UNDERSTANDING USER INTENTS IN ONLINE HEALTH FORUMS

BY

THOMAS QIAN-YUN ZHANG

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Adviser:

Professor Chengxiang Zhai

ABSTRACT

Online health forums provide a convenient way for patients to obtain medical information and connect with physicians and peers outside of clinical settings. However, the large quantities of unstructured and diversified content generated on these forums make it difficult for users to digest and extract useful information. Understanding the intents of people who post on these forums would enable the retrieval of relevant information from existing threads which would in turn allow users to more effectively find answers to their medical needs. In this paper, we derive a taxonomy of intents to capture user information need in online health forums, and propose novel pattern based features to classify original thread posts according to their underlying intents. Since no dataset existed for this task, we employ three annotators to manually tag a dataset of 1,200 HealthBoards posts spanning four topics. Experimentation finds that pattern based features are highly capable of identifying user intents in forum posts, reaching a precision of 75%. In addition, we achieve comparable classification performance by training and testing on posts from different forums, thereby showing the robustness of our method. Finally, we run our trained classifier on a MedHelp dataset to analyze the distribution of intents of different topics in the forum.

To my parents and brother, for their love and support.

ACKNOWLEDGMENTS

First and foremost, I would like to express my sincere appreciation to my adviser, Professor Chengxiang Zhai, whose knowledge, insight, and attention to detail have helped make this thesis what it is. His continuous support and encouragement have been a constant source of motivation and have allowed me to pull through the tough times.

I would like to extend my utmost gratitude to Hyun Duk Cho, who was both my closest colleague and mentor over the past year. His invaluable words of advice on all things research have been instrumental in helping me settle in as a first-year graduate student. In addition, I would like to thank Son Nguyen and Josh Friedman, who labeled the entire experimental dataset, and Adam Szmelter and Niteesh Chitturu, who labeled data that justified my formulation of the intent taxonomy.

I am also deeply appreciative to Jump Trading and the Siebel Foundation for their generous scholarships that have gone a long way in supporting my research. I would especially like to thank Illinois alumnus Steve Yi, the vice-president of Jump Trading, for his initial push for the establishment of the Jump Scholars program, and Thomas Siebel, the chairman of the Siebel Foundation, for his establishment of the Siebel Scholars program.

Finally, I am eternally grateful to my parents and brother, who have shown me unconditional love and support no matter the circumstances.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Motivation

The spread of Health 2.0 [2] technologies in the last decade has made the Internet a popular place to learn about health matters. A recent Pew survey [3] reports that 80% of web users searched for health information online, and of these, 6% have contributed to health related discussions. Many of these discussions can be found in online medical forums such as HealthBoards[1], MedHelp[2], and Wellescent[3] which provide very cost-effective ways for users to learn about health related issues outside of clinical care settings. On these forums, users can post their problems and obtain advice from both peers and health care professionals, or simply browse relevant threads. Forums are particularly valuable in the sense that they contain first hand experiences, which often have richer content than that offered by any single expert. For example, [4] finds that many physicians are unaware of the numerous alternative and complementary treatment medications found in forums discussions. This unique benefit is further confirmed in a recent study [5] that shows patients offer expertise that differs significantly from that offered by health professionals.

As the popularity of health forums continues to grow, more research is needed to better connect users with the vast quantities of information present on these forums. The key to this connection is not merely the *transmission* of information from the sender (i.e. the forum) and the receiver (i.e. the user), but instead the *exchange* of information between the two parties [6]. To facilitate this exchange, knowledge about the receiver's information need is indispensable. In the current structure of health forums, users often post questions

[1] http://www.healthboards.com/boards
[2] http://www.medhelp.org/forums/list
[3] http://wellescent.com/

to their problems despite the fact that similar discussions may have occurred in the past, in which case it would be helpful for them to read those related threads. Understanding the motivations and intentions behind users' posts would enable forums to identify such similar threads, find relevant information in those discussions, and even recommend users with comparable issues. These features would greatly enhance the utility and usability of online health forums.

In addition to improving forum functionality, knowledge of post intents would have positive ramifications for existing works that deals with information extraction from health forums. The intent of an original thread post is directly responsible for the type of content that we would expect to find in subsequent posts in the thread. Therefore, knowing the intent of the original thread posts would allow irrelevant threads to be filtered out, leading to more efficient identification of information from health forums. For example, [7] uses information from health forum posts to model the trustworthiness of a medical claim. If we understand what a particular claim is about (e.g. effectiveness of a medication), we can focus on threads where the original post wants to know about treatment options regarding that particular medication, since replies would likely contain information about . Similarly, [8] uses health forum messages to conduct Comparative Effectiveness Research (CER), which compares the benefits and harms of different prevention and treatment methods. In this case, threads where the original poster asks about treatment and side effects would contain the most information regarding those matters.

1.2 Challenges

To our knowledge, no previous work has sought to understand user intents behind original health forum thread posts. In this thesis, we frame this problem as a classification problem to make the task more tractable. However, the novelty of the problem means that no intent taxonomy exists for health forum posters. Coming up with such a taxonomy is difficult because the definition of intent (i.e. user information need) is quite ambiguous. For example, intents in general purpose questions have been broadly characterized as users seeking subjective or objective information [9, 10]. On the other hand, intents in medical

questions have been much more specifically classified into a generic taxonomy of health questions [11]. Designing the taxonomy is therefore an important and necessary first step in identifying intents from medical forum posts.

The novelty of this problem presents numerous other challenges. For starters, no annotated datasets exist, so the creation of a new dataset will require significant effort. Next, a large number of machine learning classifiers exist, so it is important to pick the one that is most suited for this problem. Once the classifier is picked, discriminative features will need to be selected that will accurately depict the characteristics of posts for each intent in the taxonomy. Finally, the method should ideally generalize to posts across all health topics, meaning that the classifier should be able to identify intents from posts from topics that it has not seen during training. All of these challenges will need to be addressed in order to solve this problem.

