# Understanding User Intents in Online Health Forums

Thomas Zhang, Jason H.D. Cho, Chengxiang Zhai
Department of Computer Science, College of Engineering
University of Illinois at Urbana-Champaign, Urbana, IL, 61801
{zhang156, hcho33, czhai}@illinois.edu

## 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, large quantities of unstructured and diversified content generated on these forums make it difficult for users to digest and extract useful information. Understanding user intents would enable forums to more accurately and efficiently find relevant information by filtering out threads that do not match particular intents. In this paper, we derive a taxonomy of intents to capture user information needs in online health forums, and propose novel pattern based features for use with a multiclass support vector machine (SVM) classifier to classify original thread posts according to their underlying intents. Since no dataset existed for this task, we employ three annotators to manually label a dataset of 1,200 HealthBoards posts spanning four forum topics. Experimental results show that SVM with pattern based features is highly capable of identifying user intents in forum posts, reaching a maximum precision of 75%. Furthermore, comparable classification performance can be achieved by training and testing on posts from different forum topics (e.g. training on allergy posts, testing on depression posts). Finally, we run a trained classifier on a MedHelp dataset to analyze the distribution of intents of posts from different forum topics.

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information filtering; J.3 [**Life and Medical Sciences**]: Health

## General Terms

Design, Experimentation, Human Factors

## Keywords

User Intent Classification, Forum Intents, Online Health Forums, Support Vector Machines, Pattern Based Features


*BCB'14,* September 20–23, 2014, Newport Beach, CA, USA.

http://dx.doi.org/10.1145/2649387.2649445

## 1. INTRODUCTION

The spread of Health 2.0 [35] technologies in the last decade has made the Internet a popular place to learn about health matters. A recent Pew survey [19] 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, [17] 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 [20] 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. In present-day health forums, users often start new threads to ask questions despite the fact that similar discussions may have taken place in the past. Users would then patiently wait for responses while the answers that they are looking for may already be on the forum. An example of this type of behavior is shown in Figure 1. Here, we see that a question posted in 2012 that has received no replies as of 2014 has in fact been answered eight years prior. The author has wasted valuable time and energy both constructing the post and waiting for a response when he or she could have simply obtained answers from past forum discussions. Had we known the intent of that post, we could have recommended to its author a set of similar threads that match both its intent and content. Doing so would have offered the author a chance to find the information that he or she was looking for before deciding whether or not to start a new thread. This example clearly demonstrates how knowledge of intents can be used to connect users with relevant information in online health forums more efficiently.

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

**Figure 1: Two HealthBoards threads asking about treatments for cholinergic urticaria. In this example, we see that an unanswered post may have already been answered in a previous thread.**


Knowing the intents of original thread posts would also assist a number of existing works that aims to retrieve information from health forum content. The intent of the first post in a thread sets the topic of that thread and determines what type of information users would expect to find in the subsequent posts. Applications can therefore use this knowledge to reduce the sizes of search spaces by filtering out threads with intents that are less likely to contain relevant information, resulting in more efficient run times and more accurate results. We can illustrate this claim by hypothetically incorporating intents to several works that utilizes health forum data. For example, Vydiswaran et al. [36] assessed the trustworthiness of disease and treatment claims by searching for all relevant evidence documents that supports each claim from an online health forum corpus, and scoring the claims by combining features from those documents. Had they had access to thread intents, they could have simply conducted the search on threads with "treatment" intent from their corpus of documents. Similarly, Cho et al. [12] conducted Comparative Effectiveness Research (CER), defined as how well patients respond to treatments, by extracting treatment sentiments from over 130K online health forum posts to model treatment effectiveness. With knowledge of intents, they could instead run their algorithm only on posts from threads with"treatment" and "adverse effects of treatment" intent which would dramatically reduce the search space and arguably improve the quality of the results obtained. Finally, Jiang et al. [22] designed a clustering system to organize and integrate patient drug outcomes by splitting the health forum comments into "comment units", and classifying each unit into one of two groups belonging to an outcome cluster, each of which is determined by an expert comment: similar opinion (with the expert comment), or opposite opinion (from the expert comment). Much like in the case of Cho et al., we claim that the comment units can be extracted from messages in threads with "treatment" or "adverse effects of treatment" intent since these posts are mostly likely to contain information about patient drug outcomes. Furthermore, we know that threads with "treatment" intent would contain content that have largely positive sentiment, while threads with "adverse" intent would contain mostly negative sentiment. This gives us a general idea of which group (similar or opposite opinion) the thread units would mostly be classified into given the sentiment of the expert comment.

