50 posts per manual annotation session.

Sentences were coded according to double annotation followed by an adjudication step from one dedicated adjudicator throughout the annotated dataset. Three coders were hired for the annotation, all female native English speakers with undergraduate degrees. To train for the annotations, coders practiced annotating the 439 sentences (37 posts) referred to above using the annotation guidelines. Inter-annotator agreement with gold-standard topic annotation was monitored throughout training, and training was terminated when a coder had achieved a 0.6 Kappa (agreement statistic) with the gold-standard annotation [Cohen and Others, 1960]. Note that given the large number of potential labels in the schema and the fact that each sentence can be labeled according to multiple topics, this is a particularly stringent training constraint. Afterwards, each batch of posts was assigned two coders and was doubly annotated at the sentence level. Finally, the adjudicator went through all posts, resolved differences between coders and made final decisions over sentence topic labels.

Table 5.2 shows distributions and example sentences for different topics in the manually-annotated dataset. Treatment and Miscellaneous sentences are the most frequent topics in our annotated dataset, whereas Alternative Medicine and Test topics are the least prevalent. The high number of Miscellaneous sentences is explained by the fact that most posts start with greetings and end with encouragements, blessings, and signatures (all categorized as Miscellaneous in our coding).

5.2.2 Evaluation

Here we report classification performance in F scores of different classifiers on sentence-level classification with 5-fold cross validations, since our original topic annotation is carried out at sentence-level. We report results in table 5.3. We found that CNN outperforms other model significantly in almost all topics. Labeled LDA, although relying only on the raw content without feature engineering, performs roughly on par with logistic regression and support vector machine which leverage complex features as presented in Figure 5.1.

Topic	**#Sentences**	**Example**
ALTR	302	I tried everything to no avail & in desperation had acupuncture.
DAIL	600	I use virgin organic coconut oil on my skin and all organic cosmetics, shampoo, conditioner, laundry detergent, household cleaner, the works!
DIAG	1127	My cancer was a 1.2 cm mucinous bc in a duct, with low growth rate.
FIND	1195	I don't feel faint or anything- it just feels weird- anyone else out there had this happen?
HSYS	864	I don't know where you are located, but I would start with the Cancer Treatment Centers of America.
MISC	1956	Hope this helps, cheers
NUTR	608	I am staying on a bland diet, eating every 2 hours, and forcing fluids, but am worried about tomorrow based on what happened last time.
PERS	1011	He has a family history of very high triglycerides.
RSRC	568	I just did internet research and here is a good site with information on Curcumin
TEST	295	When I went in for my second mammogram on Dec. 18th, the radiologist told me I had to go get a biopsy based upon the mammogram.
TREA	2078	I'm just curious about other warriors experience with herceptin.
ALTR,NUTR	113	I read that cinnamon capsules could help with lowering glucose and ldl in our blood.
HSYS,TREA	104	After dealing with the insurance company for weeks.....she finally started taking the Xeloda last month.

Table 5.2: Topic labels and the number of manually annotated sentences according to each topic. For each topic, an example of manually annotated sentence is provided. The table also includes two examples with multiple labels.

	L-LDA	**MaxEnt**	**SVM**	**CNN**
Micro	54.4	55.8	58.3	65.4
ALTR	9.2	9.4	30.7	35.5
DAIL	30.1	28.8	46.4	48.1
DIAG	58.8	60.2	65.3	67.1
FIND	50.1	50.9	60.0	60.3
HSYS	45.4	41.1	55.3	57.7
MISC	76.2	75.8	71.4	78.1
NUTR	57.3	58.6	68.4	72.8
PERS	24.4	26.5	47.7	47.8
RSRC	48.0	48.3	55.2	61.1
TEST	27.6	26.1	47.9	52.6
TREA	65.7	66.0	64.2	73.6

Table 5.3: Topic classification performance measured by F score on different topic categories, with four machine learning classifiers.

5.3 Tool 2: A sentiment classifier

5.3.1 Data annotation

Data annotation of sentiment was also carried out on the BC dataset (section 3.1). The process of sentiment annotation is similar with that of topic annotation. However, unlike the multi-label topic annotation across 11 topic categories, sentiment annotation is a process of binary choice which is much simpler for annotators.

A random sample of 1,226 posts from the dataset was manually annotated by two annotators according to the sentiment polarity they conveyed overall[Bo Pang and Lillian Lee, 2006]. To ensure annotators chose a polarity, we restrained the annotation to positive or negative only (no neutral), and provided guidelines and examples to the annotators. Overall, a post was considered positive if its author conveyed typical positive emotions, like joy, happiness, gratitude, as well as curiosity, whether intellectual or towards other participants. Conversely, a post was considered negative if it conveyed negative emotions, such as anger, anxiety, sadness, and hopelessness. Disagreements between the two annotators were adjudicated, resulting in a dataset of 1,226 posts annotated as either positive or negative sentiment.

The manual sentiment annotation of the 1,226 yielded nearly perfect inter-annotator agreement (Cohen's Kappa of 0.798). After adjudication and resolving disagreements, 859 out of 1,226 posts were annotated as positive, and 367 were annotated as negative. Following are examples of two positive and two negative posts:

Positive label *The recovery from my lumpectomy was easy. Really. Nowhere near as difficult as I imagined. Very little pain at all. I never needed any pain meds after surgery. Good luck.*

Positive label *I'm so happy you're feeling better!! Strange, but hey, that's our life these days.*

Negative label *I had a mastectomy about three weeks ago and will be starting chemo at the end of the month (Dec. 27th). I wake up every morning anxious and scared. When does this go away?*

	P (pos.)	R (pos.)	F (pos.)	P (neg.)	R (neg.)	F (neg.)
MaxEnt	87.1	86.5	86.8	51.4	56.8	53.7
SVM	65.3	71.6	68.4	58.5	58.3	58.4

Table 5.4: Sentiment classification performance measured by precision, recall, and F score for positive and negative sentiment, with SVM and logistic regression.

Negative label *Just had a 6month followup with my onc. My second round of scans came out clean. However in 3 months I will be doing bloodwork for tumor markers. She didn't discuss it with me and I don't know what it is about. I understand my cancer is aggressive, but what am I not understanding here? :(*

5.3.2 Evaluation

The classification performances of the three classifiers are given in Table 5.4. To demonstrate the effectiveness of machine learning models, performance of a baseline system is also given, which simply classified all posts as positive. The best performing system was logistic regression (MaxEnt). Both MaxEnt and SVM tended to classify posts as positive, caused by the uneven distribution of positive and negative samples in the training set. For logistic regression, once the threshold of prediction was calibrated towards favoring negative (i.e., a post is classified as negative once the predicted probability was lower than 0.6 rather than 0.5), the F score of negative was dramatically improved. Fortunately, in our following application to the entire dataset, we are more concerned with probabilities rather than discrete labels, since our modeling was based on the average likelihood of various groups of posts being positive or negative, rather than number of predicted positive and negative instances.

We analyzed the impact of individual features on the MaxEnt classifier, which assigns a weight to each feature after training, indicative of its discriminative power for the given task. Among all features, keywords representing negative emotions in our emotion lexicon (weight +2.7) had the strongest correlation with positive emotion, while negative emoticons (weight -1.9) were most correlated with negative emotion. On the contrary, bag of words (weight 0.003) and number of exclamation marks (weight 0.03) were borderline features,

suggesting similar distributions of these features in positive and negative samples.

5.3.3 Exploring sentiment classification on heterogeneous OHC data

As an exploration of how portable our sentiment classifier is, the tool is also evaluated on the BCC dataset, which consists of heterogeneous texts from multiple OHCs as well as from an expressive writing intervention (see section 3.3 for details of the BCC dataset). Specifically, we compare our machine learning based system with a well-established baseline: dictionary matching with LIWC [Pennebaker *et al.*, 2001]. Traditionally, health psychologists have used LIWC as the main tool for emotion analysis of text.

An annotated dataset was created based on a sampling of the BCC dataset. Originally, 20 types of emotions were considered in the annotation, including interest, fear, affection, gratitude, and so on. For the sake of application of our sentiment classifier, annotators were asked to merge emotion categories into three main types: positive emotion, negative emotion, and indirect emotion. Details of the coding process can be found in the original paper [Bantum *et al.*, 2016]. We manually coded emotion in 39,367 sentences from 476 posts. Of these, 31,872 (81%) sentences were classified as not containing emotion, 6,342 (16.1%) were classified as positive emotion, 971 (2.4%) were classified as negative emotion, and 182 (0.5%) were classified as indirect emotion. It is noteworthy that each sentence could be associated with more than one emotion, making the task a multi-label classification rather than binary choice between positive and negative. As such, instead of one classifier for positive/negative, two binary classifiers (positive or not, negative or not) were relied on for the experiment.

The baseline system includes classifiers straightforwardly use dictionary matching from LIWC. We used dictionaries from LIWC 2007 and 1997, compiling lists of words representing positive and negative emotions. For example, the classifiers code a sentence with positive emotion if any words in the sentence can be found in the list of words representing positive emotion compiled from LIWC. As such, each sentence could possibly be coded as both positive and negative, which is valid given that our task is multi-labeled.

The second system relies on the sentiment classifier we described above. We relied on the same set of features as ones we used for the BC study described in the previous section.

The only difference is that two classifiers, one for positive and one for negative, were trained.

class	**liwc: p**	**ours: p**	**liwc: r**	**ours: r**	**liwc: f**	**ours: f**
Positive	18.0 (1.3)	83.4 (1.9)	34.2 (1.6)	64.8 (1.9)	23.5 (1.4)	72.9 (1.9)
Negative	12.8 (2.2)	36.6 (2.8)	85.1 (2.8)	40.6 (2.6)	22.2 (2.4)	38.5 (2.7)

Table 5.5: Comparison of LIWC and our classifier on BCC dataset. p: precision. r: recall. f: f score.

Emotion classification results for LIWC are provided in Table 5.5. Of all the sentences classified by LIWC as representative of positive emotional expression, 18% were in agreement with human coders, and only 12.8% of sentences classified as negative emotional expression were in agreement with human coders. LIWC successfully identified 34.2% of all sentences containing positive emotional expression and 85.1% of all sentences containing negative emotional expression.

Our machine-learning based classifier significantly outperformed LIWC with respect to precision for both positive and negative emotional expression (83.4% agreement with coders for positive emotion and 36.6% for negative emotion; see Table 5.5) and recall of positive emotional expression (64.8% agreement with coders). LIWC outperformed machine learning with respect to recall of negative emotional expression (LIWC captured 85.1% of instances here), with machine learning agreement with raters occurring in 40.6% of instances, while this was the best category of prediction for LIWC. Overall F-scores were significantly higher for our tool for both positive (72.9) and negative (38.5) emotional expression.

In order to further examine the portability of our classifier, we also evaluated classification performance of our tool on the BCC dataset, by using BC dataset as training data (see table 5.6). It is expected that the classifier trained on BCC dataset outperformed its counterpart trained on BC. However, changing the training data did not change the fact that our machine learning based yield better results than the dictionary matching method. The experiments show that although our classifier was originally designed for sentiment classification on the breast cancer forum data, it can successfully identify positive and negative emotions for heterogeneous OHC texts other than ones from the BC forum. The results

suggest that our tool has certain level of portability across types of text, given the same type of domain knowledge underneath the content.

Training	**BC: p**	**BCC: p**	**BC: r**	**BCC: r**	**BC: f**	**BCC: f**
Positive	71.3 (1.5)	83.4 (1.9)	64.2 (1.6)	64.8 (1.9)	67.6 (1.5)	72.9 (1.9)
Negative	29.1 (2.5)	36.6 (2.8)	38.7 (2.6)	40.6 (2.6)	33.3 (2.5)	38.5 (2.7)

Table 5.6: Comparison of classification performance on BCC dataset, by using BCC and BC dataset as training data respectively.

5.4 Tool 3: Debate and stance detectors

We create our debate and stance annotation on the BC dataset which is described in section 3.1. To evaluate our tools, we construct a gold standard dataset focusing on a particular type of content: discussions of complimentary and alternative medicine (CAM) (see Chapter 10 for details). CAM is usually not accepted by medical establishment and thus can be controversial among patients with respect to its effectiveness, which easily triggers debates in OHCs. Although we focus on CAM discussions in our data annotation and evaluation, the tools described below are applicable to detecting any types of debates given necessary training data.

5.4.1 Data annotation

To assemble a gold standard of posts with debate information, we relied on a manual annotation process. We asked two annotators to determine two types of information for each post in the gold standard: if the post is involved in a debate, and whether the post is for or against alternative medicine usage. The annotation process started with a pilot annotation of 50 posts, in which the annotators made sense of the task by deciding which types of debates of interest to identify, which led to a consensus on three types of debates to be considered: CAM debate (debates over effectiveness/impact/side effects of certain CAM usage), BC debate (debates over other cancer-related topics), and other conflicts

amongst members. The two annotators then annotated 100 posts each to calculate inter-rater agreement, and the remaining part of dataset is coded by single annotator only.

For the first annotation task, deciding whether a post is involved in a debate is heavily dependent on the context of the post: how its author interacts in this post with other posts in the thread, and what the general topic of the thread is. As such, to construct our gold standard, we sampled posts from entire threads rather than individual posts throughout the community. For threads with a reasonable number of posts, the annotators annotated all posts in the thread. However, for the giant threads, which often occur in such communities, the annotators annotated the first 180 posts in the thread. Overall, 1,066 posts within 73 threads were annotated. As previously mentioned, we are interested in controversial topics which trigger debates involving opposing opinions, rather than treatment options that are comprehensively accepted and mostly uncontested. As such, a debate in our definition must involve different stances from different participants, and should have some degree of opposing interactions. In other words, a post simply stating an opinion but not disagreeing explicitly or implicitly with anothers opinion, as well as receiving no opposing responses from other persons, would not be considered a debate post, even if it represents a stance on the issue.

For the second task, stance identification, only posts identified in the previous step as CAM-related debates were considered. A "con CAM" stance was annotated, when the posts author opposes the usage of the specific CAM under discussion, are suspicious of its effectiveness, or concerned about its side effect. Any other opinion, including willingness to try a CAM, defending its effectiveness, or describing the outcome objectively, was considered as a pro CAM stance.

The two annotators reached an inter-rater agreement measured by Cohens Kappa 0.68 on the 100 double-annotated posts with respect to debate identification [35]. Disagreements were then resolved. Out of the 1,066 annotated posts, 174 were coded as debates. Specifically 97 were coded for debates about CAM, 37 for debates about other breast cancer related topics, and 40 coded general conflicts. Figure 5.2 shows an example of a series of debate posts in a thread with context, and Table 5.7 gives examples of debate discourse out of context for the three types of debates, respectively.

Figure 5.2: An example debate in thread. Green and blue posts were published by two users engaged in the debate respectively. Grey posts are not engaged in the debate, but provide context. User names are removed from the text and replaced by X, Y, and Z.

Type of debate	Example post
CAM	Laetrile is snake oil and potentially dangerous. it is illegal to sell it as a cancer treatment because there is zero evidence to so much as suggest that it has any efficacy.
Breast cancer related	X, Y is correct. Please read all parts of your link. It clearly states that dcis can be any size.
Other	X, no offense taken and I usually agree with you on the harmless/lonely bit. However, there were some truly over the top comments made that needed to be addressed, IMHO.

Table 5.7: Example posts annotated as three types of debates (presented here out of their thread context). User names are removed from the text and replaced by X and Y.

The inter-rater agreement of stance identification between the annotators was 0.77. After resolving disagreements, 97 posts were annotated as in CAM related debates, 67 were annotated as supporting and 30 against CAM usage.

5.4.2 Evaluation

Table 5.8 lists the precision, recall, and F measures for the different models for the binary classification of a post into debate vs. non-debate. The baseline always classifies a post as debate. The thread classifier only considers thread-level features, while the Thread+Post model consider thread- and post-level features. To investigate the value of different lexical features and lexicons, we looked at how the binary classification performs when they are used jointly with thread- and post- level features. For each experiment, we calculated 95% confidence intervals by considering the 5 individual folds as 5 re-samplings from the dataset and assuming that the performance scores are normally distributed. The confidence intervals can be used to measure whether differences amongst systems are indeed significant.

The system using only thread features performed poorly as expected, since all posts in one thread would have identical thread level feature values, making it impossible for the classifier to make post-level distinctions. However, the system outperforms the baseline,

	Precision	**Recall**	**F**
Baseline	16.3	100	29.6
Thread	26.1 (4.7)	73.4 (4.5)	38.2 (4.7)
Thread + Post	60.6 (4.0)	83.9 (3.7)	70.4 (3.9)
All	64.6 (3.5)	89.6 (3.7)	75.1 (3.7)

Table 5.8: System performance for binary debate classification with different types of features. The baseline system simply classifies everything as debate.

	Precision	**Recall**	**F**
Non-debate	71.4 (5.7)	79.1 (5.1)	75.1 (5.4)
CAM	58.0 (6.8)	73.9 (6.7)	65.0 (6.7)
Breast cancer related	43.4 (8.4)	41.3 (9.1)	41.9 (8.7)
Other	55.1 (7.7)	59.4 (7.8)	57.2 (7.8)

Table 5.9: System performance for 4-class debate classification with all features combined.

indicating that thread-level features are still somewhat informative. The system relying on all features combined yielded the best performance, but differences amongst systems, except the one using only thread level ones, are not significant, primarily because of the relatively small sample size.

Another set of classifiers, which were trained with 4 types of annotated debates (including non-debate), were also evaluated. Table 5.9 shows detailed performance for each class by using all features combined. Since decomposing binary into 4-class makes the dataset sparser and the task more challenging, it is reasonable that accuracies of prediction drop for all categories compared with the binary result.

Table 5.10 shows the performance of stance classification (pro vs con) on the gold-standard CAM-related debate posts. Like for the previous experiments, the different models are cross validated, and evaluation is reported through precision, recall, and F score for the con-CAM class. The baseline system simply classified everything as con-CAM. It is interesting that systems performance using only thread-level features is identical to the

	Precision	**Recall**	**F**
Baseline	30.9	100	47.2
Thread	50.1 (6.4)	44.7 (7.0)	47.3 (6.8)
Thread + Post	67.7 (5.4)	63.0 (5.7)	65.3 (5.6)
All	69.6 (5.8)	70.6 (5.7)	70.1 (5.7)

Table 5.10: System performance for binary stance classification with different types of features. Precision, recall, and F are calculated for the con-CAM class. The baseline system classifies everything as con-CAM.

baselines, which suggests that thread-level features add no information in distinguishing post-level stances.

5.5 Tool 4: An attribution classifier

5.5.1 Data annotation

We used the ASD dataset, described in section 3.2, to train the attribution classifier. Five types of attributions were considered in the manual annotations, with descriptions given in table 5.11. In general, the labels were designed to reflect whom the treatment is tied to. In online health community text, an entity of a treatment does not necessarily indicate an actual history of usage. For instance, in "The doctor suggested to put my son on risperdal", although the mention "risperdal" is associated with the patient (my son), it is not clear whether the drug is actually prescribed or taken. Therefore, in order to support subsequent user modeling applications in which we establish a treatment catalogue for each user, in the annotation schema we distinguish mentions of treatments attributed to patients which do and do not indicate actual usage or usage history.

A randomly sampled 500 posts were extracted and split into two sub-sets, with 50 posts overlapping (i.e. first set from post 1 to post 275, second set from post 225 to post 500). Two annotators were asked to 1) identify mentions of treatments (entities) from text, and 2) annotate the attribution label for each mention. It is noteworthy, however, that the

Attribution label	**Description**
Patient	Mention of treatment which indicates an actual usage or usage history of the patient, usually child of the user in this particular forum, of interest.
Patient general	Mention of treatment tied to the patient but does not indicate actual usage.
Caregiver	Mention of treatment tied to the caregiver of the patient, usually the user herself.
Others	Mention of treatment tied to specific individuals other than the caregiver or the patient. Can be other members in the community, or other people in the author's real life.
General	Mention not tied to a specific individual.

Table 5.11: Attribution labels for treatment mentions and their descriptions.

annotators were asked to classify attributions locally, without considering context which may shift the attribution of a mention. For example, in "The doctor suggested to put my son on risperdal.....My son tried risperdal and...", the first mention of risperdal should be labeled as patient general, even if following context actually indicates an actual usage of the same drug. In the annotation for this task, we did not consider co-references, e.g. pronouns which refer to treatments.

The annotation started with each annotator coding the overlapping part of the two sets, on which we tracked inter-rater agreement. Our annotators reached a Kappa of 0.77. Disagreements were resolved, and the remaining parts of the two sets were coded by the two annotators independently. In total, 4,264 mentions of treatments were identified. Among them, 434 were annotated as patient general, 1830 as patient, 210 as others, 95 as caregiver, and 1635 as general.

5.5.2 Evaluation

Three separate sets of evaluations were carried out for this task. Since the model we rely on, the conditional random fields (CRF), handles term identification and term classification jointly, it is necessary to evaluate these two separate steps in a explicit way. As such, the first set of evaluations is to examine how well the classifier can detect treatment entities, regardless of their attribution labels. The second set of evaluations takes attributions into consideration and evaluates the end-to-end performance of the method on the task. Finally, in our particular scenario of application in which we aim at building up treatment catalogue for each patient, we are more interested in one attribution label, the "Patient" class. Therefore, one additional evaluation is also carried out in which only two attribution labels are considered, "Patient" and "non-Patient". The "non-Patient" class is simply the aggregation of all attribution labels other than "Patient". For each set of evaluations, we report the performance of CRF model with different sets of features, ranging from basic lexical ones to syntactic features and information from context posts. In addition, we implemented a baseline system, which relies on keyword matching based on the "treatment" lexicon we collected.

Table 5.12 lists performance measured by F score for the treatment identification. All CRF-based systems, no matter what the features are, outperform the baseline significantly. However, syntactical features and features representing contextual information do not help the system performance. It seems to suggest that regardless of the treatment attributions, lexical features alone (including ones based on lexicons) are sufficient to identify the treatment mentions for CRF model.

Performance of the end-to-end evaluation of joint treatment mention detection and attribution classification is given in table 5.13. A true positive in this evaluation is a recognized treatment mention with both boundary and attribution correctly identified. As a result, it is a more challenging task since either an incomplete boundary or an incorrect attribution label will make the prediction counted as an error. The overall micro averaged F score is around 50 to 60, which varies by different feature sets. Syntactic features and context features, which represent global information from the whole sentence and whole post, are decisive in this task. Compared with the standalone evaluation of treatment

	Precision	**Recall**	**F**
Baseline	78.2	56.5	65.6
lexical+semantic	82.1	83.6	82.9
lexical+semantic+syntax	81.4	83.8	82.6
lexical+semantic+syntax+context	81.0	83.5	82.2

Table 5.12: System performance for binary treatment mention detection with different types of features. The baseline system relies on keyword matching from the "treatment" lexicon created based on the unsupervised lexicon expansion method.

identification, it seems to suggest that syntactic and contextual features are helpful for attribution classification, but not entity recognition.

Across the five attribution categories, our method is able to classify General and Patient better than the other three. This is primarily because of the distributions of these attributions in the training and test datasets - Patient and General are the most dominant attributions which provide more information for the classifier to learn from. Fortunately, in our downstream application in this thesis, building up treatment profiles for patients, only treatment mentions with Patient attribution will be used. To see if excluding other attributions from the dataset to make the classification as a binary choice (Patient vs. non-Patient) can help boost the accuracy of identifying mentions attributed to Patient, we carried out an additional evaluation in which General, Other, Patient-general, and Caregiver were merged into one class. The performance is given in table 5.14. Compared with table 5.13, accuracy of identifying Patient is boosted for around 4-5 percent, although the dataset, feature, and model keep exactly the same ones. The results suggest that properly formulating the task and setting up the target categories make significant difference in this type of tasks.