1.3 Contributions

In this thesis, we address all of the challenges brought forth by Section 1.2. The main contributions are threefold. First, we derive an intent taxonomy to capture the information need of health forum users who start new threads. This taxonomy is formulated from an existing study [12] of the intents of web users who seek health information online and an existing taxonomy of common clinical questions asked by doctors about patient care [1]. Second, we design a set of novel pattern based features and show that a support vector machine (SVM) classifier can use them to classify original thread posts by intent with precisions upwards of 75%. Third, we demonstrate that the classifier is capable of classifying posts from health topics not seen during training with high accuracy, thus proving the feasibility of our method to generalize to posts across all health topics.

CHAPTER 2

RELATED WORK

Previous research on developing intelligent medical question answering (QA) systems identified question understanding, framed as a classification problem, as a necessary first step. In particular, Yu et el. [13] made use of supervised learning approaches to classify questions based on the Evidence Taxonomy proposed by Ely et al. [14] and later on general topics [15], and found that including concepts and semantic types from the Unified Medical Language System (UMLS) as additional features can enhance classification results. Kobayashi and Shyu [11] classified questions into taxonomies by the Family Physicians Inquiries Network (FPIN) and the generic taxonomy proposed by Ely et al. [1], and showed that augmenting UMLS concepts and semantic types with standard parsing representations improves classification performance. Slaughter et al. [16] investigated semantic patterns of health consumers' questions and physicians' answers, and found that semantic relationships can lead to clues for creating semantic-based QA techniques. These studies all demonstrate the feasibility of using semantic features to classify health questions in medical QA systems. In this thesis, we show that semantic features can also be used to classify original thread posts in health forums according to their underlying user intent.

Subjective understanding of user intents has also been extensively studied in the context of general Community Question Answering (CQA) services. Categorizing questions into different semantic classes impose constraints on potential answers so that they can be used in later stages of the question answering process. Prominent works in this area include the novel CQA question taxonomy developed by Liu et al. [17] which expand upon Broder's taxonomy of web search queries to include both informational and social categories, the three-level question taxonomy proposed by Zhang et al. [18] that make use of interrogative patterns, hidden user intentions, and specific answer expectations to model user information

need, the semi-supervised co-training system introduced by Li et al. [9, 10] which exploits the association between questions and answers to predict whether a user is seeking subjective or objective information, and the ensuing work by Chen et al. [19, 20] which adds a new social category to Li's taxonomy and proposes a classification method using only features extracted from questions. However, these studies are wholly insufficient for our purposes as their methods uses content found on general CQA, and thus do not leverage the unique semantic information that can be found on more domain-specific platforms such as medical forums. In addition, the proposed taxonomies in these studies are irrelevant to the health domain and thus cannot be used to describe the intents of medical forum users.

In addition to questions, web search engine queries have also been used to study user intents. Cartright et al. [21] explored information goals and patterns of attention in web exploratory health search (EHS) through analysis of search sessions. They identify EHS sessions, extract different intentions persisting as foci of attention from those sessions, and demonstrate how this knowledge can be used to better understand EHS behavior and support health search on the web. Similarly, other works such as [22, 23, 24] have also used interaction logs to study web search behavior, but none have focused on identifying medical query intent. In general purpose search, Broder's seminal work [25] finds that user query goals can be classified into a trichotomy of web search types: information, navigational, and transactional. Subsequent works such as [26, 27, 28, 29] show that various automatic learning-based approaches can be used to produce solid predictive performance in classifying queries. Here we note several key differences between query and forum data. First, queries often consist of discrete keywords whereas forum posts are formulated in natural language, reflecting the discrepancy between their intended audiences. Second, users who type queries typically possess some specific "need" [25] whereas the intents of users who post on forums may not be as clear-cut. Our method solves these problems by taking into account the characteristics of forum post content as well as the intrinsic intents of users in online health forums.

CHAPTER 3

TAXONOMY

3.1 Motivation

Ely et al. [1] developed a taxonomy of doctor's questions about patient care consisting of 64 generic question types. Their taxonomy aims to completely capture the information need of doctors during patient visits. Boot and Meijman [30] investigated the feasibility of using this taxonomy to classify health questions asked by the public. In the process, they ran into many difficulties, most of which are attributed to the difference in information need and question specificity between patients and professionals. For example, there exists no suitable category in Ely's taxonomy for questions about standard medical knowledge (e.g. What can I expect during treatment x?"), despite them being among the most popularly asked questions by patients. In addition, patients often tend to ask vaguer questions than doctors would due to their lack of medical knowledge. This in turn becomes problematic during classification since Ely's taxonomy contains categories with very similar meanings (e.g. "What is the cause of symptom x?" and "Could this patient have condition y?").

3.2 Formulation

Boot and Meijman's study confirms the need for a new taxonomy specifically designed for the general public. Choudhury et al. [12] examined the intents of 197 survey respondents who seek health information online using search engines. They find that the most common motivations are to identify treatment options, diagnose health conditions, understand health conditions or procedures, and understand medications, in that order. Going back to the taxonomy proposed by Ely et al., we find that the top 10 most commonly asked generic

Table 3.1: Top 10 generic questions by primary care doctors from [1].

Rank	Question
1	What is the drug of choice for condition x?
2	What is the cause of symptom x?
3	What test is indicated in situation x?
4	What is the dose of drug x?
5	How should I manage condition x (not specifying diagnostic or therapeutic)?
6	What is the cause of physical finding x?
7	How should I treat condition x (not limited to drug treatment)?
8	What is the cause of test finding x?
9	Could this patient have condition x?
10	Can drug x cause (adverse) finding y?

questions by doctors (shown in Table 3.1) can be clustered into groups with related intents. The clustering is as follows. (2), (6), (8), and (9) are reduced into the intent class "What is the cause of symptom, physical finding, or test finding x?". (1), (4), (5), (7) are reduced to the intent class "How should I manage or treat condition x?" (note that (1) and (4) are essentially questions pertaining to treatment). (10) becomes its own standalone class, and (3) is discarded because it corresponds to questions that only doctors would ask. We next propose two additional classes to add to this taxonomy: "Combination." to account for multiple intents, and "Story telling.", to account for ambiguous or no intent.