To our knowledge, no previous work has sought to identify user intentions from original health forum thread posts. Framing this problem is especially challenging since the definition of intent is quite vague. In this paper, we cast this problem as a classification problem to make the task more tractable. However, since this is a new task, no existing intent taxonomy or datasets exist for this problem. Therefore, we first derive a taxonomy of user intents from existing medical literature and create a new labeled data set for evaluation. For classification, we use a supervised learning method, and propose a set of novel pattern based features specific to the content found in health forums. Experimental results show that a support vector machine (SVM) classifier using pattern based features can achieve a precision of 75%, thus demonstrating the feasibility of our method to automatically identify intents from original health forum posts.

The rest of the paper is organized as follows. The next section surveys relevant past work in both the health and general domains. Section 3 motivates and derives the intent taxonomy. Section 4 formally presents the problem, while Section 5 and 6 introduces the classification framework and feature set, respectively. Section 7 presents the data, details the experimental setup, and summarizes the evaluation results. Finally, in Section 8, we apply our method to datasets from two different forums to analyze the distribution of intents for several forum topics.

## 2. RELATED WORK

Much research have been done on medical question answering (QA) systems. Many of these works have identified question understanding, framed as a classification problem, as a necessary and important first step in the implementation of such systems. For example, Yu et el. [38] made use of supervised learning approaches to classify questions based on the Evidence Taxonomy proposed by Ely et al. [15] and later on general topics [37], and found that including concepts and semantic types from the Unified Medical Language System (UMLS) as additional features can enhance classification results. Later, Kobayashi and Shyu [24] classified questions into taxonomies by the Family Physicians Inquiries Network (FPIN) and the generic taxonomy proposed by Ely et al. [16], and showed that augmenting UMLS concepts and semantic types with standard parsing representations improves classification performance. Last but not

least, Slaughter et al. [33] investigated semantic patterns of health consumers' questions and physicians' answers, and found that semantic relationships can indeed lead to clues for creating semantic-based QA techniques. These studies all demonstrate semantic based question classification approaches in medical QA systems, and we will show in this paper how similar approaches can be used to classify original thread posts in online health forums.

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. [29] 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. [39] 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. [27, 28] 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. [9, 10] which adds a new social category to Li's taxonomy and proposes a classification method using only features extracted from questions. However, all of these studies are insufficient for our purposes as their methods make use of 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 health 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 health forum users.

In addition to questions, previous research on user intents have also focused on web search engine queries. Cartright et al. [8] explored information goals and patterns of attention in web exploratory health search (EHS) through analysis of search sessions. They identified EHS sessions, extracted different intentions persisting as foci of attention from those sessions, and demonstrated how this knowledge can be used to better understand EHS behavior and support health search on the web. Similarly, other works such as [4, 34, 2] 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 [7] found that user query goals can be classified into a trichotomy of web search types: information, navigational, and transactional. Subsequent works such as [21, 26, 3, 23] show that various automatic learning-based approaches can be used to produce solid predictive performance in classifying queries. Much like questions, queries differ from forum posts in several ways. First, queries often consist of discrete keywords whereas forum posts are formulated in natural language, reflecting the discrepancy between their intended audiences. Second, search queries typically reflect some specific underlying "need" [7] whereas the "need" behind forum posts may not be as clear. These key differences mean that we must take into account both the structural properties of forum posts as well as the needs of their authors while trying to characterize their intents.