5.6 Effectiveness of feature engineering across tools

A wide range of features have been leveraged in the pragmatic analyses, including general and domain-specific features. General features refer to those that are general across different communities for different disease, while domain-specific features require domain knowledge

	micro	**cg**	**gen**	**other**	**pt**	**pt-gen**
lexical+semantic	55.4	18.2	56.0	37.0	61.6	19.2
lexical+semantic+syntax	56.1	18.1	57.4	36.8	61.7	20.9
lexical+semantic+syntax+context	62.3	18.2	64.1	51.6	66.8	34.1

Table 5.13: System performance (F score) for joint treatment detection and attribution classification with different types of features. cg: caregiver; gen: general; pt: patient; pt-gen: patient-general

	Precision (pt)	**Recall (pt)**	**F (pt)**
lexical+semantic	61.0	56.6	58.7
lexical+semantic+syntax	63.0	58.9	60.9
lexical+semantic+syntax+context	68.7	64.8	66.7

Table 5.14: System performance for mentions with Patient attribution with different types of features, when all other types of attributions are merged into one as non-patient.

about a disease or a community, which can be either from knowledge bases or extracted from OHC content in certain ways. From the linguistic perspective, several levels of feature representations, including lexical, syntactical, and semantic ones are exploited across different tools. Lexical features focus on individual words and phrases, capturing token level information such as word form, word stem, part-of-speech, and so on. Syntactic features describe syntactic roles of words and phrases, such as what the sentences' subject and predicate are, and positions of words in syntax trees. Semantic features represent meaning of words or sentences, including those salient terms representing domain knowledge from lexicons and thesaurus.

Two specific dimensionality reduction techniques are exploited to capture relevant information, to reduce dimension of the feature spaces, and to produce abstractions over content: topic modeling and word embedding. The two techniques are common-used unsupervised ways to overcome the sparsity issue of traditional bag-of-words representation, while preserving important semantic information. We use the results obtained by applying these two

	Topic	**Sentiment**	**Debate**	**Attribution**
Thread-level meta-information	NA	∘	∘	X
Post-level meta-information	X	X	∘	X
Lexical	∘	∘	∘	∘
Syntactical	NA	X	∘	∘
Semantic	∘	∘	∘	∘
Context posts	∘	X	∘	∘
Word embedding	∘	NA	∘	X
Topic modeling	∘	NA	∘	NA

Table 5.15: Effectiveness of different features in different pragmatic tasks for OHC content. ∘: feature effective in the tool. X: feature ineffective in the tool. NA: feature not applied in the tool.

algorithms on the dataset as additional features for our tools.

To obtain an overview of effectiveness of features across different pragmatic tasks for OHC content, we aggregate our results of evaluating different tools and compare how different features perform in different tasks, as an additional technical guide to future research. Table 5.15 summarizes effectiveness of all features across different tools. A feature is regarded as effective if the following condition is satisfied: the system performance of all features combined and the system performance of all features except the one being investigated are significantly different. Statistical significance test is carried out by running bootstrap re-samplings over the dataset using 5-fold cross validation.

Lexical features representing basic word-level information, as well as semantic information representing domain knowledge, are effective across all tools. Post-level meta-information, such as timestamp and author ID, is only helpful in identifying debates in the four pragmatic tasks, since debate detection as a task modeling member interaction is sensitive to how promptly members communicate with other members. Information from context posts is also critical to tasks that heavily depend on interactions among members, such as sentiment analysis and debate detection. On the contrary, attribution classification

and topic modeling depend more on content of post of interest, where context information show insignificant effectiveness.

Word embedding and topic modeling were exploited in tools other than sentiment analysis, and are proved to be helpful in topic classification and debate detection, which seems to suggest that these features only help in identifying theme-related variables.

Part III

Content Analysis for Modeling Members in Online Health Communities

Introduction

Equipped with computational tools and resources created in the previous part of the thesis, our next step toward characterizing online health communities is to build up multi-dimensional descriptions of members based on automated content analysis, consisting of what topics members discuss, what sentiment they express, what treatments they discuss and adopt, etc. Particularly, we are interested in discovering longitudinal patterns of how these member variables change through time, and in identifying correlations among these characteristics. By aggregating the descriptions, we are able to establish a mini modeling of characterization for each member, which is the primary aim of this part of the thesis. With respect to our framework, in this part of the thesis we focus on to what extent content analysis can help identify and model critical member characteristics (un-greyed parts in figure 9.2) .

Traditionally, as we discussed in section 2.3, content analysis and member characterization were investigated separately. Content analysis usually relied on qualitative review of sampled posts, and member characterization was carried out by collecting subjective data from patients directly in a controlled research setting. Due to relatively small sample size in traditional studies of online peer support groups and existence of confounding factors, linking patterns discovered in content to member characteristics in a statistical significant sense can be difficult. The small sample size also prevented researchers from identifying representative characteristics of a patient population [Zhang *et al.*, 2016a]. Public online health communities, instead, provided large cohorts of members along with massive amounts of user-generated content. The bottleneck would then become proper tools that can support such large-scale analytics. In this thesis, based on the computational tools and resources we built, we are able to identify certain member-specific characteristics for all members, and to study longitudinally the trajectories of how these variables change across different patient populations. We do not attempt to cover every aspect of members in this thesis, and will focus on three important variables of member characteristics and content that are also widely investigated in previous research but mostly by non-computational methods (see Chapter 2 for evidence from literature review): topic of discussion, sentiment expression, and disease treatment profile.

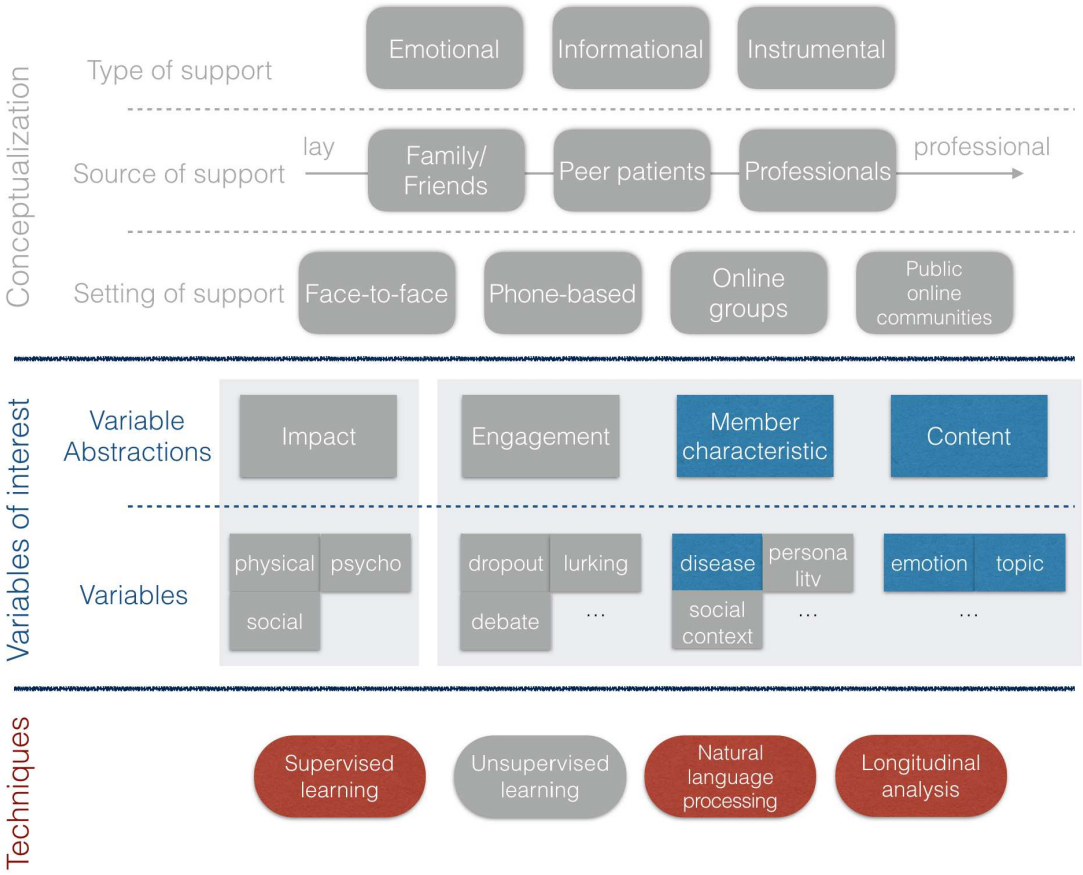


Figure 5.3: Variables of interest discussed in this thesis. Colored elements are the foci of this part of thesis.

In the next chapter, we present what topics of discussions are prevalent in an online health community and how prevalence of topics changes through time among breast cancer forum members, based on the application of our topic classification tool created previously on the entire BC dataset [Zhang *et al.*, 2016c]. In chapter 7, relying on the sentiment analysis tool, a longitudinal study is presented in which we investigate how sentiment of users changes through time as they participate longer in the community [Zhang *et al.*, 2014]. In Chapter 8, we present how catalogues of treatment can be created for members and how the frequencies of a treatment differ in discussion and in real practice [Zhang and Elhadad, 2016b]. In chapter 9, we aggregate above member characteristics and build a joint visualization which considers correlations among these variables, as our preliminary effort toward user modeling.

Chapter 6

Trajectory of topics discussed

In this chapter, we focus on one particular content-related variable, topic of discussion, and investigate at scale what topics members discuss in OHCs, and how topics of discussions change through time as members keep participating. As we discussed in Chapter 2, modeling topic of discussion is one basic step toward understanding OHC content and hence member behaviors. Topic of discussion may be correlated with other variables, such as members' disease severities, which is another issue to explore in this chapter. Specifically, we take the breast cancer forum as an example on which we carry out static cross-sectional and longitudinal topic analyses. We apply the computational tool introduced in section 5.2 to the entire BC dataset, to answer following research questions:

1 What are the most prevalent topics in discussions in the breast cancer forum?

2 Are there any differences of topic prevalence among users of different disease severities (e.g. cancer stages)?

3 How do members' foci of topics change through time, as members participate longer in the community?

The topic classifiers introduced in section 5.2 are able to identify which topics are associated with each post, with respect to the eleven topic categories designed for analyzing OHC content. By applying the best classifier, the one based on CNN, we obtain multi-label topic assignments for all post in the entire BC dataset. All following analyses will be

based on the output of the CNN classifier. For each of the analysis, we take one particular factor into account: whether the post is initializing a discussion or relying to other's post. Previous studies indicate that members seek support by initializing discussions and provide support by replying and giving feedbacks [Zhang *et al.*, 2014; Kim *et al.*, 2012b; Qiu *et al.*, 2011a], which necessitates the distinction between initial and reply posts in our analysis.

6.1 General prevalence of topics

Prevalence of all topics at post level is given in Table 6.1. The most prevalent topic is personal (PERS) among all posts, with 24.6% of posts labeled as such, followed by treatment (TREA, 24.6%) and diagnosis (DIAG, 9.3%). The least prevalent topics are alternative medicine (ALTR, 0.2%) and test (TEST, 1.0%). Specific to initial posts of threads, diagnosis is significantly more dominant than other topics, while popular topics among reply posts such as personal and finding are almost not found among initial posts.

In general, clinically relevant topics such as treatment, diagnosis, and finding are more prevalent than non-clinical ones, with one exception of PERS among all posts. Topic distribution in the entire BC dataset is more skewed that that in the annotated dataset, because the annotated dataset was sampled toward collecting more posts of rare topics such as alternative medicine (ALTR). Distribution of topics among initial posts is more uneven, suggesting that a significant amount of threads initialized by members focus on cancer diagnosis.

6.2 Topic prevalence stratified by cancer stage

In the breast cancer forum, many users self-reported disease information in their member profiles, including cancer diagnoses and treatment histories. These profile information show up in signatures when authors post, which is available to the public. One particular information that is mostly structured and easy to be extracted is cancer stage. Out of all 57,424 authors in the dataset we crawled, 17,950 (31.3%) have their cancer stage information available in signatures. Among them, 2,325 are stage 0 (total number of posts: 170,610), 5,968

All posts	**ALTR**	**DAIL**	**DIAG**	**FIND**	**HSYS**
	0.2	7.4	9.3	6.3	7.8
	NUTR	**PERS**	**RSRC**	**TEST**	**TREA**
	3.9	24.9	1.7	1.0	24.6
Initial posts	**ALTR**	**DAIL**	**DIAG**	**FIND**	**HSYS**
	0.0	0.8	46.4	1.4	7.1
	NUTR	**PERS**	**RSRC**	**TEST**	**TREA**
	0.6	8.0	2.7	0.1	22.9
p values	**ALTR**	**DAIL**	**DIAG**	**FIND**	**HSYS**
	0.54	0	0	0	0.002
	NUTR	**PERS**	**RSRC**	**TEST**	**TREA**
	0.011	0	0	0.040	0

Table 6.1: Percentages of all topics at post level based on automated topic classification, for all posts and initial posts respectively. Differences were measured by t-tests and p-values are reported.

are stage I (total number of posts: 600,500), 5,907 are stage II (total number of posts: 661,990), 2,447 are stage III (total number of posts: 229,955), and 2,438 are stage IV (total number of posts: 460,313).

Topic distributions of posts published by members of different cancer stages are given in Figure 6.1 for all posts and Figure 6.2 for initial posts. Statistical tests (multi-variate and univariate t-tests) were also carried out between numbers of different stages. Most visible differences in the two figures are statistical significant, given relatively large sample size. Stage 0 users focus more on cancer diagnosis and health systems, which are typical topics at early time of cancer journeys. Stage IV members, counter-intuitively, discuss more about personal lives but significantly less about treatment and clinical findings. This seems to suggest that stage IV members rely on the forum to exchange emotional more than informational support with their peers. Most differences found among all posts are even amplified among initial posts. One particular pattern among initial posts is that members with stage information, in general, posts significantly less about diagnosis than other members in initial posts. One explanation might be that many of the initial posts discussing diagnosis are published by new members to the community, many of whom only posted a few times which are asking questions about whether certain signs they found indicate cancer.

6.3 Topic trajectory of users

Armed with topic labels for each post in the dataset, we conducted the following longitudinal analyses to take timestamp into account. The primary objective for our analysis was to assess if participation in the community has an impact on topic of discussion. We compared distributions of topics of posts published in different periods of time with respect to users registration date, and tracked their changes. As such, each data point we consider is the average frequency of a topic within all posts in a given time slice (e.g., all posts published by their authors after 3 weeks of their joining the community). To visualize the changes in topic distributions through time, we plotted in addition to the individual data points fitted curves. To show both short-term and long-term changes, three measures of time progression

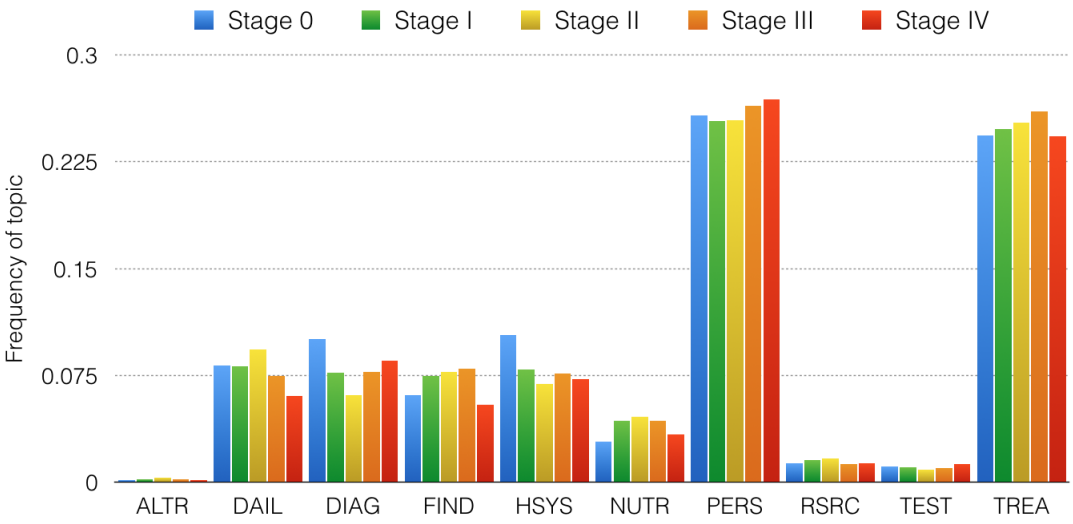


Figure 6.1: Frequencies of topics of all posts, stratified by cancer stages of authors.

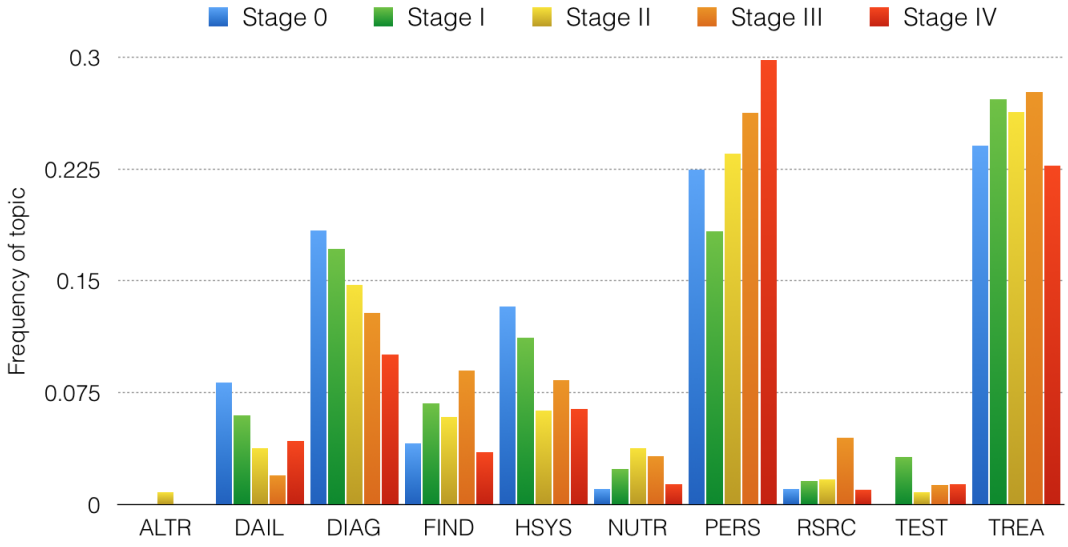


Figure 6.2: Frequencies of topics of initial posts, stratified by cancer stages of authors.

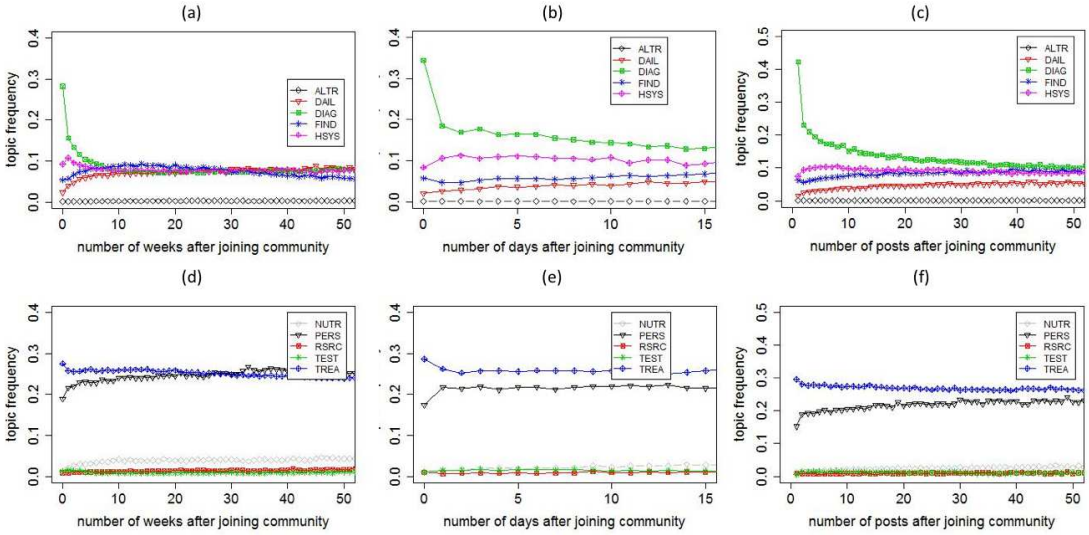


Figure 6.3: How topic frequencies of all posts change through time after members join the community. X axes represents the time point after members' first activity. Y axis is the average topic frequency of all posts that are published in the corresponding time. Units of x axes in (a)(d), (b)(e), and (c)(f) are weeks, days, and post orders, respectively.

are used (represented as x-axis): post, day, and week. In addition, we split our analysis by considering all posts (Figure 6.3) and initial posts of discussions (Figure 6.4) separately.

Several patterns are identified among all posts. First, diagnosis is the most dominant topic at early stages of participation, especially in first posts and first days. Second, prevalences of some topics such as personal (PERS), daily matters (DAIL), and nutrition (NUTR) grow steadily, while prevalences of diagnosis (DIAG) and treatment (TREA) decline as members stay longer in the community. Third, frequencies of health systems (HSYS) and findings (FIND) increase at the beginning, but slide after reaching the peaks. Finally, alternative medicine (ALTR), laboratory test (TEST), and resources (RSRC) are unpopular topics throughout members' participation. The results suggest that members' focus shifted from informational support, represented by clinically concentrated topics such as diagnosis and treatment, to emotional support, represented by personal focused on topics such as nutrition and daily lives.

Initial posts of discussions show simpler patterns. Frequency of diagnosis, as the most

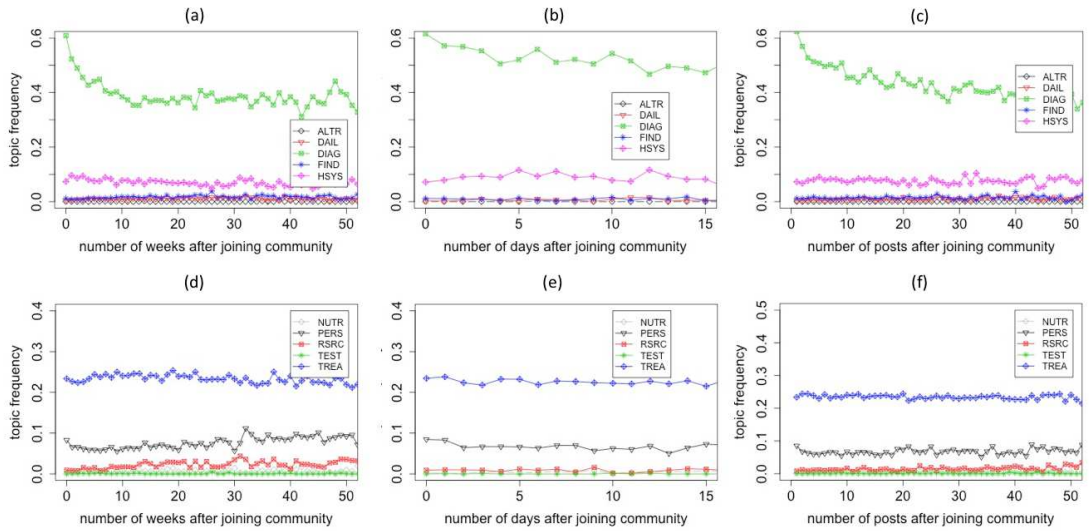


Figure 6.4: How topic frequencies of initial posts change through time after members join the community. X axes represents the time point after members' first activity. Y axis is the average topic frequency of all posts that are published in the corresponding time. Units of x axes in (a)(d), (b)(e), and (c)(f) are weeks, days, and post orders, respectively.

prevalent topic among initial posts, declines as members stay longer. Frequencies of other topics do not show clear patterns of changes.

6.4 Summary of findings

A wide range of topics are discussed in the online health community, ranging from clinically relevant ones such as diagnosis and treatment to more daily matters such as nutritional supplements and stories in personal lives. In the breast cancer forum, personal and treatment are the most dominant topics, possibly representing a mix of emotional support and informational support being exchanged. When it comes to posts that initializing discussions, cancer diagnosis is the most prevalent topic. Topics representing more personal or daily issues barely show up in initial posts, although they are quite dominant among other posts.