Note that our formulated taxonomy classes approximately match the most common user motivations found by Choudhury et al. Here, we argue that "What is the cause of symptom, physical finding, or test finding x?" maps to diagnosing health conditions, "How should I manage/treat condition x?" maps to identifying treatment options, and "Can drug/treatment x cause (adverse) finding y?" maps to understanding medications and procedures. This mapping effectively validates the ability of our taxonomy to capture the intents of users who seek health information online.

3.3 Summary

In this section, we explain the proposed taxonomy as it pertains to original thread posts in health forums. Specifically, for each class in the taxonomy, we describe the user intent in

layman's terms, and provide an example of a post for that class.

Manage: **How should I manage or treat condition X?** Posts with this intent ask about general information related to treatment options or ways to manage or alleviate symptoms of certain conditions.

Example: Hello ive found out through many self test that i have depression i know i should see a councler but i feel i shouldn't i dont want to tell my parents because they think im a happy person i just dont know what to do at this point does anyone else know how i should get through this?

Cause: **What is the cause of symptoms, physical finding, or test findings X?** Posts with this intent generally ask for diagnosis regarding health symptoms.

Example: My husband has been waking up with a slight stuffy nose that he says feels like pressure at times and has a slight headache. He has some drainage that goes down his throat and he says that he has some congestion. Does this sound typical of allergies? The weather has been really changing alot here and he wanted to know if that was all allergy related. I wasn't sure. :wave:

Adverse: **Can drugs or treatments X cause (adverse) finding Y?** Posts with this intent typically ask about the side effects of drugs or treatments, including withdrawal effects.

Example: I hear people takling about how certain nasal sprays has steroids in them which could be bad for you if you continue to take it regularly. Are these the OTC nasal sprasy? I assume astelin, flonase, nasonex and other prescription nasal sprays are okay to take regularly?

Combo: **Combination ($\geq$2 of manage, cause, or adverse findings).** Posts in this category possess multiple intents that usually come from multiple medical inquiries.

Example: i have had a constant pain in my chest and sometimes my neck. What is happening? and right now im having a pain in the center of my chest and shortness of breath

and my heads kind of spinning what should i do?

Story: **Story telling, news, sharing or asking about experience, soliciting support, or others.** Posts in this category usually consists of experience or news sharing, personal rants to garner support, or off-topic inquiries.

Example: Everyone will lie to me. Everyone wants stuff from me and gives little in return. The ONLY person I can count on is me. Living again on klonopin. Thank God for it. But I feel like a zombie. Like I am not really here. Scared to be here though. All I want to do is turn the ac way up, get my room dark as possible, crawl in bed with my dogs and sleep. Thanks for listening.

CHAPTER 4

PROBLEM FORMULATION

Define O as an original thread post with intent c_i from a taxonomy of intents $C = \{c_1, ..., c_k\}$, and let $S = (s_1, ..., s_n)$ denote the sentence representation of O. We classify O as some $c_j \in C$ using S as evidence. O is correctly classified if and only if $j = i$.

Note that this formulation does not identify all intents from posts with multiple intents, which have a lot of overlap between them and in general are more difficult to identify. We decide to use this simplified formulation to focus on designing features for classifying posts into single classes. Posts with multiple intents (i.e. *Combo* posts) will be considered to be correctly classified if one of its intents match the predicted intent. Identification of multiple intents will be left for future work.

In addition, note that this formulation excludes using thread titles for classification. After manually annotating all 1193 posts in our evaluative dataset for (1) if the title is discriminative (i.e. it signifies a clear intent) and (2) given (1) is true, if the title agrees with the intent of the post (i.e. the post signifies the exact same intent of that in the title, and no other intents), we find that 137/1193 posts possess discriminative titles, and that a large fraction of all posts (31/137, 22.63%) exhibit conflicting intents between themselves and their titles. From this fact, we conclude that titles should not be used in classification due to the propensity of their intents to disagree with those of the posts.

CHAPTER 5

METHODOLOGY

Our classification method is based on the classic supervised learning framework. To do so, we design various features to capture clues in posts that will help in identifying their intents. We then apply these features to our dataset to construct a feature representation of each post, and separate these representations into discrete training and test sets. Finally, we train a classifier using the training set and evaluate on the test set. For each post in the test set, the classifier computes a score for each class, and the class with the highest score is assigned to that post.

As our goal is to study the effectiveness of features in classification, we decide to use the popular Support Vector Machine (SVM) classifier. We assume that our choice of features will generalize well to other classifiers, leaving experimentation with different classifiers as future work.

5.1 Support Vector Machines

Support vector machines first introduced in [31] are binary classifiers that construct hyperplanes to separate training instances belonging to two classes. SVMs maximize the separation margin between this hyperplane and the nearest training data points of any class. The larger the margin, the lower the generalization error of the classifier. SVMs can efficiently perform both linear and non-linear classification, and have shown to have good performance on high dimensional data. In our experiments, we employ the LIBSVM [32] implementation with a RBF kernel, and train classifiers using a one-versus-all multiclass approach.

5.2 A Hierarchical Classifier

Our hierarchical classifier makes use of a sequence of two cascading SVM classifiers using pattern and word features (features are described in more detail in Chapter 6). The first classifies posts that match at least one pattern feature into one of *Manage*, *Cause*, *Adverse* intent classes (*Pattern Classifier*), while the second classifies all posts that do not match any pattern features into one of *Manage*, *Cause*, *Adverse*, *Story* intent classes using word features (*Word Classifier*). The hierarchical classifier classifies all posts in the dataset which allows us to compare its overall performance in comparison with that of the baseline word classifier.

CHAPTER 6

FEATURES

In this chapter, we first describe a baseline of unigram word features and then propose four novel pattern based feature sets that may be useful for the classification task.

6.1 Word Features

The baseline features are based on the traditional bag-of-words model [33], a simplifying representation used in natural language processing (NLP) and information retrieval (IR). In this model, text is represented as a set of its words, disregarding grammar and word order but keeping multiplicity. This model is often used in methods of document classification, where the occurrence of each word in the document is weighted by some scheme and used as a feature for training a classifier.