# 3. INTENT TAXONOMY

## 3.1 Motivation

Ely et al. [16] developed a taxonomy of doctor's questions about patient care consisting of 64 generic question types. Their taxonomy aims to completely capture information needs of doctors during patient visits. Boot and Meijman [5] investigated the feasibility of using this taxonomy to classify health questions asked by the general public. In the process, they found many differences between the information needs of 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?") due to the fact that they are rarely asked by doctors, yet these questions are frequently asked by patients. In addition, patients often tend to ask more ambiguous questions than doctors would due to their lack of expertise in health related matters. Classification using Ely's taxonomy would in turn become problematic since the taxonomy contains categories with very similar meanings (e.g. "What is the cause of symptom x?" and "Could this patient have condition y?"). From these differences we can clearly see that it is inappropriate to use a taxonomy designed for doctor's questions to characterize the intents of forum users.

## 3.2 Derivation

### *3.2.1 Intuition*

Boot and Meijman's study raises the need for a new taxonomy designed for the general public. For our purposes, we want a taxonomy that captures the intents of online health forum users, specifically, the intents of original forum posts composed by these users. To our knowledge, no previous work has been done in this area, but Choudhury et al. [14] identified the intents of online users who search for general purpose health information. If we assume that these users have roughly the same intents as online health forum users, we can derive an intent taxonomy from the original taxonomy of doctor's questions proposed by Ely et al. to generate a one-to-one mapping to the user intents discovered by Choudhury et al. Our ability to generate this mapping effectively validates the correctness of our classes in capturing the majority of intents of online health forum users. Finally, we can add several additional intent classes specific to health forums to complete the taxonomy.

### *3.2.2 Clustering*

Ely et al. [16] presented a list of the top 10 most commonly asked generic questions by doctors (shown in Table 1). On close inspection, we find that these questions 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?" ((1) and (4) are essentially questions pertaining to treatment). (10) becomes its own standalone class, and (3) is discarded because it refers to questions that only doctors would ask.

### *3.2.3 Expansion*

We make two observations with regards to health forums. First, we notice that some health forum posts contain multiple medical inquiries corresponding to more than one in-

**Table 1: Top 10 generic questions by primary care doctors from [16].**

| 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? |

tent. We therefore introduce a "Combination" class that corresponds to such posts. Second, we find that it is common for users to ask for or share health related experiences or news, post personal stories aiming to garner emotional support from the forum community, or post off-topic messages. For such posts, we propose a "Story Telling" class. As we shall see, these particular types of posts tend to show up in certain forums more often than others.

#### 3.2.4 Mapping

Choudhury et al. [14] examined the intents of 197 survey respondents who seek health information online using search engines. They found that the most common motivations of these users behind their searches are, in decreasing order, identifying treatment options, diagnosing health conditions, understanding health conditions or procedures, and understanding medications. We note that our formulated taxonomy classes from Section 3.2.2 more or less match the most common user motivations. 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?" to understanding medications and procedures. This mapping verifies our derivation of an intent taxonomy for online health forum users from an existing taxonomy of doctor's questions.

### 3.3 Summary

Our derivation in Section 3.2 yields five relatively broad categories of intent, which we now describe in more detail.

***Manage:* How should I manage/treat condition X?**

*Description:* Information regarding treatment options; management of long-term illnesses; illness prevention.

*HealthBoards 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 to get through this?

***Cause:* What is the cause of symptoms/physical findings/test findings X?**

*Description:* Diagnosis of physical findings or test results, including statistics (e.g. high blood pressure readings).

*HealthBoards 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/treatments X cause (adverse) finding Y?**

*Description:* Negative side effects of drugs or treatments (e.g. heart surgery), including short/long term health risks, effects of dosage, and withdrawal effects.

*HealthBoards 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).**

*Description:* Multiple inquiries of two or more of the three intents above.

*HealthBoards 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.**

*Description:* Asking/sharing experience or news; personal comments to garner sentimental or emotional support; off-topic content.

*HealthBoards 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.

## 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 our 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 matches the predicted intent. Identification of multiple intents will be left as future work.

In addition, note that our formulation excludes using thread titles for classification. After manually examining all 1,192 posts in our evaluative dataset to see (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 137/1,192 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.