Cancer stage plays a role in deciding members' topics of discussions. Early stage members, many of whom are newcomers to the community, care more about diagnosis related information. Stage 0 members, in particular, focus on whether certain signs indicate cancer. They also exchange anecdotes about their experiences with healthcare providers when being diagnosed. Late stage members, such as stage IV members, usually have stayed in the community for longer time. For these members, seeking information is no longer the major motivation of participation; on the contrary, they established closer relationships with their online peers, and disclose more personal information and support each other emotionally. It it noteworthy, however, that cancer stage information extracted from signatures may be inaccurate, since members may not report stage change timely. Also, it is naturally the case that members with late stages are more likely to be long time users, which makes length of membership an important confounder in considering differences between members of different stages.

Finally, we found that members shifted their focus in participation, from clinically relevant topics to more casual topics as they participate longer and longer. This coincides with the difference between cancer stages, and supports that the difference is caused by length of participation more than cancer stage. Putting all the findings together, we may get a more

complete picture of OHC participation with respect to topics: as members stay longer in the community, and build up closer relationship with their peers, they tend to disclose more personal information, discuss more private stories, and exchange more support emotionally; meanwhile, they seek help less but provide more, and shifted their interest from cancer diagnosis to cancer treatment.

Synthesizing above discoveries, the difference between initial and reply posts becomes somewhat expected [Qiu *et al.*, 2011b] and may be explained as follows: initial posts are more likely to be posted by new members, who ask questions and seek help more often than old members, while old members mostly provide help and reply to others' requests; meanwhile, these new members are more likely to be newly diagnosed patients focusing on cancer diagnosis, while a lot of active old members are in sessions of treatment and they exchange personal stories more often with their familiar peers. However, this cannot explain why clinical finding and health system are not prevalent among initial post, which needs to be further investigated future work.

Chapter 7

Trajectory of sentiment expressed

In this chapter, like what we did for topics in the previous chapter, we seek to understand the effect of changes in post sentiment overall through sustained participation in a community, toward understanding members sentiment expression as a member characteristic. We seek to answer the following research questions: (1) does member participation in the community over different periods of time have an impact on the member posts sentiment? And (2) do the following factors contribute to changes in posts sentiment: age of members, cancer stage of members, duration of membership, and amount of posting affect?

Our automated sentiment analysis tool outputs for each post a predicted probability of being positive, or sentiment score. The sentiment scores are useful, because they allow us to compare posts against each other. As such, the scores are not absolute representation of sentiment, but rather enable us to rank posts according to their sentiment polarity. The best performing classifier, the logistic regression (described in section 5.3.2), was applied to the entire dataset based on the model trained with the 1,226 annotated samples. The average sentiment score of the entire dataset was 0.785 (0.210 standard deviation). For the initial posts, the average sentiment score was 0.695 (0.263 standard deviation). In general, our research aligned with previous work on other online health communities that found initial posts to be less positive [Qiu *et al.*, 2011b].

Armed with such sentiment score for each post in the dataset, we conducted the following analyses. The primary objective for our study was to assess if participation in the community has an impact on sentiment. We thus compared average sentiment scores of

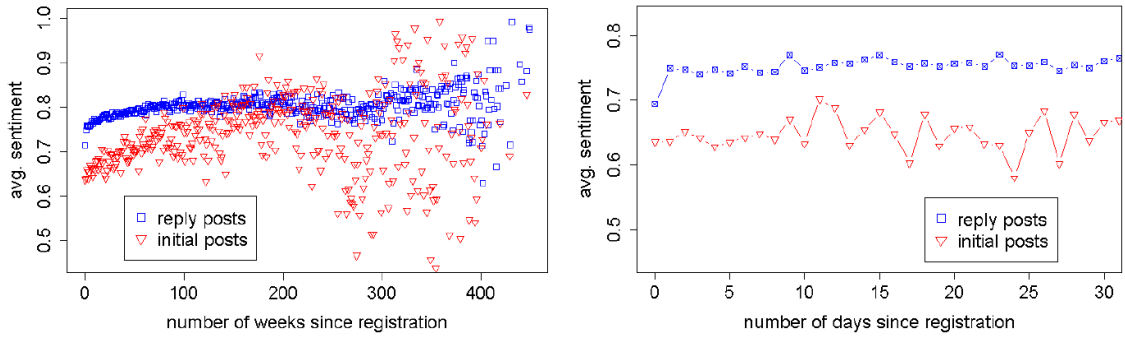
posts published in different periods of time with respect to users registration date, and tracked changes of sentiment. As such, each data point is the average sentiment of all posts in a given time slice (e.g., all posts published by their authors after 3 weeks of their joining the community). To visualize the changes in sentiment through time, we plotted in addition to the individual data points a fitted curve.

For our second research question, we considered three factors (age of members, cancer stage of members, and amount of posting) in both static and longitudinal analyses to examine their impact on post sentiment. In the static analysis, members were stratified by age/stage/amount of posting, and average post sentiments were calculated for each group. Statistical tests (ANOVA and TukeyHSD [Winer *et al.*, 1971]) were carried out to detect differences across groups. In the longitudinal analysis, sentiment scores were compared across stratified groups and duration of participation in the community to identify the patterns of sentiment change across members from different groups through time. All p-value were adjusted for multiple comparisons with the Bonferroni correction.

7.1 Longitudinal analysis of sentiment change

In order to examine the impact of participation through time in online discussion on sentiment overall, we plotted how sentiment scores changed through time, as computed since members registration date. The registration dates of users were provided in the profile information of metadata. Figure 7.1 shows the average sentiment scores of posts that were published after membership creation at both weekly (a) and daily (b) intervals. For example, the left-most blue data point in Figure 7.1(a) represented the average sentiment score of all reply (i.e., non-initial) posts published by all users respectively within one week of their joining the community.

Figure 7.1(a) indicates that, for both responding and initial posts, sentiment gets more and more positive through at least 100 weeks (2 years) of participation, with such changes most significant right after joining the community. Members, in their first days joining the community, publish posts, which are significantly more negative than later on. This is particularly true for initial posts, suggesting that newcomers to the community (likely newly



(a) (b)

Figure 7.1: Sentiment changes by length of membership at the time of posting, by number of weeks in (a) and number of days in (b). A colored point at (x, y) in the graph represents that the average sentiment score of all posts published by all users in the xth week (a) or day (b) after their registration is y.

diagnosed patients) express more anxiety and concerns than later in their questions to the community. Figure 7.1(b) provides a more granular view over the sentiment changes in the first 30 days of participation in the community, confirming that reply posts are significantly more positive than initial posts, and the increase of sentiment of initial posts does not happen until later on, at least 1 month into participation in the community. We do note a drastic increase in sentiment from posts published on the first day of joining the community to the later days, when looking at all posts (replies and initial posts combined).

In our dataset, the average length of membership of all users was 2 years 5 months (around 120 weeks); therefore, most of posts published after 200 weeks were written by a small portion of long-time users. We found that most of them were stage IV patients and showed a slight sentiment decline between 200 and 300 weeks. Topics of these posts were primarily about chemotherapy or metastasis/recurrence. While this set of posts is indeed homogeneous in sentiment and topic, it is difficult to assess the value of the analysis on such a small sample for the posts written by members who have been more than four years active in the community.

In order to obtain a more concrete understanding of how sentiment changed through sustained participation in the community, we grouped posts into nine groups, considering

both short-term and long-term periods of participation. The nine groups were posts published within one day of registration, 1-3 days, 3 days to 1 week, 1 to 2 weeks, 2 weeks to 1 month, 1 to 3 months, 3 months to 1 year, 1 to 2 years, and more than 2 years since registration. An ANOVA test was carried out for the groups, for all posts and initial posts respectively, followed by a TukeyHSD test to illustrate the significances of differences between all possible group pairs. ANOVA test showed significant difference among groups in both cases (p values ¡¡ 0.001). Post distribution, average sentiment scores, and p values compared with previous category given by TukeyHSD test are listed in Table 7.1. In this table as well as following tables, all posts represent initial posts and reply posts. Results showed same pattern as Figure 7.1, and demonstrated that the dramatic sentiment change after the first day was statistical significant in the case of all posts, while we could only see long term (3 months and then 1 year) significant changes for initial posts.

		<1d	**1-3d**	**3d-1w**	**1-2w**	**2w-1m**	**1-3m**	**3m-1y**	**1-2y**	**>2y**
All	sentiment	.693	.748	.745	.753	.756	.766	.782	.800	.804
	# posts	8,369	4,203	4,361	6,235	9,906	32,302	89,304	60,944	75,781
	p value	N/A	<0.001	1.000	1.000	1.000	0.025	<0.001	<0.001	0.577
Initial	sentiment	.636	.642	.637	.656	.644	.664	.685	.728	.760
	# posts	3,304	732	734	1,064	1,487	3,842	8,085	5,134	6,641
	p value	N/A	1.000	1.000	1.000	1.000	1.000	0.032	<0.001	<0.001

Table 7.1: Post distribution, average sentiment scores, and p values compared with previous category returned by TukeyHSD test, for all posts and initial posts respectively. The first p value for ¡1d is not available since there is no previous category to compare sentiment to. P values are adjusted for multiple comparisons with the Bonferroni correction.

7.2 Impact of member's age on sentiment

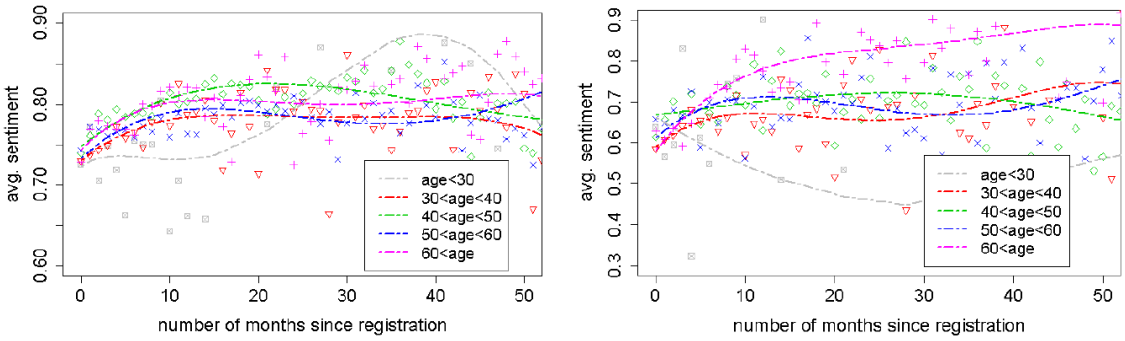
The posts in the dataset were published by 12,819 users, while a total of 14,919 user profiles were filled at least partially in the online breast cancer community and there were about 60,000 members overall. This meant that a very large majority of members were so called lurkers [Setoyama *et al.*, 2011], who never published anything but were likely to browse some of the posts. Behavior of lurkers was beyond the scope of this study. Rather, we

focused on members who had posted content. Among all non-lurkers, 1,211 provided date of birth in their profiles. Members born between 1960 and 1970 were the most dominant at the time of data collection, and the average age of all users were 47.5 (standard deviation 9.6 years), an older mean than in some other online health communities, such as weight loss forums [Hwang and Ottenbacher, 2010].

Age group (# users)		**<30 (38)**	**30-40 (198)**	**40-50 (485)**	**50-60 (358)**	**60+ (132)**
All	sentiment	0.742	0.768	0.793	0.778	0.791
	# posts	278	6,417	22,180	14,479	4,217
Initial	sentiment	0.614	0.643	0.681	0.681	0.744
	# posts	54	841	1,873	1,323	339

Table 7.2: Average sentiment scores and number of posts published by different age groups, for all posts and initial posts respectively. This analysis is restricted to posters who provided date of birth in their profile only, 1,211 members overall.

To study whether age affected sentiment, we considered members who disclosed their date of birth, and grouped them into 5 groups: below 30 years old, between 30 and 40, between 40 and 50, between 50 and 60, and above 60 years old. There were 47,571 posts in the dataset published by members with date of birth information. We calculated averaged post sentiment scores, and carried out statistical tests for the groups. Table 7.2 shows numbers of posts published by each age group and average sentiment score of posts of each group. The ANOVA test showed significant differences among groups for both all posts and initial posts. For all posts, TukeyHSD test found that difference between all pairs of groups were significant, except between <30 and 30-40, <30 and 50-60, and between 40-50 and 60+. For initial posts, differences between <30 and all other groups were not significant. We suspect that this is caused by the very low number of members in the age group <30, as expected in a community for a disease that affects older women predominantly. Members older than 60 showed markedly more positive sentiment than younger members, especially while publishing initial posts to start new threads. These facts might be explained by previous psychological finding of effects of older age on lower levels of psychological distress [Singer *et al.*, 2007; Hoffman *et al.*, 2009a].



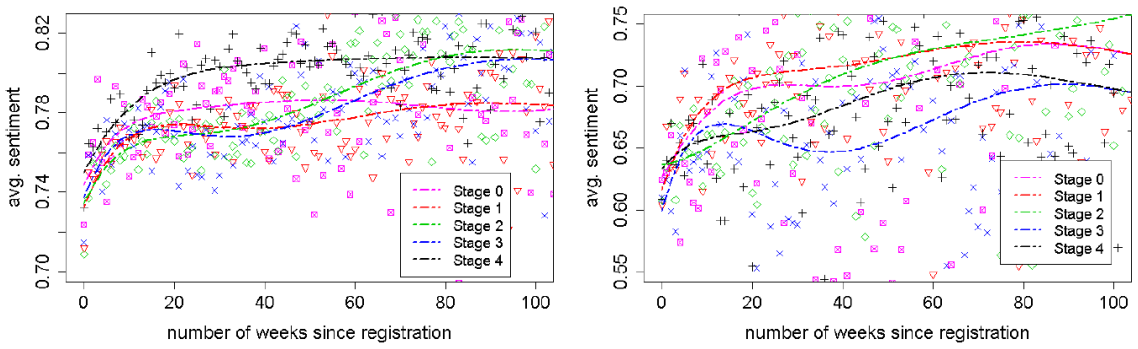
(a) all posts (b) initial posts

Figure 7.2: Sentiment changes by length of membership at the time of posting for different age groups, for (a) all posts and (b) initial posts. A colored point at (x, y) in the graph represents that the average sentiment score of all posts (a) or initial posts (b) published by users in corresponding age group in the xth month after their registration is y. Polynomial curves fitting each group were drawn for the sake of visualization.

To illustrate ages impact on longitudinal sentiment, sentiment changes over time after registration for different age groups were plotted, along with polynomial curves fitting each set of points to visualize the tendencies (Figure 7.2). Keeping in mind the very low sample size for members ¡30 years old, we do not attempt to interpret their longitudinal sentiment changes. For all other groups, however, the general trend observed earlier holds true independently of age: the longer the members participate in the community, the more positive their posts are on average. The observation that older members (>60 years old) post more positive posts, especially initial posts is visible as well on the plots.

7.3 Impact of member's cancer stage on sentiment

In our dataset, 4,602 users (who published 172,566 posts) had self-reported cancer stage information. Among them, 442 members were stage 0 patients, 1,407 were stage I, 1,544 were stage II, 650 were stage III, and 559 members were stage IV. Table 7.3 provides numbers and average sentiment scores of posts published by members in different stages. Although there were significantly fewer stage IV patients than stage I and II patients,



(a) all posts (b) initial posts

Figure 7.3: Sentiment changes by length of membership at the time of posting for different cancer stage groups, for (a) all posts and (b) initial posts. A colored point at (x, y) in the graph represents that the average sentiment score of all posts (a) or initial posts (b) published by users in corresponding cancer stage in the xth month after their registration is y. Polynomial curves fitting each group were drawn for the sake of visualization.

they published many more posts and formed the most active cancer stage group in breast cancer forum23. Moreover, stage IV patients were the most positives posters in term of the emotion expressed through the reply posts they wrote, but not initial posts. For all posts, comparisons between stage 0, stage I, and stage II, returns non-significant results according to adjusted p values. For initial posts, only the differences between stage I and stage III and between stage II and stage III were significant.

Cancer stage (# users)		**0 (442)**	**I (1,407)**	**II (1,544)**	**III (650)**	**IV (559)**
All	sentiment	0.775	0.771	0.776	0.782	0.796
	# posts	9,229	36,422	39,398	27,806	59,711
Initial	sentiment	0.675	0.690	0.687	0.661	0.675
	# posts	820	3,344	4,218	2,534	4,829

Table 7.3: Average sentiment scores and number of posts published by patients in different stages, for all posts and initial posts respectively.

Figure 7.3 illustrates longitudinal sentiment of different cancer stage groups. Not only were the stage IV users the most positive, but they also showed the fastest change towards

positive after registering in the breast cancer forum. However, these findings were specific to reply posts. These findings indicate that stage IV users seek support through starting threads with negative posts, but are very active in providing emotional support to their peers, through posting positive replies.

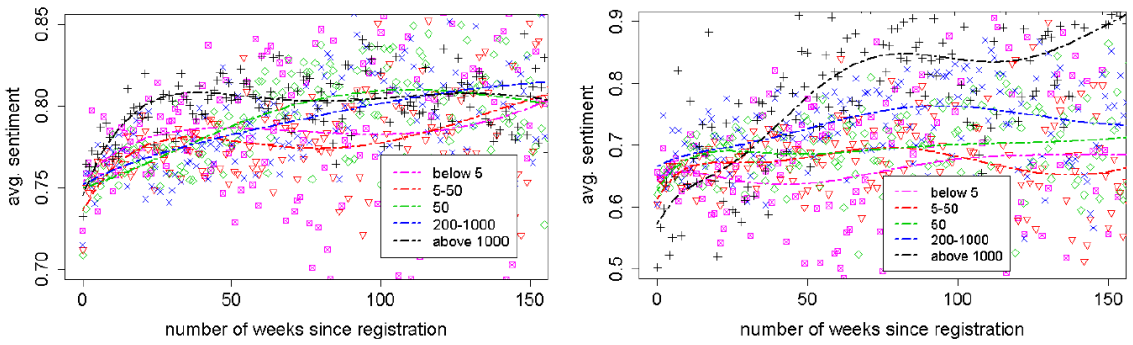
7.4 Impact of member's posting activity on sentiment

The last factor we considered was the amount of posting by each individual. Table 7.4 groups members into 5 groups by number of posts, listing the distributions and average sentiment of each group. There were 8,247, 3527, 757, 255, and 24 profiles in the 5 groups respectively. Although members who published less than 50 times wrote only 20% of all posts, approximately half of the initial posts were authored by these members. This suggests that new members tend to seek information and support while long-time members provided information and support more than they requested it. All differences of sentiment scores between groups, including both all posts and initial posts, were significant, except between group of <5 and 5-50 for initial posts.

post number (# users)		**<5 (8,247)**	**5-50 (3,527)**	**50-200 (757)**	**200-1000 (255)**	**1000+ (24)**
All	sentiment	0.727	0.754	0.779	0.806	0.817
	# posts	16,725	36,422	73,951	102,466	39,944
	avg # post	2.0	10.3	97.7	401.8	1664.3
Initial	sentiment	0.657	0.658	0.683	0.730	0.828
	# posts	4,565	9,445	7,399	6,635	2,990
	avg # post	0.6	2.7	9.8	26.0	124.6

Table 7.4: Average sentiment scores, number of posts published by patients, and number of posts published per user by frequency of posting, for all posts and initial posts respectively.

Figure 7.4 illustrates how sentiment changed over time for different groups of members with different posting activity count. In general, active members (i.e., with more posts authored) were likely to gain sentiment improvement faster and more significantly. It is particularly interesting to note that although members posting more than 1,000 times throughout their time in the community, and who were long-time users, had a significantly higher sentiment score in average, their sentiments were as negative as other members when



(a) all posts (b) initial posts

Figure 7.4: Sentiment changes by length of membership at the time of posting for different groups of posting amount, for (a) all posts and (b) initial posts. A colored point at (x, y) in the graph represents that the average sentiment score of all posts (a) or initial posts (b) published by users grouped by their number of posts in the xth month after their registration is y. Polynomial curves fitting each group were drawn for the sake of visualization.

they just joined the forum, especially for their initial posts. The pattern seen in Table 7.4 and Figure 7.4 seems to suggest that long-time users, who suffered from cancer but benefited from hearing from their peers online at early stages of participation, changed their roles in the forum later and acted as information and support providers more than requesters. Such role change should be another important outcome of online discussion participation.

7.5 Summary of findings

Our study results suggest that members may benefit from sustained participation in a breast cancer community with respect to the sentiment they convey through their posts. To our best knowledge, it is the first time longitudinal trajectories of member sentiment are found in online health communities. At the early stages of participation, sentiment of users usually increased significantly, and the rate of improvement dropped after several weeks, followed by a slower positive sentiment increase which could last for as long as several years. Our study also showed that compared with reply posts, initial posts of threads were more emotionally negative, especially at the beginning of participation. Sentiment increases of initial posts were more dramatic but long term. A qualitative analysis over the forum data showed

that newcomers of the forum were more likely to be newly diagnosed or post-treatment patients. For most of them, going online was the choice when some of their needs, either informational or emotional, could not be met in other settings such as family and hospitals. As a result, we found a large amount of posts with strong negative sentiments, especially initial posts, published by newcomers asking various questions about cancer symptoms, medication use and side effects, and choices of therapeutic method, which were the issues usually brought up by individuals with little cancer or treatment experiences. In contrast, long-time members were more likely to be cancer survivors or patients who were recovering or being treated as a routine part of their lives. It is likely they were more experienced, empowered, and acted more as informational and emotional support providers rather than requesters, and were expressing more encouragement and empathy in the threads in which they participated. The different patterns of reply posts and initial posts also suggested that people immersed themselves quickly into the discussion by learning to encourage others and provide information through replying, but were still concerned about their own issues.

Our study examined three factors impacts on sentiment and sentimental changes: age, cancer stage, and amount of posting. We showed that all three factors had an impact on the members sentiment on average. Statistically significant differences were found for every stratified group. For age, we found that users older than 60 years old showed the most positive sentiment, especially while publishing initial posts. There were no significant differences between longitudinal aspects of different age groups. With respect to cancer stage, although there were significantly fewer stage IV patients than any other stage, they published many more posts and formed the most active cancer stage group in the breast cancer forum. They showed the fastest change towards positive sentiment after registering in the breast cancer forum. They also were the most positive in their replies, while the most negative in their initial posts. The last factor, amount of posting, also made a difference. Members who published less than 50 posts, mostly newcomers and lurkers, were responsible for only 20% of all posts, but around half of the initial posts were authored by these users, which indicated that new users and lurkers tended to seek information and support while long-time members provided information and support more than requested it. Long-time members, who suffered from cancer but benefited from hearing from their peers online at

early stages, later changed their roles in the forum later and acted more as information and support providers.

Chapter 8

Catalogue of treatments used

Members of OHCs exchange social support, both informational and emotional, in online communications. Information related to disease treatment, such as medications, therapeutic protocols, and surgeries, are particularly prevalent in online health discussion, as suggested by previous research (see Chapter 2 for examples) as well as by our discussions of topic analysis described in chapter 6. One research question pertaining to member characteristics that may be of interest to both patients and clinical researchers is what treatments are actually adopted by online health community members in real lives. This information is critical to patients because they care about what drugs or protocols their peers use, and such information exchange is central to their decision making. For health researchers, public online health communities provide massive cohorts of patients who are potential subjects of post-market research. For example, researchers may be interested in what treatments are actually consumed by patients, in contrast to what are suggested by established clinical guidelines. However, it is usually difficult to conduct such research outside of clinical setting, which provides an opportunity to rely on content analysis to solve the problem.

In this chapter, as the third example of how content analysis can be leveraged to characterize members of OHCs, we aim to build catalogues of treatments for users in OHCs. We solve the problem by analyzing content of user posts, extracting evidence of actual adoption of treatment, and creating profiles of treatments for users. Results of this study can be further used to compare the list of treatments as extracted from the community and establishing prevalence of use from clinical guidelines, to study the gap between clinical

expectation and patients' actual practice. In this chapter, we will also take temporal information into consideration, investigating how members' perceptions and usages of treatments change through time.

