CLUSTERING AND COMPARING INFORMATION EXTRACTED FROM PERSONAL HEALTH MESSAGES

BY

YUNLIANG JIANG

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2012

Urbana, Illinois

Doctoral Committee:

Professor Bruce R. Schatz, Chair
Professor Jiawei Han
Associate Professor Chengxiang Zhai
Assistant Professor Qiaozhu Mei, University of Michigan at Ann-Arbor

Abstract

The development of Web 2.0 techniques has led to the prosperity of online communities, which spread to various domains and areas in our daily life. When it comes to the *medicine and healthcare* domain, a series of good online services such as Yahoo! Groups, WebMD and MedHelp, offer patients and physicians a good platform to discuss health problems, *e.g.*, diseases and drugs, diagnoses and treatments, which also provide a large volume of data for researchers to analyze and explore. However, some nature of the personal messages, *e.g.*, unclean, unstructured and isolated from clinical practice, hinders users' effective digestion of information in the front end and challenges the data analysis in the back end. In such a scenario, the objective of my thesis is to apply the advanced data mining, information retrieval and natural language processing techniques to effectively analyze and re-organize the rich source of personal health messages from online medical communities, in order to satisfy patients' information need and support physicians' clinical practice.

Specially, in the first part of the dissertation, I introduce an SVM-based multi-class classification method which utilizes term-appearance, lexical and semantic features to effectively classify health messages sampled from our unique dataset of Yahoo! Health Groups into three categories: *News*, *User Comments* and *Spam*; in the second part, I depict a comprehensive system with an extensive evaluation framework to organize and cluster patient outcomes utilizing topic model, which groups large collections of personal comments into a series of topics, guided by expert comments; in the third part of the dissertation, I address a novel and promising topic: Comparative Effectiveness Research (CER) hypothesis prediction, by presenting a

study which evaluates patients' opinions on different treatments by machine enabled sentiment analysis or human analysts utilizing our MedHelp dataset. By suggesting three different methods to *compare* such opinions, reliable conclusions about the patients' preference on different treatments can be drawn consistently, which imply the effectiveness of the treatments. Furthermore, the study is also extended to demographic analysis to explore the preference in specific group of people, representing population cohorts.

To my family and my friends.

Acknowledgements

At the beginning, I thank all the people who have helped, supported me during my Ph.D. study since 2007. You gave me the power and confidence to achieve today's goal.

First and foremost, I wish to express my most sincere gratitude and appreciation to my advisor, Professor Bruce Schatz for his guidance, support and encouragement during my Ph.D. study. He accepted me as his Ph.D. student at the time I switched my research topic from traditional web search to medical informatics. During the years we spent together, his academic knowledge, vision, enthusiasm always guide me to the correct direction. This thesis would not have been possible without his help.

Also I would like to thank my doctoral committee members, Professor Jiawei Han, Professor Chengxiang Zhai and Professor Qiaozhu Mei, for their valuable face-to-face guidance on my study and research, as well as constructive suggestions on this dissertation. I thank Doctor Richard Berlin for his professional medical motivation and correction.

I owe sincere gratitude to my closest colleagues and long-term friends, Cindy Xide Lin, Vera Qingzi Liao and Shuyi Chen. They offer valuable help and support to my research and my career in industry. I would also thank all my colleagues in Database and Information System (DAIS) group, and all my friends in Champaign.

I thank United States Department of Agriculture (USDA) National Research Initiative (NRI), grant 2009 - 35302 - 05285, Institute for Genomic Biology at University of Illinois at Urbana-Champaign, and State Farm Doctoral Scholar to support the major work of the dissertation.

I am indebted to my father Longchun Jiang, and my family for their care, love and support all the time.

Last but not least, I would like to dedicate this dissertation to my dear mother Guanglin Song. Although she passed way two years ago, she always lives in my heart. She is very brave and optimistic while facing the disease, which inspires me to overcome any difficulty and challenge in my research study, as well as my life. No matter where I am, she is there, giving me greatest power and love.

Table of Contents

List of Tables . . . . . **ix**

List of Figures . . . . . **x**

CHAPTER 1 Introduction **1**

CHAPTER 2 Related Work **7**

CHAPTER 3 Multi-class Classification in Online Personal Health Messages **10**
3.1 Problem Definition 12
3.2 Methodology 14
3.3 Experiment and Result 17
3.3.1 Data and Setup 17
3.3.2 Experiment 1: Choosing Proper Term-appearance (TA) Feature 18
3.3.3 Experiment 2: Choosing Proper Combination of Feature Spaces 19
3.3.4 Experiment 3: The Performance of Detecting User Comments (C) . . . 20
3.4 Conclusion 22

CHAPTER 4 Designing and Evaluating a Clustering System for Integrating Patient Drug Outcomes **23**
4.1 Problem Definition 27
4.2 System Architecture 28
4.3 Methodology 30
4.3.1 Pre-processing: Filtering the Messages 30
4.3.2 Clustering: Outcome Selection and Integration 30
4.3.3 Post-processing: Separating Similar and Opposite Opinions 33
4.4 Experiments and Results 34
4.4.1 Data and Setup 34
4.4.2 Annotation Framework 35
4.4.3 Clustering Results and Analysis 37
4.4.4 Case Study 41
4.4.5 Advantages and Limitations of Model 44
4.5 Conclusion and Future Work 48

CHAPTER 5 Comparative Effectiveness Research(CER) Hypothesis Prediction in Personal Health Messages 49
5.1 Problem Definition 53
5.1.1 Setup 53
5.1.2 Design 55
5.1.3 Direct Comparison by Same Author 57
5.1.4 Indirect Comparison by Same Author 59
5.1.5 Indirect Comparison in Overall Case 60
5.1.6 Demographic Analysis 61
5.2 Results 63
5.2.1 Experiment A: Direct Comparison by Same Author 63
5.2.2 Experiment B: Indirect Comparison by Same Author 66
5.2.3 Experiment C: Indirect Comparison in Overall Case 69
5.2.4 Experiment D: Demographic Analysis 71
5.3 Limitations 78
5.4 Conclusion and Future Work 80

CHAPTER 6 Conclusion and Summary 81

References 84

List of Tables

3.1 Message Examples 13
3.2 The Result of Term-appearance feature 18
3.3 The Result of different feature selection 19
3.4 The Result of detecting user comments C 20

4.1 The performance of the clustering result for all the drugs 39
4.2 The performance of distinguishing O_{i_sim} and O_{i_opp} 41
4.3 Sample results of O_m for Clonazepam 44
4.4 The outcome results for Clonazepam with expert comments 46

5.1 The general statistics of breast cancer sub-forums 54
5.2 The general statistics of depression sub-forums 54
5.3 The keywords for each treatment 56
5.4 The results of Direct Comparison by author 63
5.5 The chi-square test results of Direct Comparison by author 64
5.6 The proportion test results of Direct Comparison by author 65
5.7 The results of Indirect Comparison by author 66
5.8 The chi-square test results of Indirect Comparison by author 67
5.9 The proportion test results of Indirect Comparison by author 68
5.10 The results of Indirect Comparison in overall 69
5.11 The two-sample proportion test results(positive) in overall 70
5.12 The two-sample proportion test results(negative) in overall 71
5.13 The attitude to Chemo, Breast Cancer 72
5.14 The attitude to Radiation, Breast Cancer 72
5.15 The attitude to Hormonal, Breast Cancer 72
5.16 The attitude to Meditation, Depression 72
5.17 The attitude to SSRI, Depression 72
5.18 The attitude to SNRI, Depression 72
5.19 The attitude to TCA, Depression 73
5.20 The preference of inner-group of Breast Cancer 74
5.21 The preference of cross-group for breast cancer 75
5.22 The preference of cross-group for depression 75

List of Figures

1.1 Three dimensional elements of problems 2
1.2 Dimension decomposition of three problems 5

4.1 The architecture of the integration system 28
4.2 Annotation Interface 36
4.3 Systerm User interface 42

CHAPTER 1

Introduction

The development of Web 2.0 techniques has led to the prosperity of online communities, which spread to various domains and areas in our daily life. When a person plans to buy an electronic product, she would like to view other customers' reviews from shopping websites. When a user tries to trade on some specific stock, she would like to know other traders' comments from online stock discussion board. Web forums in specific area provide valuable information in support of users' product judgment by exposing them to others' recent experiences. Though subjective, they reflect the most direct and comprehensive opinions from actual people using actual products.

When it comes to the *medicine and healthcare* domain, the situation is similar: online bulletin boards and chat groups, such as Yahoo! Groups[1], WebMD[2] and MedHelp[3]. offer patients and physicians a good platform to discuss health problems, *e.g.*, diseases and drugs, diagnoses and treatments. These online user discussions also provide rich material for case studies, which are a standard approach to medicine. From case studies of drug outcomes, physicians could know how well a drug works, what its outcomes are, including side effects, and patients could know the experience from similar patients, whether the drug is effective, under what conditions.

However, online medical discussions have limitations hindering users' effective digestion of information – they usually contain millions of unstructured messages. They do offer a

[1] http://groups.yahoo.com/
[2] http://www.webmd.com/
[3] http://www.medhelp.org/

convenient platform for communication about medical issues while is though lack of direct connection with clinical practise and evaluation. In addition, various of online medical discussions are addressed by different groups of people thus the topics discussed are diverse. Thus, it is appealing to effectively *analyze* and *re-organize* the rich source of personal health messages by applying advanced data mining, information retrieval and natural language processing techniques, in order to satisfy patients' information need and support physicians' clinical practise.

Many interesting and promising problems are related to the topic of personal health message analysis, and re-organization. For our observation, each potential problem consists of three dimensional elements: data sources, objects and techniques, which is shown in Figure 1.1.

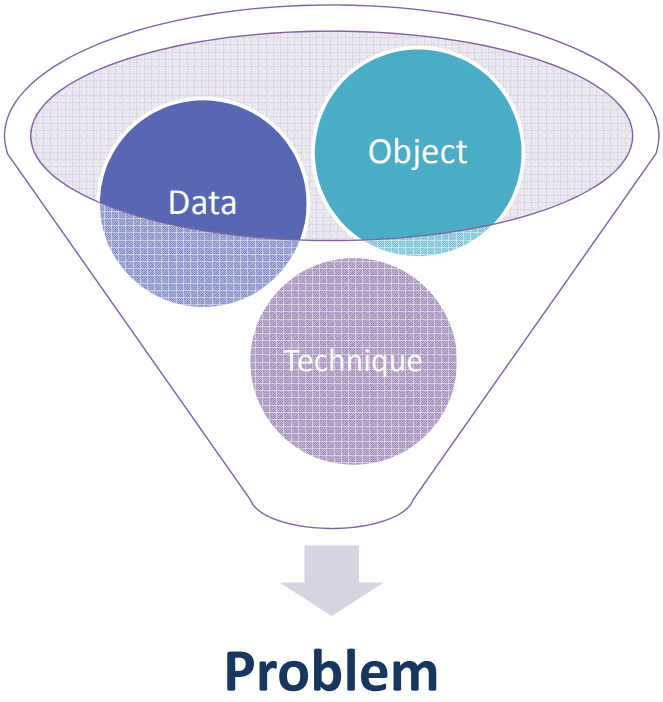


Figure 1.1: Three dimensional elements of problems

Data sources: Many online services offer specific healthcare forums and discussion boards. For example: patientslikeme [4], MedHelp[5], Yahoo! Groups in Health and Wellness, Yahoo! Answers [6] in Health Category, while people can freely discuss and post healthcare-related messages on social network communities such as Twitter [7] or Sina weibo [8]. Both of the sources

[4]http://www.patientslikeme.com/
[5]http://www.medhelp.org/
[6]http://answers.yahoo.com/
[7]https://twitter.com/
[8]http://www.weibo.com/

can be retrieved and applied in format of text messages. Also, personal health messages also include voice messages.

Objects: Sometimes users such as patients and pharmacists focus on *drugs*: what is the main purpose of each specific drug? Does it have severe side effects? While users' topics are also organized by *diseases* in many forums. People may share the treatment information, or just experiences about the same disease they may suffer. Furthermore, the messages can also be organized by treatments, clinical trials, diagnoses or even symptoms.

Techniques: Many applicable techniques in data mining, information retrieval and natural language processing make it possible to build bridges between conceptual model in computer science to real medical problems. Such techniques vary according to different problems and purposes.

Online personal health messages are not always clean: not all messages are relevant and useful for our learning tasks, *i.e.*, a huge amount of advertisements can be found on the web forum. Even they are, they may contribute in different ways: some are expertise; some are comments from patients, *etc*. If a user searches for a specific drug, there are usually thousands of comments or reviews returned. The user has difficulty in digesting and understanding the information quickly – she has to select the useful posts from the pool and filter out the spam. In such a scenario, it is appealing to **classify these messages**, not only for the reason of improving the effectiveness (*e.g.*, filter out spam) but also to benefit the efficiency (*e.g.*, decrease the number of message to process).

After successfully classifying the messages and acquiring the useful ones, another and even more challenging problem arises: *how to help users to understand and digest such unstructured messages?* In online medical forum, each comment talks about several topics in one piece of plain text. These messages must be partitioned into parts, and these parts must be grouped together according to what topic category they each belong to. By doing this we will have a coherent view on different aspects of the medical issue based on all the information available

from our source. Our purpose is to **integrate and cluster the drug-based outcomes** from a large number of unstructured online messages, into meaningful groups according to the topics, in order to aid users navigate through the vast information pool and satisfy their information need.

For the purpose of clinical practice, recently **Comparative Effectiveness Research**(CER), which is defined and sponsored by Institute of Medicine [1], draws researchers' attention to analyze and compare different interventions and strategies in clinical trials. After successfully integrating the patients' outcomes, CER can be conducted to compare the effectiveness of different drugs or treatments on some specific disease. From the result we can answer the question: *which treatment is more effective and favored by more patients?* (*e.g.*, Aspirin is more effective than Ibuprofen to treat migraine headache). Based on CER's definition, the real pragmatic trials are required thus conclusions can not be made by only analyzing text-based personal messages. However, the comparative results generated from personal messages can either become the proof of existing CER results, or the hypothesis of future CER topics.

Let us re-consider the three dimensional elements of problems. Figure 1.2 shows the details of decomposition of the above three problems we proposed. From Figure 1.2 we observe that various of data sources can be covered, multiple medical objects can be addressed and plenty of techniques can be explored and applied by solving the proposed problems, which ensures that the topic is adequately studied. However, for the time and technique limits, some of the components will not be addressed, *e.g.*, the process of voice health messages, the extraction of symptoms and diagnoses, *etc*, which leaves the topic as an open pool for future exploration.

Organization: In the first part of the dissertation, I will introduce an SVM-based multi-class classification method which utilizes term-appearance, lexical and semantic features to effectively classify health messages sampled from our unique dataset of Yahoo! Health Groups into three categories: *News*, *User Comments* and *Spam*.

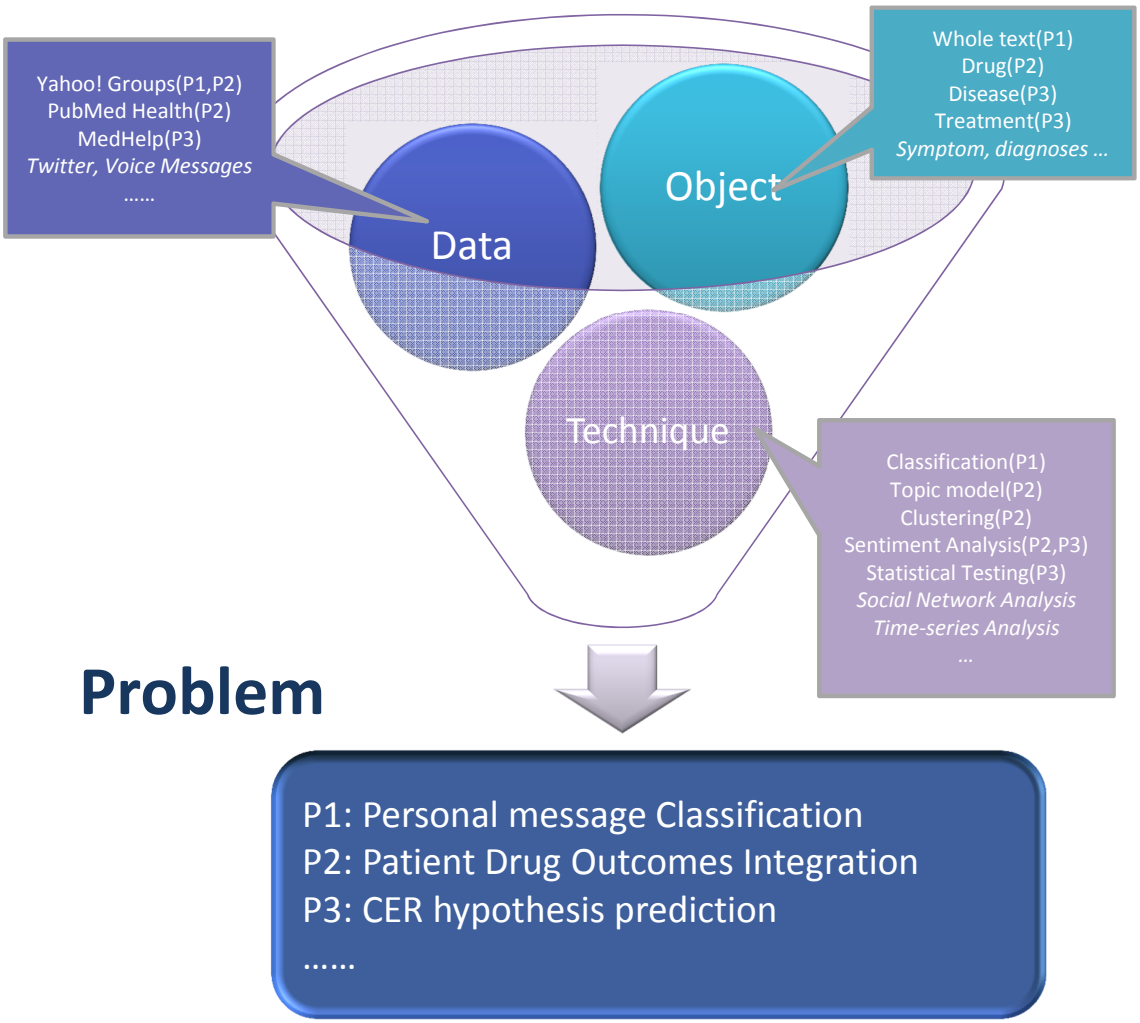


Figure 1.2: Dimension decomposition of three problems

In the second part of the dissertation, I will present a comprehensive system to organize and cluster patient outcomes utilizing topic model, which groups large collections of personal comments into a series of topics, guided by expert comments. A prototype implementation is built to extract situational evidences by digesting user comments provided by patients. Ten drugs, belonging to two groups (*specialized* and *generalized*) have been sampled and tested from the same dataset of Yahoo! Health Groups. Also an extensive evaluation of the clustering results has been performed by medical school students, to validate the good performance of the system.