# 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 will 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 will train a classifier using the training set and evaluate on the test set. For each post in the test set, the classifier will compute a score for each class, and the class with the highest score will be 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 [6] 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 [11] 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 Section 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 our dataset which allows us to compare its overall performance with that of the baseline word classifier.

# 6. FEATURES

The main technical challenge in supervised learning is to design appropriate features. In this section, we will first describe our baseline of standard unigram word features and then propose four novel pattern based feature sets to aid us in classification.

## 6.1 Word Features

Our baseline features are based on the traditional bag-of-words model [31], 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 purposes, we use standard unigram word features weighted with TF-IDF [32], 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 the word in the corpus (inverse document frequency), which helps to control 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 observed recurring sentence patterns in original thread posts from different intent classes and found that they are excellent 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 would have significant discriminative power in identifying 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, middle, or end of a sentence). A pattern is "matched" if every $s_i, 1 \leq i \leq |S|$ matches a token within a single sentence in that same order. For all pattern features, we use binary weights (i.e. 1 if a pattern matches, 0 if it doesn't), due to the fact that it is rare for a pattern to be matched more than once in a post.

### 6.2.3 MetaMap

The Unified Medical Language System[4] (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) developed MetaMap [1], a highly configurable program to map biomedical text to Metathesaurus concepts and their associated semantic types. As we shall see in the next section, we will use the MetaMap API to replace post phrases with their semantic group labels.

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

**Table 2: 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 |

#### 6.2.4 Data Preprocessing

We construct four different data representations for each original thread post in our dataset. The first consists of tokenizing and lowercasing the sentences. The second and third consist of stemming and POS tagging the data from the first, respectively. The fourth first involves feeding the original thread posts into the MetaMap API to generate phrase to semantic type mappings (Table 2 shows the subset of Metathesaurus semantic types considered). Next, the types are mapped to their corresponding semantic groups and all mapped phrases in the posts are replaced by these groups. Finally, the output is tokenized and lowercased.

#### 6.2.5 Pattern Identification

To capture the intuition from Section 6.2.1, we carefully compile a list of patterns that we think are most representative of the *Manage*, *Cause*, and *Adverse* intent classes. We can divide this list 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 have increasing discriminative power in classification.

**LSP** These patterns contain 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..."), meaning that they are more likely to match posts from 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 otherwise 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).

#### 6.2.6 Lack of *Story* Patterns

It is difficult to identify pattern base features for posts with *Story* intent due to large variations in content. This limitation, however, is acceptable from an information retrieval perspective. Recall from Section 3.3 that posts with *Story* intent consist mostly of story telling, sharing experience, or soliciting emotional support, none of which can be directly answered by content from another thread. This means that only direct responses to these posts is useful for their authors, and that knowing that a post has *Story* intent will not necessarily help us. Therefore, identification of such posts is not crucial and we will leave them as future work.

# 7. EVALUATION

## 7.1 Data

Since our classification task is novel, 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 ideally a larger dataset is preferred, 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 wanted 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 Labeling

**Four-Way Agreement.** First, we employ four humans to label 75 posts from our dataset according to our five-class taxonomy. The humans consist of two medical students and two computer science master's students. Since it is impractical to have the medical students label our entire dataset, we compare labeling differences between humans with health domain expertise with those that don't. To make this comparison, we use Fleiss' kappa [18], a measure for assessing agreement reliability between a fixed number of raters, and found $\kappa$ to be 0.67, indicating substantial agreement per Landis and Koch [25]. This result proves that we can rely on the computer science students to label analogously to the medical students.

**Inter-Annotator Agreement.** Next, we evaluate the labeling agreement between the two computer science students using Cohen's kappa [13], a measure for assessing agreement reliability between two raters, and found $\kappa$ to be 0.665 (56/75, $\approx$ 74.67% of labels match), indicating substantial agreement per Landis and Koch [25]. Given the fuzzy nature of the task at hand, this $\kappa$ value is certainly satisfactory.