The task is not as trivial as simply extracting mentions of treatment from user posts. The most challenging part lies in the fact that many of the mentions are not attributed to the patients. For example, users in OHCs may discuss related scientific findings about a treatment, in which a large number of treatment names may occur. Such mentions do not indicate any actual usage of the drugs, and therefore should be excluded in the catalogues.

We rely on communities for autism in this chapter. Autism is a common condition among population, and usually starts developing in early childhood. Unlike communities for most other diseases where participants are primarily patients or survivors, autism communities' members are mostly parents of autistic children. In autism forums, members primarily discuss their children's diagnoses and treatments, but also rely on the community functions to exchange information and support about themselves. As such, in order to build up catalogues of treatment for patients of interest (i.e. autistic children), treatment mentions attributed to the patients and attributed to caregivers of the patients must be distinguished.

In chapter 5, we described a tool that can identify mentions of treatments, and classify these mentions by their attributions. Relying on this tool, treatment indicating actual usage of treatments by community members (or their children in the case of autism) can be marked off from mentions attributed to other ones or general mentions which do not associate with anyone. All results presented in this chapter is based on the application of the tool on the entire ASD data set.

8.1 Creating treatment catalogues for members

In total, 164,335 mentions of 3,981 different treatment terms were identified in the entire ASD data set. In average, around every three posts have one treatment mention. *Patient*, which represents that a mention is attributed to patients of interest, is the most dominant attribution label, with 79,778 mentions of 3,552 treatment terms identified. Since some of the terms may refer to the same treatment (e.g. chelation and chelating), actual number

of treatment identified may be less. 71,1102 mentions of 3,622 treatment terms, 7,783 mentions of 1,142 treatment terms, 5,297 mentions of 915 treatment terms, and 275 mentions of 176 treatment terms are identified for attribution *General*, *Patient-general*, *Other*, and *Caregiver*, respectively. Detailed definitions of these attributions can be found in section 5.5.

The original top ten most frequent treatment terms with corresponding numbers of mentions for each attribution class are given in table 8.1. The lists contain common treatment options for autism patients, as well as alternative therapies. Prevalence of the same treatment in different attribution classes may differ. For instance, although chelation is the most prevalent treatment discussed in the forum, and is particularly popular when attributed to general discussions (See the class *General* in table 8.1), number of mentions of chelation which attributed to the users' actual usage is not that dominant. It is interesting that alternative therapies, such as probiotics and vitamins, are used by patients in the forum almost as frequently as conventional drugs such as Risperdal. Moreover, it is surprising that almost all the top ten terms identified for each attribution class are indeed either treatment options or nutritional supplements, with only one false positive appeared in the list of Caregiver (cab). Given the broad coverage of treatments identified, the high precision of the top term lists indicate a successful application of the computational method to identify treatment terms. However, some of the terms identified in specific attribution classes are questionable. In particular, treatments attributed to caregivers in current result are mostly treatment options for autism, which are likely to be caused by incorrect classifications.

After identifying attributions of treatment mentions, for each user we are able to create a treatment catalogue in which all treatments attributed to their kids are recorded. We obtain this by simply aggregating all treatment mentions whose attribution are *Patient*, in all the posts of individual users in the forum. As such, we are able to create treatment catalogues for 3,635 members. Among them, 2,301 have tried more than one treatment according to the identification. Distributions of number of users, by number of used treatment, is given in figure 8.1. Most of the members have tried multiple treatments, which is consistent with the fact that parents tried various treatments as well as supplements for their autistic children, since autism is complex and hardly be curable with standard conventional protocols.

Term	**Frequency**	**Term**	**Frequency**
Patient		Patient-general	
chelation	4935	chelation	1259
probiotics	2498	probiotics	389
zinc	2011	chelating	210
enzymes	1705	speech therapy	99
melatonin	1425	probiotic	98
special education	1287	activated charcoal	77
antibiotics	1283	nystatin	75
speech therapy	1245	melatonin	73
early intervention	1061	calcium	70
magnesium	889	early intervention	66
Caregiver		Other	
chelation	16	probiotics	424
progesterone	7	chelation	408
probiotics	7	probiotic	163
cod liver oil	5	chelating	150
chelator	4	melatonin	121
cab	4	enzymes	117
molybdenum glycinate chelate	4	zinc	114
sensory integration	4	risperdal	80
aloe vera	3	charcoal	77
pyridoxine hydrochloride	3	homeopathy	76
General			
chelation	8341		
vitamin	1418		
early intervention	1268		
probiotics	1267		
special education	1153		
chelator	910		
vitamins	886		
melatonin	877		
homeopathy	862		
thimerosal	801		

Table 8.1: Top 10 treatment with number of mentions for the five attribution classes, identified in the ASD data set.

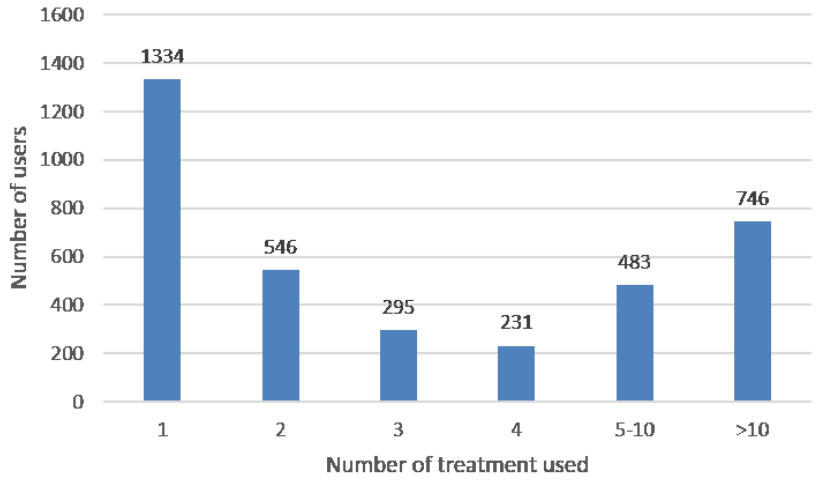


Figure 8.1: Distributions of number of users, by number of used treatment. The x axis is the number of used treatment identified, and the y axis is the number of users.

Table 8.2 shows the top ten treatments most used by members in the autism communities. The difference between this table and the *Patient* columns in table 8.1 is that multiple mentions of a treatment of the same attribution posted by one user will only be counted once in this table. As such, numbers in table 8.2 represents the true prevalence of treatment adoptions among autism forum users, rather than frequencies of keywords. Chelation, as a controversial therapy which lacks sufficient scientific evidence of effectiveness, attracts a lot of discussions in the forums, according to its general frequency. However, it only ranks 3rd as the most used treatment by patients. On the contrary, probiotics as a nutritional supplement, and speech therapy as a well-established psycho-social therapy for autistic children, are more popular in real practice.

8.2 Longitudinal analysis of treatment catalogues of members

Following the rationale of longitudinal analysis for topic and sentiment, in this section we investigate how frequencies of treatment mentions change through time, and how the patterns differ across attribution types. Specifically, treatment mentions of attribution type

Term	Number of users
probiotics	819
speech therapy	565
chelation	520
early intervention	475
special education	395
melatonin	391
antibiotics	381
enzymes	352
zinc	332
vitamins	283

Table 8.2: Top 10 treatment by number of users, identified in the ASD data set.

Patient and ones of other types are considered separately. We illustrate how frequencies of mentions change through time in weeks and in days since members joining the community in Figure 8.2.

In general, no clear pattern can be identified for each treatment. Unlike topic and sentiment, members do not necessarily focus on certain treatment at the beginning stage of participation. In the long run, members keep discussing treatment options throughout their participation, with no decline in frequencies of mentions of any terms significantly.

In terms of frequencies of mentions attributed to patients, it was expected that such mentions should occur more frequently at the initial stage of participation, when members join the community and introduce conditions and current treatment adoptions of their autistic children. However, such pattern is not found in our analysis. On the contrary, frequencies of mentions attributed to patients fluctuate with total frequencies, and maintain substantial percentages throughout members' participation. One possible explanation is that members try different treatment options for their children at different times, and keep updating about their effectiveness in the forums.

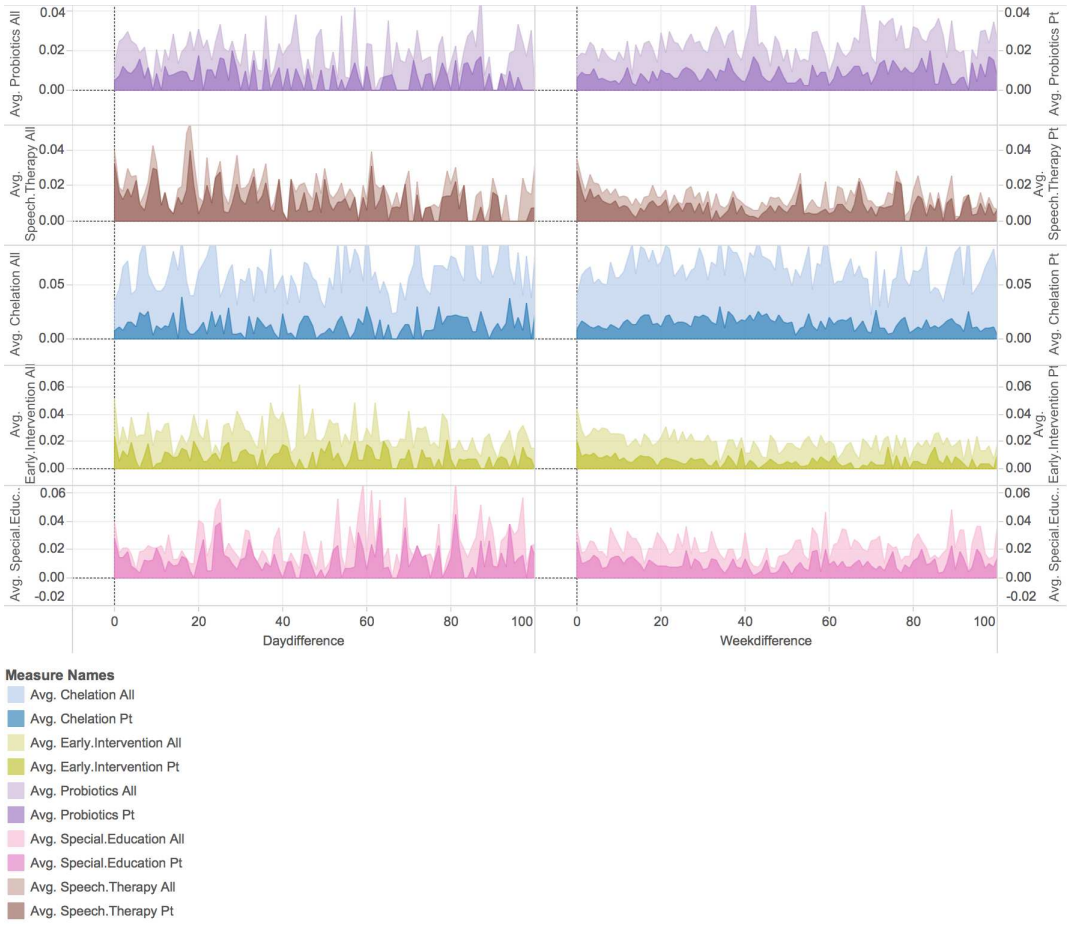


Figure 8.2: Changes of frequencies (mention per post) of top five treatments in autism communities, since members joining the community. Two separate X-axes represent views in weeks (right) and in days (left), respectively. Variables (measure names) ending with "all" represent total frequencies of mentions of corresponding treatment, regardless of their attribution types. Variables ending with "pt" represent frequencies of mentions of attribution type *Patient*.

8.3 Summary of findings and future work

The results suggest that abundant treatment options, ranging from conventional therapies to alternative ones, are discussed in the autism forums. Mentions of treatments are attributed to different stakeholders of autism care such as patients and caregivers. In the autism forums, most of the treatment discussions are attributed to autistic children of community members. Although not all mentions of treatment indicate actual adoption, around 90% of treatment mentions attributed to patients (autistic children) represent an ongoing treatment or a history of usage. Specifically, members keep updating status of their kids as they are treated, in which massive amount of treatment mentions occur. Members of the autism forums also discuss therapies frequently on issues like scientific evidence of effectiveness and information received from health professionals and online sources. A small proportion of treatment mentions are attributed to the caregivers themselves as well as other people in the community or in their real lives.

We notice that some of the treatments, such as chelation, are discussed prevalently in the communities. However, they are not necessarily options that are mostly taken in practice. Within the top treatment list that represents actual usage, non-chemical psychosocial interventions such as speech therapy and special education are popular, although they are not necessarily the most popular ones under discussion. Our results provide a clear evidence that users' perceptions, and hence actual adoptions, of treatment may not be accurately reflected by popularities in discussions, not to mention merely frequencies of certain keywords. More broadly, the results remind us that when connecting online content to members' real life actions in a quantitative way, hidden information (e.g. attributions) of content must be taken into account to avoid mis-interpretation of results.

Longitudinally, we found that members discuss treatment therapies with quite constant frequency in the communities throughout their participation. No clear pattern could be identified in terms of how sustained participation affects frequencies of discussions of certain treatments. Moreover, frequencies of mentions attributed to patients fluctuate with total frequencies, and maintain substantial percentages in all the mentions throughout members' participation, which is somewhat counter-intuitive. In the future work, it is therefore an interesting question to explore how and why members keep mentioning treatment attributed

to their autistic children throughout their participation.

The most important building block of future work following this study is to compare the list of treatments discovered in OHCs automatically by the computational tool with established clinical guidelines. For example, while effectiveness of chelation is still under investigation by researchers [Davis *et al.*, 2013], it already becomes a rather popular choice among autism community members. It is therefore critical to further quantify how broad the gap is between established guideline and patients' actual practice. The future work will contribute to understanding how information support and consumption in OHC affect members' decision makings regarding disease management, and hence how OHC participation makes physical and psychological impact.

Chapter 9

Toward a User Modeling of OHC Members

9.1 Putting things together: how much do we capture about OHC members?

In the previous chapters, relying on computational tools created, we investigate three content-related variables of OHC members: topic of discussion, sentiment expression, and treatment usage, respectively. Our overall goal is to establish multi-dimensional descriptions of members based on content they author in OHCs, and to identify patterns, correlations, and trajectories associated with these characteristics. By applying our computational tools on user-generated content, we are able to identify for each user their interests in different topics, their sentiment expressions, and treatment they discussed and used. By aggregating results across members, we find several interesting patterns, both static and longitudinal, identified through automated content analysis at scale. Trajectories of member participation are also detected, which generate interesting hypotheses for future research to examine.

First, we characterize members' interests in different topics successfully in a breast cancer community. We find that members show particular interest in discussing diagnosis-related topics in the community at the initial stage of participation. As they participate longer, their interest shifted more or less from ones that are closely related to cancer con-

ditions to ones that are more casual and daily-matter concentrated. Members of different disease severities focus on different topics, indicating an correlation between disease profile and topics. The findings may convey critical messages to health researchers in terms of informational support, suggesting that in order to maximize benefits of online social support, communities should deliver different themes of information to different members at different times.

Second, we characterize members' sentiment expression through posting in communities. Similar with what we find for topics, one particular pattern is identified at early stage of participation: members experience a rapid increase toward positive in sentiment expression at the beginning of participation. This does not necessarily mean that every member has benefited from community participation; however, it clearly points out the importance of support intervention for new members of a community. Members of different ages and different cancer stages also show different patterns in sentiment expression, suggesting that personalizations should also be made in social support interventions by considering disease status. One confounder of the study is that sentiment expressed through content may not truly represent emotion or psychological wellness of members. However, the findings are still informative signals that can be validated further by health researchers in future research.

One particular factor of posting, whether a post is initializing a discussion or replying to other posts, is taken into consideration in our analyses of sentiment and topic. The importance of distinguishing the two lies in the fact that most initial posts represent support seeking while most reply posts represent support providing [Zhang *et al.*, 2014]. Our analyses suggest that such difference in posting motivation is associated with differences in content. For example, initial posts of OHCs are more likely to be discussing disease diagnosis, and their sentiment expressions tend to be more negative. Our results further suggest that it may be needed for future research to consider these two types of posts separately in content analysis.

Finally, we create catalogues of treatments for members in the communities, by identifying entities of treatment and classifying attributions of these mentions. In this study, we show one particular challenge of connecting online content with reality: messages conveyed through content may not necessarily indicate the happening of corresponding events

in real lives. In the case of treatment attribution, we found that discussing a treatment in community does not always indicate the action of using the treatment. Luckily, our study demonstrates that it is possible to rely on automated content analysis to overcome the issue by identifying attributions of extracted information and further filtering out information that is not associated with real actions. Similar issues also exist in studying other characteristics (e.g. connecting sentiment expression to actual emotion), and should be the focus of future work.

As we discussed in chapter 2, there are many other characteristics that are critical to understanding OHC members, such as disease profile, social status, and personalities. We are unable to cover all the issues in this thesis. Our purpose is to show how computational tools can be used in the identification of such characteristics, and how patterns and trajectories can be found by using longitudinal analysis. Studies presented in this part of thesis are examples of how we successfully breakdown the machine-unreadable content of OHC written in natural language, and transform the narratives to discrete dimensions of characteristics toward modeling individual members of OHCs.

9.2 Visualizing member characteristics: how do they correlate?

One research question worth exploring is how the variables identified, topic, sentiment, and treatment, correlate with each other in a longitudinal standpoint. In this section, we envision a multi-dimensional, longitudinal visualization of members characteristics that can help make sense of such correlations. Figure 9.1 shows a prototype of such visualization. Sentiment is used as the base variable in this example. We use frequencies of four topics in the breast cancer forum, diagnosis, treatment, personal, and nutrition, and frequency of a popular breast cancer treatment, Tamoxifen, as variables in consideration. For each variable, we plot changes of its frequency through time, and how the Pearson correlation between the variable and the base variable (sentiment in this example) changes through time. Ideally, the visualization tool should allow changing the base variable easily, so that any comparisons can be made.

In this particular example, we found that topics such as diagnosis are always negatively correlated with positive sentiment expression, and the negative correlations are quite significant at the beginning of participation. However, nutrition as a less clinically relevant topic is correlated with positive sentiment, and its frequency grows as members participate longer. Tamoxifen, as a popular treatment option for estrogen receptor positive breast cancer, is mentioned frequently throughout members' participations, and its frequency is negatively correlated with sentiment.

It is noteworthy that such visualization is not only able to capture aggregated patterns, but also can be used to make sense of characteristics of individual members, by simply replacing aggregated frequencies with frequencies of variables in one member's content. In our future work, we will make an interactive implementation of this visualization and make it available to the research community.

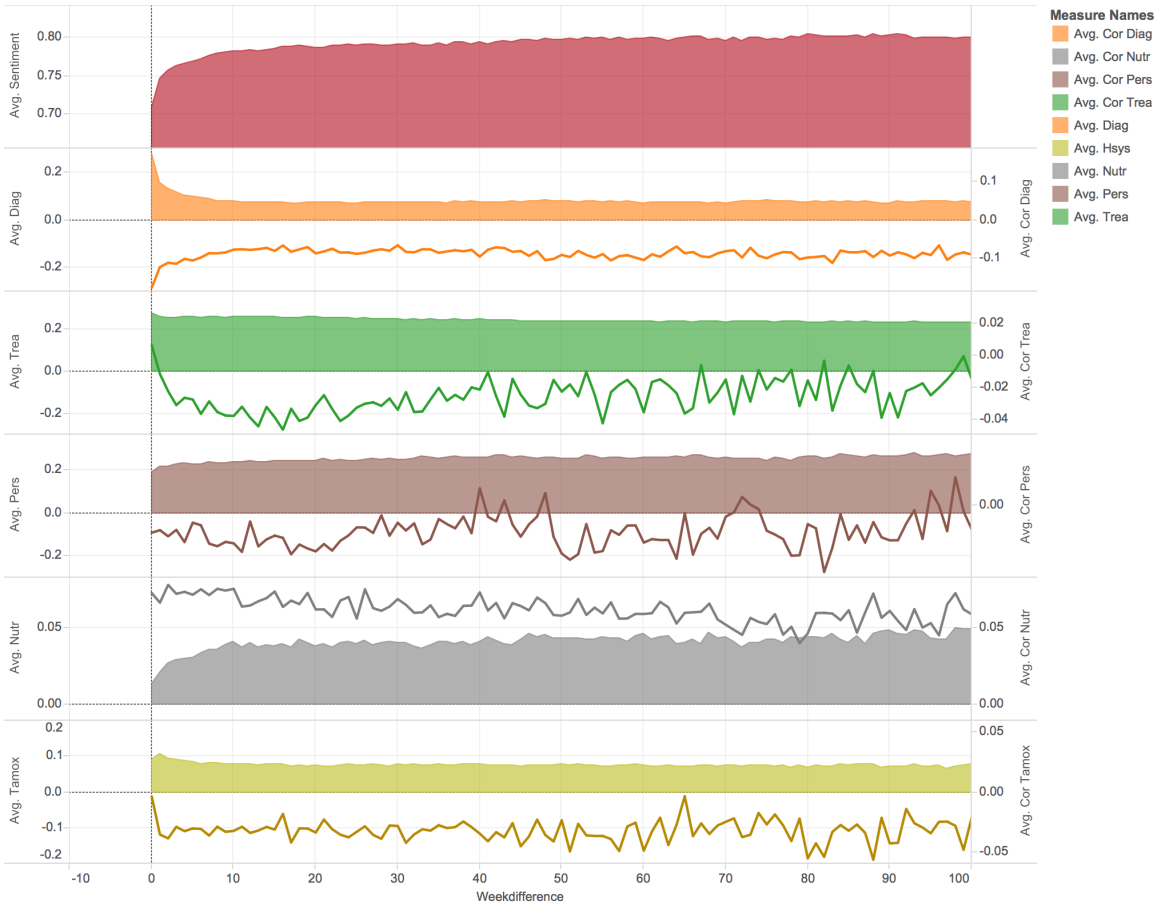


Figure 9.1: A joint longitudinal view of different member characteristics. Sentiment is used as the base variable in this example. Other variables are compared with the base variable by calculating Pearson correlations. Colored areas represent frequencies (scores in the case of sentiment) of different variables, and lines represent changes of correlations between these variables and the base variable.

Part IV

Characterizing Member Engagement in Online Health Communities

Introduction

In the previous chapter, we investigated how content analysis could be used to model characteristics of individual OHC members at scale, leveraging computational methods and tools we created. According to our framework described in Chapter 2, another important variable describing OHCs is users' engagement, which includes posting activities (e.g. initializing discussion vs. replying to others posts), lurking, debates on certain issues, decision of staying or withdrawing, etc. In the discussions of topic and sentiment, we have taken initial v.s. reply into account from the perspective of post content, and have identified interesting distinctions between these two types of posts. Such distinctions of engagement between support reception through initializing discussions or asking questions and support providing through replying are also studied previously, suggesting that both types of activities are crucial to attaining optimal benefits for patients [Zhang *et al.*, 2014; Kim *et al.*, 2012a; Han *et al.*, 2011; Namkoong *et al.*, 2013].

Other variables of user engagement are also critically related to member characteristics and interactions among members. For example, members' decisions of dropping-out or their stances toward certain issues may depend on personal beliefs and their disease status. As such, our efforts of identifying member characteristics from content in the previous chapters have the potential to substantially help the analysis of user engagement. In this chapter, based on the member characteristics identified in the previous chapters by using computational methods, we investigate two important user behaviors pertaining to engagement: debate and dropping-out. With respect to our framework, in this part of thesis we focus on the building block of Engagement in the community characterization meta-layer (figure 9.2).