In the third part, I will address a novel and promising topic: comparative effectiveness research (CER) hypothesis prediction. Different from the traditional analysis from EMR data set, we present a study which evaluates patients' opinions on different treatments by machine enabled sentiment analysis or human analysts. By suggesting three different methods to *compare* such opinions, we can draw the reliable conclusion about the patients' preference on different treatments consistently, which can also imply the effectiveness of the treatments. Furthermore, the study is also extended to demographic analysis to explore the preference in specific group of people. As a case study, we utilize MedHelp dataset and focus on two common diseases: breast cancer and depression to compare the effectiveness of their treatments, respectively.

The conclusions and summaries are given in the last chapter.

CHAPTER 2

Related Work

Nowadays, many applications utilize *formal* medical literatures to extract and integrate useful information. Such applications include: generating text summaries [2], topic modeling [3], mining predictive rules [4], search query optimization [5], mining drug-AE(adverse effects) associations [6], evaluation by propensity-score matching [7, 8], and BeeSpace Navigator [1] which builds an interactive system for functional analysis of biological literatures [9, 10, 11]. *etc.* Compared with the formal literatures, *informal* medical messages which are generated by large number of online users are more unstructured and noisy, which challenges the information extraction.

Instead of formal and structured literatures, some research papers apply natural language processing techniques on unstructured *clinical notes*, such as document clustering [12], medication information extraction [13, 14], term characteristics analysis [15], abbreviation analysis [16], social-history information detection [17], patient identification [18] *etc.* Compared with these, personal medical messages are more informal and contain more useless information, which challenges the data processing. Meanwhile, the topic diversity of the online personal messages reflects various responses and opinions from real physicians and patients.

There have been only a few studies using informal medical sources: Crain *et al.* [19] and Zhang [20] worked on consumer medical search by using Yahoo! Answer health messages, while Yang *et al.* did a solid query log analysis [21] based on the Electronic Medical Record Search Engine (EMERSE) [22]. Chee *et al.* have published several papers based on Yahoo!

[1] http://www.beespace.illinois.edu/

Groups messages, such as tracking users' sentiments [23] and prediction of adverse drugs [24]. Using the same dataset, we perform a comprehensive information extraction task and apply a series of text mining techniques.

Classification in short text messages has been addressed by some researchers. Hidalgo *et al.* [25] and Cormack *et al.* [26] detected spam messages from the data pool while different classification approaches were applied such as Naive Bayes, SVM, decision trees, *etc*. Munro [27] utilized the sub-word variation to classify medical clinic notes written in Chichewa language. Volkova *et al.* [28] tried to classify the sentences extracted from animal-disease-related text into "suspected" or "confirmed" group. Compared with such classification work, the problem we proposed addresses how to select the feature space which can recognize the characters of different web medical messages, as well as solving a multi-class classification problem rather than binary classification.

Topic models such as Probabilistic Latent Semantic Analysis (PLSA) [29] and Latent Dirichlet Allocation (LDA) [30] have been applied to text mining problems [31, 32, 33, 34, 35, 36] with good results. For example, Lu *et al.* applied PLSA model to integrate product aspects [37] and assign rating [38] from online product reviews, while Kandulaweb *et al.* [39] utilized LDA model to discover diabetic-related medical materials. In the above work, limited evaluations upon a small number of manual post-labeling were applied. Undertaking semi-supervised PLSA model, a convincing evaluation strategy based on a large-scale gold standard data produced by professionals is effective to evaluate the performance of the model.

Comparative Effectiveness Research (CER) identifies what treatment works best for which patients under what circumstances. The conclusions can aid consumer, patients, and care providers' decisions on which diagnostic or treatment option to use. Lately, a number of studies have taken the initiatives to operationalize the concept of CER in medical research ([40, 41, 42]). For example, Sox and Greenfield [43] introduced the CER definition, high-

priority topics, methodologies of CER as well as the translation from research findings to the patient care, which give a comprehensive guide for the further CER.

Some studies aims at outlining the general framework of CER research. For example, Ratner *et al* [44] as IOM committee chairs reported the priority directions of CER field and Luce *et al* [45] discussed the better research methods on CER. Trill [46] , Garber and Tunis [47] discussed the relation between CER and personalized medicine. Meanwhile increasing number of research papers published have explored CER for specific condition. For example, CER in chronic obstructive pulmonary disease(COPD) [48], asthma [49], Raiology [50] [51], colorectal cancer [52], heart disease(IHD) [53], *etc.*

Comparative Effectiveness Research (CER) also increasingly draws researchers' attention in medical informatics area [54]. Djulbegovic [55] discussed the importance of applying data mining techniques in CER. Bhavnani *et al.* analyzed the clinic trail network on depression [56] and the relation with CER. Roy and Hennesy applied Bayesian model to compare the effectiveness among the drug classes on hypertension [57], *etc.* However, to the best of our knowledge, few studies have explored how to detect the treatment effectiveness by mining personal messages, which could be a highly promising area for informing clinical trial and other formal CER methods.

Comparative study has been a frequently visited topic by researchers in text mining area. Jindal and Liu did solid research work on detecting the comparative sentences and relations from text documents [58], [59], as well as the sentiment analysis in text [60], [61]. Based on such fundamental work, some researchers did comparative study by analyzing the online product reviews, such as product comparison and recommendation [62], [63], opinion visualization [64], generating comparative summaries [65], judgement aggregation [66], *etc.* My dissertation utilizes some of the techniques in the above paper and apply them on the personal health messages.

CHAPTER 3

Multi-class Classification in Online Personal Health Messages

In this section, I propose an effective approach to process and digest the raw data about healthcare – to classify all the messages into three main categories: News (N), User comment (C) and Spam (S). Based on such results, different analysis can be applied on specific category data in purpose.

Particularly, an SVM-based multi-class classification method is applied to separate the messages into three pre-defined categories. The core step of SVM classification is to select the proper feature spaces. We carefully choose three kinds of feature spaces: *term appearance*, *lexical* and *semantic* feature. Based on that, a tri-class SVM classifier has been trained upon the health messages sampled from our unique dataset of Yahoo! Groups under Health and Wellness.

We apply 10-fold cross-validation to verify the quality of the classifier and design several experiments to choose the best combination of the feature spaces. Evaluation results show that our classifier can achieve a high accuracy (90.15%) while all three kinds of features are applied. Our SVM-based multi-class classification approach with careful feature-space selection can successfully classify the health messages into different categories, especially the most useful part – user comments.

The section of work is complete and published in Proceeding of 2011 International workshop on Web Science and Information Exchange in the Medical Web (MedEX 2011), co-located with 2011 International Conference on Information and Knowledge Management (CIKM 2011) [67].

Organization. Section 3.1 defines the problem, as the solution of which an SVM-based multi-class classifier is proposed in Section 3.2. We then present experiments and results on Yahoo! Groups data in Section 3.3. Finally, we make conclusion in Section 3.4.

3.1 Problem Definition

According to the observation on our unique dataset from Yahoo! Groups, containing $12M$ personal messages from $27K$ public groups in Health and Wellness, we can specify all the messages into three categories:

- News (N): Community members may quote some news articles in their posts, whose content is about the latest scientific discovery or FDA announcement of medical drug or treatment. Such messages are usually very long, having several paragraphs, and written by scientific journalists. This kind of messages can be used to extract medical entities, event and expert opinions [68] or to compare the language usage with other sources [69].

- Comment (C): User comments reflect direct opinions from actual patients and physicians on specific healthcare topics. They have proper lengths and most of the parts are informational. In Chee's work using the same dataset [70], this kind of messages was actually applied to evaluate the sentiment of user opinions. Also an opinion integration system could be built upon such messages.

- Spam (S): Most advertisements [71], posted by human or robots, should be eliminated for our purpose. They are often short, and appear repeatedly, which makes it easy to identify them automatically.

Table 3.1 shows a snippet of each category of messages, which are extracted from the real Yahoo data.

To automatically detect each message's category, we propose an SVM-based multi-class classification method. To train an accurate classifier where the core step is feature selection, we carefully select three kinds of feature spaces, namely *term-appearance, lexical* and *semantic* feature. Based on the feature space, we translate and label the health messages sampled from

...An FDA drug advisory panel deliberating about the safety of Sanofi-Aventis' marketing application for its weight reducing drug, Zimulti rejected the application 14 to 0. They did so on the basis of the evidence presented. The panel was informed by an medical reviewer, Dr. Amy G. Egan, that: "The potential market for this drug and the continued uncertainty about its risks, both known and unknown, lead to our concern about the use of this drug in the general population. ...	...I am on Meridia. I started taking it six days ago continuously now. I had tried it a month or so ago, and got worried and anxious about the side effects, so I stopped taking it after 3 days, but now I am glad that I restarted it. I am 5'6" @ 36 years old, (almost 37) and started out at 263 pounds. I am a binge eater. I used to skip breakfast and then binge the rest of the day. I was not exercising or eating for health. My LDL's were sky high, and my HDL were low. I have high cholesterol etc. However, I have NORMAL blood pressure and resting pulse, which is important in starting Meridia (sibutrimine) ...
(a). News (N)	(b). User comment (C)

... Everything is 70% off for this week only! Get for men's health Buy Valium for CHEAP Get Xanax for Anti-Anxiety. Buy Meridia online for weight loss Get Ambien to help you sleep We have all the products for your needs: Angina, Arthritis, Antibiotic, Anxiety disorder ...
(c). Spam (S)

Table 3.1: Message Examples

our unique dataset of Yahoo! Groups with Health and Wellness and train a tri-class SVM classifier with a RBF kernel upon it. In Section 3.2, we will introduce the method in details.

3.2 Methodology

We apply SVM classification [72] to classify all the messages into three classes (N, C, S). Support vector machines (SVM) are a group of related supervised learning methods which are used for classification task analyzing data and recognizing patterns. More paretically, a support vector machine builds a *hyper-plane* or multiple *hyper-planes* (when there are multiple classes) in a relatively high-dimensional space, in order to make the distance from hyper-planes to the nearest training data as far as possible, to minimize the error. Compared with other potential approaches like manually labeling, rule-based parsing, SVM has relatively high accuracy and can handle high-dimensional data automatically. The main steps are as follows:

- Choose the data pool and label messages in the pool as news (N), user comment (C), and spam (S).
- Select the proper feature spaces and transfer each message into a vector, where each dimension refers to one feature.
- Separate the pool into training and testing set. Train an SVM classifier on the training set, which automatically studies the weight of each dimension and find the most discriminative margin.
- Test the classifier on the testing set.

The most essential step is the feature selection. Our approach explores three kinds of feature spaces: *term appearance, lexical* and *semantic* feature.

Term-appearance feature (TA): Some text classification work [24, 73] has already explored such feature as term frequency (TF): for each message i's feature vector $v(i)$, the value of dimension j: $v(i, j)$ refers to the number of appearances of the term j in message i. In the paper we use one additional representative format – word distribution (WD) which represents

the probability of term's appearance in a message. ***i.e.***,

$$v(i,j) = \alpha \cdot \frac{freq(i,j)}{length(i)} + (1-\alpha) \cdot p(j,B) \tag{3.1}$$

In Formula, $freq(i,j)$ refers to the frequency of term j in message i, $length(i)$ means the total number of terms in message i, and $p(j,B)$ shows the probability of term j in the whole text collection B. Actually word distribution (WD) is a normalized version of term frequency (TF). We will test which one suits our classifier better in Section 3.3.

Should all the terms or only high-TA terms be included in the feature space? Intuitively, many terms with low TA may involve noises since their appearances are randomly scattered across multiple messages if all the terms are included. However insufficient feature size may miss some discriminative terms if only a few high-TA terms are considered. Both cases would damage the accuracy. Thus we will discuss the range of term choosing in the experiment.

Lexical feature (LE): This feature space includes a series of lexical statistics of one message, from which we know the structural information e.g., terms, sentences, paragraphs, even punctuation. From Instruction Section we have the sense that lexical features may be discriminative among three types of messages. Applying lexical feature would distinguish them well. Here we list several lexical features:

- The number of terms/unique terms
- The ratio of unique terms to total number of terms
- The number of sentences/paragraphs
- The average length of sentences/paragraphs
- The number/percentage of uppercase terms

Semantic feature (SE): These features analyze the semantic meaning of one message, which need to "understand" the content, including:

- The number/percentage of drug/treatment names
- The number/percentage of medical terms
- The number/percentage of sentences with medical terms
- The number/percentage of paragraphs with medical terms
- The percentage of positive/negative/subjective sentences

Since different types of messages always express different content by purpose, semantic features would be potentially good discriminators although capturing them is more challenging than lexical features. We generate drug/treatment names from a professional drug website [1] as well as medical terms from wikipedia [2], and extract them from each message. For sentiment analysis, we utilize a sentiment analysis tool [60] which is capable of determining the sentiment polarity of each sentence, *i.e.*, positive or negative opinions. We train a tri-class SVM classifier and design a series of experiments to select the proper feature spaces and test the performance in Section 3.3.

[1] http://www.drugs.com/
[2] http://en.wikipedia.org/

3.3 Experiment and Result

3.3.1 Data and Setup

We utilize our unique dataset which is segmented from Yahoo! Groups with Health and Wellness data to build the SVM classifier and test. The dataset consists of $27,290$ public groups and $12,519,807$ messages in total, crossing seven years and covering various topics. The data has been applied in Chee's work [70, 24]. All the experiments run on a 4TB-disk, 4GB-RAM, and 10-core server.

We randomly choose 1254 messages from the mixed collection as our data pool. After manually labeling them we acquire 293 pieces of news (N), 329 user comments (C) as well as 632 pieces of spam (S). Since our data comes from public groups which are open to all web users, spam (S) takes a much larger portion than that in private groups where only registered group members can communicate with each other in order to prevent other individuals or robots from sending advertisements and irrelevant messages. From this we can see the personal messages are really dirty – a huge number of data is useless and should be eliminated.

We then apply 10-fold cross-validation [74] to verify the quality of the classifier. The labeled data is randomly partitioned into 10 equal-sized subsets. Each time one single set is selected as testing set for evaluation and the rest 9 subsets are used as training set. The whole process will be repeated 10 times to make sure each subset be tested exact once. We use accuracy as the evaluation metric, defined as:

$$\text{Accuracy} = \#\text{ correctly classified messages}/\#\text{ of tested sentences} \tag{3.2}$$

3.3.2 Experiment 1: Choosing Proper Term-appearance (TA) Feature

As we discussed in Method Section, Word distribution (WD) and Term frequency (TF) are two formats of Term appearance (TA) feature. We would choose one of them which shows higher performance. Meanwhile, we also need to determine the range of terms choosing in the experiment.

We rank all the terms appearing in the data pool according to TF, choose the terms appearing at least k times and calculate TF and WD of selected terms to form the feature space, respectively. Apparently, a larger k refers to a smaller number of terms chosen.

k	Feature Space Size (# of terms chosen)	Accuracy (TF)	Accuracy (WD)
3	6211	78.29%	80.71%
10	2846	78.03%	80.03%
20	1878	**80.59%**	**83.11%**
50	927	80.08%	82.79%
100	507	79.13%	82.54%
200	248	77.29%	81.46%
500	93	75.38%	78.57%

k means the terms which appear at least k times in the pool

Table 3.2: The Result of Term-appearance feature

Table 3.2 shows the evaluation results of different choices of TA feature spaces. We observe that generally the classifiers using word distribution (WD) outperform than ones using term frequency (TF). This is reasonable since Word distribution (WD) is the normalized version of Term frequency (TF) by the text length and WD even considers tuning the influence of stop words. Thus WD describes the text form more precisely.

When it comes to the range of terms chosen, as we expected, over-sized or under-sized feature space hurts the performance. Finally the Term-appearance (TA) feature space will reach the best performance when Word distribution (WD) is applied and terms appearing at

least 20 times (k=20) are chosen. This will become the standard of TA feature setting in the following experiments.

3.3.3 Experiment 2: Choosing Proper Combination of Feature Spaces

In this section, we explore three kinds of features: spaces: *Term appearance (TA), lexical (LE)* and *semantic (SE) feature*. How to utilize such features? Solely or combinedly? In this experiment we test the performance of classifiers using each kind of feature respectively, and different combinations among them. The results are still present by accuracy and compared with the baseline which utilizes Term appearance (TA) solely. Table 3.3 shows the performance of each possible feature selection:

Feature space	Feature Space Size	Accuracy	Change
Term appearance (TA)	1878	83.11%	/
Lexical (LE)	14	84%	+1.07%
Semantic (SE)	21	78.60%	-5.43%
TA + LE	1892	85.20%	+2.51%
TA + SE	1899	86.46%	+4.03%
LE + SE	35	87.86%	+5.71%
TA + LE + SE	1913	**90.15%**	**+8.47%**

Table 3.3: The Result of different feature selection

From the result we observed the following facts:

- Lexical (LE) and semantic (SE) feature can fairly capture the differences among the messages, *i.e.*, to distinguish each kind of message from other, even with a much smaller feature space size.

- To utilize the combination of feature space will improve the performance of using them solely. Using two features improve the baseline by 2.51% (TA + LE), 4.03% (TA + SE), and 5.71% (LE + SE), respectively.

- The whole combination of three kinds of features overwhelms any other selection. The accuracy reaches 90.15%, which improves the baseline a lot (8.47%).

3.3.4 Experiment 3: The Performance of Detecting User Comments (C)

Compared with News (N) which is difficult to parse, Spam (S) which is almost useless, User comment (C) reflects the most opinions from actual patients and doctors. Thus, user comments would become our target message category to be directly applied in the future. In this experiment, we will repeat what we did in Section 3.3.4 instead of changing the evaluation metrics to precision, recall, and F-score of user comments (C), defined as:

$$\text{Precision} = \#\text{ of messages correctly classified as C}/\#\text{ of user comments C} \tag{3.3}$$

$$\text{Recall} = \#\text{ of messages correctly classified as C}/\#\text{ of messages classified as C} \tag{3.4}$$

$$\text{F-score} = 2 \cdot \text{Precision} \cdot \text{Recall}/(\text{Precision} + \text{Recall}) \tag{3.5}$$

Feature space	Feature Space Size	Recall of C	Precision of C	F-score
Term appearance (TA)	1878	85.10%	84.75%	84.92%
Lexical (LE)	14	85.71%	86.05%	85.88%
Semantic (SE)	21	82.07%	82.85%	82.46%
TA + LE	1892	87.74%	86.98%	87.36%
TA + SE	1899	87.53%	87.75%	87.64%
LE + SE	35	90.58%	90.08%	90.33%
TA + LE + SE	1913	**93.31%**	**93.92%**	**93.61%**

Table 3.4: The Result of detecting user comments C

From Table 3.4 we will not only make the same observation as Section 3.3.4: the whole combination acquires best performance, but also notice that our feature selection can detect user comment (C) more precisely since the recall value (actually it is the accuracy of detecting

user C) is higher than the overall accuracy in every case. The best performance to detect user comment C is 93.61% by F-score while three feature spaces are fully combined. This fact makes sense since user comments (C) are more consistent with the feature spaces while news (N) and spam(S) are more diverse.