**Gold Labeling.** To create a gold standard for this dataset, we employ the same two computer science students from the

**Table 3: 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 4: 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 |

agreement experiments to label the dataset with our proposed five-class taxonomy. We employ a third computer science student to label *Combo* posts with at least two classes from {*Manage*, *Cause*, *Adverse*}. The final distribution of gold standard labels for five-class labeling and *Combo* labeling are shown in Tables 3 and 4, respectively.

## 7.2 Experimental Setup

In this section, we will first describe the experimental setup to compare the performances of the pattern classifier using different combinations of pattern features. Next, we will explain how we setup experiments to compare the performances of the word classifier baseline with our hierarchical classifier using both standard 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 in our 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 our 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 our 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.

**Table 5: 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** |

### 7.2.3 4-Fold Forum Cross Validation

The previous section describes the usual way of performing 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 our 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 (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

Our experimental results are summarized in Tables 5-7. In this section, we will 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 5 compares the performance of the pattern classifier for each feature set combination. Unsurprisingly, we find that as we added 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 our pattern classifier) increases. The accuracy of the classifier, however, remains relatively constant. We picked the 6th feature space for use in the rest of our experiments because it gives us the highest number of matches without sacrificing performance.

### 7.3.2 Cross Validation

In this section, we summarize our cross validation results.

**Table 6: Baseline cross validation results.**

| | 5-Fold CV | | | 4-Fold Forum CV | | |
|---|---|---|---|---|---|---|
| Intent | P | R | F1 | P | R | F1 |
| Manage | 58.25 | 62.85 | 60.46 | 54.34 | 61.15 | 57.55 |
| Cause | 61.92 | 59.75 | 60.82 | 61.20 | 53.65 | 57.18 |
| Adverse | 39.47 | 29.41 | 33.71 | 35.29 | 24.24 | 28.74 |
| Story | 39.54 | 40.74 | 40.13 | 37.31 | 41.08 | 39.10 |
| Overall | **53.44** | | | **50.59** | | |

**Hierarchical classifier improves over baseline word classifier.** From Tables 6 and 7 we see that the hierarchical classifier yields an 8.5% improvement (an overall performance increase from 53.44% to 57.63%) over the baseline for 5-fold cross validation, and a 10.4% improvement (from 50.59% to 55.87%) over the baseline for 4-fold forum cross validation. We found these results to be statistically significant per McNemar's test [30] at 0.05-level. Note that this modest performance increase could have been higher had the hierarchical word classifier performed as well as the baseline.

**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%. The low recall is attributed to a low number of posts matching at least one pattern feature (recall that our pattern classifier only classifies posts that match at least one pattern feature). However, we argue that a high precision, low recall classifier is acceptable since we would 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 precisions that approach the labeling agreement upper bound.** Recall from Section 7.1.1 that the observed agreement between the two labelers is $\approx 74.67\%$. This number tells us that the performance of our classifier is restricted to roughly 75% as it is theoretically not possible to achieve precision higher than the agreement. Indeed, we see from Table 5 that the precisions obtained from the pattern classifier do approach the agreement, proving that the classification performance have reached the upper bound.

**Pattern classifier achieves comparable performance in 4-fold forum cross validation.** From Table 7 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 6 and 7 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 test posts that do not match any patterns possessing more ambiguous intents than those that do, and are therefore harder to classify.

**Table 7: Hierarchical classifier CV results. * indicates statistical significance at $\alpha = 0.05$ against corresponding baseline cross validation 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 | 44.62 | 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*** | | |

**Word classifier fails at identifying *Adverse* and *Story* intents.** The unigram classifier performs poorly on posts with *Adverse* and *Story* intent. We believe this low performance is due to the *Adverse* intent class containing very few data points (resulting in the classifier not having enough posts to learn from), and posts with *Story* intent inherently possessing ambiguity in its intents and having little to no distinguishing word features.