Identifying debates, and hence participants' stances in debate, is an important task in opinion analysis. Debates are particularly popular in online communities for political purposes [Somasundaran and Wiebe, 2009], but can also be intense in certain online health communities [Zhang *et al.*, 2016b]. Researchers also found that compared with offline communities, it is more difficult for online community members to interpret others tone and emotion in the absence of physical and non-verbal cues, which might lead to conflicts to quickly escalate [Friedman and Currall, 2003]. Debates in OHCs usually surround con-

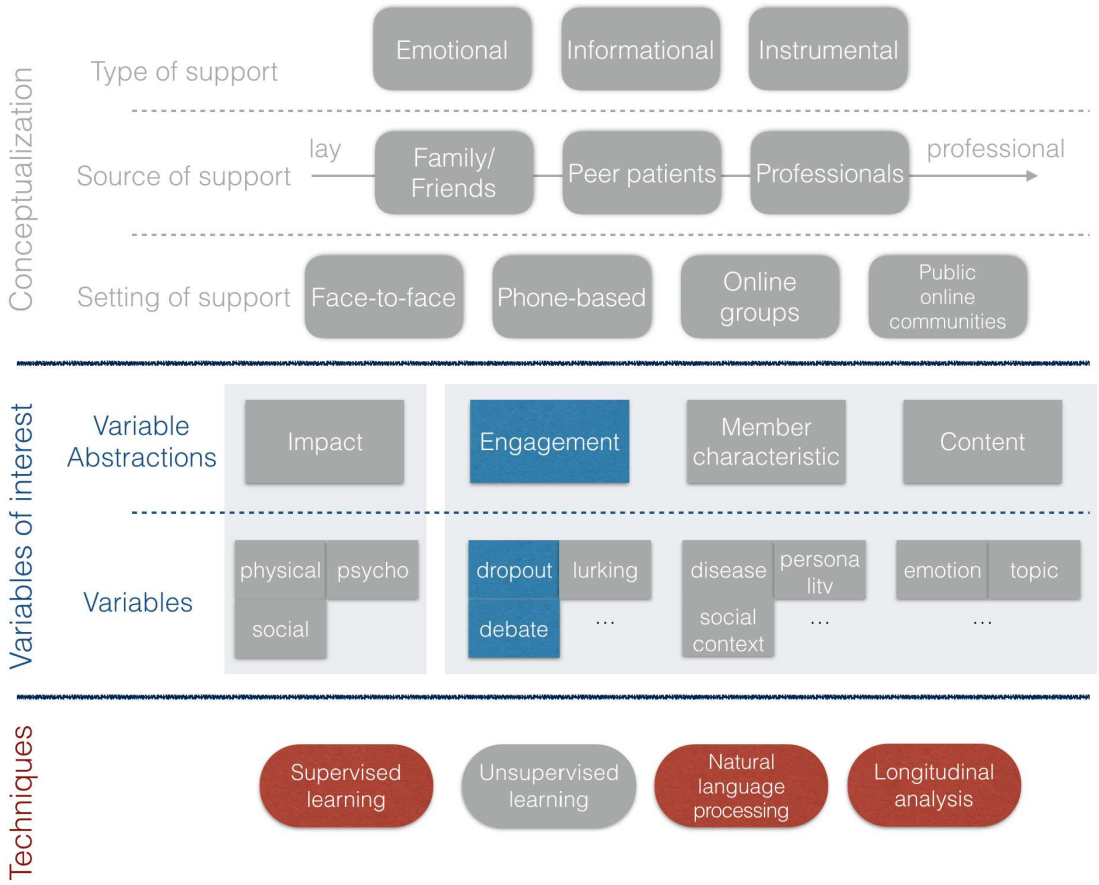


Figure 9.2: Variables of interest discussed in this thesis. Colored elements are the foci of this part of thesis.

troversial health issues, such as effectiveness of certain therapies, (dis)satisfaction towards health providers, and causes of diseases [Zhang *et al.*, 2016b]. Detecting debates and stances in OHCs in an automated fashion are helpful, because 1) it helps community creators and managers to prevent as well as to respond quickly to outbreaking conflicts, and to cultivate peaceful discussion environment; 2) it helps health researchers and epidemiologists to quickly locate controversial health issues online, and to make sense of online opinions on these issues. In the next chapter, we present an exploratory study of identifying and characterizing debates and stances for issues of alternative and complimentary medicine (CAM) from an online health community. We present how computational methods and tools we created can help automate this challenging task.

The second study we present, pertaining to members' engagement, is to identify and characterize dropout [Zhang and Elhadad, 2016a]. Dropping-out, which refers to when an individual abandons an intervention, is common in Internet-based studies as well as in online health communities. Community facilitators and health researchers are interested in this phenomenon because it usually indicates dissatisfaction towards the community and/or its failure to deliver expected benefits. Dropout is also a critical issue which may undermine a community's activeness. Traditionally, dropping-out of members can only be investigated in tightly controlled research settings, in which questionnaires and surveys are the major instruments to identify causes of dropout. Recent years have witnessed research progress in utilizing quantitative approaches to study dropout. For example, Wang and colleagues did a survival analysis on a breast cancer forum, showing that users who received emotional support are more likely to keep participating while users who received informational support are more likely to drop out [Wang *et al.*,]. In this thesis, equipped with various member characteristics identified for users, we are able to investigate correlations between a wider range of factors and the phenomenon of dropping-out quantitatively and longitudinally at scale. For example, we are able to identify if sentiment changes of users are correlated with the decision of dropping-out. In chapter 11, three variables were investigated: sentiment, topic, and user interactions, with respect to how they correlate with dropping-out.

Chapter 10

Identifying and characterizing debates in OHCs

10.1 Introduction: detecting CAM-related debates from an OHC

In this chapter, we introduce how debates about one certain issue, complementary and alternative medicine (CAM), can be detected from and online health community, and characteristics of these debates.

Complementary and alternative medicine (CAM) is widely used by populations worldwide in concert with conventional evidence based medicine, particularly for treating and managing chronic diseases and life-threatening illnesses [Barnes *et al.*, 2009; Hyodo, 2004; Xue *et al.*, 2007; Molassiotis, 2005]. Yet, impact of CAM usage has been controversial, and the motivations of CAM usages have been diverse. For example, it is reported that the majority of alternative medicine users appear to be doing so, not so much as a result of being dissatisfied with conventional medicine, but largely because they find these healthcare alternatives to be more congruent with their own values, beliefs, and philosophical orientations toward health and life [Astin, 1998]. As such, patients may take CAM following personal beliefs and sometimes without informing their care providers [Furlow *et al.*, 2008a], which may bring uncertainties in disease managements. For healthcare practitioners and

researchers, it is therefore critical to gain a deeper insight into how CAM therapies are perceived and used by patients.

Previous studies revealed that many patients are critical of and skeptical about the efficacy of modern medicine and believe that treatment should concentrate on the whole person and greater knowledge of the physiology of the body [Furnham and Forey, 1994]. Recent research has also focused on attitudes of physicians and patients toward CAM relying on different study instruments, many of which found incongruent views on effectiveness [Furlow *et al.*, 2008b; Lapi *et al.*, 2010; Adams *et al.*, 2011]. Most of these studies are based on rigorous study designs on sampled populations, in which subjects are asked to respond to survey instruments or participate in focus groups.

Because CAM usage is linked to personal beliefs and because most of CAMs are not adopted by the medical establishment, one research question for this work is to which extent peer-to-peer CAM-related discussions contain conflicting opinions on CAM adoption and/or efficacy. A secondary set of questions pertain to identifying which specific CAM therapies are more likely to trigger debate amongst patients, and what are the stances of patients overall toward these controversial CAMs.

Our overall objectives are therefore (i) to detect instances of debates about CAM in a community; (ii) to classify patients' stances toward these therapies; and (iii) to identify which specific CAM therapies are more likely to trigger debates in the community. Our study is carried out in an automated and quantitative fashion relying on computational methods we created, and aims to complement perspectives obtained through qualitative methods.

Critical to this objective is a tool that can precisely locate CAM-related debates in the different posts of a community, and identify the stances of the different debate participants towards the CAM under discussion. We rely on the machine learning debate detector and stance classifier we developed in Chapter 5, which enable us to identify all CAM-related debate posts, along with the stances of the participants throughout the entire forum. In this chapter, we focus on debate posts identified by our classifier through a qualitative analysis to study prevalence of these debates and to characterize which alternative treatments trigger debates more often than others.

10.2 Manual analysis of debate posts

The 4-class classifier based on support vector machine using all features combined, introduced in section 5.4, was applied to all the 25,013 posts in 396 threads in the alternative medicine sub-forum of the BC data set. Among them, 5,714 posts in 187 threads were identified as in debate, in which 3,166 posts in 116 threads were CAM-related, 1,144 posts in 78 threads as breast cancer related, and 1,404 posts in 81 threads as other types of debates such as conduct of rules. The stance classifier was then applied to the 3,166 posts identified in previous step as CAM-related debates. 950 of them were identified as opposing CAM usage, which means that around 2/3 posts in CAM related debates are in supportive stances.

We carried out a manual analysis to identify which specific CAM therapies are under dispute frequently in the community. We randomly sampled 500 posts from the 3,166 CAM-related debate posts (from 116 threads), as identified by our classifier. To ensure that each thread is represented in the sampled set and to get around over-sampling posts from massively long threads, we made sure that at least one post from each thread was sampled, in accordance with the length of the different threads. This resulted in a total of 523 sampled posts.

Two annotators coded the sampled posts as (i) not debate (i.e., the classifier miscategorized the sampled post as debate); (ii) not CAM-related (e.g.., posts with a debate, but about rules of conduct in the community, or any topic not directly related to CAM); (iii) general CAM debate (e.g., debate post about chosing CAMs as an alternative to chemotherapy); or (iv) specific CAM therapies or groups of therapies (e.g., nutritional supplements). Because some specific therapies had a very high number of threads discussing them, they were assigned their own code (e.g., Gerson diet was kept a separate code from the more general diet code). Appendix B provides detailed list of concepts we use in coding and how they correspond to specific topics or therapies. The stance classification was also applied to the sampled posts. At the end of this process, we thus can assess in our sample of posts (i) what CAM therapies are prevalently under debate; and (ii) the participants' stances towards these treatments.

10.3 Prevalence of therapies in debate posts

Out of the 523, 118 of the posts were coded as non-debate ones, and 78 of them were coded as debate but not CAM-related (46 about cancer cause, 16 about cancer diagnosis, and 16 trolling or rules of conduct in the community). The breakdown of the remaining 327 posts coding is provided in Table 10.1. In addition to the different therapies and their prevalence in the sample posts, Figure 10.1 illustrates the prevalence of pro and con posts for each group.

Code	**Example**	**# posts**
CAM	General CAM v.s. conventional discussions; Effectiveness and use of CAM v.s. chemotherapy	135
Gerson therapy	Effectiveness and scientific validity of Gerson therapy	44
Diet	Effectiveness and/or practice of diets for cure, prevention, and management of breast cancer therapy (gluten free, low carb, hormone free meal, vegan, Ayurvedic, etc.)	42
Supplements	Any supplement whose purpose is not to control estrogen	33
Laetrile	Laetrile or food/supplement that contains laetrile	27
Estrogen control	Therapies/supplements to control estrogen, including DIM, soy, natural replacements for tamoxifen, bioidentical hormones, etc.	24
TCM	Use and effectiveness of Traditional Chinese Medicine for cancer management	12
Med marijuana	Use of medical marijuana for cancer management	5
Issels	Issels treatment	2
Colonics	Colonics treatments	1

Table 10.1: CAM Therapies identified through for the manual coding, and number of posts identified for each therapy group in the sampled posts.

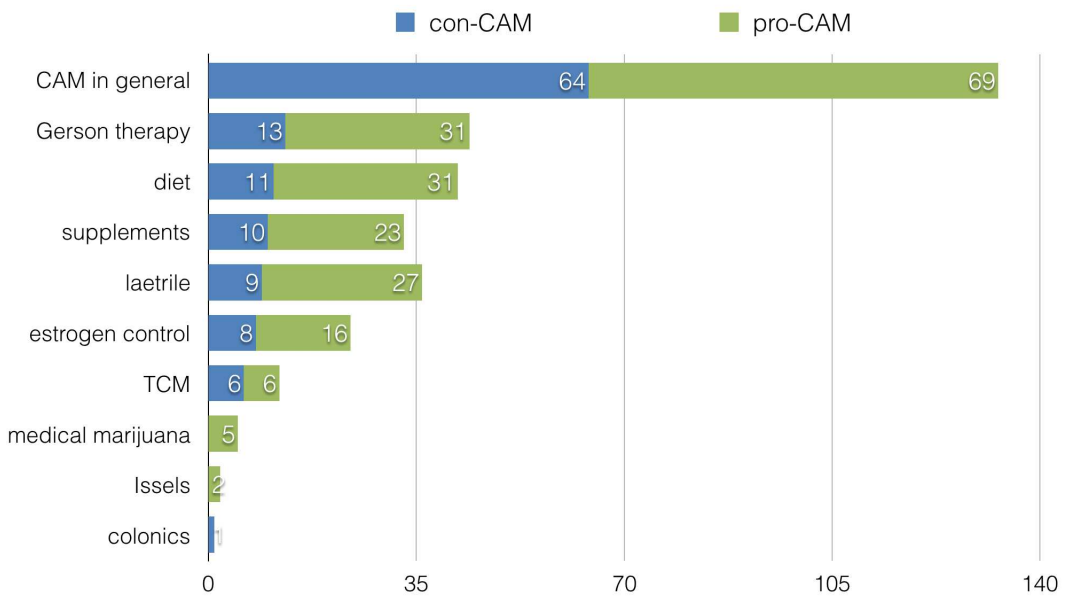


Figure 10.1: Stances of posts on CAM usage clustered by topics. X axis represents the numbers of posts in pro-CAM and con-CAM stances, respectively.

A large proportion of debates are amongst proponents of CAM therapy and their opponents, on issues such as effectiveness of CAM as a general alternative to conventional treatments like chemotherapy, as well as in addition to conventional treatments. Although all posts in this analysis were from the alternative medicine sub-forum, which is presented to the breast cancer community as a safe place to discuss alternative medicines, there were still a significant number of con-CAM posts present in the sample. Many of the specific alternative treatments, such as Gerson therapy and laetrile, also attract a large amount of debates in the forum, mostly about the scientific validity of the therapies.

10.4 Comparing with non-CAM posts

One interesting question worth exploring is whether CAM as a controversial topic is more likely to trigger debates than other cancer related issues. To investigate, we applied our debate classifier to the entire breast cancer forum which consists of more than 3 million posts. Results indicates that more than 500 thousand (563,231/3,283,016, 17%) posts were iden-

tified as debate. Compared with the ratio in CAM sub-forum (5,414/25,013, 22.8%), lower proportion of debate posts were found in other sub-forums. However, since our classifier is trained completely on data from the alternative medicine sub-forum, it may underestimate the ratio of debate in the other forums.

10.5 How are these debates triggered?

The most prevalent type of debates is about effectiveness, scientific validity, and usage of alternative therapies in general. Many of such posts are published in threads initiated by newly diagnosed patients or patients suffering from side effects of conventional treatments, who are looking for evidence that supports CAM usage. Debates escalate particularly quickly in discussions when someone considers completely replacing conventional medicine with CAM. Members may be in an opposing stance on such opinions, although many of them in this sub-forum are supposed to be users and hence supporters of CAM. This is consistent with a previous research finding that members of online health communities are able to self-correct misleading opinions [Esquivel *et al.*, 2006]. Similarly, debates can be triggered often when CAMs are perceived by some users as a standalone treatment of cancer, instead of common perception of CAM as complimentary ways of relieving side effects brought by conventional therapies, such as pain, fatigue, and hot flashes, and to help improve quality of life. Although previous research suggests that CAM use can no longer be regarded as an "alternative" or unusual approach to managing breast cancer given its increasing popularity [Boon *et al.*, 2007], our study suggests that many patients, even practicing alternative therapies themselves, are still rather rational and cautious with CAM usage. A small group of firm anti-CAM users, which are sometimes treated by other users as trolls, were also identified. Sometimes CAM supporters respond to these persons in a quite drastic way, such as in following post: "I will never understand why women who do not have breast cancer feel the need to post on a breast cancer board. Why? Consider yourselves lucky....you dont have cancer! Go live your life!."

Chapter 11

Identifying and characterizing dropouts in OHCs

11.1 Introduction

Critical to studying OHCs' impact on their members is characterizing and understanding the patterns of participation in a community. Researchers have studied whether users actively participate or lurk [Setoyama *et al.*, 2011], as well as when they decide to withdraw from the community permanently [Eysenbach, 2005]. Lurking—the phenomenon of users browsing the content but not actively participating in discussions— has been shown to correlate with lower perceived social support and diminished emotional benefits when compared to active participation in a community [Setoyama *et al.*, 2011; van Uden-Kraan, 2008; Mo and Coulson, 2010; Han *et al.*, 2014]. Dropping-out— i.e., stopping participation or leaving the community altogether— when studied across members indicates the level of activity in an OHC. For instance, Eysenbach and colleagues reported that the phenomenon of attrition (or dropout) is particularly common in online-based interventions [Eysenbach, 2005], with more than 90% of study subjects quitting throughout Internet-based studies. In the case of OHCs, understanding factors associated with dropping-out might help identify opportunities for more targeted support of members, and more generally identify for which members participation in an OHC is beneficial and for which it is not. Wang and colleagues examined how type of information received affect users' choices between staying and leaving,

and suggested that informational support is positively correlated with dropping-out while emotional support is positively correlated with staying active in the community [Wang *et al.*,]. Zhang suggested that information and small group interactions, like emotions, also play a key role in retaining users [Zhang, 2015]. Sadeque and colleagues proposed a supervised model to predict dropping-out, and found that factors like time since last activity were predictive [Sadeque *et al.*, 2015]. To date, however, it is still unclear which other factors of individual members are moderating dropping-out from online health communities, such as topic of discussions, users' sentiment expressions, and interactions among users.

In previous chapters of this thesis, longitudinal analysis was leveraged in online health community research to investigate how participation affects sentiment of users and topic of discussions (see Chapter 6 and 7). In this chapter, we carry out a series of static and longitudinal analyses, which take topic of discussions, sentiment expression, and user interactions as variables of interest. We explore if and how these factors correlate with users' decisions of dropping-out. Because there is no explicit marker for any participant to convey that a member has dropped out of the community, we explore different approaches to determining that a member dropped out. To explore factors in context of dropping-out, we leverage established machine-learning-based methods for sentiment analysis and topic classification of a given member's posts. We hypothesize that dropout members discuss more disease-specific topics, express more negative sentiments, and interact with other members less actively than the members who stay active in the community. Furthermore, we hypothesize that characteristics of dropping-out can be detected by investigating patterns of changes of these factors.

We rely on the BC dataset in the analysis of this chapter. The basic workflow of our analysis is as follows. First, we identify members that have dropped out from the community, i.e. users who had history of active participation in community, but have been inactive for a certain amount of time. We collect a set of member characteristics for each members throughout their history of participation in the community. We compare distributions of each variables between dropout members and other members. The longitudinal analyses focus on dropout members to investigate if any patterns of changes of variables exist *before* they drop out the community, with respect to their sentiment expressions, topics of

discussions, and interactions with other members.

11.2 Identifying dropout members

Identifying which members in a public community dropped out is not a trivial task. In practice, it is impossible to determine with absolute certainty a dropout member from a public community solely based on changes of posting activity, since an inactive user can always return to the community and resume participation. Moreover, in many communities, like our community of interest in this study, there is no publicly available information about members and their loggin patterns; and as such the only available information relates to their posting activity. Thus, a member could withdraw from posting content, but still act as a lurker.

To identify the cohort of dropout members for our study, we explored different heuristics. We defined a user in the breast cancer forum as a dropout member, if she has posted more than n times in the community (i.e. had some history of posting activity), but has been inactive for at least t years at the time of data collection. The first cut-off is to ensure that users we identify are users who participated in the community discussion meaningfully, instead of one-time information seekers or users who just chimed in a limited number of discussions without real information or support exchanges with other members. The second threshold is to exclude members that may return to the community in the near future, as we assume that users who have been inactive for longer time are less likely to return.

In this particular study, n and t were experimentally set as 10 posts and 3 years. As such, for the remaining part of this chapter, dropout members refer to those who have posted more than 10 times in the community, and whose most recent post was before January 2012 (three years before January 2015).

6,338 dropout members were identified using our definition, corresponding roughly to 11% of all users that have posting history in the breast cancer forum. When accounting for all users who have posted more than 10 times (i.e., "meaningfully active") in the community, the dropout members amounted to 42% of these 15,199 users. The identified dropout members posted 570,932 posts in total in the breast cancer forum, with each one posting

t cut-off	1	2	3	4	5	6	7	8	9	10
# of dropout members	13,997	9677	6,338	3,864	2,311	925	210	76	32	11

Table 11.1: Number of dropout members identified as the cut-off t changes.

90.1 posts in average. The average posting number is roughly the same as the average across all users posted more than 10 times (91.8). 195 out of these 6,338 dropout members have been highly active in the forum, with each of them posted more than 500 times. These "super-users", although relatively small in number, contributed to roughly 45% of posts identified.

In our method, the most tricky part is to choose the t cut-off, which represents the minimal length of inactiveness for a member to be considered as dropped out. A larger t would definitely bring a set of dropout members with higher precision, but may excludes eligible dropouts incorrectly. Given the fact that most members joined the community in recent years and the forum was getting increasingly popular, a large t would lead to a small sample size. As such, the problem becomes a precision-recall trade-off, and our task is to finds the best value that balance the two properly. The oldest posts of our data set date back to Sep 2004, which is roughly 10 years before the data collection. To see how the cut-off impacts the sample size, we show in table 11.1 number of dropout members identified by setting t from 1 to 10. It can be seen that sizes of samples shrink rapidly when larger t is used.

The major false-positives of our method are the users that return to the community after long time inactiveness. To quantify the prevalence of these comebacks, we designed a sanity check experiment in which we calculate the percentage of users who have been inactive for more than 3 years in the community, anytime in the history, but return to the community after the long break, over the total number of users who have been active for more than 3 years. The number we get is 1.2%, which suggests a relatively good precision of our identification method.

11.3 Longitudinal analysis for dropout members

Three specific variables were studied to examine if they are correlated with dropout: topic of discussion, interaction with other members, and sentiment expression. These three variables are important building blocks of OHC content and member characteristics, and have been investigated in a wide range of previous studies [Civan and Pratt, 2007; Zhang *et al.*, 2014; Wang *et al.*, 2015]. Sentiment and Topic, in particular, has been discussed in previous chapters of this thesis, and our following analysis will be based on results from those chapters.

Our research hypotheses are as follows:

1. Dropout members are more likely to discuss certain topics such as cancer treatments and their side effects, and show certain patterns in topic transitions, before they drop out. These topics and topic transitions may indicate end of cancer treatment journeys, which are usually followed by withdraw of participation.

2. Dropout members receive inadequate social support from other members. They ask questions and seek support more often than other members, but receive less responses. These may indicate lower levels of social support reception leading to lower senses of benefits and belonging, which are vital to self-perceived effectiveness of community usage [Høybye *et al.*, 2005].

3. Dropout members express more negative sentiment in general, or in their final stage of participation, which indicates a declining level of satisfaction towards community participation.

Dropout and topics

To study how topics of discussions correlate with dropping-out, topics of posts must be identified. In this study, topics of posts were identified using the supervised machine-learning tool based on convolutional neural networks (CNN) using the eleven topic schema we introduced before.

To characterize the impact of discussed topics, for each user (either dropout or non-dropout members), we aggregate numbers of topics of all posts authored by the user, and average the topic numbers by the total number of the user's posts. As such, a eleven-

dimensional distribution of topics can be established for a member in the forum, representing frequencies of topics discussed by the user.

Armed with distributions of topics for all users in the community users, we first did a multivariate t-test to examine the difference of topic distributions between posts of dropout members and posts of other members. For each topic, we then carried out a univariate t-test, adjusted by Bonferroni correction due to multiple comparisons, between the dropout members and other members in the community to test if a significant difference exists. These two static analyses identify the distributional differences between topics of discussions between dropout members and other members.