The high accuracy of our classifier ensures the quantity of the personal messages to be processed in our following problem: drug outcome integration, which will be discussed in Section 4.

3.4 Conclusion

In this chapter, I propose a multi-class SVM-based classification method by proper feature selection upon web personal healthcare messages where term-appearance, lexical and semantic features are applied. Experiments show that our approach could successfully distinguish healthcare messages in medicine forum as news, user comments, and spams, which (90.15%) outperforms the baseline case by 8.47% in accuracy. Future work includes improvement of the accuracy of the current result as well as further utilization of recognized news and comment messages.

CHAPTER 4

Designing and Evaluating a Clustering System for Integrating Patient Drug Outcomes

In this section, I propose a comprehensive system to organize and integrate patient outcomes utilizing semantic analysis, which groups large collections of personal comments into a series of topics. A prototype implementation was built to extract situational evidences by filtering and digesting user comments provided by patients. Our methods do not require extensive training or dictionaries, while categorizing comments based on expert opinions from standard source, or patient-specified categories. This system has been tested with sample health messages from our unique dataset from Yahoo! Groups, containing 12M personal messages from 27K public groups in Health and Wellness. We have performed an extensive evaluation of the clustering results with medical students. Evaluated results show high quality of labeled clustering, promising an effective automatic system for discovering patient outcomes from large volumes of health information.

As I introduced in Section 1, unlike product reviews, medical discussion messages are unstructured, *i.e.*, each comment talks about several topics in one piece of plain text. These messages must be partitioned into parts, and these parts must be grouped together according to what topic category they each belong to. By doing this we will have a coherent view on different aspects of the medical issue based on all the information available from our source. Our purpose is to re-construct and integrate a large number of unstructured online messages, into meaningful groups according to the topics, in order to aid users navigate through the vast information pool and satisfy their information need.

We designed a prototype model for clustering patient outcomes by effectively digesting large volumes of personal health messages. First, the useful *user comments* are retained, while *news* and *advertisements* which are noise for our purpose, are filtered out by an Support Vector Machine (SVM)-based classifier. In the main step, similar topics are grouped from sentences appearing in different messages by Probabilistic Latent Semantic Analysis (PLSA) topic model, where the topic categories can be guided by standard outcome descriptions from expert sources. In addition to identifying the sentences which are *similar* to(agree with) expert opinion into the corresponding topic, the model also clusters the sentences which are *opposite* to(disagree with) expert opinion into the same topic. In other words, for each outcome provided by experts, the system automatically identifies the sentences which provide positive support for the expert opinion and those provide negative support. The process organizes and integrates all the messages from online medical discussions in a practical way, relevant to particular persons in particular situations.

We have implemented a prototype interactive system for text mining of health messages. This system has been tested with sample messages from our unique dataset from Yahoo! Groups, which contains 12M personal messages from 27K public groups in Health and Wellness. This outcome research utilizes deeper processing of natural language, such as SVM and PLSA, than our previous studies on drug reactions with the same dataset [23, 24]. Our methods do not require extensive training nor dictionaries. In addition, they allow users to specify their own topics for digesting. Therefore, our methods provide general and powerful solutions to mine health messages.

We have evaluated the prototype system with a sample set of drugs using a sample cohort of medical students. 5000 sentences relevant to 10 representative drugs were randomly selected, and automatically clustered into topics extracted from PubMed Health database [1], a well known expert source for drug information. The accuracy of these clustering results was eval-

[1] http://www.nlm.nih.gov/pubmedhealth/

uated by medical students in the College of Medicine at the University of Illinois in Urbana. By comparing the automatically generated clustering results to the ones generated by these professional annotators, it is shown that our topic clustering methods produce highly accurate results. We also statistically prove that all judges were consistent in classifying the sentences and thus have produced a valid gold standard for our evaluation.

Guided by the standard expert opinions extracted from PubMed Health, our topic clustering provides robust automatic classification of patient-reported drug outcomes. That is, our system can automatically classify patient outcomes, which describe patients' experience and result of using a particular drug, often using layman language, into standard categories derived from PubMed Health, with high accuracy. Drugs used for our evaluation can be divided into two classes: *specialized* and *generalized*. The first class treats a particular medical condition (*e.g.,* Metformin), while the second class includes over-the-counter drugs (*e.g.,* Ibuprofen) and commonly-prescribed-drugs (*e.g.,* Heparin). The results show the accuracy of clustering specialized drugs is higher than that of generalized drugs. This is reasonable since specialized drugs often have a focused range of treatments and side effects, which makes patients' outcome description more specific and consistent. In addition, we also observe that the clustering methods work better for more common drugs, possibly because users are likely to be more knowledgeable about drugs they encounter often.

We also show that our system can explore outcomes not included in the standard expert source. In this particular experiment, we have computed an additional cluster that groups together sentences not closely associated with any of the standard clusters. By examining this additional cluster, we discover some patient comments concerning serious side-effects or other treatments, but not discussed in the standard outcome description on PubMed Health. By referring to the medical literature, we are able to confirm many of these patient-provided outcomes have been recorded as possible results of using the particular drug. Patient-reported

outcomes can be an important supplementary source of information, even when automatically extracted from health messages.

The section of work is complete and is published in American Medical Informatics Association Annual Symposium 2012, which is the the worlds premier scientific meeting for biomedical and health informatics [75].

Organization. Section 4.1 formally describes the problem definition, followed by the architecture of the system in Section 4.2. I will introduce the solution of each step in Section 4.3, including (1). Pre-processing: Filtering the messages, (2). Clustering: Outcome selection and integration and (3). Post-processing: Separating similar and opposite opinions, and then present the evaluation setting, experiment and results on Yahoo! Groups data in Section 4.4. Finally, I will make conclusion and discuss the future work in Section 4.5.

4.1 Problem Definition

For one particular drug, we collect all the related health messages from Yahoo! Groups, denoted as M, in which N is the set of all the *news*, C is the set of all the *user comments* (our target), and S is the set of all the *spam* such as advertisements. Clearly, we have $M = N \cup C \cup S$.

After successfully extracting C, we split it into a set of meaningful sentences, denoted as D. Each sentence $d \in D$ is called a **comment unit**, which would potentially present one side of outcomes. Our target goal is to group all the comments into m meaningful **outcome clusters** $O_1, O_2, ...O_m$, given the collection D.

Here are several key concepts to be introduced:

- **Expert comment** e_i: To better cluster the outcomes, semi-supervised PLSA model [29] is applied. Expert comments aim to offer the prior knowledge for PLSA and guide the topic of each O_i. For each O_i we have one expert comment e_i. Compared with the user comments, the expert comments are more well-written, professional and semantically vertical to each other. We collect the set of expert comments E (formed by $e_1, e_2, ...e_{m-1}$) for each drug, from the PubMed Health database of U.S. National Library of Medicine[2]. The reason why we choose $m-1$ expert comments is that we want to create some groups of opinions with prior expert knowledge ($O_1, O_2, ...O_{m-1}$) as well as another group of opinions whose topics are beyond the expert's (O_m).

- **Similar opinion** O_{i_sim} and **Opposite opinion** O_{i_opp}: Each outcome O_i ($1 \leq i \leq m-1$) consists of a group of comment units D_i and is associated with one expert comment e_i. Some of the comments represent similar or relevant opinion with e_i, which form O_{i_sim}. Others reflect different or opposite opinions from e_i, though they still talk about the same topic. We call such collection O_{i_opp}.

[2]http://www.nlm.nih.gov/pubmedhealth/

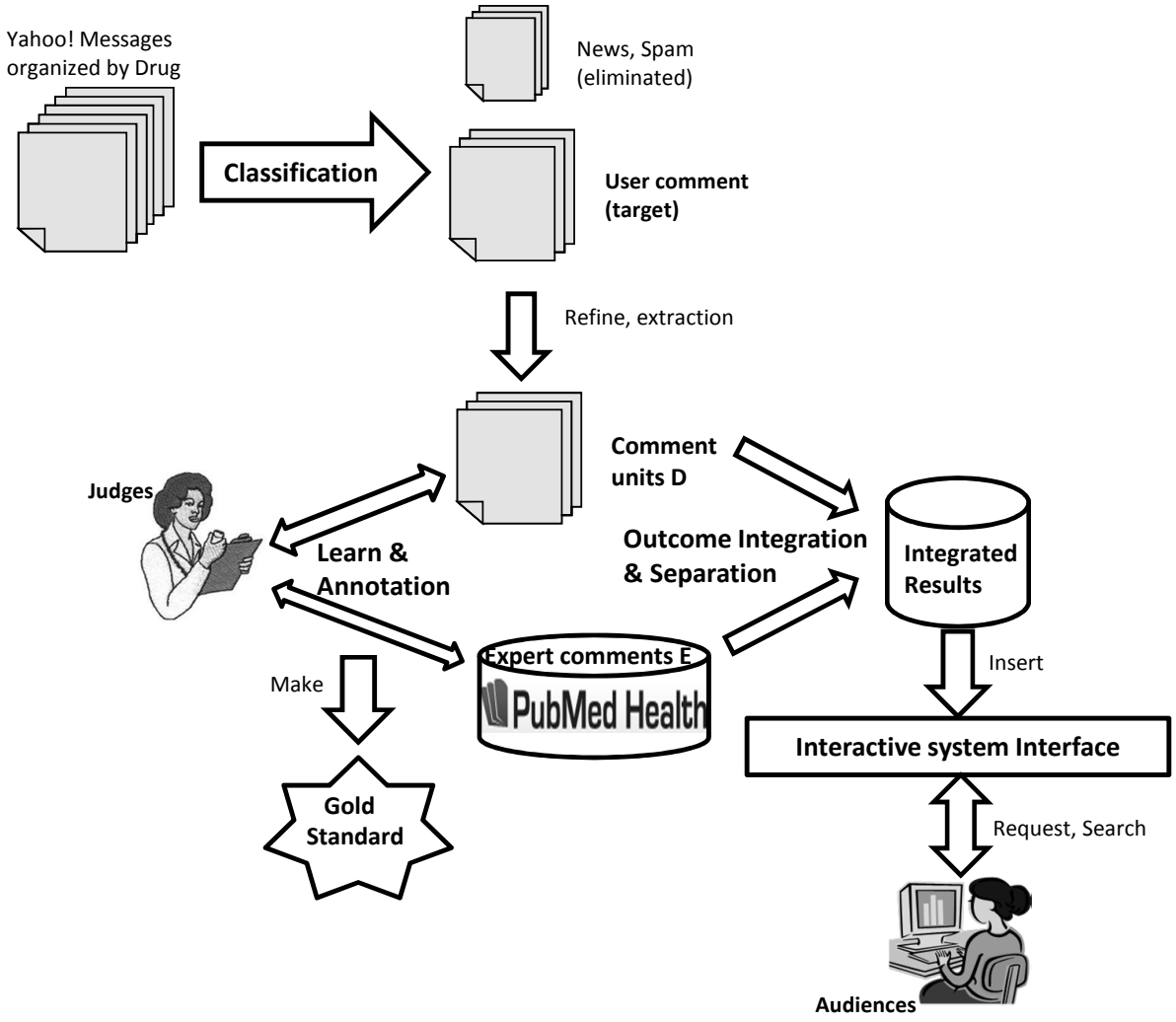


Figure 4.1: The architecture of the integration system

4.2 System Architecture

Figure 4.1 illustrates the whole design and implementation process of our system.

Messages from Yahoo! Groups are firstly organized by drug and then *classified* into three categories: N, C, S. N and S are eliminated while the the collection of *comment units (D)* is extracted from *user comments (C)*, together with the *expert comments (E)* as input. On the one hand, such comment units (D) will be *re-organized and integrated* into several meaningful outcomes $O_1, O_2, ...O_m$ by our topic model, and presented to audiences via our interactive

system interface. On the other hand, judges will *annotate* the extracted data (D and E). Their annotation results will become a gold standard to evaluate the performance of our model. In Section 4.3 we will introduce the key components of our system in details.

4.3 Methodology

As we introduced before, there are several steps to complete the whole task.

- **Data Pre-processing** (step 1): The input is the whole collection M. We will separate it into three categories: News (N), Comment (N) and Spam (S). C is our target while N, T will be filtered out. C is then split into a set of comment units D.

- **Data Clustering** (step 2): Given D and $m-1$ expert comments $e_1, e_2, ...e_{m-1}$, we will generate m outcome clusters $O_1, O_2, ..., O_m$. Each cluster O_i refers to one meaningful drug outcome, either guided by expert opinion e_i ($1 \leq i \leq m-1$), or contributing to "other opinions" ($i = m$).

- **Data Post-processing** (step 3): For each cluster with prior expert knowledge O_i, we will split it into O_{i_sim} – expressing the similar opinion to e_i, as well as O_{i_opp}, which shows the opposite opinion.

4.3.1 Pre-processing: Filtering the Messages

To distinguish the messages M, we applied the multi-class SVM-based classifier we have trained and implemented in Section 3, The evaluation result shows that we can precisely acquire the user comments (C) for each drug.

4.3.2 Clustering: Outcome Selection and Integration

To achieve our core step of the system: grouping the comments into reasonable and discriminative clusters, where each cluster represent one main outcome of the drug, semi-supervised PLSA model [29] is applied. We would introduce the model first and then describe the integration process.

PLSA model

In PLSA model, we consider each comment unit $d \in D$ is generated from a mixture of $m+1$ multinomial component models. One component model is the background model θ_B which dismisses the affect of non-discriminative (*i.e.*, stopwords) words and the rest are m latent theme topic models (saying $\Theta = \{\theta_1, \theta_2, ...\theta_m\}$), each of which captures one topic. Each comment unit (*i.e.* sentence) d can then be regarded as a sample of the following mixture model:

$$p_d(w) = \lambda_B p(w|\theta_B) + (1-\lambda_B)\sum_{j=1}^{m}[\pi_{d,j} p(w|\theta_j)] \tag{4.1}$$

Here $\pi_{d,j}$ refers to a document-specific mixing weight for the j-th aspect. w is a word. By applying Expectation Maximization (EM) algorithm [76], a method for finding maximum likelihood estimates of parameters in statistical models, all the parameters and results can be computed and updated using the following formulas:

$$p(z_d, w, j) = \frac{(1-\lambda_B)\pi_{d,j}^{(n)} p^{(n)}(w|\theta_j)}{\lambda_B p(w|\theta_B) + (1-\lambda_B)\sum_{j'=1}^{k}\pi_{d,j'}^{(n)} p^{(n)}(w|\theta_{j'})} \tag{4.2}$$

$$p(z_d, w, B) = \frac{\lambda_B p(w|\theta_B)}{\lambda_B p(w|\theta_B) + (1-\lambda_B)\sum_{j'=1}^{k}\pi_{d,j'}^{(n)} p^{(n)}(w|\theta_{j'})} \tag{4.3}$$

$$\pi_{d,j}^{(n+1)} = \frac{\sum_{w\in V} c(w,d) p(z_{d,w,j})}{\sum_{j'}\sum_{w\in V} c(w,d) p(z_{d,w,j'})} \tag{4.4}$$

$$p^{(n+1)}(w|\theta_j) = \frac{\sum_{d\in D} c(w,d) p(z_{d,w,j})}{\sum_{w'\in V}\sum_{d\in D} c(w',d) p(z_{d,w',j})} \tag{4.5}$$

To better cluster m outcomes, we can enroll some *prior knowledge* by extending the basic PLSA based on expert comments $e_1, e_2, ..., e_{m-1}$ (*i.e.,* semi-supervised). For each outcome cluster O_i ($1 \leq i \leq m-1$), since we have already acquired the expert comment e_i, we can build a unigram language model $\{p(w|e_i)\}$ and incorporate it into the above formulas,

Formula 4.5 will turn into Formula 4.6 as below:

$$p^{(n+1)}(w|\theta_j) = \frac{\sum_{d \in D} c(w,d) p(z_{d,w,j}) + \mu p(w|e_j)}{\sum_{w' \in V} \sum_{d' \in D} c(w',d') p(z_{d',w',j}) + \mu} \tag{4.6}$$

Note, for cluster O_m, there is no prior knowledge since we expect to discover some additional opinions rather than the experts'.

Integration Progress

To build m meaningful clusters by applying semi-supervised PLSA model, there are several steps described below:

- Build the prior knowledge. For each cluster O_i ($1 \leq i \leq m-1$), we have already acquired an expert comment e_i from PubMed Health database of U.S. National Library of Medicine. Based on it, we estimate $\{p(w|e_i)\}$ by Maximum Likelihood as the prior estimator. Here only adjectives, adverbs, verbs and nouns are considered in the estimator since they are the terms which express the opinions.

- Given such prior knowledge and the set of the comment units D, we could estimate the topic models $\{\theta_1, \theta_2, ...\theta_m\}$ by the formulas above.

- For each comment unit $d \in D$, we assign it to the most suitable cluster by the following formula:

$$\underset{j}{argmax}\, p(d|\theta_j) = \underset{j}{argmax} \sum_{w \in V} c(w,d) p(w|\theta_j) \tag{4.7}$$

- For each opinion O_i, we generate a topic model θ_j as well as a bunch of corresponding comment units D_i. the terms which has high probability $p(w|\theta_j)$ in θ_j as well as featured comment unites can represent such topic.

We apply the above process on the real data and will show the result in Section 4.4.

4.3.3 Post-processing: Separating Similar and Opposite Opinions

Now in each cluster O_i, there are a couple of assigned comment units D_i. For the cluster which has a prior expert comment, we will split it into Similar opinion O_{i_sim} and Opposite opinion O_{i_opp}, by applying semi-supervised PLSA model (creating two clusters with e_i as one O_{i_sim}'s prior knowledge while O_{i_opp} has no prior knowledge).

To create such two clusters, the straightforward approach is to build one cluster's prior estimator $\{p(w|e_i)\}$ by the typical way: *all the adjectives, adverbs, verbs and nouns* are considered in the estimator since they are the terms which express the opinions, and leave another cluster's prior as empty.

This approach has one limitation that it does not address the sentiment meaning. Look at the following two sentences: "I took some Aspirin to treat the back pain and it works well". "I took some Aspirin to treat the back pain, but bad effect." They have nearly the same vocabularies to address the same topic but with opposite opinions. For such sentences with similar lexical structure, a better way to distinguish them is to detect the sentiment terms. *i.e.,* "well" and "bad", which are key points to express the *positive* opinion and *negative* opinion, respectively. Start with this observation, we propose another approach to build the prior estimator $\{p(w|e_i)\}$ where *only positive/negative terms* in e_i are considered. Since all the sentences in O_i have already been considered to have the same topic with e_i. It is more appropriate to focus on the sentiments while splitting O_i into O_{i_sim} and O_{i_opp}.

In Section 4.4, we will implement each approach respectively and compare their performances.

4.4 Experiments and Results

4.4.1 Data and Setup

We utilize our unique dataset which is segmented from Yahoo! Groups with Health and Wellness data. The dataset consists of $27,290$ public groups and $12,519,807$ messages in total, crossing seven years and covering multiple topics. All the experiments run on a 4TB-disk, 4GB-RAM, and 10-core server.