For our experiments, we use unigram word features weighted with tf-idf [34], a numerical statistic intended to reflect how important a word is to a document in a corpus. The tf-idf value increases proportionally to the number of times a word appears in a document (term frequency), but is offset by the frequency of that word in the corpus (inverse document frequency), which helps to account for the fact that some words are generally more common than others.

6.2 Pattern Features

6.2.1 Motivation

During the data labeling process, we observe recurring sentence patterns in posts from different intent classes and find that they are great indicators of user intent. For example, finding the pattern "what could X be..." in a post signifies strong *Cause* intent, but finding "what can X do..." would suggest more *Manage* intent. These observations lead us to believe that patterns could have significant discriminative power in identifying forum post intent.

6.2.2 Formal Definition

We define a pattern to be a sequence of slots $S = (s_1, ..., s_n)$, $|S| \geq 1$, where each slot must be filled by a token from one of four types: *Lowercase* (LT), *Stemmed* (ST), *Part-of-Speech* (POST), and *Semantic Group* (SGT). Patterns may or may not allow additional non-matching tokens between their slots. The relative position of a pattern may also be specified (i.e. start of a sentence, middle, or end). A pattern is matched if every $s_i, 1 \leq i \leq |S|$ matches a token within a single sentence in the same order. For all pattern features, we use binary weights (i.e. 1 if a pattern matches, 0 if it doesn't), because it is rare for a pattern to be matched more than once in a post.

6.2.3 Pattern Identification

To capture the intuition from Section 6.2.1, we look carefully into frequent patterns in the dataset and manually compile a list of patterns that we think are most representative of the *Manage*, *Cause*, and *Adverse* intent classes. Unfortunately, we are unable to identify a rich enough set of features for *Story* posts due to the large variations in their content. Finally, we remove repeated patterns and merge the lists into one feature set.

6.2.4 MetaMap

The Unified Medical Language System[1] (UMLS) Metathesaurus, the largest thesaurus in the biomedical domain, provides a representation of biomedical knowledge consisting of more than one million concepts classified by semantic type and relationships among the concepts. To make it easier for users to integrate this knowledge into their applications, the National Library of Medicine (NLM) has developed MetaMap [35], a highly configurable program to map biomedical text to Metathesaurus concepts and their associated semantic types. In this paper, we use the MetaMap API to replace text phrases in health forum posts by their semantic group labels, as described in the next section.

6.2.5 Data Preprocessing

We construct four different data representations for each original thread post in the dataset. These representations are used to construct the feature sets in Section 6.2.6. The first consists of tokenizing and lowercasing the sentences. The second and third consist of stemming and POS tagging the data from the first. To construct the fourth representation, we feed the original thread posts into the MetaMap API to generate phrase to semantic type mappings. Only a small subset of the semantic types within the Metathesaurus is considered (shown in Table 6.1). We then map these semantic types to their corresponding semantic groups, replace all mapped phrases by these groups, and tokenize and lowercase the data.

6.2.6 Feature Sets

We divide the pattern based feature set into four discrete sets, each containing patterns with a different mix of token types: (1) patterns with LT and ST tokens (LSP), (2) patterns with LT, ST, and POST (POSP), (3) patterns with LT, ST, and and SGT (SGP), and (4) patterns with all four token types (ALL). In general, we characterize patterns from these sets to roughly have increasing discriminative power in classification.

[1] http://www.nlm.nih.gov/research/umls

Table 6.1: UMLS semantic types considered, with their corresponding semantic groups.

Group Abbv	Group Name	Type Name
CHEM	Chemicals & Drugs	Steroid
CHEM	Chemicals & Drugs	Pharmacologic Substance
CHEM	Chemicals & Drugs	Antibiotic
CHEM	Chemicals & Drugs	Clinical Drug
PROC	Procedures	Therapeutic or Preventive Procedure
PROC	Procedures	Health Care Activity
PROC	Procedures	Diagnostic Procedure
DISO	Disorders	Disease or Syndrome
DISO	Disorders	Pathologic Function
DISO	Disorders	Sign or Symptom
DISO	Disorders	Neoplastic Process
DISO	Disorders	Acquired Abnormality
DISO	Disorders	Congenital Abnormality
DISO	Disorders	Mental or Behavioral Dysfunction

LSP These patterns consist of only lowercase and stemmed tokens. Some patterns in this set are very specific to a particular intent (e.g. "...what can cause..."), while others are more general (e.g. "how does..."), which means they are more likely to match posts belonging to different intent classes.

POSP These patterns contain both lowercase and stemmed tokens and part-of-speech (POS) tags. We use POS tags to replace certain words in the pattern (e.g. "...how to ⟨VERB⟩..."), which allows for more flexible matching.

SGP These patterns contain both lowercase and stemmed tokens and semantic group labels. Replacing medical terminology with more general labels saves us from having to explicitly enumerate every possibility. For example, the pattern "...if ⟨CHEM⟩ works..." replaces all patterns where "⟨CHEM⟩" is some drug or medication.

ALL These patterns are the most expressive because they contain the richest mix of token types (e.g. "...⟨CHEM⟩ makes ⟨PRP⟩ feel...", where ⟨PRP⟩ replaces a personal pronoun).

CHAPTER 7

EVALUATION

7.1 Data

The novelty of the classification task means that there is no existing dataset available. As a result, we create a new dataset consisting of a collection of 1,200 original thread posts from HealthBoards. Although a larger dataset would be more ideal, we settle for 1,200 posts due to limited resources for data labeling. These posts are evenly divided between four topics: allergies, breast cancer, depression, and heart disease. We split the dataset between four topics because we want to have good mix of posts from both major and minor health disorders. Next, we filter out all posts with empty or incomplete content, ending up with 1,192 posts.

7.1.1 Gold Labeling

To create a gold standard for this dataset we employ two humans to label the dataset with our proposed 5-class taxonomy. We employ a third human to label *Combo* posts with at least two classes from {*Manage*, *Cause*, *Adverse*}. The final distribution of gold standard labels for 5-class labeling and *Combo* labeling are shown in Tables 7.1 and 7.2, respectively.