Finally, we examined how the averaged frequencies of topics change through time for dropout members before they actually quit the community from a longitudinal standpoint, to investigate whether certain patterns of changes could be detected.

The multivariate t-test between the topic distributions of posts contributed by dropout members and other members respectively yielded a result which supports a difference with p-value less than 10^{-16}. Average prevalence of each topic for the two types of members is given in table 11.2 with corresponding p-values based on the univariate t-tests. We did not include MISC in the table because it is a default topic category only given to those posts which are not assigned any topics otherwise. We used 0.001 as the threshold of p-value for significance. Five topics amongst all ten show significant differences in average numbers between dropout members and other members. Specifically, dropout members posted more relevant to diagnosis and treatment, but less about nutrition and daily matters. The hypothesis that dropout members discuss more about treatment and diagnosis than other members is thus supported.

	ALTR	**DAIL**	**DIAG**	**FIND**	**HSYS**	**NUTR**	**PERS**	**RSRC**	**TEST**	**TREA**
Dropout	0.002	0.059	0.099	0.063	0.081	0.034	0.274	0.013	0.009	0.053
Others	0.002	0.074	0.093	0.063	0.078	0.039	0.279	0.017	0.010	0.046
p-value	0.226	<0.001*	<0.001*	0.953	0.030	0*	0.162	0*	0.247	0*

Table 11.2: Average prevalence of topics (per post) in posts of dropout members and other members. P-values are calculated by a t tests adjusted by Bonferroni correction. We use 0.001 as the threshold of p-value for significance.

Figure 11.1 shows how topic frequencies change through time as dropout members approach the time point of withdrawing. The way we illustrate the changes is as follows. For each topic category, we plotted change of its average frequencies in all posts that were published a certain length of time before their authors' respective dropout time. We used week, days, and post orders as three different measures to show both long term and short term effects. For example, a point (1, 0.3) in Figure 2(a) or 2(d) represents that the average frequency of the corresponding topic of all posts that are published in the final week of their authors' participation is 0.3. Except for an trend for a higher frequencies of DIAG and HSYS posts in the final weeks, no significant changes of topic frequencies were identified before members' dropping-out.

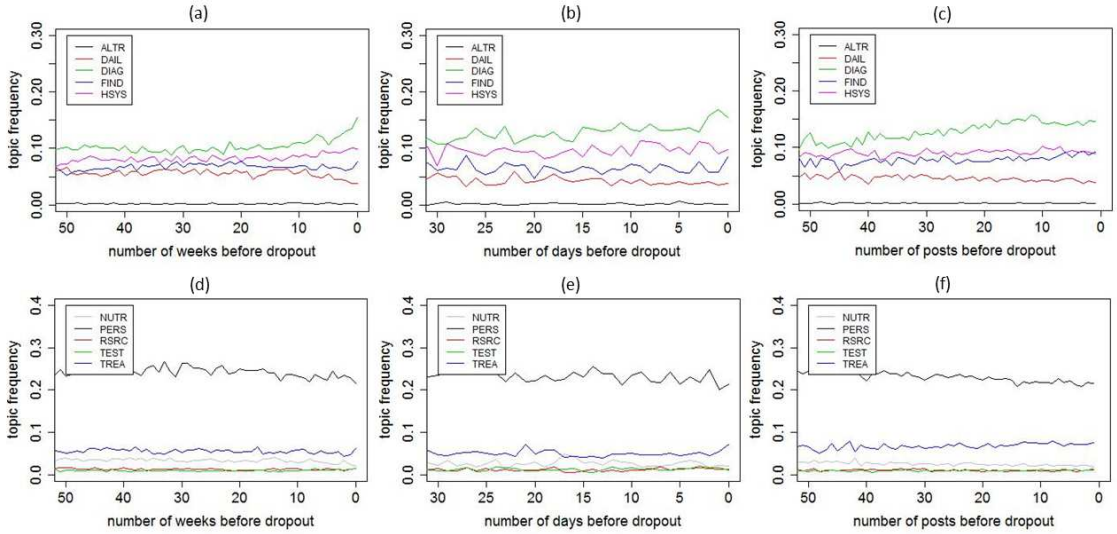


Figure 11.1: How topic frequencies change through time before members' dropping-out. X axes, which are in reserve order, represent the time point before members' dropping-out. Y axis is the average topic frequency of all posts that are published in the corresponding time. Units of x axes in (a)(d), (b)(e), and (c)(f) are weeks, days, and post orders, respectively.

Dropout and interaction

Member interaction is the primary medium of exchanging social support, which can be complex in online health communities [Davison *et al.*, 2000; Biyani *et al.*, 2014]. In this study, we considered two basic aspects of user interactions: number of initial posts versus

number of reply posts, and average number of responses received from other members in the community. As previously discussed, initial posts are those posts initializing threads of discussions, which are usually question asking or help seeking which represent needs of support requesting. Previous research has reported that initial posts are vital part of interactions amongst members, and are usually more negative emotionally [Zhang *et al.*, 2014]. Reply posts, usually representing support providing, are those posts responding to the initial posts, which can exert positive influence on the discussion originator (i.e., author of the initial post) [Zhao *et al.*, 2012]. As such, the ratio of number of initial posts to the number of reply posts can be seen as how often the user seek support from others rather than actively provide support to others. Average number of responses received when initializing discussions, on the other hand, represent how much social support in average members receive from other ones. Previous studies have suggested that support providing and receiving may have different effects on perceived benefits [Namkoong *et al.*, 2010; Han *et al.*, 2011].

For each member, we counted the number of their initial posts, the number of their reply posts to other member's threads, and the number of responses received from other people when initiating a thread. We then calculated the two measures described above, and examined how these numbers differ between dropout members and other members. We relied on a Chi-square test (for initial vs. reply) and t test (for number of replies). Like for the topics, we also examined how these numbers change longitudinally before members' dropping-out.

121,193(3.9%) of all posts in the forum are initial posts of threads. Among them, 31,277 were posted by dropout members, which are 5.5% of all dropout member publications. However, the Chi-squared test indicates no significant difference between dropout members and other members in terms of ratio of initial to reply posts, with a p-value over 0.9. Across the entire forum, an initial post can receive 24.4 replies in average. Dropout members, in particular, can receive an average number of 23.7 replies throughout their community engagement when initializing discussions. A t-test between the numbers of dropout members and other members indicate no significance with p value 0.69. As such, the hypotheses that dropout members receive less reply from other people and that post initial posts more often

in the community are both rejected.

In contrast, the ratio of initial posts increases towards dropout time (Figure 11.3). It is particularly significant from a longer term standpoint, where the ratio of initial posts dramatically increase from around 5% to over 10% in the last 10 weeks of participation before dropping-out. We carried out a supplementary t-test, in which we compare all posts in final 10 weeks and posts before 10 weeks in terms of the initial/reply ratio, and indeed found a significant difference between the two with p value less than 0.001. Short term changes can also also be observed, particularly in the final 5 days. Meanwhile, in term of number of replies received, a landslide can be observed in the week view, which roughly accompanies temporally the ratio increase of initial posts.

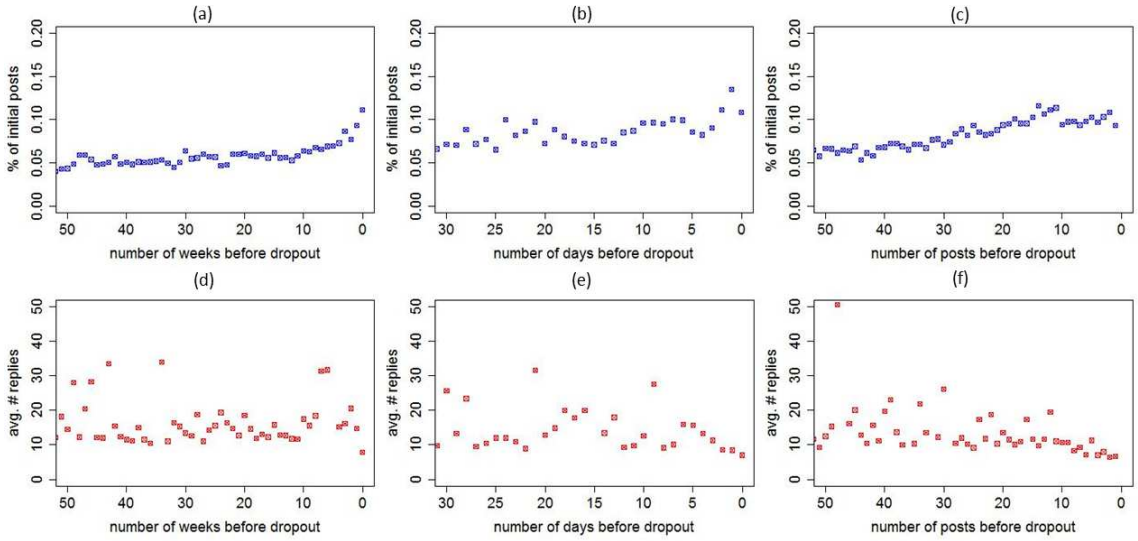


Figure 11.2: How percentage of initial posts and number of replies change through time before members' dropping-out. X axes, which are in reserve order, represent the time point before members' dropping-out. Units of x axes in (a)(d), (b)(e), and (c)(f) are weeks, days, and post orders, respectively.

Dropout and sentiment

Sentiment expression reveals how positive the author's emotion is when posting. We rely on the sentiment analysis results introduced in chapter 7. Based on the sentiment scores of posts, we first identified if a significant difference exists between the averaged sentiment scores of posts published by dropout members and posts published by other members, by

doing a t-test. Second, we illustrated how sentiment of posts changed through time as dropout members approached the time point when they withdrawn from the community, to see if a decline of sentiment actually happened as suggested by our hypothesis.

The average sentiment score (probability of being positive) for all posts in the community is 0.786, while the average sentiment score of dropout member authoring is 0.788, with no significant difference according to a statistical t-test. Longitudinally, an insignificant decline of sentiment can be observed from the week view, but no other patterns can be found. Although we found a tendency of posting more initial posts in the final stage of participation in the previous analysis, no patterns of sentiment change is visible when initial posts and reply posts are considered separately. In contrast to our expectation, dropout members not necessarily express more negative emotion in discussion, and no significant changes of sentiment can be detected before they drop out.

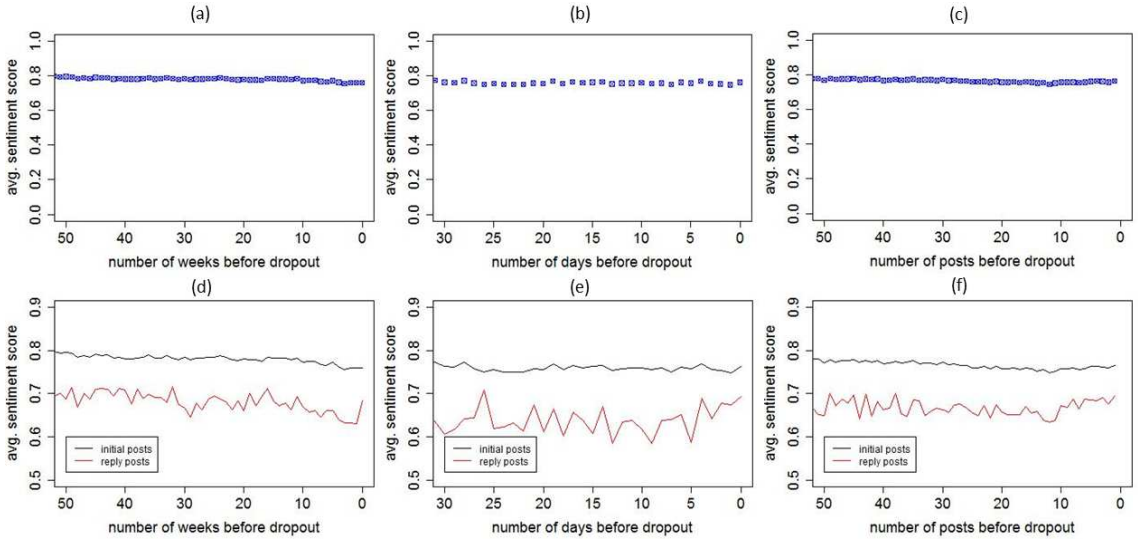


Figure 11.3: How average sentiment score changes through time before members' dropping-out. X axes, which are in reserve order, represent the time point before members' dropping-out. The first three figures show the average score of posts including both initial and reply, and the last three figures distinguish the two. Units of x axes in (a)(d), (b)(e), and (c)(f) are weeks, days, and post orders, respectively.

11.4 Summary of findings

Our first hypothesis that dropout members are more likely to discuss certain topics is supported by our experimental results. We find that dropout members tend to discuss more about disease diagnosis and treatment, but less about daily issues and nutrition. Topics of treatment and diagnosis are common in posts that tell stories of one's cancer journey, or that describe cancer treatment experience. On the contrary, more daily-matter issues like exercises and nutrition are less focused by these users. Not many significant patterns of topic changes are identified longitudinally, except increased frequencies of health system and diagnosis in the final weeks before dropping out. This seems to suggest that although dropout members are more interested in certain topics in general, they do not necessarily shift their focus drastically throughout their participation. The increasing frequency of DIAG is interesting, however. One possible explanation may be that many dropout members were patients who were diagnosed with cancer recurrences or metastasis, which may be followed by the deterioration of the disease.

Our second hypothesis, with respect to user interactions with other members, is partially supported by the results. We originally expected that dropout members receive less replies from other members, which represents a lower level of social support received from other users, and that dropout members post initial posts more often, which represents that they are more likely to be information seekers rather than social support providers. Previous research in online social support groups suggested that emotional support providing is an important motivation of participation [Chung, 2013] and is beneficial to the providers themselves socially [Rodgers and Chen, 2005], which is a factor that are expected to be negatively correlated with attrition.

However, in our static analyses, no significant differences are identified in the static analysis between dropout members and other members with respect to number of replies received, or retio of initial posts to number of reply posts. The result may have two possible explanations. The first is that neither of the two measures can truly represent the degree of social support exchange in online health communities, and the other is that OHC users, particularly BC forum users, are different from online social support group members studied in previous research in how they perceive and understand benefits.

Although static comparison finds no difference, longitudinally we indeed find a rather significant increased ratio of initial posts at the end of user participation, as well as an insignificant drop of numbers of received replies, which is consistent with findings in the previous research that number of replies is important predictor of dropout [Sadeque *et al.*, 2015]. The change is particularly dramatic in the final few weeks from the week view, and in the final 5 days from the day view.

This result, along with results from the static analysis of interactions as well as from previous analyses of topics, possibly shows a more complete picture of dropping-out: dropout members, in terms of support seeking and support providing, are identical to other community members in most of the times throughout their participations; however, certain events, which may be from the real lives of the users such as recurrence of cancer, trigger online behavioral changes and make the users seek much more support than before. At this moment, if these members don't receive adequate support, dropout may eventually happen.

Our final hypothesis that users express increasingly negative emotions in posts are not supported by our analysis. No significant difference is found between dropout members and other members, and no clear patterns can be identified longitudinally. The results contradict findings in previous research that usages of emotional keywords are associated with dropping out [Sadeque *et al.*, 2015], possibly because keywords of emotions cannot truly represent sentiment. Synthesizing the sentiment and interaction results seems to suggest that changes at the end of participation are mostly peaceful in sentiment, with no evident clue emotionally.

What we learned from our topic analysis that dropout members focus more on diagnosis and treatment related themes also reminds us that users may drop out of the community because of death. Their escalated interest in diagnosis and treatment related issues may just be a signal of cancer metastasis, or unsuccessful treatments which may be followed by deterioration of the disease. These members leave the community not because of dissatisfaction towards community usage, and should usually be excluded in the attrition analysis. Similar to the issue of returning of inactive users, in public online communities there is no way to accurately identify dead members.

To investigate how much this confounder impacts our results, we extract cancer stage

information from user signatures, exclude cancer stage IV users, and replicate all analyses. The rationale is that stage IV users are the ones most likely to leave the community because of death, while stage 0 to stage III breast cancer are believed to have quite high 5-year survival rate. These supplementary analyses show identical findings as we demonstrated previously, and the exclusion of stage IV users does not impact the results. It is noteworthy, however, that the result does not indicate the nonexistence of impact of dead members on our study since not all users have accurate profile information in signatures.

Part V

Conclusions and Future Work

Chapter 12

Conclusions and Future Work

We conclude this thesis by first summarizing the main contributions of our work. In this thesis, we created a series of computational tools using natural language processing and machine learning to facilitate content analysis of online health communities at scale. Subsequent studies in the thesis rely on these computational tools and resources to solve specific research problems for online health communities. In particular, we focus on characterizing community members from a social support standpoint, and studying longitudinally patterns of changes of members characteristics and member engagement. The thesis contributed to both research fields of informatics and health psychology from different perspectives. In this chapter, we will also discuss the main limitations of our work and propose directions for future work.

12.1 Contributions

12.1.1 To health researchers

In this thesis, informatics techniques, particularly computational approaches based on natural language processing and machine learning, are exploited in studying online health communities. In contrast to the traditional health interventions of OHCs [Campbell *et al.*, 2004; Hoey *et al.*, 2008], our studies focus on large-scale public online communities, where researchers access content of massive number of patient users but have no control over the underlying design choices of the communities. Our methods are able to characterize com-

munity members from different aspects effectively and efficiently in an automated fashion. We show that computational tools and resources created in this thesis can be straightforwardly used to study OHCs at scale. Successful collaborations between health psychologists and informaticists are also presented in this thesis [Bantum *et al.*, 2016], demonstrating informatics techniques' potential in helping facilitate psychological research of OHCs. We believe that this is an exciting and unprecedented time for OHC research: informaticists and health researchers can join forces and study together the role of online social support and patient health through meaningful collaborations and complementary.

Tools to facilitate research of large-scale OHCs. Previous interventions through online health communities have been carried out in tight experimental setup with full control of the research setting and accessing necessary information to answer research questions of member characteristics and to identify outcomes. Such interventions are usually small in scale, in contrast to the large cohort of users seeking support in public online health communities [Zhang *et al.*, 2016a]. In this thesis, we made it possible for health researchers to study public OHCs and their members at scale, by providing tools of content analysis that can automatically extract information and knowledge from large OHCs with no or little manual work. Particularly, our tools are able to locate salient concepts discussed by users [Zhang and Elhadad, 2013; Elhadad *et al.*, 2014] and hence characterize members [Zhang and Elhadad, 2016b; Zhang *et al.*, 2014; Zhang *et al.*, 2016c] and their engagement behaviors [Zhang *et al.*, 2016b; Zhang and Elhadad, 2016a], which are vital building blocks in studying online social support. For example, our tool for debate detection is able to quickly discover controversial topics under debate in a community and help researchers make sense of opinions on such health issues. This type of automated discoveries acts as an information compression, reducing significantly the amount of content to be manually processed and consumed for humans, and thus has the potential to save significant amount of time for researchers.

Our tools also enable health researchers to discover interesting questions worth exploring about public OHCs for future work, which can be further investigated through traditional interventional methods. For example, in the study we presented in Chapter 7, we found that users in a breast cancer community tend to express more positive sentiment as they

participate longer. Although this does not necessarily indicate that participation brought benefit, it is just not possible to discover such pattern without the automated sentiment analysis tool that can analyze emotion expression of all members in such a massive forum all at once. Our study is also the first time that longitudinal patterns of sentiment change in online health communities are investigated. For another example, in Chapter 6 we discovered that members discuss disease diagnosis when they just join the community, which is consistent with previous findings [Owen *et al.*, 2004b]; however, their topics of discussions shift from clinically relevant ones to more personal ones as they stay longer. Such patterns can be important guidance for health researchers in designing optimal interventions to deliver social support, and can also be valuable hypotheses to be examined in future clinical research.

Tools to discover hidden knowledge of users in traditional interventions. One additional contribution of this thesis is the application of the methods not only to large-scale public OHCs, but also to traditional online peer support groups created by health researchers. Although such groups may not be large in number of participants, content generated by users can still be massive [Gustafson *et al.*, 2002]. In that sense, the methods we created in this thesis also have the potential to facilitate content analysis of these online peer support groups. The study we presented in section 5.3.3 shows that sometimes tools created for public OHCs can be directly applied to online support groups, as long as the target users are of similar types with respect to their member characteristics [Bantum *et al.*, 2016]. Such tools can be used to quickly classify users, to extract information, and to discover correlations between member characteristics in future work.

Toward modeling OHC member characteristics. In this thesis, we present how characteristics of members can be identified through content analysis, based on the computational tools we created. For each member, we are able to discover what topics are discussed, what sentiment is expressed, what treatments are adopted, etc. By considering these variables longitudinally, trajectories were found for topic and sentiment changes, representing how member characteristics change through time as they participate longer in the community. We also create for each user a catalogue of treatment, in which we distinguish treatments discussed by members and treatments actually used by them. These studies

represent a novel approach to studying OHCs, demonstrate the power of large-scale analyses of OHC content, discover patterns associated with OHC usage, and generate interesting research questions to be further studied through experimental interventions.

Several longitudinal patterns with respect to member characteristics are discovered, which are possible signals of effects of online social support. In general, as members participate longer, they 1) discuss less about clinical topics like diagnosis and treatment, but more about personal lives and daily matters; 2) express increasingly positive sentiment; 3) keep updating stories of treatment usage. These findings complements significantly to some other survival analyses in identifying longitudinal patterns [Wang *et al.*, ; Qiu *et al.*, 2011b]. Initial posts and rely posts play different roles, clearly representing social support requesting and providing respectively [Qiu *et al.*, 2011b; Zhao *et al.*, 2012].

Toward modeling member engagement. User engagement and interactions are critical research questions of online health communities. While there are a number of issues regarding user activities such as lurking [Gorlick *et al.*, 2014; Nonnecke *et al.*, 2006; van Uden-Kraan, 2008; Setoyama *et al.*, 2011], in this thesis we focused on studying two important ones: debate and dropout. Based on automatic machine learning and natural language processing, debates can be identified from conversations, along with user stances toward the debated issues. Our analysis shows that certain topics are particularly controversial, and that opinions in a community can be heterogeneous. With respect to dropout, we study how different member characteristics, as identified previously by our methods, are associated with user's decision of dropping-out. To our best knowledge, it is also the first effort to study dropout of online health community at scale by considering multiple variables simultaneously from a longitudinal standpoint. We present how computational methods can be used to investigate engagement related issues, and our preliminary findings generate interesting questions for future health research.

12.1.2 To informaticists

In informatics, natural language processing and machine learning have been applied increasingly in a wide range of health and clinical applications. Most of the existing methods, however, are built upon clinical or biomedical data [Uzuner *et al.*, 2011c; Zhang and El-

hadad, 2013]. While online consumer content share many characteristics with these types of data , OHC's uniqueness makes it impractical to transplant tools created for other types of data directly [Elhadad *et al.*, 2014]. In this thesis, natural language processing and machine learning techniques are tailored to process OHC content. In particular, linguistic resources and computational tools are created to support automated content analysis. Our studies explored from both functional and technical perspectives the possible options for solving different problems, with respect to task formulation, model selection, and feature engineering.

A framework synthesized from a social support standpoint. One particular contribution of this thesis to the informatics research community is the creation of a framework which identifies important variables of interest of health researchers (especially health psychologists), from a social support perspective. The framework is useful because traditionally informatics and health psychology are two disjoint areas of studies, with insufficient realization from either side that informatics techniques can help health researchers solve various problems. This thesis creates a framework (describe in Chapter 2) based on an interdisciplinary literature synthesis, which conceptualizes online health community from a social support standpoint and identifies building blocks of OHCs for relevant research. The framework works as a guideline for informaticists, including technicians equipped with computational tools, of what problems are of interest to health researchers with respect to online peer support and online health communities. To our best knowledge, it is the first framework of conceptualization created for online health community research in particular.