# 8. DISTRIBUTION ANALYSIS

## 8.1 HealthBoards Dataset

Table 8 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 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

**Figure 2: Distribution of classified intents for all four MedHelp forums.**

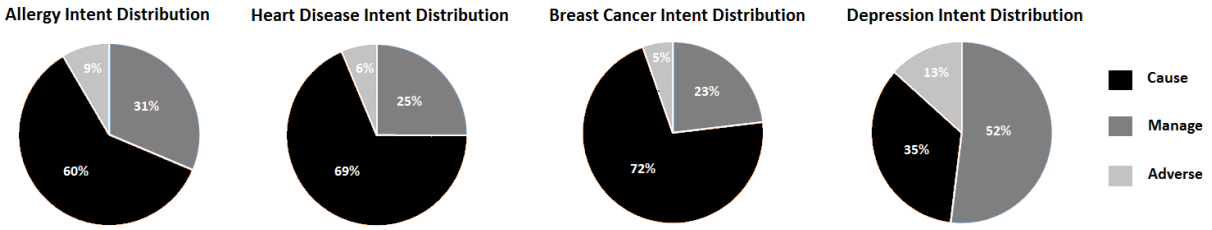


**Table 8: Distribution of gold intents from HealthBoards dataset.**

| 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 |

**Table 9: Post count for each MedHelp forum.**

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

the classifier on a collection of 61,225 posts that we crawled from MedHelp. These posts come from the same topics as those in our HealthBoards dataset (i.e. allergy, breast cancer, depression, and heart disease). Table 9 shows the total number of posts in each forum, while Figure 2 shows the distribution of classified intents for posts in all four forums. From these statistics, we can make several observations:

***Cause* make up a majority in 3/4 forums.** Allergy, breast cancer, and heart disease forum users seem to start more diagnosis related threads than any other type of thread. This follows the fact that many users use information on health forums to make preliminary diagnosis 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 fact leads us to believe that depressed patients are mostly concerned with finding ways to mitigate their symptoms.

**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%). We believe that this is due to many medications listing 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. Since allergies are relatively minor ailments, patients are more interested in asking about treatment options than obtaining an accurate diagnosis.

# 9. CONCLUSION

This paper presents a machine learning approach to identifying user intents from original thread posts from online health forums. From an information retrieval perspective, knowledge of intents are extremely important because it allows threads with certain intents to be filtered out, thereby reducing the search space. This technique can be applied to a variety of applications such as thread search and recommendation, and also benefit many existing works such as treatment trustworthiness, Comparative Effectiveness Research (CER), and drug outcome clustering.

Our main contributions in this work are threefold. First, we derived an intent taxonomy to capture information needs of online health forum users. We showed in our derivation that the classes map directly to the common motivations of users who search for health information online. Second, we identified a novel set of pattern based features to classify posts according to their underlying intent, and showed that a support vector machine classifier can use them to achieve accuracies upwards of 75% which approaches our precision upper bound. Third, we demonstrated that the performance of our classifier is capable of classifying posts from forums not seen during training with high accuracy. This proves that our classifier can be trained and tested on posts from different forum topics.

Several limitations exist within the scope of this work. First, we were unable to conduct a study of health forum user intents due to limited resources. Administrating such a study using test subjects would have been ideal as it would better justify the intents we decide to include in our taxonomy. Second, we find that the low recall of the pattern classifier obtained from Section 7.3.2 is due to an insufficient number of posts matching at least one pattern feature (thus limiting the number of posts the classifier can be run on). Therefore, to improve recall, we need to work on expanding our pattern feature set. Third, our current pattern classifier does not handle classification of posts with *Story* intent. Future work should involve coming up with ways to identify such posts with high accuracy. Fourth, we clearly see that our current classification framework is unable to extract all specific intents for posts with *Combo* intent since SVMs classify each example into a single class. Further work is needed to correctly identify all intents from forum posts. Finally, addressing all of these issues would allow us to classify all posts from a particular forum (i.e. HealthBoards) to visualize the makeup of thread intents from each forum topic. Doing so would allow us to gain a better understanding of the differences in the intents of users who post about these topics.