Inter Annotator Agreement. We have the first two labelers each label 75 posts from the dataset according to the 5-class taxonomy. A majority of the disagreement comes from labeling a post as either *Manage* or *Story*. This arises from the fact that many users that ask about "managing" problems also tend to share their own experiences and/or ask about other people's experiences, thus making classification difficult. Nevertheless, after calculating the

Table 7.1: Distribution of 5-class gold labels.

Forum	Manage	Cause	Adv.	Combo	Story
Allergies	90	99	15	37	58
Br. Cancer	79	94	14	18	92
Depression	112	35	45	34	73
Heart Diso.	63	108	11	41	74
Total	344	336	85	130	297

Table 7.2: Distribution of gold labels for *Combo* posts. MC, MA, CA, and MCA correspond to the different combinations of *(M)anage*, *(C)ause*, and *(A)dverse*.

Forum	MC	MA	CA	MCA
Allergies	27	5	0	4
Breast Cancer	16	2	0	1
Depression	16	12	4	3
Heart Disorder	35	4	1	0
Total	94	23	5	8

observed agreement and Cohen's Kappa [36], we find that the labeling results match quite well (56/75, $\approx 74.67\%$), with $\kappa = 0.665$, indicating substantial agreement per Landis and Koch [37]. Given the fuzzy nature of the task at hand, this κ value is certainly satisfactory.

7.2 Experimental Setup

In this section, we first describe the experimental setup to compare the performances of the pattern classifier using different combinations of pattern features. Next, we explain how we setup experiments to compare the performances of the word classifier baseline with the hierarchical classifier using both 5-fold cross validation and 4-fold forum cross validation.

For all of our experiments, we exclude *Combo* posts for training because we want only the most discriminative data from the training set. However, we use every post in the dataset for testing. We consider a *Combo* post to be correctly classified if its predicted class label matches at least one of its gold labels. Otherwise, we pick the first gold label in the order of *Manage*, *Cause*, and *Adverse* and consider the post to be misclassified for that particular class.

7.2.1 Feature Space Selection

This experiment aims to find a combination of pattern features that gives the best performance by evaluating our pattern classifier over six different pattern feature set combinations: (1) LSP (baseline), (2) LSP+POSP, (3) LSP+SGP, (4) LSP+ALL, (5) LSP+PSOP+SGP, and (6) LSP+POSP+SGP+ALL. We choose to evaluate only these feature space combinations because the others are not large enough for classification. For each feature space, we perform 5-fold cross validation by training our classifier using only *Manage*, *Cause*, *Adverse* posts that match at least one pattern from four folds, and test using posts from the last fold.

7.2.2 5-Fold Cross Validation

This experiment evaluates the performance of each individual classifier in the hierarchical setup separately and compares the overall performance of the hierarchical classifier with that of the baseline word classifier. We construct five equally sized folds from the dataset and perform standard 5-fold cross validation on both the word classifier (baseline), and the hierarchical classifier. Baseline cross validation involves training the word classifier using *Manage*, *Cause*, *Adverse*, and *Story* posts from four folds, and testing using posts from the last fold. Hierarchical cross validation involves first cross validating the pattern classifier by training it on *Manage*, *Cause*, and *Adverse* posts from four folds and testing it using the posts from the last fold. We then cross validate the word classifier by training it on four classes (excluding *Combo*) and using it to classify posts that do not match any patterns.

7.2.3 4-Fold Forum Cross Validation

The previous section describes standard cross validation. However, we would also like to evaluate the performance of our classifier when it is tested on posts from forums that it has not been trained on. This experiment evaluates the capacity of the classifier to predict the intents of posts from forums not seen in training, which is akin to how the classifier will likely be used in real life scenarios. To do so, we separate the posts from the four forums

Table 7.3: Performance of pattern classifier using different feature set combinations.

No.	Feat. Space	Tot.	Cor.	P	R	F1
1	LSP(BL)	364	263	72.25	29.39	41.78
2	(1)+POSP	427	321	**75.18**	35.87	48.57
3	(1)+SGP	422	306	72.51	34.19	46.47
4	(1)+ALL	366	263	71.86	29.39	41.72
5	(2)+SGP	479	356	74.32	39.78	51.82
6	(5)+ALL	481	361	75.05	**40.34**	**52.47**

(allergies, breast cancer, depression, and heart disease) into four folds. we then evaluate the performance of our classifier by performing 4-fold forum cross validation (i.e. training the classifier using posts from three forums and testing it on the last forum).

7.3 Results

The experimental results are summarized in Tables 7.3-7.5. In this section, we first describe the results from our investigation of the performance of various feature spaces, then explain the cross validation results in more detail.

7.3.1 Feature Space Selection

Table 7.3 compares the performance of the pattern classifier for each feature set combination. Unsurprisingly, we find that as we add more pattern sets into our feature space, the total number of posts that match at least one pattern (and therefore will be able to be classified by the pattern classifier) increases. The accuracy of the classifier, however, remains relatively constant. We pick the 6th feature space for use in the rest of the experiments because it gives the highest number of matches without sacrificing performance.

7.3.2 Cross Validation

In this section, we summarize our cross validation results.

Table 7.4: Baseline cross validation results.

	5-Fold CV			4-Fold Forum CV		
Intent	P	R	F1	P	R	F1
Manage	58.25	62.85	52.48	54.34	61.15	57.55
Cause	61.92	59.75	61.28	61.20	53.65	57.18
Adverse	39.47	29.41	37.80	35.29	24.24	28.74
Story	39.54	40.74	42.16	37.31	41.08	39.10
Overall	**53.44**			**50.59**		

Table 7.5: Hierarchical classifier CV results.