We have trained an SVM-based tri-class classifier with an RBF kernel on the real data and tested it. Evaluation results [67] show that our classifier can achieve 90.15% overall accuracy as well as 93.61% by F-score of detecting user comments (C), which indicates that our approach could successfully distinguish messages' categories, especially user comments.

Since we are targeting personal medical information, we choose to design the system with outcomes of specific drugs. From the data pool we select 10 drugs and each appears more than 1000 times. Their relatively high frequency of appearance may ensure that sufficient personal messages can be processed. Half of the drugs are *Prescription - Variety Medical Conditions* **(Prescript-VMC)** drugs: Metformin, Clonidine, Gabapentin, Clonazepam and Oxaliplatin, and five are *Pain Relief (may be OTC) or Anti-coagulation* **(PRAC)** medications: Aspirin, Heparin, Ibuprofen, Hydrocodone and Naproxen. For each drug, we collect all the messages containing it, process them by our classifier and get the user comments C, split C and collect sentences D which either contain the drug name, or are next to the sentences containing the drug. These sentences D potentially represent users' diverse opinions on the drug.

We were aware that a more straightforward approach is to compare Prescription drugs with out-the-counter (OTC) drugs. However, OTC drugs tend to contain more noisy data and we noticed that, after pre-processing, many OTC drugs simply do not have enough information on this forum of proper length and diversity for our evaluation purpose. In this case, we set two groups as Prescript-VMC drugs with different *specific* treatments, and PRAC with more

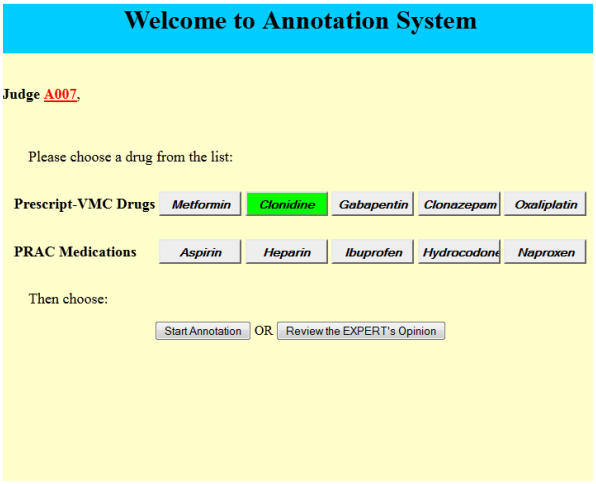
general treatments, such as some OTC drugs (Aspirin, Naproxen, Ibuprofen) as well as drugs with similar treatments (Hydrocodone and Heparin).

Due to the uniqueness of our data source – Yahoo! Groups, it is difficult to apply the traditional evaluation approaches by comparing to a gold standard like TREC medical informatics [77]. Thus, a comprehensive evaluation system which contains a large-scale of professional-labeled sentences as our gold standard should be built and applied.

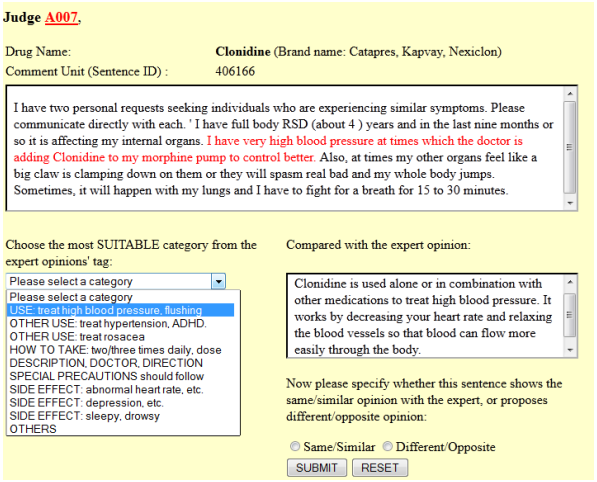
4.4.2 Annotation Framework

We have built an interactive web-based database system to support the evaluation process. 500 comment units (*i.e.,* sentences) are randomly generated for *each* drug and stored in the database (Note, for some drug like Naproxen, the total number of available comment units is slightly more than 500). For a specific drug, the professional judge needs to get familiar with the pre-defined expert comments (8-10 per drug), each of which is associated with a given tag, and then enter the actual annotation. Each time one comment unit d is given together with its context in the actual personal message. The judge needs to understand the meaning of d and assign it to the most suitable cluster (recognized by the tag of the corresponding expert opinion) it belongs to, or to "other" cluster if no prior expert opinion matches. After that, the judge is also asked to determine whether d shows similar or opposite opinion with the chosen expert's. The annotation of one comment unit is then finished and the result will be stored in the database. Figure 4.2 simplifies the interface for the annotation process.

Similar to Blake's work [78], we design a two-step annotation process. The purpose of the first step, a *pilot study*, is to validate the design of annotation process, including the instruction and defined categories, is easy to understand and unambiguous for human annotators. If the pilot study shows there is no large variance of understanding about the annotation process among the annotators, we will proceed to the *main study* to complete the actual annotation.



(a) Choose the drug



(b) Label a sentence

Figure 4.2: Annotation Interface

Pilot study: In this experiment, three graduate students (majors are computer science, nutrition and bioinformatics) and one medical school student were enrolled. We assigned each of them 100 identical sentences (50 for Prescript-VMC and 50 for PRAC, randomly generated, covering all common clusters) and they labeled the sentences independently following the instructions. Then we collect their annotations and test the inter-rater agreement by using Fleiss' kappa [79]. Fleiss' kappa is widely used for assessing the reliability of agreement between a

fixed number of judges when assigning categorical ratings to a number of items or classifying items.

From computation, the four judges' kappa-value reaches 0.84. According to the interpretation of the kappa statistic [80], this result shows almost-perfect agreement among the judges, which proves that our annotation process is designed with minimum ambiguousness.

Main study: To reach our initial goal, a gold standard should be made by professional annotators upon all the sentences. In the main study, we invited 10 medical school students who are well-trained, experienced, and familiar with information about drugs and treatments thus qualified as our standard maker. For time and quality concern, it is impossible to ask one judge to label all the sentences so we randomly split the whole annotation task to 10 judges. To test the inter-judge reliability, we repeated the process of pilot study – assigning them 50 identical sentences, which are randomly selected from the sentence pool and cover all common clusters. The Fleiss' kappa value reaches 0.81, which is considered almost perfect agreement.

The above results indicate that there is no significant variance among all the annotators, and prove that, by comprehending the task instruction, our judges are well-trained enough to provide generally consistent gold standard. Therefore, we could confidently apply the gold standard to evaluate our clustering results. The whole annotation process lasted half a month and a random follow-up check was executed by two other medical school students afterward. The output of such well-designed and professional-enrolled annotation framework is good enough to become not only the gold standard of our integrated system, but also a useful resource for further research.

4.4.3 Clustering Results and Analysis

Now we have the set of comment units D for each drug, as well as the expert comments E extracted from PubMed Health database of U.S. National Library of Medicine. The number

of expert comments for each drug is based on the content of the description in the database. Thus it varies from 8 to 10 for different drugs. According to the semi-supervised PLSA model introduced before, we assign each comment unit to the suitable cluster with prior expert knowledge or "other outcome" cluster without prior knowledge and compare the results to the gold standard data.

To measure the quality of our clustering results, we utilize the following measurements: accuracy, precision, recall and F-score, which are defined by:

$$\text{Accuracy} = \#\text{ of correctly clustered sentences}/\#\text{ of total sentences} \tag{4.8}$$

$$\text{Precision} = \#\text{ of correctly clustered expert sentences}/\#\text{ of total expert sentences retrieved} \tag{4.9}$$

$$\text{Recall} = \#\text{ of correctly clustered expert sentences}/\#\text{ of total "real" expert sentences} \tag{4.10}$$

$$\text{F-score} = 2 \cdot \text{Precision} \cdot \text{Recall}/(\text{Precision} + \text{Recall}) \tag{4.11}$$

In Formula 4.9 and 4.10, "expert sentence" means the sentence which is assigned into a cluster associated with one expect opinion ($O_1, ...O_{m-1}$), regardless of annotated or auto-retrieved, and "real expert sentences" means expect sentences in gold standard. In other words, the measurements precision, recall and F-score ignore the potential affect by "other opinion" cluster (O_m). Note, our measurements only evaluate the correctness of clusters, not the correctness of the sentences though "incorrect" user comments do exist.

Drug	Accuracy	Precision	Recall	F-score
Metformin	0.675	0.706	0.661	0.683
Clonidine	0.705	0.783	0.696	0.737
Gabapentin	0.665	0.669	0.663	0.680
Clonazepam	0.770	0.766	0.740	0.753
Oxaliplatin	0.655	0.709	0.653	0.680
Prescript-VMC(standard deviation)	**0.694**(0.05)	**0.726**(0.05)	**0.683**(0.04)	**0.707**(0.04)
Aspirin	0.725	0.768	0.708	0.737
Heparin	0.580	0.635	0.563	0.597
Ibuprofen	0.600	0.634	0.586	0.609
Hydrocodone	0.620	0.665	0.616	0.640
Naproxen	0.575	0.591	0.569	0.580
PRAC(standard deviation)	**0.620**(0.06)	**0.659**(0.06)	**0.608**(0.05)	**0.632**(0.06)
Overall	**0.657**	**0.693**	**0.646**	**0.670**

Table 4.1: The performance of the clustering result for all the drugs

Table 4.1 shows the performance of the clustering result by each drug, each category (prescript-VMC or PRAC) and overall. we can observe the following facts:

- Our semi-supervised PLSA model can achieve a relatively high performance of clustering. *i.e.,* the overall accuracy is 0.657 and the overall F-score is 0.670, considering the large number of clusters (9 to 11 per drug).

- Our result also shows that F-score is higher than the corresponding accuracy in all cases. The paired t-test [81] between F-score and accuracy for the 10 drugs is also significant

(p-value < 0.05). It indicates that cluster with prior knowledge could effectively improve the performance than that of no prior knowledge.

- Compared with PRAC medications, Prescript-VMC drugs perform better in all cases. *i.e.*, accuracy: 0.694 *v.s.* 0.620, F-score: 0.707 *v.s.* 0.632, *etc.* We also conduct t-tests between PRAC and Prescript-VMC for each of the measurements. It shows, compared to PRAC, the recall and F-score of Prescript-VMC are significantly higher(p-value < 0.05), and the accuracy(p-value $= 0.06$) and recall(p-value $= 0.10$) are marginally significantly higher. The results further confirm our conclusion that the clustering on Prescript-VMC outperforms that of PRAC. This makes sense since people can relatively easily describe the outcome of Prescript-VMC drugs since they may have more *specific* treatments, more *strict* usage and *easier-described* side effects.

- Some interesting phenomena: among all the Prescript-VMC drugs, Oxaliplatin, a cancer chemotherapy drug, is probably the most uncommon one since patients are not likely to know it unless they are facing colorectal cancer. In contrast, among all the PRAC, Aspirin is the most popular one since it is a well-known pain-relief that most people have encountered or heard of. Our results show that the performance of Oxaliplatin is the *worst* among all the Prescript-VMC while Aspirin is the *best* among all the PRAC. It reveals that: the more people get familiar with a drug, the more accurate that people can describe its outcomes, thus the better the model achieves the performance.

Also for all the correctly clustered "expert sentences", we split each group O_i into two sub-groups O_{i_sim} – showing the similar opinion with the expert's e_i, and O_{i_opp} – showing the opposite opinion. Two different strategies to form the prior estimator $\{p(w|e_i)\}$ are applied and compared. Table 4.2 shows the accuracy of two approaches compared to the gold standard. Approach 1 refers to that all meaningful terms are considered while Approach 2 refers to that

only sentiment terms are considered. We utilize the technique and open source introduced in Hu and Liu's sentiment analysis work [82].

Accuracy	Approach 1	Approach 2	Change
Prescript-VMC	0.820	0.831	+1.4%
PRAC	0.792	0.811	+2.4%
Overall	0.806	0.821	+1.9%

Table 4.2: The performance of distinguishing O_{i_sim} and O_{i_opp}

From Table 4.2 we observe that both of the approaches reach a high accuracy to determine the similar or opposite opinion (overall accuracies are above 0.80), which indicates that semi-supervised PLSA model can successfully solve such problem. Furthermore, compared to the traditional way to build estimator (Approach 1), the novel way where sentiment analysis is highly addressed (Approach 2) performs better across Prescript-VMC, PRAC, and overall case.

4.4.4 Case Study

Interface

Eventually, our prototype system is able to provide clustered and integrated drug information from personal health messages to enable examination through a user-friendly interface [3], whose composite format for the entire session is shown in Figure 4.3. Users can choose the specific drug and outcome that they are interested in, and the system will respond to the request by displaying the topic word distribution, the corresponding expert comment and all the comment units which belong to this outcome. Each comment unit is labeled by "similar" or "opposite" and users can also click the link of the corresponding PID to read its complete context.

[3]Currently the system only works on the 10 drugs used for the evaluation task

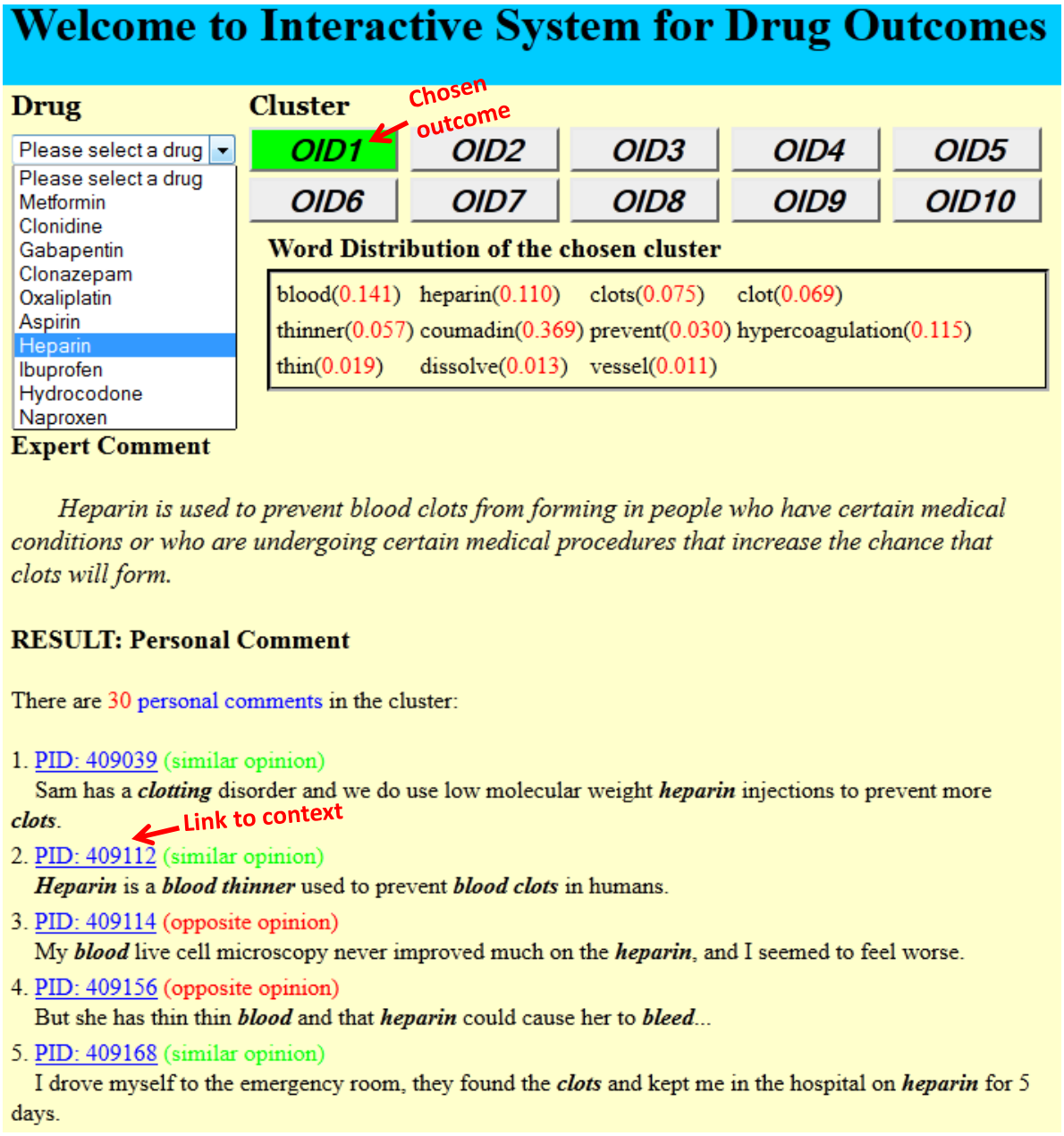


Figure 4.3: Systerm User interface

Example of Clonazepam

Table 4.4 shows the outcome integration results with expert comments, for the drug Clonazepam, which achieves the best performance out of 10 drugs. (accuracy: 0.770, F-score: 0.753). In the table, "Topic model" column shows the most common terms in this outcome as well as its probability. From this column we expect users could easily conceptualize the particular cluster at a glance. The third and fourth column show the number of "Similar Opinions" and "Opposite Opinions" for the corresponding outcome (denoted by s), respectively, as well as one sample personal sentence (for the space limit).

From Table 4.4, we build 8 clusters for Clonazepam, where each of them focuses on one meaningful semantic outcome, guided by an expert comment. For example, Outcome ID 1 talks about the main treatment of Clonazepam – to treat seizures, which we know from the topic model and the expert comment e_1. 59 comments express the *similar/same* opinion, while other 16 show *different/opposite* opinions on the same topic. From each row, we could easily understand the general user experience and how common this experience are among users regarding the particular outcome while taking Clonazepam: how do they feel about this drug? Do they agree or disagree with the expert's? Similarly, the rest outcomes such as side effects, dosage are shown in the following rows.

Such information is scattered in the huge amount of messages and impossible to integrate by hand without our system. Reading such information, audiences may find our system can effectively collect, organize and integrate the drug-based information and display it in a well-readable way.

New discovery of "other outcome"

Note for each drug, we also generate an additional cluster O_m, *i.e.,* "other outcome" which includes information mentioned by online users but not in standard description from expert comments. Table 4.3 shows some sentences (31 sentences in total, of which 8 relate to mouth burning) in O_m for Clonazepam.

He said klonopin would help the burning sensation so I tried it and it did.
The Klonopin is the only drug that helped me with the burning taste.
We reintroduced the klonopin in a dropper full of water
sublingually and eventually stabilized.
...

Table 4.3: Sample results of O_m for Clonazepam

We examine this cluster for each drug and find some interesting opinions. *e.g.,* 34 comments report that Metformin is also used to treat obesity. For Clonazepam, 8 sentences show that Clonazepam may help to relief stomatodynia (burning mouth syndrome). For Aspirin, 5 sentences say that taking Aspirin causes eye problem, such as bursting eye vessels. For Heparin, 6 sentences show the concern that Heparin may cause severe bleeding to death. *etc.*

Although such "other outcomes" are not mentioned in expert comments, *i.e.,* not recorded as standard outcome of the drug, they are discussed by actual users. In fact, formal medical literatures have mentioned each of the above additional outcomes – Metformin [83], Clonazepam [84], Aspirin [85] and Heparin [86]. Therefore, our system can effectively *discover such "new outcome" from the clinical experiences as reported directly by the patients*, which will provide supplemental information for the drug's standard description.