Annotated corpora. High-quality annotated data is vital to building computational tools for OHC research. First, all tools need to be evaluated upon a benchmark with gold-standard answers; in practice, these gold-standards are usually provided by human experts. Second, annotated data is necessary for methods based on supervised machine learning, in which knowledge is learned from examining correlations between data and corresponding annotations. In this thesis, based on content from public online health communities, we provide multi-dimensional annotations for posts: topic of discussion, sentiment, debate, and attribution of treatment mentions (see Chapter 5). All annotations are created with rigorous quality control including double annotation, inter-rater agreement tracking, and

disagreement adjudication. The annotated datasets are built for two types of public online health communities respectively, breast cancer forum and autism forums, with different emphases. Our annotations can be used as either training data or evaluation ground truth in future research involving computational methods and tools.

An unsupervised method for lexicon discovery. One critical step of understanding textual content is to identify salient names, terms, concepts, and hence build lexicons of such terms for further usage. Since types of terms of interest may vary in different applications for different communities, we provide in this thesis an unsupervised method to automatically recognize named entities of interest, and to create lexicons by categorizing these terms (see chapter 4). The tool does not rely on annotated data, and can be adapted to any applications to identify different types of entities. Our tool is the first general-purpose unsupervised tool to identify biomedical terms from text and create domain-specific lexicons. In addition, we demonstrate the difference between using bag of words and word embedding in distributional semantics, as another technical contribution to understanding the lexical semantics of OHC content.

A supervised method for various pragmatical tasks. Understanding OHC content requires not only lexical semantics, but also pragmatics of the conversations. In this thesis, a series of problems, including topic classification, sentiment analysis, debate detection and stance identification, and attribution learning, are solved by a supervised learning pipeline. This pipeline depends on the annotated datasets we created, as well as the lexicons built from our unsupervised approach described above. We demonstrate the feasibility of using supervised machine learning to identify topics, sentiment, debates, and attributions of entities. From a technical standpoint, we also show that convolutional neural network can be a superior choice in the multi-label classification of topics [Zhang *et al.*, 2016c], and that Markov-based CRF model is effective in the sequential learning task of attribution learning [Zhang and Elhadad, 2016b].

Effective features for OHC content analysis. In this thesis, we rely on a wide range of features in the supervised learning tasks, some of which are unique to online health communities. Two particular features were found to be successful across tasks, context and lexicons. As platforms for peer interactions, context information is found to be critical

in content analysis. For instance, in deciding whether positive or negative sentiment is expressed in a post, sentiment conveyed in previous posts can be important. Lexicons, which include salient terms that are important to the domain of interest, are also helpful. Terms in lexicons can be particularly decisive in identifying topics of discussion of a post, for example. Although different communities for different diseases could have completely different sets of salient keywords, our unsupervised lexicon creation method is able to solve this problem by collecting community-specific lexicons automatically. Our study also adds knowledge to machine learning based studies in information extraction from online communities, such as from Twitter [Bian *et al.*, 2012].

12.2 Limitations

There are several limitations, in general, of studies in this thesis, which can be roughly classified into two categories: technical limitations and functional limitations. Technical limitations refer to system's incapability of solving problems with sufficient accuracy, random errors, over-fitting, or lack of portability. For example, for all the supervised machine learning tools, our evaluations show that performance of the systems are usually around %70 to %90, which are far from perfect. Since system predictions are usually carried out at scale, it is impossible to validate results on every sample manually. Actually, getting rid of manual work for all samples is precisely the point of using computational methods. Although different metrics of evaluations help us understand weaknesses of the systems to some extent, a complete solution of correcting errors is not available. This also explains why computational methods standalone are great weapons to attack the scalability obstacles, and are ideal in discovering patterns at scale and in generating hypotheses, but are not sufficient to provide persuasive explanations.

Another important technical limitation of the methods in this thesis is portability of the methods. Supervised learning tools, in particular, depend critically on the availability of annotated data, which needs to be created case by case. We made some efforts in this thesis to make the lexicon creation tool completely unsupervised and portable, but are unable to provide unsupervised solution for every task. As such, it may require additional work before

applying models trained on one dataset (and as a result, "over-fitted" to a certain domain) to another community of a different genre, although one of our studies (sentiment analysis on BC and BCC datasets) show that it is possible to transplant the method as long as the target users are from the same patient population.

Functional limitations, on the other hand, refer to those limitations brought by the fact that all studies presented in this thesis are retrospective, which is primarily because of the lack of control over study subjects in the case of public OHC. All datasets collected for this thesis are from existing content of public OHCs. Without controlling the OHC environment and research setting, we can only study **correlations** between different variables, instead of **causations**. For example, in the sentiment study, we indeed find that sentiment of users is getting more and more positive while members participate longer, but without accessing dropout members we are unable to tell if the change is caused by a true impact of community usage, or just because unhappy members left. This also explains why impact of community usage is not comprehensively discussed in this thesis and we only focus on the characterization parts in the framework. To study psycho-social impact of participation, rigorous study protocols need to be followed, such as a control group, randomization in sampling, a prospective design, etc. Nevertheless, the retrospective and large-scale analysis presented can compliment traditional study designs to gain knowledge of OHCs from a different perspective. Functional limitations are difficult to be tackled just by the informatics researchers, and need to be solved relying on collaborations with health researchers who carry out experimental clinical research.

12.3 Future work

We believe that informatics techniques, particularly computational methods, have enormous potential in studying online health communities. Although we acknowledge that studies in this thesis are retrospective and thus are not sufficient to explain actual impact of participation, it does not mean that these methods cannot be used in prospective research. On the contrary, we believe that rigorously designed clinical studies could benefit a lot from equipping computational tools. For example, imagine a randomized controlled trial of online peer

support in which participants are observed and tracked; computational methods introduced in this thesis can be leveraged to effectively and simultaneously keep track of how topic of discussion changes in the group, how emotion of participants develops, whether certain debates are triggered, etc. The tools could significantly save time for health researchers to monitor these variables and to consume content, provide much more abundant multi-variate descriptions of users, and more timely aid managers of groups to intervene the discussions when necessary. It may also be possible for health researchers to study several outcome measures at the same time with little manual work, and to identify patterns that are unexpected. As such, leveraging computational methods in clinical research settings will be a significant and promising part of the future work. In the future, we will primarily seek to 1) continue making more computational tools for OHC content analysis, and improve existing tools in terms of their accuracy, robustness and portability; 2) apply the methods in more heterogeneous OHCs to find meaningful patterns; 3) to use the methods to study the impact of participation in a experimental study design. Our future work will be more depending on collaborations with health researchers, especially those working on online social support interventions and those interested in gaining knowledge from user-generated content from online health communities.

Part VI

Bibliography

Bibliography

[Adams *et al.*, 2011] Jon Adams, Chi-Wai Lui, David Sibbritt, Alex Broom, Jon Wardle, and Caroline Homer. Attitudes and referral practices of maternity care professionals with regard to complementary and alternative medicine: an integrative review. *Journal of Advanced Nursing*, 67(3):472–483, 2011.

[Andrenucci and Sneiders, 2005] Andrea Andrenucci and Eriks Sneiders. Automated question answering: Review of the main approaches. In *null*, pages 514–519. IEEE, 2005.

[Aramaki *et al.*, 2011] Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. Twitter catches the flu: detecting influenza epidemics using twitter. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1568–1576. Association for Computational Linguistics, 2011.

[Astin, 1998] J a Astin. Why patients use alternative medicine: results of a national study. *JAMA (Chicago, Ill.)*, 279(19):1548–53, 1998.

[Bantum *et al.*, 2016] Erin O'Carroll Bantum, Noemie Elhadad, Jason Owen, and Shaodian Zhang. Machine learning for identifying emotional expression in text: Improving the accuracy of established methods. *Manuscript*, 2016.

[Barnes *et al.*, 2009] P. M. Barnes, B. Bloom, and R. L. Nahin. Complementary and Alternative Medicine Use among Adults and Children: United States, 2007. (12), 2009.

[Batenburg and Das, 2014] Anika Batenburg and Enny Das. Emotional coping differences among breast cancer patients from an online support group: a cross-sectional study. *Journal of medical Internet research*, 16(2):e28, January 2014.

[Beaudoin and Tao, 2008] CE Beaudoin and CC Tao. Modeling the impact of online cancer resources on supporters of cancer patients. *New Media & Society*, 10(2):321–344, 2008.

[Bender *et al.*, 2011] Jacqueline L Bender, Maria-Carolina Jimenez-Marroquin, and Alejandro R Jadad. Seeking support on facebook: a content analysis of breast cancer groups. *Journal of medical Internet research*, 13(1):e16, January 2011.

[Bender *et al.*, 2013] Jacqueline L Bender, Joel Katz, Lorraine E Ferris, and Alejandro R Jadad. What is the role of online support from the perspective of facilitators of face-to-face support groups? A multi-method study of the use of breast cancer online communities. *Patient education and counseling*, 2013.

[Bian *et al.*, 2012] Jiang Bian, Umit Topaloglu, and Fan Yu. Towards large-scale twitter mining for drug-related adverse events. In *Proceedings of the 2012 international workshop on Smart health and wellbeing*, pages 25–32. ACM, 2012.

[Biyani *et al.*, 2014] Prakhar Biyani, Cornelia Caragea, Prasenjit Mitra, and John Yen. Identifying Emotional and Informational Support in Online Health Communities. *anthology.aclweb.org*, 1(1):827–836, 2014.

[Blank *et al.*, 2010] Thomas O Blank, Steven D Schmidt, Stacey A Vangsness, Anna Karina Monteiro, and Paul V Santagata. Differences among breast and prostate cancer online support groups. *Computers in Human Behavior*, 26(6):1400–1404, 2010.

[Blei *et al.*, 2003] DM Blei, AY Ng, and MI Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[Bo Pang and Lillian Lee, 2006] Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 1(2):91–231, 2006.

[Bodenreider, 2004] Olivier Bodenreider. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270, 2004.

[Boon *et al.*, 2007] Heather S Boon, Folashade Olatunde, and Suzanna M Zick. Trends in complementary/alternative medicine use by breast cancer survivors: comparing survey data from 1998 and 2005. *BMC women's health*, 7(1):1, 2007.

[Børøsund *et al.*, 2014] Elin Børøsund, Milada Cvancarova, Shirley M Moore, Mirjam Ekstedt, and Cornelia M Ruland. Comparing effects in regular practice of e-communication and web-based self-management support among breast cancer patients: preliminary results from a randomized controlled trial. *Journal of medical Internet research*, 16(12), 2014.

[Bouma *et al.*, 2015] Grietje Bouma, Jolien M Admiraal, Elisabeth GE de Vries, Carolien P Schröder, Annemiek ME Walenkamp, and Anna KL Reyners. Internet-based support programs to alleviate psychosocial and physical symptoms in cancer patients: A literature analysis. *Critical reviews in oncology/hematology*, 95(1):26–37, 2015.

[Brown *et al.*, 1990] Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Fredrick Jelinek, John D Lafferty, Robert L Mercer, and Paul S Roossin. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85, 1990.

[Campbell *et al.*, 2004] H Sharon Campbell, Marie Rose Phaneuf, and Karen Deane. Cancer peer support programs-do they work? *Patient education and counseling*, 55(1):3–15, October 2004.

[Cappiello *et al.*, 2007] Michelle Cappiello, Regina S Cunningham, M Tish Knobf, and Diane Erdos. Breast cancer survivors: information and support after treatment. *Clinical nursing research*, 16(4):278–93; discussion 294–301, November 2007.

[Castleton *et al.*, 2011] Kimra Castleton, Thomas Fong, Andrea Wang-Gillam, Muhammad A Waqar, Donna B Jeffe, Lisa Kehlenbrink, Feng Gao, and Ramaswamy Govindan. A survey of internet utilization among patients with cancer. *Supportive Care in Cancer*, 19(8):1183–1190, 2011.

[Chapman *et al.*, 2001] W.W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, and B.G. Buchanan. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310, 2001.

[Chung, 2013] Jae Eun Chung. Social Networking in Online Support Groups for Health: How Online Social Networking Benefits Patients. *Journal of health communication*, April 2013.

[Civan and Pratt, 2007] Andrea Civan and Wanda Pratt. Threading together patient expertise. In *AMIA Annual Symposium Proceedings*, volume 2007, page 140. American Medical Informatics Association, 2007.

[Cohen and Others, 1960] Jacob Cohen and Others. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[Cohen *et al.*, 2000] S. Cohen, L.G. Underwood, and B. Gottlieb. *Social Support Measurement and Intervention: A Guide for Health and Social Scientists*. 2000.

[Collobert *et al.*, 2011] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

[Das and Faxvaag, 2014] Anita Das and Arild Faxvaag. What influences patient participation in an online forum for weight loss surgery? A qualitative case study. *Interactive journal of medical research*, 3(1):e4, January 2014.

[Davis *et al.*, 2013] Tonya N Davis, Mark OReilly, Soyeon Kang, Russell Lang, Mandy Rispoli, Jeff Sigafoos, Giulio Lancioni, Daelynn Copeland, Shanna Attai, and Austin Mulloy. Chelation treatment for autism spectrum disorders: A systematic review. *Research in Autism Spectrum Disorders*, 7(1):49–55, 2013.

[Davison *et al.*, 2000] Kathryn P Davison, James W Pennebaker, and Sally S Dickerson. Who talks? the social psychology of illness support groups. *American Psychologist*, 55(2):205, 2000.

[Davison, 2000] KP Davison. Who talks? The social psychology of illness support groups. *American Psychologist*, 2000.

[Della Pietra *et al.*, 1995] Stephen Della Pietra, Vincent Della Pietra, and John Lafferty. Inducing features of random fields. *arXiv preprint cmp-lg/9506014*, 1995.

[Dennis, 2003] CL Dennis. Peer support within a health care context: a concept analysis. *International journal of nursing studies*, 2003.

[Elhadad *et al.*, 2014] Noémie Elhadad, Shaodian Zhang, Patricia Driscoll, and Samuel Brody. Characterizing the sublanguage of online breast cancer forums for medications, symptoms, and emotions. In *AMIA Annual Symposium Proceedings*, volume 2014, page 1123. American Medical Informatics Association, 2014.

[Esquivel *et al.*, 2006] A Esquivel, F Meric-Bernstam, and EV Bernstam. Accuracy and self correction of information received from an internet breast cancer list: content analysis. *BMJ*, 332(7547):937–9, April 2006.

[Eysenbach *et al.*, 2004] Gunther Eysenbach, John Powell, Marina Englesakis, Carlos Rizo, and Anita Stern. Health related virtual communities and electronic support groups: systematic review of the effects of online peer to peer interactions. *BMJ*, 328(May):1–6, 2004.

[Eysenbach, 2005] Gunther Eysenbach. The law of attrition. *Journal of medical Internet research*, 7(1):e11, January 2005.

[Fiszman *et al.*, 2000] M. Fiszman, W.W. Chapman, D. Aronsky, R.S. Evans, and P.J. Haug. Automatic detection of acute bacterial pneumonia from chest x-ray reports. *Journal of the American Medical Informatics Association*, 7(6):593–604, 2000.

[Fox and Duggan, 2013] Susannah Fox and M Duggan. Health online 2013. *Health*, 2013.

[Friedman and Currall, 2003] Raymond A Friedman and Steven C Currall. Conflict escalation: Dispute exacerbating elements of e-mail communication. *Human relations*, 56(11):1325–1347, 2003.

[Friedman and Silver, 2007a] Howard S Friedman and Roxane Cohen Silver. *Foundations of health psychology*. Oxford University Press, USA, 2007.

[Friedman and Silver, 2007b] HS Friedman and RC Silver. *Foundations of health psychology*. 2007.

[Friedman *et al.*, 1994] C. Friedman, P.O. Alderson, J.H.M. Austin, J.J. Cimino, and S.B. Johnson. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174, 1994.

[Friedman *et al.*, 2002] Carol Friedman, Pauline Kra, and Andrey Rzhetsky. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35(4):222–235, August 2002.

[Furlow *et al.*, 2008a] Mandi L Furlow, Divya A Patel, Ananda Sen, and J Rebecca Liu. Physician and patient attitudes towards complementary and alternative medicine in obstetrics and gynecology. *BMC complementary and alternative medicine*, 8:35, 2008.

[Furlow *et al.*, 2008b] Mandi L Furlow, Divya A Patel, Ananda Sen, and J Rebecca Liu. Physician and patient attitudes towards complementary and alternative medicine in obstetrics and gynecology. *BMC complementary and alternative medicine*, 8(1):1, 2008.

[Furnham and Forey, 1994] Adrian Furnham and Julie Forey. The attitudes, behaviors and beliefs of patients of conventional vs. complementary (alternative) medicine. *Journal of clinical psychology*, 50(3):458–469, 1994.

[Gabrilovich and Markovitch, 2007] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611, 2007.

[Gorlick *et al.*, 2014] Amanda Gorlick, Erin O'Carroll Bantum, and Jason E Owen. Internet-based interventions for cancer-related distress: exploring the experiences of those whose needs are not met. *Psycho-oncology*, 23(4):452–8, May 2014.

[Gustafson and Hawkins, 2001] DH Gustafson and Robert Hawkins. Effect of computer support on younger women with breast cancer. *Journal of General Medicine*, pages 435–445, 2001.

[Gustafson *et al.*, 2002] David H Gustafson, Robert P Hawkins, Eric W Boberg, Fiona McTavish, Betta Owens, Meg Wise, Haile Berhe, and Suzanne Pingree. CHESS: 10 years of research and development in consumer health informatics for broad populations, including the underserved. *International journal of medical informatics*, 65(3):169–77, November 2002.

[Han *et al.*, 2011] Jeong Yeob Han, Dhavan V Shah, Eunkyung Kim, Kang Namkoong, Sun-Young Lee, Tae Joon Moon, Rich Cleland, Q Lisa Bu, Fiona M McTavish, and David H Gustafson. Empathic exchanges in online cancer support groups: distinguishing message expression and reception effects. *Health communication*, 26(2):185–97, March 2011.

[Han *et al.*, 2014] Jeong Yeob Han, Jiran Hou, Eunkyung Kim, and David H Gustafson. Lurking as an Active Participation Process: A Longitudinal Investigation of Engagement with an Online Cancer Support Group. *Health communication*, (February 2014):37–41, December 2014.

[Harris and Harris, 1991] ZS Harris and Z Harris. A theory of language and information: a mathematical approach. 1991.

[Hartzler and Pratt, 2011] Andrea Hartzler and Wanda Pratt. Managing the personal side of health: how patient expertise differs from the expertise of clinicians. *Journal of medical Internet research*, 13(3):e62, January 2011.

[Hawn, 2009] Carleen Hawn. Take two aspirin and tweet me in the morning: how Twitter, Facebook, and other social media are reshaping health care. *Health affairs (Project Hope)*, 28(2):361–8, 2009.

[Hø ybye, 2005] MT Hø ybye. Online interaction. Effects of storytelling in an internet breast cancer support group. *Psycho-oncology*, 220(July 2004):211–220, 2005.

[Hoey *et al.*, 2008] Louisa M Hoey, Sandra C Ieropoli, Victoria M White, and Michael Jefford. Systematic review of peer-support programs for people with cancer. *Patient education and counseling*, 70(3):315–337, 2008.

[Hoffman *et al.*, 2009a] Karen E Hoffman, Ellen P McCarthy, Christopher J Recklitis, and Andrea K Ng. Psychological distress in long-term survivors of adult-onset cancer: results from a national survey. *Archives of internal medicine*, 169(14):1274–1281, 2009.

[Hoffman *et al.*, 2009b] KE Hoffman, Ellen P. McCarthy, Christopher J. Recklitis, and Andrea K. Ng. Psychological distress in long-term survivors of adult-onset cancer: results from a national survey. *Archives of Internal Medicine*, 169(14):1274–1281, 2009.

[Houston *et al.*, 2014] Thomas K Houston, Lisa A Cooper, and Daniel E Ford. Internet support groups for depression: a 1-year prospective cohort study. *American Journal of Psychiatry*, 2014.

[Høybye *et al.*, 2005] Mette Terp Høybye, Christoffer Johansen, and Tine Tjørnhøj-Thomsen. Online interaction. effects of storytelling in an internet breast cancer support group. *Psycho-Oncology*, 14(3):211–220, 2005.

[Høybye *et al.*, 2010] MT Høybye, Susanne Oksbjerg Dalton, I Deltour, PE Bidstrup, K Frederiksen, and C Johansen. Effect of internet peer-support groups on psychosocial adjustment to cancer: a randomised study. *British journal of cancer*, 102(9):1348–1354, 2010.

[Hripcsak *et al.*, 1995] G. Hripcsak, C. Friedman, P.O. Alderson, W. DuMouchel, S.B. Johnson, P.D. Clayton, et al. Unlocking clinical data from narrative reports: a study of natural language processing. *Annals of Internal Medicine*, 122(9):681, 1995.

[Huh *et al.*, 2013] Jina Huh, Meliha Yetisgen-Yildiz, and Wanda Pratt. Text classification for assisting moderators in online health communities. *Journal of Biomedical Informatics*, 46(6):998–1005, 2013.

[Hwang and Ottenbacher, 2010] KO Hwang and AJ Ottenbacher. Social support in an Internet weight loss community. *International journal of medical informatics*, 79(1):5–13, 2010.

[Hyodo, 2004] I. Hyodo. Nationwide Survey on Complementary and Alternative Medicine in Cancer Patients in Japan. *Journal of Clinical Oncology*, 23(12):2645–2654, 2004.

[Kim and Shin, 2013] Juhyeon Kim and Hyunjung Shin. Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. *Journal of the American Medical Informatics Association : JAMIA*, 20(4):613–8, July 2013.

[Kim *et al.*, 2003] J-D Kim, Tomoko Ohta, Yuka Tateisi, and Junichi Tsujii. Genia corpusa semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):i180–i182, 2003.

[Kim *et al.*, 2004] J.D. Kim, T. Ohta, Y. Tsuruoka, Y. Tateisi, and N. Collier. Introduction to the bio-entity recognition task at jnlpba. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, pages 70–75, 2004.

[Kim *et al.*, 2012a] E Kim, JY Han, TJ Moon, and B Shaw. The process and effect of supportive message expression and reception in online breast cancer support groups. *Psycho-oncology*, 21(5):531–540, 2012.

[Kim *et al.*, 2012b] Eunkyung Kim, Jeong Yeob Han, Tae Joon Moon, Bret Shaw, Dhavan V Shah, Fiona M McTavish, and David H Gustafson. The process and effect of supportive message expression and reception in online breast cancer support groups. *Psycho-oncology*, 21(5):531–40, May 2012.

[Kim *et al.*, 2013] Sojung Claire Kim, Dhavan V Shah, Kang Namkoong, Fiona M McTavish, and David H Gustafson. Predictors of Online Health Information Seeking Among Women with Breast Cancer: The Role of Social Support Perception and Emotional Well-Being. *Journal of Computer-Mediated Communication*, 2013.

[Kim, 2014] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[Klein and Manning, 2003] Dan Klein and Christopher D Manning. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 423–430. Association for Computational Linguistics, 2003.

[Klemm *et al.*, 1998] Paula Klemm, Melanie Hurst, Sandra L Dearholt, and Susan R Trone. Gender differences on internet cancer support groups. *Computers in nursing*, 17(2):65–72, 1998.

[Kramer *et al.*, 2004] Adam D. I. Kramer, Susan R. Fussell, and Leslie D. Setlock. Text analysis as a tool for analyzing conversation in online support groups. *Extended abstracts of the 2004 conference on Human factors and computing systems - CHI '04*, page 1485, 2004.

[Kraut and Fiore, 2014] RE Kraut and AT Fiore. The role of founders in building online groups. *CSCW*, pages 722–732, 2014.

[Kuhn *et al.*, 2010] Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology*, 6(1), 2010.

[Lafferty *et al.*, 2001] John Lafferty, A McCallum, and FCN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[Lamb *et al.*, 2013] Alex Lamb, Michael J Paul, and Mark Dredze. Separating fact from fear: Tracking flu infections on twitter. In *HLT-NAACL*, pages 789–795, 2013.