	Pattern Classifier					
	5-Fold CV			4-Fold Forum CV		
Intent	P	R	F1	P	R	F1
Manage	75.51	36.72	49.41	72.59	34.96	48.38
Cause	73.53	43.86	54.95	72.72	42.75	53.17
Adverse	80.85	40.86	54.29	71.70	40.86	48.41
3-Class	**75.05**	40.34	52.47	**72.55**	38.99	50.72
	Word Classifier					
	5-Fold CV			4-Fold Forum CV		
Intent	P	R	F1	P	R	F1
Manage	45.63	52.51	48.83	44.16	54.75	48.89
Cause	47.15	48.92	48.02	48.13	41.85	44.77
Adverse	25.93	15.22	19.18	19.23	10.87	13.89
Story	45.85	43.46	45.29	45.42	43.85	44.62
Overall	45.85			44.59		
	Overall Performance					
	5-Fold CV			4-Fold Forum CV		
Intent	P	R	F1	P	R	F1
Manage	58.71	65.26	61.81	56.05	64.55	60.00
Cause	61.72	66.67	64.10	62.66	62.34	62.50
Adverse	60.81	48.39	53.89	54.43	46.24	50.00
Story	47.28	38.05	42.16	45.42	38.38	41.61
Overall	**57.63**			**55.87**		

Hierarchical classifier achieves a slight improvement over baseline word classifier. From Tables 7.4 and 7.5 we can see that the hierarchical classifier yields a 4.19% improvement (57.63% vs 53.44%) over the baseline for 5-fold cross validation, and a 5.28% improvement (55.87% vs 50.59%) over the baseline for 4-fold forum cross validation. The reason for this underwhelming result is due to the word classifier in the hierarchical setup dragging down overall performance by performing much poorer than the baseline classifier.

Pattern classifier achieves high precision but low recall. Perhaps the most important result is the performance of the pattern classifier, which achieves precisions of 75.05% and 72.55% for 5-fold cross validation and 4-fold forum cross validation respectively, albeit with relatively low recalls of 40.34% and 38.99%. However, we argue that a high precision, low recall classifier is acceptable for our task, since we would much rather predict the intent of fewer posts with high accuracy than more posts with lower accuracy. Further work is needed to handle classification of posts that do not match patterns.

Pattern classifier achieves comparable performance in 4-fold forum cross validation. From Table 7.5 we see that the pattern classifier achieves comparable performance when it is trained exclusively on posts from three forums and tested on the last forum with that from training and testing on all four forums. This result demonstrates the ability of our method to generalize to posts from forums that are not represented in the training set, and allows us to claim that the pattern classifier can accurately identify the intents of posts across different forum topics.

Word classifier in hierarchical setup performs worse than baseline word classifier. We see from Tables 7.4 and 7.5 that the word classifier in the hierarchical setup performs much worse than the baseline word classifier. This clearly demonstrates that the word classifier is unable to handle classification of posts that do not match patterns. We believe that this performance drop is due to the fact that test posts that do not match any patterns possess more ambiguous intents than those that do, and are therefore harder to classify.

Word classifier fails at identifying *Adverse* and *Story* intents. The unigram classifier performs poorly on posts with *Adverse* and *Story* intent. The *Adverse* intent class contains very few data points, while posts with *Story* intent inherently display either highly ambiguous intent or no intent, and possess little to no distinguishing word features. These factors help explain the low performance of the classifier.

CHAPTER 8

DISTRIBUTION ANALYSIS

8.1 HealthBoards Dataset

Table 8.1 shows the intent distribution of gold labels from our HealthBoards dataset for the *Manage*, *Cause*, and *Adverse* classes in comparison with that of gold labels from posts matched by our pattern classifier from Section 7.2.2. From the data, we see that the distribution of gold labels from posts that match at least one pattern is very similar to that of gold labels from all posts in the dataset. We can extend this fact to make a general claim that posts that match at least one pattern from *any* health forum will have a distribution very close to that of the entire forum corpus.

8.2 MedHelp Dataset

Since our hierarchical classifier does not give good enough performance, we cannot use it to classify unlabeled posts. Instead, we train a 3-class (*Manage*, *Cause*, and *Adverse*) pattern classifier using our HealthBoards dataset and run the classifier on a collection of 61,225 posts that we crawled from MedHelp. These posts come from MedHelp forums corresponding to the same topics as those in our HealthBoards dataset (i.e. allergy, breast cancer, depression,

Table 8.1: Distribution of gold intents from HealthBoards dataset.

Topic	Total	% Total	Match	% Match
Manage	344	44.97	165	46.61
Cause	336	43.92	150	42.37
Adverse	85	11.11	39	11.02
Total	765	100.00	354	100.00

Figure 8.1: Distribution of classified intents for all four MedHelp forums.

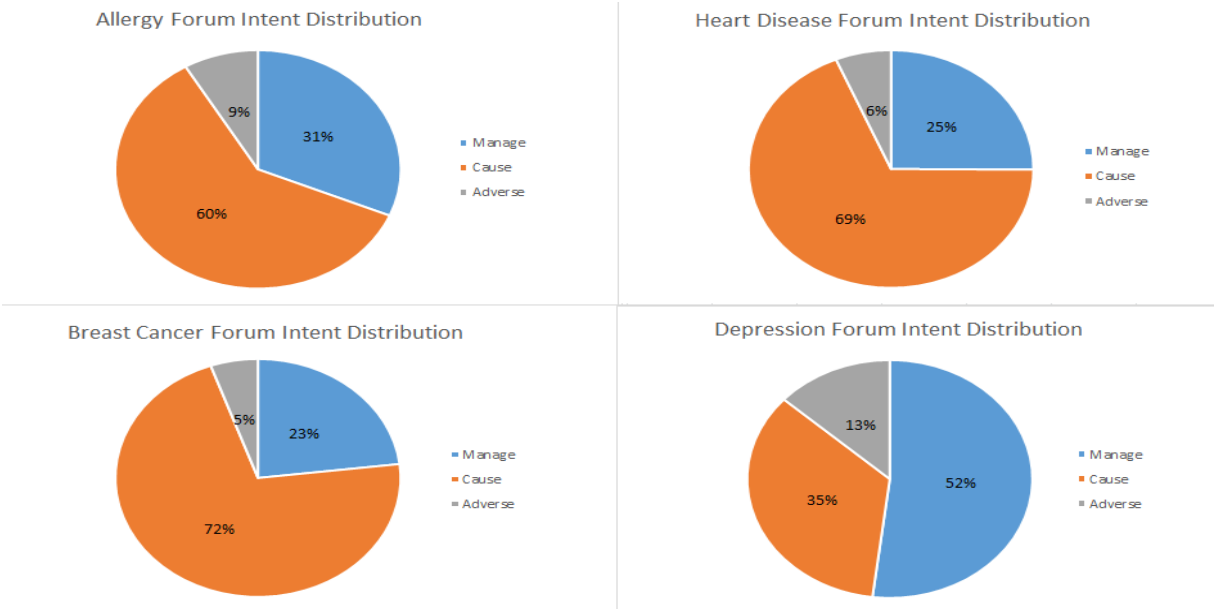

Table 8.2: Post count for each MedHelp forum.