4.4.5 Advantages and Limitations of Model

One advantage of our system over other simple statistic methods relies on its capacity of capturing the coherence of terms (*e.g.*, appositive, synonym). The PLSA model is able to detect such connection between two comments which contain relevant, but not identical information since they share the similar contexts. Take one of the Clonazepam's outcomes (OID 4) as an example, the expert comment is "Follow ... and ask your *doctor* or pharmacist to explain any

part you do not understand." The 18 similar sentences include not only "I would like to discuss with my *Doctor*..." which capture the word "doctor" exactly, but also "I first tried Klonapam prescribed by *Dr.* Cheney", and "you need to be monitored by a *physician*", which capture the word "dr", "physician".

OID	Topic model	Expert comment	Similar Opinions	Opposite Opinions
1	seizures(0.10) panic(0.09), attacks(0.07) seizure(0.06), brain(0.05) activity(0.02)	Clonazepam is used alone or in combination with other medications to control certain type of seizures. It is also used to relieve panic attacks and works by decreasing abnormal electrical activity in the brain.	[s=59] She has only had a handful of seizures since then, Klonopin seems to control her seizures well	[s=16] Shy, Klonopin did not seem to contribute to my brain fog.
2	disorder(0.05), restless(0.04), plmd(0.03) dystonia(0.03) movement(0.02) mental(0.02)	Clonazepam is also used to treat symptoms of akathisia that may occur as a side effect of treatment with anti-psychotic medications and to treat PLMD , dystonia, and acute catatonic reactions)	[s=26] Klonopin works really well for Periodic Limb Movement Disorder, or any other med in the benzo class.	[s=12] Tried clonazepam for stress induced issues but it was too strong for me
3	times(0.09), mg(0.08) daily(0.07), three(0.06) bedtime(0.04) tablet(0.03)	Clonazepam comes as a tablet to take by mouth. It usually is taken one to three times a day with or without food. Take clonazepam at around the same time(s) every day.	[s=33] Zach has been on Klonopin .5mg three times a day for years	[s=9] Please do not take more than 8 Klonopin tablets a month
4	doctor(0.03), ask(0.02) prescription(0.02) dr(0.02), explain(0.01) pharmacist(0.01)	Follow the directions on your prescription label carefully, and ask your doctor or pharmacist to explain any part you do not understand.	[s=18] Patricia, Klonopin is the best medication to take , but you need to be monitored by a physician.	[s=6] I know they can't prescribe the klonopin but a recommendation would be helpful.
5	allergic(0.07) pregnant(0.06) allergic(0.06), myoclonus(0.05) pregnancy(0.02)	Before taking clonazepam, tell your doctor if you are allergic to clonazepam, tell your doctor if you are pregnant.	[s=24] Most anti-seizure meds aren't allowed to be taken while you are pregnant,like Klonopin	[s=20] Sydnie has never had any allergy from the klonopin so we have been pretty pleased with it!
6	anxiety(0.21), anti(0.04) depression(0.03) mood(0.02) emotional(0.02) suicide(0.02)	Report any new or worsening symptoms such as mood or behavior changes, or if you feel agitated, irritable, hostile, aggressive, or have thoughts about suicide or hurting yourself.	[s=55] I have been worried about taking the Klonopin for the anxiety and sleeplessness because I have this history of depression.	[s=26] I just started taking Klonopin a couple of months ago for my anxiety.
7	tired(0.05), redness(0.04) rash(0.03), eye(0.02) breathing(0.02) liver(0.02)	Call your doctor if you have a serious side effect such as: tiredness, shallow breathing; unusual eye movements; stomach problem, liver or kidney problem, redness, abnormal weight	[s=23] The klonopin just made me tired and that kind of made me feel more out of control.	[s=8] Only on the left side, and it is more like a rash than a redness
8	addiction(0.14) abuse(0.13) addictive(0.07) highly(0.02) pregnancy(0.02)	Clonazepam may cause someone drug abuse or addiction.	[s=31] Unfortunately, Klonopin is a very addictive drug.	[s=9] Klonopin is a little addictive, but does help when you need it

Table 4.4: The outcome results for Clonazepam with expert comments

Another advantage of the model is the flexibility of setting expert comments. Users can follow our example, *i.e.*, extracting E from a professional drug database, or define and input the expert comments by themselves as prior knowledge of PLSA model. The clustering results will vary according to different expert comments. This customizable design could cater to users' various information need.

PLSA model requires to manually set a *fixed* number of clusters and lacks a way to *dynamically* determine the proper number of clusters. We will try to solve the problem in the future work. Furthermore, our model can analyze each independent drug effectively by embedding its *specific* expert comments. However, a *unified* expert-comment setting strategy should be designed and implemented while extending our system to the universal drugs or even treatments.

To sum up, our outcomes system is accurate for clustering standard outcomes and effective for discovering novel outcomes, while is fully automatic with text processing. Thus by using this interactive system as the core engine,we believe a national wide surveillance system for drug-related outcomes is highly feasible.

4.5 Conclusion and Future Work

In this chapter, I describe a useful system we built to cluster and integrate drug-based medical information. PLSA model and sentiment analysis techniques are applied in the system. We design a large-scale and professional-quality annotation framework, the output of which is good enough to be the gold standard to test the performance of the model. The experiment results with high accuracy and F-score show that our system can successfully organize the online medical information in a meaningful way. Users can request and search the well-organized personal opinions on different types of drugs via our prototype system, which could satisfy not only physicians', but also patients' information need.

In the future, we plan to explore and analyze different online healthcare resources, such as twitter[4] – a more general social network where people discuss their health problem in a casual way, or MedHelp[5] – a more professional medical forum providing well documented user demographic information (the utilization of MedHelp is in Section 5). We also expect to compare the up-to-date contents and language features across different online medical discussion platforms.

From a system perspective, we will design and implement a flexible expert-comment setting strategy which offers the choice of unified standard from expert resource or patient-oriented prior knowledge. Building upon the clustering engine described in this paper, a wide range of drugs and treatments can be automatically analyzed from patient-specified personal health messages, and an effective interactive system can be built upon their integrated results.

[4]http://www.twitter.com
[5]http://www.medhelp.org/forums/list/

CHAPTER 5

Comparative Effectiveness Research(CER) Hypothesis Prediction in Personal Health Messages

In this section, I will introduce our latest project – conducting comparative effectiveness research (CER) hypothesis by analyzing personal health messages.

According to Institute of Medicine, comparative effectiveness research (CER) is defined as the generation and synthesis of evidence that compares the benefits and harms of alternative methods to prevent, diagnose, treat, and monitor a clinical condition or to improve the delivery of care [1]. By answering the core question of "which treatment works best for whom and under what circumstances", the research results could assist consumers, care providers, researchers and policy makers to make informed health care decisions. In the year 2009, US$ 1.1 billion of government funding was earmarked for CER research. Since then, increasing research effort and organizational leaderships have emerged national-wide to push forward this research field.

Similar to other medical research, CER research embraces the methodology of clinical trial. Difficulties, however, are posed by the nature of CER research, as it focuses on the comparisons "for whom" and "under what circumstances". Given the numerous demographical and conditional variables and their potential combinations, it is simply impossible to exhaustively include all of them in the comparative testing in expectation of opportunistic results. However, if the researchers could possibly identify variables with highest potential to see a difference, they may prioritize them by selectively including, thus largely decreasing the number of variables in the clinical trial. In this paper, we argue that mining opinions about alternative treatments from large repository of online health messages posted by real users could be a promising way to serve that purpose.

The popularity of Web 2.0 technology has made internet a common platform for people to express, share and discuss their opinions. As a result, the Internet provides rich resource for researchers and analysts to mining public opinions, which, in many fields, has proved itself to be an equivalent, or even preferred approach to other time and labor consuming methods such as polling or longitude data logging.

Health related topic is one of the most discussed topics on the internet. There exist many social media services dedicated to health related discussions. Also many health information websites include components that allows user generated contents such as user review system for medication website. Frequently people share their first-hand experience with various treatments or offer suggestions to other patients who seek for opinions. By aggregating these individual messages regarding specific treatment, one could potentially get rich data from a sample that is potentially well distributed in the population, given that nowadays Internet is frequently used by people of various age, gender, race, income, geographical location, etc. Therefore, it is possible to compare multiple treatments by comparing the collective opinions on each of them. The results could supplement the conclusions of formal CER research, or serve as basis for forming hypothesis for clinical research.

The great quantity of health messages available on the internet enables various forms of medical analysis, which can be extremely time consuming or even impossible with traditional data collecting methods. For example, Yahoo Groups! contains more than 100 million personal messages spanning over 20 years that are relevant to health and medical domain. MedHelp contains over 10 million messages covering more than 100 medical topics, *etc*. Compared with electronic medical record (EMR), though possibly less accurate, the scale of online health messages makes it possible to carry on population based medical-related research, such as CER prediction.

Furthermore, online health messages enable data collection by individual case, which could provide contextual information about the particular patient, including his or her demographical

information, social information and health profile. These kinds of information are valuable for informing medical research, especially CER research, for which individual difference is the core research focus. This individual contextual information could be derived either explicitly, by directly extracting from user filled profile, or implicitly, by tracking historical user generated content or behavioral data. In this paper we focus on the former, although the same technique could be applied as long as individual specific contextual information is identifiable.

Nowadays, It is a common practice that users fill out a profile page on social media website, which often specifies the basic demographical information. Utilizing this information enables us to aggregate opinions by demographical variables, which seems to be especially relevant and useful for informing CER research. For example, if one could find evidence from the user opinions that only a certain age group of users favor treatment A more than treatment B, then this observation could serve as the hypothesis for clinical trial and comparative testing among this particular group could be prioritized among all age groups.

In this paper, we present a study that, by first extracting treatment specific opinions from MedHelp [1], a large and professional dataset of online health messages, then evaluating these opinions by either machine enabled sentiment analysis or human analysts which targets to the treatments of two common diseases: breast cancer and depression, we were able to generate evidence that could be used by CER research. By suggesting three different methods (*i.e.* direct comparison by the same author, indirect comparison by the same author and indirect comparison in overall case) to conduct the opinions comparison, we attempted to draw reliable conclusions that are strengthened by multiple evidences. Specifically, from the personal profile page, we were able to track the demographic information of the opinion source, including age, gender and geographical location. We proved that, by aggregating user opinions based on these variables, we were able to identify demographical characteristics for which candi-

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

date treatments are most likely to show differences. Theses results could serve as preliminary evidence for "which treatment works best for whom".

The section of work is complete and is submitted to the journal of ACM Transactions of Management Information Systems (ACM-TMIS), Special Issue for Smart Health and Well Being, to appear in 2013.

Organization. Section 5.1 defines the problems, objectives and setup of the work. Section 5.2 shows all the experiment results as well as statistical testing and is followed by the discussion of limitation in Section 5.3. Finally, I will make conclusion and discuss the future work in Section 5.4.

5.1 Problem Definition

5.1.1 Setup

We choose all posts on MedHelp [2] as our dataset. MedHelp is a on online forum where patients and providers discuss various health related topics. As its name suggests, commonly people posts questions regarding specific health problem and seek for help from other patients or providers. The site listed more than 100 registered providers who frequently visit the site and preside a number of "expert forums" to answer patients' questions. By 2009, this site was getting over 6 million unique visitors monthly, and contained more than 10 million messages. Compared to other social media sites that are less dedicated to health discussion, such as Yahoo Answer! or Twitter, MedHelp contains significantly more focused and informative discussions contributed by experienced patients and care providers, while less noised data such as spamming and irrelevant discussions. Therefore, we believe this site offers us a rich and relatively clean dataset for extracting and comparing opinions regarding different treatments.

We choose *breast cancer* and *depression* as our CER objects. Both diseases are fairly common and raise great public attention in US. More than $200,000$ American women are diagnosed annually with breast cancer [87] and millions of American adults are diagnosed annually with depression. Particularly, the effectiveness comparisons among different treatments for breast cancer and depression are listed as top 100 CER candidates by Institute of Medicine of the National Academy [88].

Disease 1: Breast Cancer

There are two specific sub-forums on MedHelp which are related to breast cancer: Breast Cancer and Breast Cancer Expert. Table 5.1 shows some general statistics of the sub-forums.

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

Sub-forum name	Total personal messages	Unique authors
Breast Cancer	45646	11870
Breast Cancer Expert	24860	8027
Overall	70506	17904

Table 5.1: The general statistics of breast cancer sub-forums

For breast cancer, we will compare the effectiveness between each two of the following common treatments: Chemotherapy, Radiation Therapy, and Hormonal Therapy. These are most widely-applied treatments for breast cancer and we aim to see how the actual users show preferences between: (*i.e.* Chemo *v.s.* Radiation, Chemo *v.s.* Hormonal, Radiation *v.s.* Hormonal).

Disease 2: Depression & Anxiety

There are three specific sub-forums on MedHelp which are related to depression: Anxiety, Depression and Depression Expert. Table 5.2 shows some general statistics of the sub-forums.

Sub-forum name	Total personal messages	Unique authors
Anxiety	116321	20817
Depression	40762	10340
Depression Expert	29112	10072
Overall	186195	38643

Table 5.2: The general statistics of depression sub-forums

For depression & anxiety, we will compare the effectiveness between Meditation Therapy (*e.g.*, yoga, deep breath training, etc) and the traditional medicine treatments. There are three

common medicine classes to treat depression & anxiety: Selective serotonin reuptake inhibitors (SSRIs); Serotonin and norepinephrine reuptake inhibitors (SNRIs) and Tricyclic antidepressants (TCAs). In the experiment, we will compare meditation treatment with each of the three medicine treatments, respectively.

5.1.2 Design

To analyze the opinion on the treatment effectiveness of the author [3], we need to extract the useful part of the messages which can exactly capture the opinion. On one hand, typically an author describes different things and shows multiple opinions in one single message, thus analyzing the whole message's opinion (positive, negative or neutral) may bias the opinion on one specific treatment. On the other hand, only analyzing the exact sentence where the treatment appears may miss some key information which appears in the *context* of the sentence. Therefore, we analyze the sentence containing the keywords of the specific *treatment*, as well as its treatment-keywords-free context, called *comment unit*.

The keywords of each treatment we applied are listed in Table 5.3.

[3]Here one author is referred to a patient who posted messages on the forum

treatment	Chemotherapy	Radiation	Hormonal	Meditation	SSRI	SNRI	TCA
disease	Breast Cancer	Breast Cancer	Breast Cancer	Depression	Depression	Depression	Depression
keywords	Chemotherapy chemo Abraxane, paclitaxel Adriamycin doxorubicin carboplatin Paraplatin, Cytoxan cyclophosphamide daunorubicin Cerubidine DaunoXome, Doxil doxorubicin, Ellence epirubicin, Halaven fluorouracil 5-fluorouracil, 5-FU Adrucil, Gemzar gemcitabine, eribulin, Ixempra ixabepilone methotrexate Amethopterin, Folex, Mitomycin mutamycin,Mexate mitoxantrone Navelbine Novantrone, vinorelbine, Taxol paclitaxel, Taxotere docetaxel, thiotepa Thioplex, vincristine Oncovin, Vincrex Xeloda, capecitabine AC, CMF, CEF FAC, CAF, TAC AC-T, AC	radiation radiotherapy XRT DXT	hormonal Aromatase inhibitor Aromatase inhibitors AIs Arimidex anastrozole Aromasin exemestane Femara letrozole Selective Estrogen Receptor Modulators Selective Estrogen Receptor Modulator Serms serm tamoxifen Nolvadex Evista raloxifene Fareston toremifene ERDs erd Estrogen-Receptor Downregulators Estrogen-Receptor Downregulator Estrogen Receptor Downregulators Estrogen Receptor Downregulator faslodex fulvestrant	mindfulness mindfulness-based yoga yoga-based meditation meditations deep breathing deep breath breath training breathing-training deep breath-training	Selective-serotonin-reuptake-inhibitor Selective-serotonin-reuptake-inhibitors SSRIs SSRI fluoxetine Prozac Sarafem paroxetine Paxil Pexeva sertraline Zoloft citalopram Celexa escitalopram Lexapro Symbyax	norepinephrine-reuptake-inhibitors norepinephrine-reuptake-inhibitor SNRI SNRIs duloxetine Cymbalta Yentreve venlafaxine Effexor Effexor XR desvenlafaxine Pristiq Milnacipran Dalcipran Ixel Savella Levomilnacipran F2695	Tricyclic-antidepressants Tricyclic-antidepressant TCAs TCA Amitriptyline Elavil Endep Clomipramine Anafranil Doxepin Adapin Sinequan Imipramine Tofranil Trimipramine Surmontil

Table 5.3: The keywords for each treatment

5.1.3 Direct Comparison by Same Author

To judge an author's preference on two treatments, the most straightforward approach is to detect her direct comparison (*e.g.* I prefer chemo to radiation; I think taking duloxetine is worse than yoga treatment. *etc*). We describe out first comparison method, direct comparison by same author, in this section. For each pair of treatments to be compared, we collect all the authors who have written *at least* one sentence in which *both* of the treatments were mentioned.

For each extracted comment unit and the treatment T1 and T2, one author may have one of the four cases of preference: (1). prefer T1 to T2, (2). prefer T2 to T1. (3). T1 and T2 are equally considered. (4). No direct comparison. We invited 5 medical school students who are experienced, and familiar with information about drugs and treatments, among other qualifications to annotate all the comment units. For time and quality concern, it is impossible to ask one judge to label all the sentences so we randomly split the whole annotation task to the 5 judges. We used the following coding scheme to instruct the judges.

For two treatments T1 and T2, we say T1 is preferred when:

- T1 and T2 have both been taken and T1 is clearly mentioned that it is better, or has less side effects than T2.
- T1 and T2 have both been taken and T1 is clearly mentioned that it works well, while no words for T2.
- T1 and T2 have both been taken and T2 is clearly mentioned that it works bad, while no words for T1.

we say T2 is preferred when:

- T1 and T2 have both been taken and T2 is clearly mentioned that it is better, or has less side effects than T1.

- T1 and T2 have both been taken and T2 is clearly mentioned that it works well, while no words for T1.
- T1 and T2 have both been taken and T1 is clearly mentioned that it works bad, while no words for T2.

and we say T1 and T2 are equally considered when:

- T1 and T2 have both been taken and there is no significant preference on their outcomes.
- T1 and T2 have both been taken and both of T1 and T2 are praised.
- T1 and T2 have both been taken and both of T1 and T2 are complained.

finally, T1 are T2 are considered no comparison when:

- T1 and T2 have been taken but the author says nothing about the outcome of T1 or T2.
- T1 and T2 are only mentioned in the sentence. We don't know whether they have been taken or not.