## 10. ACKNOWLEDGEMENTS

This work is also supported in part by the National Science Foundation under Grant Number CNS-1027965. We would like to thank Adam Szmelter and Niteesh Chitturu for providing their excellent medical expertise in labeling health forum posts, Son Nguyen and Josh Friedman for labeling our dataset, and Henry Lin for helping us with revision. In addition, we thank the anonymous reviewers for their insightful comments. Finally, we would also like to express our sincerest gratitude to Jump Trading and the Siebel Foundation for their generous scholarships that have supported this research.

## 11. REFERENCES

[1] A. R. Aronson. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings / AMIA ... Annual Symposium. AMIA Symposium*, pages 17–21, 2001.

[2] S. L. Ayers and J. J. Kronenfeld. Chronic illness and health-seeking information on the Internet. *Health (London)*, 11(3):327–347, Jul 2007.

[3] 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*, SPIRE'06, pages 98–109, Berlin, Heidelberg, 2006. Springer-Verlag.

[4] 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*, CHI EA '03, pages 816–817, New York, NY, USA, 2003. ACM.

[5] 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*, 25(2):175–181, 2010.

[6] 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*, pages 144–152. ACM, 1992.

[7] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, Sept. 2002.

[8] 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*, SIGIR '11, pages 65–74, New York, NY, USA, 2011. ACM.

[9] 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*, SIGIR '13, pages 973–976, New York, NY, USA, 2013. ACM.

[10] 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*, WWW '12 Companion, pages 823–828, New York, NY, USA, 2012. ACM.

[11] Chih-Chung Chang and Chih-Jen Lin. LIBSVM – A Library for Support Vector Machines, April 2013.

[12] 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*, BCB'13, pages 907:907–907:916, New York, NY, USA, 2013. ACM.

[13] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37, 1960.

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

[15] 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*, 324(7339):710, 2002.

[16] 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*, 321(7258):429–432, 2000.

[17] G. Eysenbach. The impact of the internet on cancer outcomes. *CA: A Cancer Journal for Clinicians*, 53(6):356–371, 2003.

[18] J. Fleiss et al. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[19] S. Fox. The social life of health information. *The Pew Internet & American Life Project*, May 2011.

[20] 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*, 13(3), 2011.

[21] B. J. Jansen, D. L. Booth, and A. Spink. Determining the informational, navigational, and transactional intent of web queries. *Inf. Process. Manage.*, 44(3):1251–1266, May 2008.

[22] Y. Jiang, Q. V. Liao, Q. Cheng, R. B. Berlin, and B. R. Schatz. Designing and evaluating a clustering system for organizing and integrating patient drug outcomes in personal health messages. In *AMIA Annual Symposium Proceedings*, volume 2012, page 417. American Medical Informatics Association, 2012.

[23] 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*, SIGIR '03, pages 64–71, New York, NY, USA, 2003. ACM.

[24] T. Kobayashi and C.-R. Shyu. Representing clinical questions by semantic type for better classification. In *AMIA Annual Symposium Proceedings*, volume 2006, page 987. American Medical Informatics Association, 2006.

[25] J. R. Landis, G. G. Koch, et al. The measurement of observer agreement for categorical data. *biometrics*, 33(1):159–174, 1977.

[26] 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*, WWW '05, pages 391–400, New York, NY, USA, 2005. ACM.

[27] 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*, EMNLP '08, pages 937–946, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[28] 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*, SIGIR '08, pages 735–736, New York, NY, USA, 2008. ACM.

[29] 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*, COLING '08, pages 497–504, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[30] Q. McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

[31] G. Salton and M. J. McGill. Introduction to modern information retrieval. 1983.

[32] G. Salton, A. Wong, and C.-S. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[33] L. A. Slaughter, D. Soergel, and T. C. Rindflesch. Semantic representation of consumer questions and physician answers. *I. J. Medical Informatics*, 75(7):513–529, 2006.

[34] 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*, 21(1):44–51, Mar 2004.

[35] 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*, 12(2), 2010.

[36] 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*, DMMH '11, pages 42–51, New York, NY, USA, 2011. ACM.

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

[38] 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.

[39] 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*, SMC'09, pages 1153–1158, Piscataway, NJ, USA, 2009. IEEE Press.