Allergy	Br. Cancer	Depression	Heart Dis.	Total
9,895	12,647	9,830	28,853	61,225

and heart disease). Table 8.2 shows the total number of posts in each forum, while Figure 8.1 shows the distribution of classified intents for posts in all four forums. From these statistics, we can make several observations:

***Cause* make up a strong majority in 3/4 forums.** Allergy, breast cancer, and heart disease forum users seem to start more threads looking for a diagnosis than threads with any other intent. This is possibly due to most users wanting to explore what is causing their symptoms before consulting a medical professional.

***Manage* make up a majority in depression forum.** Much in contrary to the other three forums, the depression forum contains a greater number of post with *Manage* intent than any other intent. This is possibly due to depressed patients being more concerned with finding ways to mitigate their symptoms instead of trying to figure out why they have depression.

Depression contains the greatest proportion of side effect posts. The depression forum contains a greater percentage of posts with *Adverse* intent (13%) than any other forum (allergy 9%, breast cancer 5%, heart disease 6%). This is possibly due to users being worried that their depression symptoms may be attributed to certain medications given that many list depression as a side effect.

Allergy forum contains a smaller ratio of *Cause* to *Manage* posts. The ratio of the number of posts with *Cause* intent to that of posts with *Manage* intent is smaller in the allergy forum than in the breast cancer and heart disease forums. This is possibly due to patients asking about ways to alleviate their allergy symptoms.

CHAPTER 9

CONCLUSION AND FUTURE WORKS

We have presented a supervised approach to identifying user intents from original health forum thread posts. Previous works have focused on understanding intents in questions and search engine queries which are largely different problems. The novelty of our task means that we had to derive an new intent taxonomy which we showed is able to capture the information needs of health forum users. Furthermore, we had to construct a new labeled dataset for evaluation which can be reused in future works. Our experiments demonstrated that simple unigram features cannot adequately discriminate between posts with different intents, and that pattern based features are capable of classifying posts with high accuracy. In addition, we showed that our classifier produces comparable classification performance on posts from health topics not seen during training, thus demonstrating the robustness of our method.

Future work should be directed towards four areas. First, we believe that it is possible to expand upon our proposed taxonomy to include more specific intent categories. Having finer classes will allow us to gauge a better understanding of the intents of users in health forums. Second, we can expand upon the coverage of pattern features to improve recall performance. In other words, we need to be able to accurately classify posts that don't currently match features. Third, we should look into ways to identify posts with multiple intents and *Story* posts. In particular, for posts with more than one intent, our method should be able to identify all of their intents. Finally, upon successful completion of the first three areas, we plan on using the classifier to analyze the distribution of post intents from all health forum topics. This study will give us insight into the makeup of user information needs for different medical topics.

REFERENCES

[1] J. W. Ely, J. A. Osheroff, P. N. Gorman, M. H. Ebell, M. L. Chambliss, E. A. Pifer, and P. Z. Stavri, "A taxonomy of generic clinical questions: Classification study," *British Medical Journal*, vol. 321, no. 7258, pp. 429–432, 2000.

[2] T. H. Van De Belt, L. J. Engelen, S. A. Berben, and L. Schoonhoven, "Definition of health 2.0 and medicine 2.0: a systematic review." *Journal of medical Internet research*, vol. 12, no. 2, 2010.

[3] S. Fox, "The social life of health information," *The Pew Internet & American Life Project*, May 2011. [Online]. Available: http://www.pewinternet.org/2011/05/12/the-social-life-of-health-information-2011/

[4] G. Eysenbach, "The impact of the internet on cancer outcomes," *CA: A Cancer Journal for Clinicians*, vol. 53, no. 6, pp. 356–371, 2003. [Online]. Available: http://dx.doi.org/10.3322/canjclin.53.6.356

[5] A. Hartzler and W. Pratt, "Managing the personal side of health: how patient expertise differs from the expertise of clinicians," *Journal of medical Internet research*, vol. 13, no. 3, 2011.

[6] R. G. Lee and T. Garvin, "Moving from information¡ i¿ transfer¡/i¿ to information¡ i¿ exchange¡/i¿ in health and health care," *Social science & medicine*, vol. 56, no. 3, pp. 449–464, 2003.

[7] V. V. Vydiswaran, C. Zhai, and D. Roth, "Gauging the internet doctor: Ranking medical claims based on community knowledge," in *Proceedings of the 2011 Workshop on Data Mining for Medicine and Healthcare*, ser. DMMH '11. New York, NY, USA: ACM, 2011. [Online]. Available: http://doi.acm.org/10.1145/2023582.2023589 pp. 42–51.

[8] J. H. Cho, V. Q. Liao, Y. Jiang, and B. R. Schatz, "Aggregating personal health messages for scalable comparative effectiveness research," in *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, ser. BCB'13. New York, NY, USA: ACM, 2013. [Online]. Available: http://doi.acm.org/10.1145/2506583.2512363 pp. 907:907–907:916.

[9] B. Li, Y. Liu, and E. Agichtein, "Cocqa: Co-training over questions and answers with an application to predicting question subjectivity orientation," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. [Online]. Available: http://dl.acm.org/citation.cfm?id=1613715.1613836 pp. 937–946.

[10] B. Li, Y. Liu, A. Ram, E. V. Garcia, and E. Agichtein, "Exploring question subjectivity prediction in community qa," in *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '08. New York, NY, USA: ACM, 2008. [Online]. Available: http://doi.acm.org/10.1145/1390334.1390477 pp. 735–736.