To test the inter-judge reliability, we assigned them 50 identical sentences, which are randomly selected from the sentence pool. we test the inter-rater agreement by using Fleiss' kappa [79]. Fleiss' kappa is widely used for assessing the reliability of agreement between a fixed number oef judges when assigning categorical ratings to a number of items or classifying items. For our experiment, Fleiss' kappa value reaches 0.82, which is considered almost perfect agreement.

The above results indicate that there is no significant variance among all the annotators. Therefore, we can confidently utilize the annotated results to judge the preference of each author.

Specifically, if an author has posted more than one such sentences, we will decide the author's preference based on all the sentences. If there are opposite opinions – *i.e.* In one

sentence, T1 is preferred while in another sentence, T2 is preferred, we will dismiss such author since she has inconsistent opinion. After labeling all the comment units, no such author is found, indicating that direct comparison between two treatments is always consistent by the same author.

5.1.4 Indirect Comparison by Same Author

In many cases, authors may not directly compare two treatments in one sentence but may mention their effectiveness respectively in different posts. For instance, if one author continually mentions that T1 is good for her in some posts, while she mentions that T2 has severe side effects in other posts. We can claim that this author prefer T1 to T2. Although there is no direct comparison, such indirect comparison can also reveal one author's preference.

In this section, we describe the second comparison method, indirect comparison by same author. For each pair of comparable treatments, we collect all the authors who have written sentences related to *both* of the treatments, respectively. To remove the influence of the direct comparison, the sentences analyzed by Section 5.1.3 will *not* be considered. We only count the authors who have mentioned T1 and T2 in different sentences.

For each author, we collect two sets: S1 and S2, composed by the comment units where T1 and T2 appears, respectively. Pennebaker and Campbell discovered that the words people use have strong correlation with their physical and mental health [89]. Thus, for S1 and S2, we utilize LIWC – effective sentiment analysis tool [90] to determine the opinion of the author – does she have positive, negative or neutral opinion on one specific treatment, by processing each comment unit in the set. LIWC has been utilized to compare the language usages between depressed and depressions-vulnerable students [91].

Similarly, we examined the consistency of each author's preference on one specific treatment across the whole dataset. If there are opposite opinions – *i.e.,* in one sentence, it shows

positive opinion on Treatment 1 while it shows negative opinion on the *same* treatment somewhere else, we will remove such author in the analysis since he or she has inconsistent opinion.

For all the authors who have consistent opinions on T1 and T2, we *indirectly* interpret their preferences on T1 and T2 based on the following criteria:

We say T1 is preferred when:

- The author has positive opinion on T1 and negative opinion on T2.
- The author has positive opinion on T1 and neutral opinion on T2.
- The author has neutral opinion on T1 and negative opinion on T2.

we say T2 is preferred when:

- The author has positive opinion on T2 and negative opinion on T1.
- The author has positive opinion on T2 and neutral opinion on T1.
- The author has neutral opinion on T2 and negative opinion on T1.

and we say T1 and T2 are equally considered when:

- The author has positive opinion on T1 and T2.
- The author has negative opinion on T1 and T2.
- The author has neutral opinion on T1 and T2.

5.1.5 Indirect Comparison in Overall Case

In this section,we describe the third comparison method, indirect comparison in overall case. We will indirectly compare two treatments based on the opinion of the overall population – *i.e.,* we only care how many people *favor* or *unfavor* one specific treatment. That is, we only

analyze the overall opinions for each treatment. Compared with the previous two experiments, this method may have the weakest power to reveal the comparison of two treatments since the two groups of people may differ, and some people may simply try only one treatment thus have no comparison.

Different from Experiment the previous two, in this experiment, we treat each treatment *separately* – collecting the authors who have posted sentences relative to *one* specific treatment T. LIWC is also applied to determine the opinion of each author on one treatment.

5.1.6 Demographic Analysis

In the above sections, we only consider the preference of *overall* population. However, people's preferences can be diverse according to the different demographic categories [92], *e.g.*, gender, age, region, race, income, *etc*. In the following session, we want to explore the potential difference of different demographic groups to answer the following questions: What is the preference of *two treatments*, for a specific demographic group? Is it same as the overall case, or different? How to compare the attitude of *one treatment* between two demographic groups? Is there any group significantly in higher favor of one treatment than other groups?

Some of the MedHelp users reveal detailed demographic information in their user profiles. After observing their profiles, we choose the following categories to be the representative groups:

- Gender: Male or Female. Note we only apply gender analysis on depression, not on breast cancer since the overwhelming majority of breast cancer patients are female .
- Age: We separate people into 4 groups: Age under 29, $30-49$, $50-64$, over 65.
- Region: Northeast, Midwest, South and West. This regional divisions are used by United States Census Bureau [93].

To answer the first question: what is the preference of *two treatments*, for a specific demographic group, we will do an *inner-group experiment*. For one specific group, we will calculate the portion of people who have positive/negative/neutual attitude on one specific treatment. And then comparing a pair of treatments using the approach in Section 5.1.5. The only difference between this experiment and Section 5.1.5 is the range of people investigated.

To answer the second question: how to compare the attitude of *one treatment* between two demographic groups, we will do a *cross-group* experiment. Using the information of each group's opinions on single treatment, we will compare the opinion difference between two corresponding groups upon the same treatment. *e.g.*, which group favors Radiation more? Age $30-49$ or Age $50-64$?

5.2 Results

5.2.1 Experiment A: Direct Comparison by Same Author

Table 5.4 shows the direct comparison results for each pair of treatments.

Disease	Breast Cancer	Breast Cancer	Breast Cancer	Depression	Depression	Depression
Treatment 1 Treatment 2	Chemo Radiation	Chemo Hormonal	Radiation Hormonal	Meditation SSRI	Meditation SNRI	Meditation TCA
# of authors mentioning T1 and T2, simultaneously	987	302	329	22	4	0
# of authors having no comparison	274	60	62	0	0	N/A
# of authors having comparison	713	242	267	22	4	N/A
# of authors preferring T1 percentage	62 8.70%	58 23.97%	66 24.72%	8 36.36%	3 75%	N/A N/A
# of authors preferring T2 percentage	21 2.95%	25 10.33%	23 8.61%	1 4.54%	0 0	N/A N/A
# of authors equally considering T1 & T2 percentage	630 88.36%	159 65.70%	178 66.67%	13 59.09%	1 25%	N/A N/A

Table 5.4: The results of Direct Comparison by author

From the data above, we can observe that for depression, people rarely directly compare a depression drug and meditation treatment in one sentence. In contrast, people frequently mention chemo, hormonal and radiation treatment in one sentence. It is reasonable since meditation and depression drugs have significant difference in treatment history, mechanism, *etc*, while chemo, hormonal, and radiation are usually be treated as a combination when the patient has breast cancer, thus co-mentioned frequently.

We also observe that in Table 5.4, in all available treatment pairs, The portion which prefers T1 is always larger than the portion which prefers T2 (*e.g.*, for breast cancer, when comparing

Chemo and Hormonal, 23.97% authors prefer Chemo (T1), while only 10.33% authors prefer Hormonal (T2)).

To statistically test the preference of two treatments, we first run a chi-square test [94] to test the hypothesis: the three categories(preferring T1, preferring T2, T1 and T2 equally considering) have the equal counts. The test evaluates the null hypotheses that people's preferences are equally distributed in the three categories mentioned above. Based on the chi-square definition, the chi-square test statistic is denoted as:

$$\chi^2 = \sum_{i=1}^{3} \frac{(N_i - p_i N)^2}{p_1 N} \tag{5.1}$$

In the above formula, given total count N, N_i refers the observed counts of each category which are to be compared to the expected counts $p_i N$ whereas $p_1 = p_2 = p_3 = 1/3$. Table 5.5 shows the chi-square test results for all the comparative treatments.

Disease	Breast Cancer	Breast Cancer	Breast Cancer	Depression	Depression
Treatment 1 Treatment 2	Chemo Radiation	Chemo Hormonal	Radiation Hormonal	Meditation SSRI	Meditation SNRI
N	713	242	267	22	4
N_1	62	58	66	8	3
N_2	21	25	23	1	0
N_3	630	159	178	13	1
χ^2	975.0	120.8	143.9	9.9	3.5
degree of freedom	2	2	2	2	2
p-value	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.007	0.17
judgement	significant	significant	significant	significant	not significant

Table 5.5: The chi-square test results of Direct Comparison by author

From Table 5.5 we observe that for all the treatment pairs except Meditation and TCA, there exists significantly unequal counts for some of the three categories. We therefore conduct post-hoc pairwise analysis to compare each pair of categories. For the purpose of our study, we

focus on presenting the proportion test [95] between the two groups:those preferring T1 and those preferring T2. We use p_1, p_2 to denote the portion of preferring T1 group and preferring T2 group, respectively.

Table 5.6 shows the proportion test results for all the comparative treatments.

Disease	Breast Cancer	Breast Cancer	Breast Cancer	Depression	Depression
Treatment 1 Treatment 2	Chemo Radiation	Chemo Hormonal	Radiation Hormonal	Meditation SSRI	Meditation SNRI[a]
N	83	83	89	9	4
N_1	62	58	66	8	3
p_1	0.75	0.70	0.74	0.89	0.75
N_2	21	25	23	1	0
p_2	0.25	0.30	0.26	0.11	0.25
σ	0.055	0.055	0.053	0.167	N/A
z	4.50	3.62	4.56	2.33	N/A
p-value	$< 10^{-4}$	$1.5 \cdot 10^{-4}$	$< 10^{-4}$	0.01	N/A
judgement	significant	significant	significant	significant	N/A

$\sigma = p_0 \cdot (1 - p_0)/N, p_0 = 0.5$

$z = (p_1 - p_0)/\sigma$

[a] According to the definition of proportion test, the sample should include at least 5 items in each group. Therefore the proportion test is not applied for this pair.

Table 5.6: The proportion test results of Direct Comparison by author

From Table 5.6 we observed that all the applied pairs show p_1 is significantly larger than p_2, indicating that the number of authors who preferred Treatment 1 is significantly more than that of Treatment 2, for each case.

5.2.2 Experiment B: Indirect Comparison by Same Author

From the results in Section 5.2.1, although direct comparison by the same author captures the most direct and possibly most accurate preference of the patients, we observe that the number of patients who directly compare treatments are relatively small, even none in some cases. Therefore, we conduct indirect comparison by the same author to increase the number of available cases.

Table 5.7 shows the indirect comparison results for each pair of treatments.

Disease	Breast Cancer	Breast Cancer	Breast Cancer	Depression	Depression	Depression
Treatment 1	Chemo	Chemo	Radiation	Meditation	Meditation	Meditation
Treatment 2	Radiation	Hormonal	Hormonal	SSRI	SNRI	TCA
total # of authors mentioning T1 and T2	1330	717	719	663	282	72
# of authors mentioning T1 and T2, respectively	562	529	505	654	278	72
# of authors having consistent judgement on T1 and T2	352	322	314	284	113	29
# of authors: Positive on T1; Negative on T2	15	13	21	60	20	6
# of authors: Positive on T1; Neutral on T2	45	44	39	37	21	5
# of authors: Neutral on T1; Negative on T2	40	32	37	30	8	6
# of authors preferring T1 to T2	100	89	97	124	49	17
percentage	28.41%	27.64%	30.90%	44.72%	43.36%	58.62%
# of authors: Neutral on T1; Positive on T2	46	28	37	10	9	2
# of authors: Negative on T1; Neutral on T2	22	28	20	21	10	2
# of authors: Negative on T1; Positive on T2	11	10	10	10	8	3
# of authors preferring T2 to T1	79	66	67	41	27	7
percentage	22.44%	20.50%	21.34%	14.44%	23.89%	24.14%
# of authors: Neutral on T1; Neutral on T2	106	103	85	28	7	2
# of authors: Positive on T1; Positive on T2	50	40	43	17	11	1
# of authors: Negative on T1; Negative on T2	17	24	22	71	19	2
# of authors equally considering T1 & T2	173	167	150	116	37	5
percentage	49.15%	51.86%	47.77%	40.85%	32.74%	17.24%

Table 5.7: The results of Indirect Comparison by author

Compared with Table 5.4, Table 5.7 shows there are more data available for indirect comparison of treatments in all cases, which suggested potentially increased validity of the conclusion.

Similar to Section 5.2.1, from the result of indirect comparison by the same author, We also observe that in Table 5.7, in all available treatment pairs, The portion which prefers T1 is always larger than the portion which prefers T2 (*e.g.*, for depression, when comparing Mediation and SNRI, 43.36% authors prefer Mediation (T1), while only 23.89% authors prefer SNRI (T2)).

Similarly, we start by running a chi-square test [94] to test the hypothesis: the three categories(preferring T1, preferring T2, T1 and T2 equally considering) have the equal counts. Based on the chi-square definition in Formula 5.1. Table 5.8 shows the chi-square test results for all the comparative treatments.

Disease	Breast Cancer	Breast Cancer	Breast Cancer	Depression	Depression	Depression
Treatment 1 Treatment 2	Chemo Radiation	Chemo Hormonal	Radiation Hormonal	Meditation SSRI	Meditation SNRI	Meditation TCA
N	352	322	314	284	113	29
N_1	100	89	97	124	49	17
N_2	79	66	67	41	27	7
N_3	173	167	150	116	37	5
χ^2	41.49	52.22	33.75	46.27	6.44	8.55
degree of freedom	2	2	2	2	2	2
p-value	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.04	0.013
judgement	significant	significant	significant	significant	significant	significant

Table 5.8: The chi-square test results of Indirect Comparison by author

From Table 5.8 we observe that for all the treatment pairs, the hypothesis H_0 will be rejected, *i.e.* there exists significantly unequal counts among some pair of the categories. We then conducted posthoc pairwise comparison. The tests results comparing the category of preferring T1 and the category preferring T2 are shown in Table 5.9 .

Disease	Breast Cancer	Breast Cancer	Breast Cancer	Depression	Depression	Depression
Treatment 1 Treatment 2	Chemo Radiation	Chemo Hormonal	Radiation Hormonal	Meditation SSRI	Meditation SNRI[a]	Meditation TCA
N	179	155	164	168	76	24
N_1	100	89	97	124	49	17
p_1	0.559	0.574	0.591	0.756	0.645	0.708
N_2	79	66	67	41	27	7
p_2	0.441	0.426	0.409	0.244	0.355	0.292
σ	0.037	0.040	0.039	0.039	0.057	0.102
z	1.57	1.85	2.34	6.64	2.52	2.14
p-value	0.05	0.03	0.01	$< 10^{-4}$	0.006	0.012
judgement	significant	significant	significant	significant	significant	significant

$\sigma = p_0 \cdot (1 - p_0)/N, p_0 = 0.5$

$z = (p_1 - p_0)/\sigma$

Table 5.9: The proportion test results of Indirect Comparison by author

From Table 5.9 we observe that in all the pairs p_1 is significantly larger than p_2, indicating that the number of authors who preferred Treatment 1 is significantly more than that of Treatment 2, for each case. The conclusion is consistent with the one we draw from Section 5.2.1.

5.2.3 Experiment C: Indirect Comparison in Overall Case

For indirect comparison, we focus on comparing the overall attitude expressed for each treatment, instead of detecting each author's preference between two treatments. Table 5.10 shows the descriptive results.

Disease	Breast Cancer	Breast Cancer	Breast Cancer	Depression	Depression	Depression	Depression
Treatment T	Chemo	Radiation	Hormonal	Meditation	SSRI	SNRI	TCA
# of authors mentioning T	2476	2154	1486	1454	10996	4131	473
# of authors having positive opinion on T	1043	851	505	653	2146	877	112
percentage	42.12%	39.51%	33.98%	44.91%	19.52%	21.23%	23.68%
# of authors having negative opinion on T	651	614	522	518	6499	2248	263
percentage	26.29%	28.50%	35.13%	35.63%	59.10%	54.42%	55.60%
# of authors having neutral opinion on T	782	689	459	283	2351	1006	98
percentage	31.58%	31.99%	30.89%	19.46%	21.38%	24.35%	20.72%

Table 5.10: The results of Indirect Comparison in overall

Different from Section 5.2.1 and 5.2.2, we conducted different statistical analysis to investigate the preference of the overall people:

For each pair of treatments T1 and T2, given

- $p1_{pos}$: the percentage of authors having *positive* opinion on T1
- $p2_{pos}$: the percentage of authors having *positive* opinion on T2
- $p1_{neg}$: the percentage of authors having *negative* opinion on T1
- $p2_{neg}$: the percentage of authors having *negative* opinion on T2

, if more people are positive on T1 meanwhile fewer people are negative on T1, *i.e.*, $p1_{pos}$ is significantly larger than $p2_{pos}$, meanwhile $p1_{neg}$ is significantly smaller than $p2_{neg}$, we can

conclude that the number of authors who hold positive attitude towards T1 is significantly more than that of T2.

To compare two independent portions from two samples ($p1_{pos}$ and $p2_{pos}$, $p2_{neg}$ and $p2_{neg}$), we run a two-sample proportion test. Table 5.11 and 5.12 show the two-sample proportion test result for positive and negative attitude in overall people, respectively.

Disease	Breast Cancer	Breast Cancer	Breast Cancer	Depression	Depression	Depression
Treatment 1 Treatment 2	Chemo Radiation	Chemo Hormonal	Radiation Hormonal	Meditation SSRI	Meditation SNRI[a]	Meditation TCA
N_1	2476	2476	2154	1454	1454	1454
N_2	2154	1486	1486	10996	4131	473
$p1_{pos}$	42.12%	42.12%	39.51%	44.91%	44.91%	44.91%
$p2_{pos}$	39.51%	33.98%	33.98%	19.52%	21.23%	23.68%
p	0.409	0.391	0.373	0.225	0.274	0.40
σ	0.014	0.016	0.016	0.011	0.014	0.026
z	1.80	5.08	3.39	21.79	17.4	8.20
p-value	0.036	$< 10^{-4}$	$3 \cdot 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
judgement	significant	significant	significant	significant	significant	significant

$p = (p1_{pos} \cdot N_1 + p2_{pos} \cdot N_2)/(N_1 + N_2)$
$\sigma = \sqrt{p \cdot (1-p) \cdot ((1/N_1) + (1/N_2))}$
$z = (p1_{pos} - p2_{pos})/\sigma$

Table 5.11: The two-sample proportion test results(positive) in overall

Disease	Breast Cancer	Breast Cancer	Breast Cancer	Depression	Depression	Depression
Treatment 1 Treatment 2	Chemo Radiation	Chemo Hormonal	Radiation Hormonal	Meditation SSRI	Meditation SNRI[a]	Meditation TCA
N_1	2476	2476	2154	1454	1454	1454
N_2	2154	1486	1486	10996	4131	473
$p1_{neg}$	26.29%	26.29%	28.50%	35.63%	35.63%	35.63%
$p2_{neg}$	28.50%	35.13%	35.13%	59.10%	54.42%	55.60%
p	0.273	0.296	0.312	0.564	0.495	0.405
σ	0.013	0.015	0.016	0.014	0.015	0.026
z	-1.68	-5.90	-4.24	-16.96	-12.32	-7.68
p-value	0.05	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
judgement	significant	significant	significant	significant	significant	significant

$p = (p1_{neg} \cdot N_1 + p2_{neg} \cdot N_2)/(N_1 + N_2)$
$\sigma = \sqrt{p \cdot (1-p) \cdot ((1/N_1) + (1/N_2))}$
$z = (p1_{neg} - p2_{neg})/\sigma$

Table 5.12: The two-sample proportion test results(negative) in overall

From Table 5.11 and 5.12 we observe that, in all the pairs, $p1_{pos}$ is significantly larger than $p2_{pos}$, while $p1_{neg}$ is significantly smaller than $p2_{neg}$, indicating that the number of authors who preferred Treatment 1 is significantly more than that of Treatment 2, for each case. The conclusion is consistent with the one we draw from Section 5.2.1 and 5.2.2.