[Lapi *et al.*, 2010] Francesco Lapi, Alfredo Vannacci, Martina Moschini, Fabrizio Cipollini, Maria Morsuillo, Eugenia Gallo, Grazia Banchelli, Enrica Cecchi, Marina Di Pirro, Maria Grazia Giovannini, et al. Use, attitudes and knowledge of complementary and alternative drugs (cads) among pregnant women: a preliminary survey in tuscany. *Evidence-Based Complementary and Alternative Medicine*, 7(4):477–486, 2010.

[Lepore *et al.*, 2014] Stephen J Lepore, Joanne S Buzaglo, Morton A Lieberman, Mitch Golant, Judith R Greener, and Adam Davey. Comparing standard versus prosocial internet support groups for patients with breast cancer: a randomized controlled trial of the helper therapy principle. *Journal of Clinical Oncology*, 32(36):4081–4086, 2014.

[Lewallen *et al.*, 2014] Andrea C Lewallen, Jason E Owen, Erin O'Carroll Bantum, and Annette L Stanton. How language affects peer responsiveness in an online cancer support group: implications for treatment design and facilitation. *Psycho-oncology*, 23(7):766–72, July 2014.

[Lieberman *et al.*, 2003] Morton a Lieberman, Mitch Golant, Janine Giese-Davis, Andy Winzlenberg, Harold Benjamin, Keith Humphreys, Carol Kronenwetter, Stefani Russo, and David Spiegel. Electronic support groups for breast carcinoma: a clinical trial of effectiveness. *Cancer*, 97(4):920–5, February 2003.

[Liess *et al.*, 2008] Anna Liess, Wendy Simon, Maya Yutsis, Jason E Owen, Karen Altree Piemme, Mitch Golant, and Janine Giese-Davis. Detecting emotional expression in face-to-face and online breast cancer support groups. *Journal of consulting and clinical psychology*, 76(3):517–23, June 2008.

[Lindberg, 1993] DA Lindberg. The Unified Medical Language System. *Methods of Information in Medicine*, 1993.

[Liu *et al.*, 2011] Xiaohua Liu, Shaodian Zhang, Furu Wei, and Ming Zhou. Recognizing named entities in tweets. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 359–367. Association for Computational Linguistics, 2011.

[Liu, 2007] Bing Liu. *Web data mining: exploring hyperlinks, contents, and usage data*. Springer Science & Business Media, 2007.

[Loper and Bird, 2002] Edward Loper and Steven Bird. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1*, pages 63–70. Association for Computational Linguistics, 2002.

[Magnezi *et al.*, 2014] Racheli Magnezi, Yoav S Bergman, and Dafna Grosberg. Online activity and participation in treatment affects the perceived efficacy of social health networks among patients with chronic illness. *Journal of medical Internet research*, 16(1):e12, January 2014.

[Mamykina *et al.*, 2015] Lena Mamykina, Drashko Nakikj, and Noemie Elhadad. Collective sensemaking in online health forums. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3217–3226. ACM, 2015.

[Manning *et al.*, 2008a] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

[Manning *et al.*, 2008b] Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge, 2008.

[Martin and Jurafsky, 2000] James H Martin and Daniel Jurafsky. Speech and language processing. *International Edition*, 2000.

[McCray *et al.*, 2001] A.T. McCray, A. Burgun, O. Bodenreider, et al. Aggregating umls semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, (1):216–220, 2001.

[Meier *et al.*, 2007a] Andrea Meier, Elizabeth J Lyons, Gilles Frydman, Michael Forlenza, and Barbara K Rimer. How cancer survivors provide support on cancer-related Internet mailing lists. *Journal of medical Internet research*, 9(2):e12, January 2007.

[Meier *et al.*, 2007b] Andrea Meier, Elizabeth J Lyons, Gilles Frydman, Michael Forlenza, and Barbara K Rimer. How cancer survivors provide support on cancer-related Internet mailing lists. *Journal of medical Internet research*, 9(2):e12, January 2007.

[Meier *et al.*, 2007c] Andrea Meier, Elizabeth J Lyons, Gilles Frydman, Michael Forlenza, and Barbara K Rimer. How cancer survivors provide support on cancer-related Internet mailing lists. *Journal of Medical Internet Research*, 9(2), 2007.

[Melton and Hripcsak, 2005] G.B. Melton and G. Hripcsak. Automated detection of adverse events using natural language processing of discharge summaries. *Journal of the American Medical Informatics Association*, 12(4):448–457, 2005.

[Michalski *et al.*, 2013] Ryszard S Michalski, Jaime G Carbonell, and Tom M Mitchell. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[Miller, 1995] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[Milne and Witten, 2008] David Milne and Ian H Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM, 2008.

[Mo and Coulson, 2010] Phoenix K.H. Mo and Neil S. Coulson. Empowering processes in online support groups among people living with HIV/AIDS: A comparative analysis of lurkers and posters. *Computers in Human Behavior*, 26(5):1183–1193, September 2010.

[Molassiotis, 2005] A. Molassiotis. Use of complementary and alternative medicine in cancer patients: a European survey. *Annals of Oncology*, 16(4):655–663, 2005.

[Nambisan, 2011] P Nambisan. Information seeking and social support in online health communities: impact on patients' perceived empathy. *Journal of the American Medical Informatics Association*, 18(3):298–304, 2011.

[Namkoong *et al.*, 2010] Kang Namkoong, Dhavan V Shah, Jeong Yeob Han, Sojung Claire Kim, Woohyun Yoo, David Fan, Fiona M McTavish, and David H Gustafson. Expression and reception of treatment information in breast cancer support groups: how health self-efficacy moderates effects on emotional well-being. *Patient education and counseling*, 81 Suppl:S41–7, December 2010.

[Namkoong *et al.*, 2013] Kang Namkoong, Bryan McLaughlin, Woohyun Yoo, Shawnika J Hull, Dhavan V Shah, Sojung C Kim, Tae Joon Moon, Courtney N Johnson, Robert P Hawkins, Fiona M McTavish, et al. The effects of expression: How providing emotional support online improves cancer patients coping strategies. *J Natl Cancer Inst Monogr*, 47:169–174, 2013.

[Nápoles-Springer *et al.*, 2007] Anna M Nápoles-Springer, Carmen Ortíz, Helen O'Brien, Marynieves Díaz-Méndez, and Eliseo J Pérez-Stable. Use of cancer support groups among Latina breast cancer survivors. *Journal of Cancer Survivorship*, 1(3):193–204, 2007.

[Nelson *et al.*, 2011] Stuart J Nelson, Kelly Zeng, John Kilbourne, Tammy Powell, and Robin Moore. Normalized names for clinical drugs: Rxnorm at 6 years. *Journal of the American Medical Informatics Association*, 18(4):441–448, 2011.

[Nguyen and Rosé, 2011] Dong Nguyen and CP Rosé. Language use as a reflection of socialization in online communities. *Proceedings of the Workshop on Language in Social Media (LSM*, (June):76–85, 2011.

[Nonnecke *et al.*, 2006] Blair Nonnecke, Dorine Andrews, and Jenny Preece. Non-public and public online community participation: Needs, attitudes and behavior. *Electronic Commerce Research*, 6(1):7–20, January 2006.

[ope,] Open NLP. `https://opennlp.apache.org/`.

[Overberg *et al.*, 2010] Regina Overberg, Wilma Otten, Andries de Man, Pieter Toussaint, Judith Westenbrink, and Bertie Zwetsloot-Schonk. How breast cancer patients want to search for and retrieve information from stories of other patients on the internet: an online randomized controlled experiment. *Journal of medical Internet research*, 12(1):e7, January 2010.

[Owen *et al.*, 2004a] Jason E Owen, Joshua C Klapow, David L Roth, Lisle Nabell, and Diane C Tucker. Improving the effectiveness of adjuvant psychological treatment for women with breast cancer: the feasibility of providing online support. *Psycho-oncology*, 13(4):281–92, April 2004.

[Owen *et al.*, 2004b] Jason E Owen, Joshua C Klapow, David L Roth, and Diane C Tucker. Use of the internet for information and support: disclosure among persons with breast and prostate cancer. *Journal of behavioral medicine*, 27(5):491–505, October 2004.

[Owen *et al.*, 2005] Jason E Owen, Joshua C Klapow, David L Roth, John L Shuster, Jeff Bellis, Ron Meredith, and Diane C Tucker. Randomized pilot of a self-guided internet

coping group for women with early-stage breast cancer. *Annals of behavioral medicine : a publication of the Society of Behavioral Medicine*, 30(1):54–64, August 2005.

[Owen *et al.*, 2006] Jason E Owen, Janine Giese-Davis, Matt Cordova, Carol Kronenwetter, Mitch Golant, and David Spiegel. Self-report and linguistic indicators of emotional expression in narratives as predictors of adjustment to cancer. *Journal of behavioral medicine*, 29(4):335–45, August 2006.

[Owen *et al.*, 2007] Jason E Owen, Michael S Goldstein, Jennifer H Lee, Nancy Breen, and Julia H Rowland. Use of health-related and cancer-specific support groups among adult cancer survivors. *Cancer*, 109(12):2580–9, July 2007.

[Owen *et al.*, 2008] Jason E Owen, Erin O'Carroll Bantum, and Mitch Golant. Benefits and challenges experienced by professional facilitators of online support groups for cancer survivors. *Psycho-oncology*, 18(2):144–55, March 2008.

[Owen *et al.*, 2011] Jason E Owen, Eric R Hanson, Doug A Preddy, and Erin OCarroll Bantum. Linguistically-tailored video feedback increases total and positive emotional expression in a structured writing task. *Computers in Human Behavior*, 27(2):874–882, 2011.

[Owen *et al.*, 2014a] Jason E Owen, Erin O Bantum, Amanda Gorlick, and Annette L Stanton. Engagement with a social networking intervention for cancer-related distress. *Annals of Behavioral Medicine*, 49(2):154–164, 2014.

[Owen *et al.*, 2014b] Jason E Owen, Erin OCarroll Bantum, Kevin Criswell, Julie Bazzo, Amanda Gorlick, and Annette L Stanton. Representativeness of two sampling procedures for an internet intervention targeting cancer-related distress: a comparison of convenience and registry samples. *Journal of behavioral medicine*, 37(4):630–641, 2014.

[Park *et al.*, 2015] Albert Park, Andrea L Hartzler, Jina Huh, David W McDonald, and Wanda Pratt. Automatically detecting failures in natural language processing tools for online community text. *Journal of medical Internet research*, 17(8):e212, 2015.

[Pennebaker *et al.*,] James W Pennebaker, Roger J Booth, and Martha E Francis. Linguistic Inquiry and Word Count :.

[Pennebaker *et al.*, 2001] JW Pennebaker, ME Francis, and RJ Booth. Linguistic Inquiry and Word Count (LIWC): A computerized text analysis program. *Mahwah (NJ)*, 2001.

[Portier *et al.*, 2013] Kenneth Portier, Greta E Greer, Lior Rokach, Nir Ofek, Yafei Wang, Prakhar Biyani, Mo Yu, Siddhartha Banerjee, Kang Zhao, Prasenjit Mitra, and John Yen. Understanding topics and sentiment in an online cancer survivor community. *Journal of the National Cancer Institute. Monographs*, 2013(47):195–8, January 2013.

[Qiu *et al.*, 2011a] Baojun Qiu, Kang Zhao, Prasenjit Mitra, Dinghao Wu, Cornelia Caragea, John Yen, Greta E. Greer, and Kenneth Portier. Get Online Support, Feel Better – Sentiment Analysis and Dynamics in an Online Cancer Survivor Community. *2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust*, pages 274–281, October 2011.

[Qiu *et al.*, 2011b] Baojun Qiu, Kang Zhao, Prasenjit Mitra, Dinghao Wu, Cornelia Caragea, John Yen, Greta E Greer, and Kenneth Portier. Get online support, feel better–sentiment analysis and dynamics in an online cancer survivor community. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 274–281. IEEE, 2011.

[Rabiner and Juang, 1986] Lawrence R Rabiner and Biing-Hwang Juang. An introduction to hidden markov models. *ASSP Magazine, IEEE*, 3(1):4–16, 1986.

[Ramage *et al.*, 2003] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled LDA : A supervised topic model for credit attribution in multi-labeled corpora. 2003.

[Ravert *et al.*, 2003] Russell D Ravert, Mary D Hancock, and Gary M Ingersoll. Online forum messages posted by adolescents with type 1 diabetes. *The Diabetes Educator*, 30(5):827–834, 2003.

[Rodgers and Chen, 2005] Shelly Rodgers and Qimei Chen. Internet community group participation: Psychosocial benefits for women with breast cancer. *Journal of Computer Mediated Communication*, 10:1–27, 2005.

[Rozmovits and Ziebland, 2004] Linda Rozmovits and Sue Ziebland. What do patients with prostate or breast cancer want from an Internet site? A qualitative study of information needs. *Patient education and counseling*, 53(1):57–64, 2004.

[Ruland *et al.*, 2013] Cornelia M Ruland, Trine Andersen, Annette Jeneson, Shirley Moore, Gro H Grimsbø, Elin Børøsund, and Misoo C Ellison. Effects of an internet support system to assist cancer patients in reducing symptom distress: a randomized controlled trial. *Cancer Nursing*, 36(1):6–17, 2013.

[Russell, 2013] Matthew A Russell. *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More.* " O'Reilly Media, Inc.", 2013.

[Sadeque *et al.*, 2015] Farig Sadeque, Thamar Solorio, Ted Pedersen, Prasha Shrestha, and Steven Bethard. Predicting continued participation in online health forums. In *SIXTH INTERNATIONAL WORKSHOP ON HEALTH TEXT MINING AND INFORMATION ANALYSIS (LOUHI)*, page 12, 2015.

[Salzer *et al.*, 2010] Mark S Salzer, Steven C Palmer, Katy Kaplan, Eugene Brusilovskiy, Thomas Ten Have, Maggie Hampshire, James Metz, and James C Coyne. A randomized, controlled study of Internet peer-to-peer interactions among women newly diagnosed with breast cancer. *Psycho-oncology*, 19(4):441–6, April 2010.

[Setoyama *et al.*, 2011] Yoko Setoyama, Yoshihiko Yamazaki, and Kazuhiro Namayama. Benefits of peer support in online Japanese breast cancer communities: differences between lurkers and posters. *Journal of medical Internet research*, 13(4):e122, January 2011.

[Settles, 2004] B. Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 104–107. Association for Computational Linguistics, 2004.

[Sharf, 1997] B F Sharf. Communicating breast cancer on-line: support and empowerment on the Internet. *Women & health*, 26(1):65–84, January 1997.

[Shaw *et al.*, 2000] Bret R Shaw, Fiona McTavish, Robert Hawkins, David H Gustafson, and Suzanne Pingree. Experiences of women with breast cancer: exchanging social support over the CHESS computer network. *Journal of health communication*, 5(2):135–159, 2000.

[Singer *et al.*, 2007] Jefferson Singer, Blerim Rexhaj, and Jenna Baddeley. Older, wiser, and happier? comparing older adults and college students self-defining memories. *Memory*, 15(8):886–898, 2007.

[Skeels *et al.*, 2010] Meredith M Skeels, Kenton T Unruh, Christopher Powell, and Wanda Pratt. Catalyzing social support for breast cancer patients. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 173–182. ACM, 2010.

[Somasundaran and Wiebe, 2009] Swapna Somasundaran and Janyce Wiebe. Recognizing stances in online debates. *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - ACL-IJCNLP '09*, 1(June):226, 2009.

[Stanton *et al.*, 2013] Annette L Stanton, Elizabeth H Thompson, Catherine M Crespi, John S Link, and James R Waisman. Project connect online: randomized trial of an internet-based program to chronicle the cancer experience and facilitate communication. *Journal of Clinical Oncology*, pages JCO–2012, 2013.

[Suykens and Vandewalle, 1999] Johan A K Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[Tanabe and Wilbur, 2002] L. Tanabe and W.J. Wilbur. Tagging gene and protein names in biomedical text. *Bioinformatics*, 18(8):1124–1132, 2002.

[Tuarob *et al.*, 2014] Suppawong Tuarob, Conrad S Tucker, Marcel Salathe, and Nilam Ram. An ensemble heterogeneous classification methodology for discovering health-

related knowledge in social media messages. *Journal of biomedical informatics*, 49:255–268, March 2014.

[Turian *et al.*, 2010] Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 384–394. Association for Computational Linguistics, 2010.

[Urcuyo *et al.*, 2005] Kenya R. Urcuyo, Amy E. Boyers, Charles S. Carver, and Michael H. Antoni. Finding benefit in breast cancer: Relations with personality, coping, and concurrent well-being. *Psychology & Health*, 20(2):175–192, April 2005.

[Uzuner *et al.*, 2011a] Ö. Uzuner, B.R. South, S. Shen, and S.L. DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[Uzuner *et al.*, 2011b] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):552–6, 2011.

[Uzuner *et al.*, 2011c] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association : JAMIA*, 18(5):552–6, 2011.

[Valle *et al.*, 2013] Carmina G Valle, Deborah F Tate, Deborah K Mayer, Marlyn Allicock, and Jianwen Cai. A randomized trial of a Facebook-based physical activity intervention for young adult cancer survivors. *Journal of cancer survivorship : research and practice*, 7(3):355–68, September 2013.

[van Mierlo, 2014] T van Mierlo. The 1% rule in four digital health social networks: An observational study. *Journal of medical Internet research*, pages 1–10, 2014.

[van Uden-Kraan, 2008] CF van Uden-Kraan. Self-reported differences in empowerment between lurkers and posters in online patient support groups. *Journal of medical ...*, 10(2):e18, January 2008.

[Vlahovic *et al.*, 2014] Tatiana a. Vlahovic, Yi-Chia Wang, Robert E. Kraut, and John M. Levine. Support matching and satisfaction in an online breast cancer support community. *Proceedings of the 32nd annual ACM conference on Human factors in computing systems - CHI '14*, pages 1625–1634, 2014.

[Wang *et al.*,] YC Wang, Robert Kraut, and JM Levine. To stay or leave?: the relationship of emotional and informational support to commitment in online health support groups. *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 833–842.

[Wang *et al.*, 2015] Yi-Chia Wang, Robert E Kraut, and John M Levine. Eliciting and receiving online support: Using computer-aided content analysis to examine the dynamics of online social support. *Journal of medical Internet research*, 17(4), 2015.

[Weinberg and Schmale, 1996] Nancy Weinberg and John Schmale. Online help: cancer patients participate in a computer-mediated support group. *Health & Social Works*, 1996.

[Wills, 1991] Thomas Ashby Wills. Social support and interpersonal relationships. 1991.

[Winer *et al.*, 1971] Ben James Winer, Donald R Brown, and Kenneth M Michels. *Statistical principles in experimental design*, volume 2. McGraw-Hill New York, 1971.

[Winzelberg *et al.*, 2003] Andrew J. Winzelberg, Catherine CLASSEN, Georg W. ALPERS, Heidi Roberts, Cheryl KOOPMAN, Robert E. ADAMS, Heidemarie ERNST, Parvati DEV, and C. Barr Taylor. Evaluation of an internet support group for women with primary breast cancer. *Cancer*, 97(5):1164–73, March 2003.

[Xue *et al.*, 2007] Charlie C L Xue, Anthony L Zhang, Vivian Lin, Cliff Da Costa, and David F Story. Complementary and alternative medicine use in Australia: a national population-based survey. *Journal of alternative and complementary medicine (New York, N.Y.)*, 13(6):643–650, 2007.

[Yang, 1999] Yiming Yang. An evaluation of statistical approaches to text categorization. *Information retrieval*, 1(1-2):69–90, 1999.

[Yeh *et al.*, 2005] A. Yeh, A. Morgan, M. Colosimo, and L. Hirschman. Biocreative task 1a: gene mention finding evaluation. *BMC bioinformatics*, 6(Suppl 1):S2, 2005.

[Zhang and Elhadad, 2013] Shaodian Zhang and Noémie Elhadad. Unsupervised biomedical named entity recognition: experiments with clinical and biological texts. *Journal of biomedical informatics*, 46(6):1088–98, December 2013.

[Zhang and Elhadad, 2016a] Shaodian Zhang and Noémie Elhadad. Factors contributing to dropping-out in an online health community: Static and longitudinal analyses. In *AMIA*, 2016.

[Zhang and Elhadad, 2016b] Shaodian Zhang and Noemie Elhadad. Learning attribution labels for treatment mentions in an online health community. *Manuscript*, 2016.

[Zhang *et al.*, 2014] Shaodian Zhang, Erin Bantum, Jason Owen, and Noémie Elhadad. Does sustained participation in an online health community affect sentiment? In *AMIA Annual Symposium Proceedings*, volume 2014, page 1231. American Medical Informatics Association, 2014.

[Zhang *et al.*, 2016a] Shaodian Zhang, Erin O'Carroll Bantum, Jason Owen, Suzanne Bakken, and Noemie Elhadad. Online cancer communities as informatics intervention for social support: conceptualization, characterization, and impact. *Submitted to JAMIA*, 2016.

[Zhang *et al.*, 2016b] Shaodian Zhang, Frank Chen, and Noemie Elhadad. "we make choices we think are going to save us" debate and stance identification for online breast cancer cam discussions. *Manuscript*, 2016.

[Zhang *et al.*, 2016c] Shaodian Zhang, Edoaurd Grave, Elizabeth Sklar, and Noémie Elhadad. Longitudinal analysis of discussion topics in an online breast cancer community using convolutional neural networks. In *arXiv preprint*, 2016.

[Zhang, 2015] Yan Zhang. Understanding the sustained use of online health communities from a self-determination perspective. *Journal of the Association for Information Science and Technology*, 2015.

[Zhao *et al.*, 2012] Kang Zhao, Greta Greer, Baojun Qiu, and Prasenjit Mitra. Finding influential users of an online health community: a new metric based on sentiment influence. *arXiv preprint arXiv:1211.6086*, 2012.

[Ziebland *et al.*, 2004] Sue Ziebland, Alison Chapple, Carol Dumelow, Julie Evans, Suman Prinjha, and Linda Rozmovits. How the internet affects patients' experience of cancer: a qualitative study. 328(7439):564, 2004.

Part VII

Appendices

Appendix A

Seed Term List for Treatment Identification for Autism Communities

Melatonin, Secretin, Omega3 fatty acids, Glutenfree caseinfree diet, B6magnesium, Dimethylglycine, Sulforaphane, Probiotics, Antifungal agents, Intravenous immunoglobulin, Chelation, Hyperbaric oxygen, Music therapy, Horseback riding, Transcranial magnetic stimulation, Facilitated communication, Auditory integration training, Stimulants, Alpha agonists, Alpha2adrenergic agonists, clonidine, guanfacine, Atomoxetine, risperidone, Aripiprazole, Olanzapine, Haloperidol, clozapine, quetiapine, ziprasidone, lithium, SSRI, Fluoxetine, Fluvoxamine, Sertraline, Paroxetine, Citalopram, Escitalopram, Clomipramine, Valproate, galantamine, memantine, rivastigmine, Naltrexone, Risperdal

Appendix B

Therapy Grouping for the Manual Coding of Debate Posts

CAM related coding:

1. CAM: CAM v.s. chemotherapy; CAM v.s. evidence based; CAM effectiveness; CAM
2. Gerson therapy: coffee enema, gerson therapy in general
3. Diet: ayurvedic medicine, gluten free, low carb, gluten free carb free, vegan, hormone free meat, weight control diet, fat burning
4. Supplements: black seed oil, cannabis oil, fish oil, vitamin D, vitamin B complex, Potassium
5. Laetrile: laetrile, apricot seeds, grape seeds
6. Estrogen control: DIM, soy, natural replacements for tamoxifen, bioidentical hormones
7. TCM: traditional Chinese herbal medicine, acupuncture
8. Med marijuana
9. Issels
10. Colonics

Non-CAM coding:

1. Cancer cause: fungal, root canal, smoking

2. Others: trolling, spam, cancer diagnosis, health systems