[11] T. Kobayashi and C.-R. Shyu, "Representing clinical questions by semantic type for better classification," in *AMIA Annual Symposium Proceedings*, vol. 2006. American Medical Informatics Association, 2006, p. 987.

[12] M. De Choudhury, M. R. Morris, and R. White, "Seeking and sharing health information online: Comparing search engines and social media," 2014.

[13] H. Yu, C. Sable, and H. R. Zhu, "Classifying medical questions based on an evidence taxonomy," in *Proc. AAAI'05 Workshop on Question Answering in Restricted Domains*, 2005. [Online]. Available: http://www.uwm.edu/ hongyu/publications.html

[14] J. Ely, J. A. Osheroff, M. H. Ebell, M. L. Chambliss, D. Vinson, J. J. Stevermer, and E. A. Pifer, "Obstacles to answering doctors' questions about patient care with evidence: Qualitative study," *BMJ*, vol. 324, no. 7339, p. 710, 2002. [Online]. Available: http://www.bmj.com/cgi/content/full/324/7339/710

[15] H. Yu and Y.-g. Cao, "Automatically extracting information needs from ad hoc clinical questions," in *AMIA Annu Symp Proc.*, 2008, pp. 96–100.

[16] L. A. Slaughter, D. Soergel, and T. C. Rindflesch, "Semantic representation of consumer questions and physician answers." *I. J. Medical Informatics*, vol. 75, no. 7, pp. 513–529, 2006.

[17] Y. Liu, S. Li, Y. Cao, C.-Y. Lin, D. Han, and Y. Yu, "Understanding and summarizing answers in community-based question answering services," in *Proceedings of the 22Nd International Conference on Computational Linguistics - Volume 1*, ser. COLING '08. Stroudsburg, PA, USA: Association for Computational Linguistics, 2008. [Online]. Available: http://dl.acm.org/citation.cfm?id=1599081.1599144 pp. 497–504.

[18] Y. Zhang, X. Wang, X. Wang, S. Fan, and D. Zhang, "Using question classification to model user intentions of different levels," in *Proceedings of the 2009 IEEE International Conference on Systems, Man and Cybernetics*, ser. SMC'09. Piscataway, NJ, USA: IEEE Press, 2009. [Online]. Available: http://dl.acm.org/citation.cfm?id=1732323.1732520 pp. 1153–1158.

[19] L. Chen, D. Zhang, and M. Levene, "Question retrieval with user intent," in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '13. New York, NY, USA: ACM, 2013. [Online]. Available: http://doi.acm.org/10.1145/2484028.2484129 pp. 973–976.

[20] L. Chen, D. Zhang, and L. Mark, "Understanding user intent in community question answering," in *Proceedings of the 21st International Conference Companion on World Wide Web*, ser. WWW '12 Companion. New York, NY, USA: ACM, 2012. [Online]. Available: http://doi.acm.org/10.1145/2187980.2188206 pp. 823–828.

[21] M.-A. Cartright, R. W. White, and E. Horvitz, "Intentions and attention in exploratory health search," in *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ser. SIGIR '11. New York, NY, USA: ACM, 2011. [Online]. Available: http://doi.acm.org/10.1145/2009916.2009929 pp. 65–74.

[22] S. K. Bhavnani, R. T. Jacob, J. Nardine, and F. A. Peck, "Exploring the distribution of online healthcare information," in *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '03. New York, NY, USA: ACM, 2003. [Online]. Available: http://doi.acm.org/10.1145/765891.766009 pp. 816–817.

[23] A. Spink, Y. Yang, J. Jansen, P. Nykanen, D. P. Lorence, S. Ozmutlu, and H. C. Ozmutlu, "A study of medical and health queries to web search engines," *Health Info Libr J*, vol. 21, no. 1, pp. 44–51, Mar 2004.

[24] S. L. Ayers and J. J. Kronenfeld, "Chronic illness and health-seeking information on the Internet," *Health (London)*, vol. 11, no. 3, pp. 327–347, Jul 2007.

[25] A. Broder, "A taxonomy of web search," *SIGIR Forum*, vol. 36, no. 2, pp. 3–10, Sep. 2002. [Online]. Available: http://doi.acm.org/10.1145/792550.792552

[26] B. J. Jansen, D. L. Booth, and A. Spink, "Determining the informational, navigational, and transactional intent of web queries," *Inf. Process. Manage.*, vol. 44, no. 3, pp. 1251–1266, May 2008. [Online]. Available: http://dx.doi.org/10.1016/j.ipm.2007.07.015

[27] U. Lee, Z. Liu, and J. Cho, "Automatic identification of user goals in web search," in *Proceedings of the 14th International Conference on World Wide Web*, ser. WWW '05. New York, NY, USA: ACM, 2005. [Online]. Available: http://doi.acm.org/10.1145/1060745.1060804 pp. 391–400.

[28] R. Baeza-Yates, L. Calderón-Benavides, and C. González-Caro, "The intention behind web queries," in *Proceedings of the 13th International Conference on String Processing and Information Retrieval*, ser. SPIRE'06. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 98–109.

[29] I.-H. Kang and G. Kim, "Query type classification for web document retrieval," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 64–71.

[30] C. R. Boot and F. J. Meijman, "Classifying health questions asked by the public using the icpc-2 classification and a taxonomy of generic clinical questions: an empirical exploration of the feasibility," *Health communication*, vol. 25, no. 2, pp. 175–181, 2010.

[31] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in *Proceedings of the fifth annual workshop on Computational learning theory*. ACM, 1992, pp. 144–152.

[32] Chih-Chung Chang and Chih-Jen Lin, "LIBSVM – A Library for Support Vector Machines," April 2013. [Online]. Available: http://www.csie.ntu.edu.tw/ cjlin/libsvm

[33] G. Salton and M. J. McGill, "Introduction to modern information retrieval," 1983.

[34] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.

[35] A. R. Aronson, "Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program." *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pp. 17–21, 2001. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/11825149

[36] J. Cohen, "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement*, vol. 20, no. 1, p. 37, 1960.

[37] J. R. Landis, G. G. Koch et al., "The measurement of observer agreement for categorical data." *biometrics*, vol. 33, no. 1, pp. 159–174, 1977.