5.2.4 Experiment D: Demographic Analysis

Table 5.13- 5.19 show the attitude of each demographic group for each treatment: chemo, radiation, hormonal for breast cancer; meditation, SSRI, SNRI, TCA for depression, respectively. Note in these tables, some groups' statistics is not included (*e.g.*, Age $18-29$ for Breast Cancer, Age $65+$ for Depression) since the sample size is too small.

Group	Age 30-49	Age 50-64	Age 65+	Northeast	Midwest	South	West
# of authors mentioning Chemo	700	709	433	317	315	592	399
percentage of positive	45.50%	43.22%	37.5%	39.62%	48.57%	42.13%	47.37%
percentage of negative	24.89%	26.69%	28.61%	33.02%	20.95%	24.87%	24.06%
percentage of neutral	29.61%	30.08%	33.89%	27.36%	30.48%	32.99%	28.57%

Table 5.13: The attitude to Chemo, Breast Cancer

Group	Age 30-49	Age 50-64	Age 65+	Northeast	Midwest	South	West
# of authors mentioning Radiation	484	706	378	275	313	533	407
percentage of positive	45.96%	34.04%	42.06%	42.39%	40.38%	33.90%	41.91%
percentage of negative	25.46%	30.21%	18.25%	20.65%	25.96%	31.64%	26.47%
percentage of neutral	28.57%	35.74%	39.68%	36.96%	33.65%	34.46%	31.62%

Table 5.14: The attitude to Radiation, Breast Cancer

Group	Age 30-49	Age 50-64	Age 65+	Northeast	Midwest	South	West
# of authors mentioning Hormonal	408	487	217	216	229	343	260
percentage of positive	42.65%	30.25%	27.78%	30.56%	34.21%	38.60%	35.63%
percentage of negative	30.15%	39.51%	44.44%	38.89%	36.84%	29.82%	31.03%
percentage of neutral	27.20%	30.25%	27.78%	30.56%	28.95%	31.58%	33.34%

Table 5.15: The attitude to Hormonal, Breast Cancer

Group	Male	Female	Age 18-29	Age 30-49	Age 50-64	Northeast	Midwest	South	West
# of authors mentioning Meditation	396	781	432	597	156	200	211	297	235
percentage of positive	43.33%	48.23%	44.44%	44.64%	47.44%	47%	41.90%	46.62%	41.52%
percentage of negative	36.97%	32.26%	34.26%	35.91%	28.21%	33%	38.10%	37.84%	33.05%
percentage of neutral	19.70%	19.51%	21.30%	17.45%	24.36%	20%	20%	15.54%	25.42%

Table 5.16: The attitude to Meditation, Depression

Group	Male	Female	Age 18-29	Age 30-49	Age 50-64	Northeast	Midwest	South	West
# of authors mentioning SSRI	1862	4216	1957	2967	990	1217	1248	1993	1230
percentage of positive	18.81%	21.66%	18.28%	19.82%	20.20%	21.67%	18.91%	19.08%	20%
percentage of negative	60.63%	58.23%	60.57%	59.61%	58.38%	56.98%	59.45%	59.27%	57.72%
percentage of neutral	20.55%	20.04%	21.14%	20.57%	21.41%	21.35%	21.63%	21.69%	22.28%

Table 5.17: The attitude to SSRI, Depression

Group	Male	Female	Age 18-29	Age 30-49	Age 50-64	Northeast	Midwest	South	West
# of authors mentioning SNRI	671	1707	597	1242	538	455	498	793	507
percentage of positive	20.93%	21.79%	19.80%	22.22%	21.19%	21.49%	15.26%	21.72%	24.90%
percentage of negative	57.25%	55.38%	55.70%	56.36%	55.76%	57.02%	58.23%	57.58%	54.56%
percentage of neutral	21.82%	22.84%	22.50%	21.42%	23.05%	21.49%	26.51%	20.71%	20.55%

Table 5.18: The attitude to SNRI, Depression

Group	Male	Female	Age 18-29	Age 30-49	Age 50-64	Northeast	Midwest	South	West
# of authors mentioning TCA	95	180	48	143	91	63	54	74	99
percentage of positive	24.05%	26.67%	25%	27.78%	20%	35.48%	18.52%	21.62%	17.24%
percentage of negative	56.96%	53.33%	58.33%	54.17%	66.67%	51.61%	62.96%	54.05%	65.52%
percentage of neutral	18.99%	20%	16.67%	18.06%	13.33%	12.90%	18.52%	24.32%	17.24%

Table 5.19: The attitude to TCA, Depression

Inner-group experiment

In this section we will compare the preference between two treatments of a particular demographic group, by using the indirect comparison similar to the one in Section 5.2.3: for each pair of treatments T1 and T2, given

- $p1_{pos}$: the percentage of authors having *positive* opinion on T1
- $p2_{pos}$: the percentage of authors having *positive* opinion on T2
- $p1_{neg}$: the percentage of authors having *negative* opinion on T1
- $p2_{neg}$: the percentage of authors having *negative* opinion on T2

, if more people are positive on T1 meanwhile less people are negative on T1, *i.e.*, $p1_{pos}$ is significantly larger than $p2_{pos}$ meanwhile $p1_{neg}$ is significantly smaller than $p2_{neg}$, we can conclude that the number of authors who preferred T1 is significantly more than that of T2, vice versa. While comparing the independent portions, the two-sample proportion test is applied.

Table 5.20 shows the preference of a specific demographic group between two treatments of breast cancer. Note that given the purpose of demographic analysis is to identify groups that show different preference among others, to save space, here we only present groups for which statistical analysis showed different conclusion from the trend of overall population.

Group	Age 30-49	Age 65+	Northeast	South
Treatment 1 Treatment 2	Chemo Radiation	Chemo Radiation	Chemo Radiation	Radiation Hormonal
$p1_{pos}$	45.49%	37.5%	39.62%	33.90%
$p2_{pos}$	45.96%	42.06%	42.39%	38.60%
judgement p-value	$p1_{pos}, p2_{pos}$ **not** significant different 0.88	$p1_{pos} < p2_{pos}$ 0.08	$p1_{pos} < p2_{pos}$ 0.07	$p1_{pos} < p2_{pos}$ 0.04
$p1_{neg}$	24.89%	28.61%	33.02%	31.64%
$p2_{neg}$	25.47%	18.25%	20.65%	28.82%
judgement p-value	$p1_{neg}, p2_{neg}$ **not** significant different 1.15	$p1_{neg} > p2_{neg}$ 0.002	$p1_{neg} > p2_{neg}$ $3 \cdot 10^{-4}$	$p1_{neg} > p2_{neg}$ 0.09

For the space limitation, if the preference of a specific group is the same as the overall population according to the same pair of treatments, we do not list them in the table

Table 5.20: The preference of inner-group of Breast Cancer

From Table 5.20 we can draw the following conclusion:

- people in age $30-49$ group have no different preference between Chemo and Radiation.
- In age $65+$ group, the number of people prefer Chemo is *smaller* than that of Radiation.
- In Northeast area, the number of people prefer Chemo is *smaller* than that of Radiation.
- In South area, the number of people prefer Radiation is *smaller* than that of Hormonal.

For Depression, the statistical result shows that, for each pair of treatments, the preference of every demographic group is consistent with the overall population. It indicates that mediation is indeed preferred by people to drug treatment, regardless of their sex, age, or region.

Cross-group experiment

In this section we will compare the attitude on one specific treatment between two demographic groups. We modify the strategy used in Section 5.2.3 as below:

for two demographic groups G1 and G2 and one treatment T, given

Treatment T	Chemo	Radiation	Hormonal
Group 1 Group 2	Northeast Midwest	Age 50-64 Age 65+	Age 30-49 Age 50-64
$p1_{pos}$	39.62%	34.04%	42.65%
$p2_{pos}$	48.57%	42.06%	30.25%
judgement p-value	$p1_{pos} < p2_{pos}$ 0.012	$p1_{pos} < p2_{pos}$ 0.005	$p1_{pos} > p2_{pos}$ $< 10^{-4}$
$p1_{neg}$	33.02%	30.21%	30.15%
$p2_{neg}$	20.95%	18.25%	39.51%
judgement p-value	$p1_{neg} > p2_{neg}$ $3 \cdot 10^{-4}$	$p1_{neg} > p2_{neg}$ $< 10^{-4}$	$p1_{neg} < p2_{neg}$ 0.002

Table 5.21: The preference of cross-group for breast cancer

Treatment T	Meditation	SSRI	SNRI	TCA
Group 1 Group 2	Male Female	Midwest South	Midwest West	Age 30-49 Age 50-64
$p1_{pos}$	43.33%	18.97%	15.26%	27.78%
$p2_{pos}$	48.23%	19.07%	24.90%	20%
judgement p-value	$p1_{pos} < p2_{pos}$ 0.05	$p1_{pos}, p2_{pos}$ **not** different 0.79	$p1_{pos} < p2_{pos}$ $< 10^{-4}$	$p1_{pos} > p2_{pos}$ 0.09
$p1_{neg}$	36.97%	59.46%	58.23%	54.17%
$p2_{neg}$	32.26%	59.24%	54.55%	66.67%
judgement p-value	$p1_{neg} > p2_{neg}$ 0.05	$p1_{neg}, p2_{neg}$ **not** different 0.51	$p1_{neg} > p2_{neg}$ 0.05	$p1_{neg} < p2_{neg}$ 0.03

Table 5.22: The preference of cross-group for depression

- $p1_{pos}$: the percentage of authors in G1 having *positive* opinion on T
- $p2_{pos}$: the percentage of authors in G2 having *positive* opinion on T
- $p1_{neg}$: the percentage of authors in G1 having *negative* opinion on T
- $p2_{neg}$: the percentage of authors in G2 having *negative* opinion on T

, if more people in G1 are positive and less people are negative, compared with the G2, *i.e.*, $p1_{pos}$ is significantly larger than $p2_{pos}$ as well as $p1_{neg}$ is significantly smaller than $p2_{neg}$, we can conclude that Group 1 have a more positive opinion on the particular treatment than Group 2. Again we conduct the two-sample proportion test.

Table 5.21 and 5.22 shows the preference to one specific treatment for breast cancer and depression respectively between two demographic groups. Since for each treatment there are multiple choices of the pair of demographic groups, for the space limitation, we only list group pairs that exhibit interesting results, for each treatment respectively.

From Table 5.21 and 5.22 we can observe that:

- people in Midwest *prefer* Chemo significantly more than people in Northeast.
- people in age 65+ *prefer* Radiation significantly more than people in age 50-64.
- people in age 30-49 *prefer* Hormonal significantly more than people in age 50-64.
- Women *prefer* Meditation significantly more than men.
- people in Midwest has no difference in preference to SSRI with people in South.
- people in West *prefer* SNRI significantly more than people in Midwest.
- people in age 30-49 *prefer* TCA more than people in age 50-64.

These conclusions could be understand intuitively. *e.g.* older people are more fragile to bear Chemo thus they would more prefer to relatively mild therapy such as Radiation. Younger people are more easy to the relatively novel therapy, such as Hormonal. Also the number of women who take yoga training are significantly larger than men. When it comes to a treatment of depression, women would show more interest to try meditation. While these are all reasonable inference, future research is needed to verify these statements.

From our observation we found that part of our conclusions drawn from the demographic analysis have been validated by real clinical trials. For example, The major global longitudinal study called the Early Breast Cancer Trialists' Collaborative Group (EBCTCG) [96] found that "Allocation to about 6 months of chemo reduces the annual breast cancer death rate by about 38% (SE 5) for women younger than 50 years of age when diagnosed and by about 20% (SE 4) for those of age $50-69$ years when diagnosed", which indicates that chemo is more effective to younger group (<50), compared with older group(>50). Similarly, Segal *et al* [97] found that 8 weeks of mindfulness mediation training was just as good as prolonged antidepressant treatment (SSRI, SNRI) over 18 months. Also considering the lower cost and fewer side effects

of meditation therapy, it is understandable why patients preferred meditation more, even in each demographic group.

From the demographic analysis, we can automatically predict the preference comparison among different groups, which can be used as hypotheses to be tested in clinical research.

5.3 Limitations

Our work may have following limitations:

- Preference versus Effectiveness. From the results in Section 5.2, our approach can automatically compare the preference for treatments of any group of patients, however, self reported preference may not be equal to effectiveness. For example, patient may prefer a treatment simply because it causes less pain or side effects, while the same treatment can in fact be less effective in treating the disease. In general, conclusion drawn from mining online personal health message can only be used as a supportive evidence for forming hypotheses, but not replace formal clinical trials. In Section 5.2.4, we have validated part of our conclusions with the results of clinical trials. Further research is also needed to systematically identify the relation between the patients' preference stated in personal health message and the actual treatment effectiveness. *i.e.*, validating the current hypotheses generated from personal messages by comparing with the formal medical literature or reports.

- Reliability. One advantage of online personal messages is that they are easy to acquire and utilize. However, compared with clinic notes or medical literature, not all the personal messages are reliable. *i.e.*, some of them may submit the messages with questionable motivation (e.g., commercial purpose) which do not accurately reflect patient experience, and some users may provide the incorrect profile information. Unreliable personal messages will affect the quality and validity of our results. In our current experiment, all the messages are considered having the equal reliability. Such result can be improved by techniques detecting low quality or dubious posts or personal profile, and weighted message scores (by assigning more weight to authors who provide reliable messages and profile information) Some data mining researchers have explored the approach of

truth discovery from web data [98], [99]. We can apply the techniques into the current experiment to improve the reliability in the future.

- Scalability of the data. MedHelp provides more than 100 sub-forums for patients to communicate with each other. However not all the treatments or diseases are fully discussed, which makes the quality of messages insufficient to explore. (*e.g.*, COPD forum has only 2000 personal messages, *etc*). Therefore we should consider aggregating data across multiple web sources, *e.g.* the Health and Wellness parts of Yahoo Groups that contain demographic information about unique individuals, Twitter [4] which has much more personal messages. Although it seems straightforward to apply the same techniques to other sources, future research is needed to analyze the preference consistency, information quality consistency and population bias among different sources.

[4] http://www.twitter.com

5.4 Conclusion and Future Work

In this chapter, I introduce our CER hypothesis prediction framework – automatically generating the hypotheses of patients' preferences on different treatments of *breast cancer* and *depression*, respectively. By utilizing three comparison approaches together with statistical test, we can successfully draw reliable and consistent conclusions from personal health messages. Furthermore, from the user profile information, we can also explore the demographic information and aggregate patients' preferences upon it. By applying inner-group and cross-group analysis, we can detect the differences among demographic groups and answer the question "which treatment is more effective and favored by whom?"

Based on CER's definition, the real pragmatic trials are required thus conclusions can not be made by only analyzing text-based personal messages. In the future, solid hypotheses validation is necessary by comparing our auto-generated conclusions from personal health messages with the conclusions drawn from electronic medical record or clinical trials.

In our case study, breast cancer and depression are fully investigated. In the future, more treatments and diseases will be tested by aggregating data across multiple web sources, such as Yahoo! Groups with demographic information, as well as Twitter. Thus, an interactive system can be built for patients to search and compare the effectiveness of treatments via ad-hoc queries.

CHAPTER 6

Conclusion and Summary

The development of Web 2.0 techniques has led to the prosperity of online communities, which spread to various domains and areas in our daily life. When it comes to the *medicine and healthcare* domain, a series of good online services such as Yahoo! Groups, WebMD and MedHelp, offer patients and physicians a good platform to discuss health problems, *e.g.*, diseases and drugs, diagnoses and treatments, which also provides a large volume of data for researchers to analyze and explore. However, some nature of the personal messages, *e.g.*, unclean, unstructured and isolated from clinical practice, hinders users' effective digestion of information in the front end and challenges the data analysis in the back end.

In this thesis, I apply the advanced data mining, information retrieval and natural language processing techniques to effectively analyze and re-organize the rich source of personal health messages from online medical communities, in order to satisfy patients' information need and support physicians' clinical practise. Specially, in the first part of the dissertation, I introduce an SVM-based multi-class classification method which utilizes term-appearance, lexical and semantic features to effectively classify health messages sampled from our unique dataset of Yahoo! Health Groups into three categories: *News*, *User Comments* and *Spam*; in the second part, I depict a comprehensive system with an extensive evaluation framework to organize and cluster patient outcomes utilizing topic model; in the third part of the dissertation, I address a novel and promising topic: comparative effectiveness research (CER) hypothesis prediction, by presenting a study which evaluates patients' opinions on different treatments by machine enabled sentiment analysis or human analysts utilizing our MedHelp dataset. By solving these

three problems, we conclude that personal health messages enrich the information sources which patients and physicians request the and expand the scope that traditional medical informatics research can reach.

However this thesis can not discuss the full usage of personal health messages. To further explore such scalable and diverse information, two interesting and promising directions can be studied deeply:

- Voice-based messages annotation: Currently our work is relied on *text*-based messages. However with the popularity of mobile phones, people can choose to express their health-related experiences as the format of *voice* messages. If the feature of the voice can be adequately captured, we can transplanted the text-based techniques onto voice-based messages. Some researchers have begun to analyze the usage of voice messages in improving health outcomes and processes of care [100, 101].

- Medical validation: In Section 5.2.4, we have validated part of our conclusions with the results of clinical trials. However it is not enough. Further research is also needed to systematically identify the relation between the patients' preference stated in personal health message and the actual treatment effectiveness. *i.e.*, validating the current hypotheses generated from personal messages by comparing with the formal medical literature or reports from clinical trials.

- Medical messages comparison across multiple web sources: the topic of multiple web sources comparison has been discussed in many research and application areas, such as web search recommendation [102], bursty event tracking [69], *etc*. However, in medical informatics area, this topic has not been addressed although different healthcare-related web sources exist. More scalable and reliable data can be acquired and aggregated by fully analyzing the characters of each of the web sources and investigating the correlations among them.

Last but not least, the purpose of this dissertation is to draw more and more researchers' interests to the study of personal health messages in order to offer better service and medical care to patients, eventually to everyone.

References

[1] "Congressional Budget Office Research on the Comparative Effectiveness of Medical Treatments: Issues and Options for an Expanded Federal RoleWashington, DC," *Congressional Budget Office*, 2007.

[2] Agarwal S, Yu H, "FigSum: Automatically Generating Structured Text Summaries for Figures in Biomedical Literature," *AMIA Annual Symposium Proceedings*, pp. 6–10, 2009.

[3] Arnold CW, El-Saden SM, Bui AA, Taira R, "Clinical Case-based Retrieval Using Latent Topic Analysis," *AMIA Annual Symposium Proceedings*, pp. 26–30, 2010.

[4] Batal I, Hauskrecht M, "Mining Clinical Data using Minimal Predictive Rules," *AMIA Annual Symposium Proceedings*, pp. 31–35, 2010.

[5] Sondhi P, Sun J, Zhai C, Sorrentino R, Kohn MS, "A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003," *Journal of the American Medical Informatics Association (JAMIA)*, 2012.

[6] Shetty KD, Dalal SR, "Using information mining of the medical literature to improve drug safety," *Journal of the American Medical Informatics Association (JAMIA)*, vol. 18, pp. 668–674, 2011.

[7] Austin PC, "Leveraging medical thesauri and physician feedback for improving medical literature retrieval for case queries," *Statistics in Medicine*, vol. 27(12), pp. 2037–2049, 2008.

[8] Hill J, "Discussion of research using propensity-score matching: comments on 'A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003' by Peter Austin, Statistics in Medicine," *Statistics in Medicine*, vol. 27(12), pp. 2055–2061, 2008.

[9] He X, Li Y, Khetani R, Sanders B, Lu Y, Ling X, Zhai C, Schatz B, "BSQA: integrated text mining using entity relation semantics extracted from biological literature of insects," *Nucleic Acids Research*, vol. 38, pp. 175–181, 2010.

[10] He X, Sen Sarma M, Ling X, Chee B, Zhai C, Schatz B, "Identifying overrepresented concepts in gene lists from literature: a statistical approach based on Poisson mixture model," *BMC Bioinformatics*, vol. 11:272, 2010.

[11] Sen Sarma M, Arcoleo D, Khetani RS, Chee B, Ling X, He X, Jiang J, Mei Q, Zhai C, Schatz B, "BeeSpace Navigator: exploratory analysis of gene function using semantic indexing of biological literature," *Nucleic Acids Research*, vol. 39, pp. 462–469, 2011.

[12] Patterson O, Hurdle JF, "Document Clustering of Clinical Narratives: a Systematic Study of Clinical Sublanguages," *AMIA Annual Symposium Proceedings*, p. 1099C1107, 2011.

[13] Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC, "MedEx: a medication information extraction system for clinical narratives," *Journal of the American Medical Informatics Association (JAMIA)*, vol. 17, pp. 19–24, 2010.

[14] Jagannathan V, Mullett CJ, Arbogast JG, Halbritter KA, Yellapragada D, Regulapati S, Bandaru P., "Assessment of commercial NLP engines for medication information extraction from dictated clinical notes," *International Journal of Medical Informatics*, vol. 78(4), pp. 284–291, 2009.

[15] Wu S, Liu H, Li D, Tao C, Musen MA, Chute CG, Shah NH, "Unified Medical Language System term occurrences in clinical notes: a large-scale corpus analysis," *Journal of the American Medical Informatics Association*, vol. 19, pp. 146–156, 2012.

[16] Xu H, Stetson P, Friedman C, "Methods for building sense inventories of abbreviations in clinical notes," *Journal of the American Medical Informatics Association*, vol. 15(1), pp. 87–98, 2009.

[17] Chen ES, Manaktala S, Sarkar IN, Melton GB, "A Multi-Site Content Analysis of Social History Information in Clinical Notes," *AMIA Annual Symposium Proceedings*, pp. 227–236, 2011.

[18] Wei WQ, Tao C, Jiang G, Chute CG, "A High Throughput Semantic Concept Frequency Based Approach for Patient Identification: A Case Study Using Type 2 Diabetes Mellitus Clinical Notes," *AMIA Annual Symposium Proceedings*, pp. 857–861, 2010.

[19] Crain SP, Yang S, Zha H, Jiao Y, "Dialect Topic Modeling for Improved Consumer Medical Search," *AMIA Annual Symposium Proceedings*.

[20] Zhang Y, "Contextualizing consumer health information searching: an analysis of questions in a social QA community," *Proceedings of the 1st ACM International Health Informatics Symposium*, pp. 210–219, 2010.

[21] Yang L, Mei Q, Zheng K, Hanauer DA, "Query Log Analysis of an Electronic Health Record Search Engine," *AMIA Annual Symposium Proceedings*, pp. 915–924, 2011.

[22] Hanauer DA, "EMERSE: the electronic medical record search engine," *AMIA Annual Symposium Proceedings*, p. 941, 2006.

[23] Chee B, Berlin R, Schatz B, "Measuring Population Health Using Personal Health Messages," *AMIA Annual Symposium Proceedings*, pp. 92–96, 2009.

[24] Chee B, Berlin R, Schatz B, "Predicting Adverse Drug Events from Personal Health Messages," *AMIA Annual Symposium Proceedings*, pp. 217–226, 2011.

[25] Gomez Hidalgo JM, Bringas GC, Sanz EP, Garcia FC, "Content based sms spam filtering," *Proceedings of the ACM symposium on Document engineering (DocEng)*, pp. 107–114, 2006.

[26] Cormack GV, Gomez Hidalgo JM, Sanz EP, "Spam filtering for short messages," *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM)*, pp. 313–320, 2007.

[27] Munro R, Manning CD, "Subword variation in text message classification," *Human Language Technologies: The Annual Conference of the North American Chapter of the ACL*, pp. 510–518, 2010.

[28] Volkova S, Caragea D, Hsu W, Bujuru S, "Animal Disease Event Recognition and Classification," *Proceedings of the 1st International Workshop on Web Science and Information Exchange in the Medical Web (MedEx)*, pp. 51–61, 2010.

[29] Hofmann T, "Probabilistic latent semantic indexing," *Proceedings of the 22nd annual international ACM SIGIR conference*, pp. 50–57, 1999.

[30] Blei DM, Ng AY, Jordan MI, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, pp. 993–1022, 2003.

[31] Mei Q, Ling X, Wondra M, Su H, Zhai C, "Topic sentiment mixture: modeling facets and opinions in weblogs," *Proceedings of the 18th international conference on World wide web*, pp. 171–180, 2007.

[32] Steyvers M, Smyth P, Rosen-Zvi M, and Griffiths TL, "robabilistic author-topic models for information discovery," *Proceedings of the 10th ACM SIGKDD international conference*, p. 306C315, 2004.

[33] Li W, McCallum, "Pachinko allocation: Dag-structured mixture models of topic correlations," *ICML, ser. ACM International Conference Proceeding Series, W. W. Cohen and A. Moore, Eds.*, vol. 577 - 584, p. 306C315, 2006.

[34] Dietz L, Bickel S, Scheffer T, "Unsupervised prediction of citation influences," *ICML*, pp. 233 – 240, 2007.

[35] Mei Q, Zhai C, "Discovering evolutionary theme patterns from text: an exploration of temporal text mining," *Proceedings of the 11th ACM SIGKDD international conference*, pp. 198 – 207, 2005.

[36] Gerrish S, Blei DM, "A language-based approach to measuring scholarly impact," *ICML*, pp. 375 – 382, 2010.

[37] Lu Y, Zhai C, "Opinion Integration Through Semisupervised Topic Modeling," *Proceedings of the 17th international conference on World wide web*, pp. 121–130, 2008.

[38] Lu Y, Zhai C, Sundaresan N, "Rated Aspect Summarization of Short Comments," *Proceedings of the 18th international conference on World wide web*, pp. 131–140, 2009.

[39] Kandula S, Curtis D, Hill B, Zeng-Treitler Q, "Use of Topic Modeling for Recommending Relevant Education Material to Diabetic Patients," *AMIA Annual Symposium Proceedings*, pp. 674–682, 2011.

[40] Carroll J, "Payers Step in With Real-World Comparative Effectiveness Research," *Managed Care*, vol. 20(6), pp. 23–26, 2011.

[41] Helfand M, "Comparative effectiveness research," *Medical Decision Making*, vol. 29, p. 641, 2009.

[42] Alexander GC, Stafford RS, "Does comparative effectiveness have a comparative edge?" *Medical Decision Making*, vol. 301, pp. 2488–2490, 2009.

[43] Sox HC, Greenfield S, "Comparative effectiveness research: a report from the Institute of Medicine," *Annals of Internal Medicine*, vol. 151(3), pp. 203–205, 2009.

[44] Ratner R, Eden J, Wolman D, Greenfield S, Sox H, "Initial national priorities for comparative effectiveness research," *National Academies Pr*, 2009.

[45] Luce BR, Kramer JM, Goodman SN, Connor JT, Tunis S, Whicher D, "Rethinking randomized clinical trials for comparative effectiveness research: the need for transformational change." *Annals of Internal Medicine*, vol. 151, pp. 206–209, 2009.

[46] Thill M, "CER and Personalized Medicine Face Hurdles," *Journal of Healthcare Contracting*, 2010.

[47] Garber AM, Tunis SR, "Does comparative-effectiveness research threaten personalized medicine?" *New England Journal of Medicine*, vol. 360(19), pp. 1925–1927, 2009.

[48] Krishnan JA, Mularski RA, "Acting on comparative effectiveness research in COPD," *Journal of the American Medical Association*, vol. 304(14), pp. 1554–1556, 2010.

[49] Krishnan JA, Mularski RA, Schatz M, "A call for action: Comparative effectiveness research in asthma," *Journal of Allergy and Clinical Immunology*, vol. 2011, pp. 123–127, 2011.

[50] Pandharipande PV, Gazelle GS, "Comparative effectiveness research: what it means for radiology," *Radiology*, vol. 253, pp. 600–605, 2009.

[51] Elshaug AG, Bessen T, Moss JR et al, "Addressing "waste" in diagnostic imaging: some implications of comparative effectiveness research," *Journal of the American College of Radiology*, vol. 7, pp. 603–613, 2010.

[52] Zauber AG, Lansdorp-Vogelaar I, Knudsen AB, Wilschut J, Ballegooijen MV, Kuntz KM, "Evaluating test strategies for colorectal cancer screening: a decision analysis for the U.S. Preventive Services Task Force," *Annals of Internal Medicine*, vol. 149, pp. 659–669, 2008.

[53] Bardy GH, Lee KL, Mark DB et al, "Amiodarone or an implantable cardioverter-defibrillator for congestive heart failure," *New England Journal of Medicine*, vol. 352, pp. 225–237, 2005.

[54] D'Avolio LW, Farwell WR, Fiore LD, "Comparative effectiveness research and medical informatics," *American Journal of Medicine*, vol. 123, pp. 32–37, 2010.

[55] Djulbegovic M, Djulbegovic B, "Implications of the principle of question propagation for comparative-effectiveness and "data mining" research," *Journal of the American Medical Association(JAMA)*, vol. 305(3), pp. 298–299, 2011.

[56] Bhavnani SK, Carini S, Ross J, Sim I, "Network Analysis of Clinical Trials on Depression: Implications for Comparative Effectiveness Research," *AMIA Annual Symposium Proceedings*, pp. 51–55, 2010.

[57] Roy J, Hennessy S, "Bayesian Hierarchical Pattern Mixture Models for Comparative Effectiveness of Drugs and Drug Classes Using Healthcare Data: A Case Study Involving Antihypertensive Medications," *Statistics in BioSciences*, vol. 3, pp. 79–93, 2011.

[58] Jindal N, Liu B, "Mining Comparative Sentences and Relations," *Twenty-First Conference on Artificial Intelligence (AAAI)*, pp. 1331–1336, 2006.

[59] Jindal N, Liu B, "Identifying comparative sentences in text documents," *SIGIR*, pp. 244–251, 2006.

[60] Narayanan R, Liu B, Choudhary A, "Sentiment Analysis of Conditional Sentences," *Proceedings of Conference on Empirical Methods in Natural Language Processing(EMNLP)*, pp. 180–189, 2009.

[61] Liu B, "Sentiment Analysis and Subjectivity," *Handbook of Natural Language Processing, Second Edition*, 2010.

[62] Sun J, Long C, Zhu X, Huang M, "Mining Reviews for Product Comparison and Recommendation," *Research journal on Computer science and computer engineering with applications*, vol. 39, pp. 33–40, 2009.

[63] Li S, Zhang Z, Ming Z, Wang W, Chua T, Guo J, Xu W, "Product comparison using comparative relations," *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pp. 1151–1152, 2011.

[64] Liu B, Hu M, Cheng J, "Opinion Observer: Analyzing and Comparing Opinions on the Web," *WWW '05 Proceedings of the 14th international conference on World Wide Web*, pp. 342–351, 2005.

[65] Kim H, Zhai C, "Generating Comparative Summaries of Contradictory Opinions in Text," *Proceedings of the 18th ACM International Conference on Information and Knowledge Management (CIKM)*, pp. 385–394, 2009.

[66] Kim H, Zhai C, Han J, "Aggregation of Multiple Judgments for Evaluating Ordered Lists," *Proceedings of the 32nd European Conference on Information Retrieval (ECIR)*, pp. 166–178, 2010.

[67] Jiang Y, Lin CX, Schatz B, "Multi-class classification for online personal healthcare messages," *The 2nd International Workshop on Web Science and Information Exchange in the Medical Web*, 2011.

[68] Kraus S, Blake C, West SL, "Information Extraction from Medical Notes," *Proceedings of the 12th World Congress on Health (Medical) Informatics (MedInfo)*, pp. 1662–1664, 2007.

[69] Jiang Y, Lin CX, Mei Q, "Context Comparison of Bursty Events in Web Search and Online Media," *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011.

[70] Chee B, Berlin R, Schatz B,, "Information visualization of drug regimens from health messages," *Proceedings of HEALTHINF*, pp. 282–287, 2009.

[71] Debarr D, Wechsler H, "Social Network Analysis for Spam Detection," *International Conf. on Social Computing, Behavioral Modeling, and Prediction*, pp. 62–69, 2010.

[72] Cortes C, Vapnik V, "Support-Vector Networks," *Machine Learning*, vol. 20, pp. 273–297, 1995.

[73] Liu J, Cao Y, Lin C, Huang Y, Zhou M, "Low-quality product review detection in opinion summarization," *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP)*, pp. 334–342, 2007.

[74] McLachlan GJ, Do KA, Ambroise C, "Analyzing microarray gene expression data," *Wiley-Interscience*, 2004.

[75] Jiang Y, Liao QV, Cheng Q, Berlin R, Schatz B, "Designing and evaluating a clustering system for organizing and integrating patient drug outcomes in personal health messages," *AMIA Annual Symposium Proceedings*, pp. 417–426, 2012.

[76] Dempster AP, Laird NM, Rubin DB, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society Series B (Methodological)*, vol. 39(1), pp. 1–38, 1977.

[77] Roberts PM, Cohen AM, Hersh WR, "Tasks, topics and relevance judging for the TREC Genomics Track: five years of experience evaluating biomedical text information retrieval systems," *Information Retrieval*, vol. 12, pp. 81–97, 2009.

[78] Blake C, "Beyond genes, proteins, and abstracts: Identifying scientific claims from full-text biomedical articles," *Journal of Biomedical Informatics*, vol. 43, pp. 173–189, 2010.

[79] Fleiss JL, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76(5), pp. 378–382, 1971.

[80] Landis JR, Koch GG, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, pp. 159–174, 1977.

[81] STUDENT(Gosset WS), "The Probable Error of a Mean," *Biometrika*, vol. 6(1), pp. 1–25, 1908.

[82] Hu M and Liu B, "Mining and summarizing customer reviews," *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168–177, 2004.

[83] Lee A, Morley JE, "Metformin decreases food consumption and induces weight loss in subjects with obesity with type II non-insulin-dependent diabetes," *Obesity Research*, vol. 6(1), pp. 47–53, 1998.

[84] Woda A, Navez ML, Picard P, Gremeau C, Pichard-Leandri E, "A possible therapeutic solution for stomatodynia (burning mouth syndrome)," *Journal of Orofacial Pain*, vol. 12(4), pp. 272–278, 1998.

[85] Lecuona K, "Assessing and managing eye injuries," *Community Eye Health*, vol. 18(55), pp. 101–104, 2005.

[86] Jick J, Slone D, Borda IT, Shapiro S, "Efficacy and Toxicity of Heparin in Relation to Age and Sex," *N Engl J Med*, vol. 279, pp. 284–286, 1968.

[87] "Breast Cancer Facts & Figures 2011-2012," *American Cancer Society*, 2012.

[88] "100 Initial Priority Topics for Comparative Effectiveness Research," *Institude of Medicine of the national academics*, 2009.

[89] Pennebaker JW, Campbell RS, "The effects of writing about traumatic experience," *Clinical Quarterly*, vol. 9, pp. 17–21, 2000.

[90] Pennebaker JW, Francis ME, Booth RJ, "Linguistic Inquiry and Word Count: LIWC 2007," *Lawrence Erlbaum Assoc, New Jersey*, 2007.

[91] Rude SS, Gortner E, Pennebaker JW, "Language use of depressed and depression-vulnerable college students," *Cognition and Emotion*, vol. 18(8), pp. 1121–1133, 2004.

[92] Schatz B, Berlin RB, "Healthcare infrastructure: Health systems for individuals and populations (health informatics)," *London: Springer Verlag*, 2011.

[93] "Census Bureau Regions and Divisions with State FIPS Codes," *US Census Bureau*, 2010.

[94] Pearson K, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Philosophical Magazine*, vol. 5 50(302), 1900.

[95] Sprinthall RC, *Basic Statistical Analysis: Seventh Edition*. Cambridge, MA: Pearson Education Group, 2003.

[96] Early Breast Cancer Trialists' Collaborative Group (EBCTCG), "Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials," *The Lancet*, vol. 365, pp. 1687–1717, 2005.

[97] Segal ZV, Bieling P, Young T, MacQueen G, Cooke R, Martin L, Bloch R, Levitan RD, "Antidepressant monotherapy vs sequential pharmacotherapy and mindfulness-based cognitive therapy, or placebo, for relapse prophylaxis in recurrent depression," *Archives of General Psychiatry*, vol. 67(12), pp. 1256–1264, 2010.

[98] Yin X, Han J, Yu PS, "Truth Discovery with Multiple Conflicting Information Providers on the Web," *IEEE Trans. Knowl. Data Eng.*, vol. 20(6), 2008.

[99] Tang L, Gu Q, Yu X, Han J et al, "IntruMine: Mining Intruders in Untrustworthy Data of Cyber-physical Systems," *SDM*, pp. 600–611, 2012.

[100] Krishna S, Boren SA, Balas EA, "Healthcare via Cell Phones: A Systematic Review," *Telemedicine and e-Health*, vol. 15(1), 2009.

[101] Krishna S, Balas EA, Boren SA, Maglaveras N, "Patient acceptance of educational voice messages: a review of controlled clinical studies," *Methods Inf Med*, vol. 41(5), 2002.

[102] Adar E, Weld DS, Bershad BN, Gribble SS, "Why we search: visualizing and predicting user behavior," *Proceedings of the 16th international conference on World Wide Web(WWW)*, pp. 161–170, 2007.