Understanding User Intents in Online Health Forums

Thomas Zhang, Jason H.D. Cho, and 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, 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 find and recommend relevant information to users 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 1,192 HealthBoards posts spanning four forum topics. Experimental results show that a SVM using pattern based features is highly capable of identifying user intents in forum posts, reaching a maximum precision of 75%, and that a SVM-based hierarchical classifier using both pattern and word features outperforms its SVM counterpart that uses only word features. Furthermore, comparable classification performance can be achieved by training and testing on posts from different forum topics.

I. INTRODUCTION

The spread of Health 2.0 [37] technologies in the last decade has made the Internet a popular place for patients to learn and discuss health related issues. In particular, online medical forums such as HealthBoards1, MedHelp2, and Wellescent3 have become very popular because they provide 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, [19] finds that many physicians are unaware of the numerous alternative and complementary treatment medications found in forum discussions, and [21] shows that 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 and then patiently wait for responses while the answers that they are looking for may already be on some forum. If we could somehow discern the intent of these threads, we could have recommended to the author a set of similar threads that match both his intent and the content of his post, perhaps saving him from having to indefinitely wait for a response.

Knowing the intents of original thread posts would also assist a number of existing works that aim to retrieve information from health forum content. The intent of the original post in a thread sets the topic of that thread and determines what type of information users could expect to find in the responses. 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 time and more accurate results. To illustrate this claim, we can hypothetically incorporate intents to several works that utilize health forum data. For example, Vydiswaran et al. [38] searched for all relevant posts that support claims from an online health forum corpus. If they knew the intent of every post, they could have simply conducted their search on posts with “treatment” intent. Similarly, Cho et al. [13] extracted 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 with “treatment” and “adverse effects of treatment” intent. Finally, Jiang et al. [23] organized and integrated patient drug outcomes by splitting the health forum comments into “comment units”, and classifying each unit into one of two groups. Much like in the case of Cho et al. [13], 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.

To our knowledge, no previous work has sought to identify user intents from original health forum posts. 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, and so we must 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 health forum content. Experimental results show

1http://www.healthboards.com/boards
2http://www.medhelp.org/forums/list
3http://wellescent.com/
that a support vector machine (SVM) classifier using pattern based features can achieve high precision upwards of 75%, and that a hierarchical classifier using both pattern and word features outperforms a classifier using only word features with statistical significance. In addition, we find that our classifiers give comparable performance when trained and tested on posts from different forum topics.

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 introduce the classification framework and feature set, respectively. Section 7 presents the data, describes the experiments, and summarizes the evaluation results. Finally, in Section 8, we apply our method to a MedHelp dataset to analyze the distribution of intents for several forum topics.

II. RELATED WORK

Much research has 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 al. [40] made use of supervised learning approaches to classify questions based on the Evidence Taxonomy proposed by Ely et al. [17] and later on general topics [39], 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 [26] classified questions into taxonomies by the Family Physicians Inquiries Network (FPIN) and the generic taxonomy proposed by Ely et al. [18], and showed that augmenting UMLS concepts and semantic types with standard parsing representations improves classification performance. Last but not least, Slaughter et al. [35] 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. [31] 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. [42] 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. [29], [30] 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. [10], [11] 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. Dogan et al. [16] investigated user behavior in the biomedical search engine PubMed through an analysis of user sessions and queries. Their studies revealed, among others, that the most frequent types of search are by author name, gene or protein, and disease. 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], [36], [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 [22], [28], [3], [25] 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.

As we’ve mentioned earlier, we do not know of any previous work related to user intents in the health forum space. Nevertheless, many works have utilized health forum data to accomplish other tasks in recent years, notably by [13] to conduct Comparative Effectiveness Research (CER), [23] to organize and integrate patient outcomes, [24] to classify posts into news, comments, or spam, and [9] to track trends in people’s reactions to drugs over time. These works demonstrate that health forums contain valuable information that can be exploited to advance knowledge in the health domain.

4http://www.ncbi.nlm.nih.gov/pubmed
III. INTENT TAXONOMY

A. Motivation

Ely et al. [18] developed a taxonomy of doctor’s questions about patient care consisting of 64 generic question types which aims to 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, and 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 ask 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?”). Their findings prove that it is inappropriate to use a taxonomy designed for doctors to characterize patient intents.

B. Derivation

1) Intuition: Boot and Meijman’s study raises the need for a new taxonomy designed for patients. For our purposes, we want a taxonomy that captures the intents of online health forum users, specifically, the intents of users who start new threads. To our knowledge, no previous work has been done in this area, but Choudhury et al. [15] 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 [15]. Our ability to generate this mapping validates the correctness of our classes in capturing the majority of intents of online health forum users. Later, we will complete the taxonomy by adding several additional intent classes specific to health forums.

2) Clustering: Ely et al. [18] presented a list of the top 10 most commonly asked generic questions by doctors (shown in Table I). On close inspection, we see 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) is its own class, and (3) is discarded because it refers to questions that only doctors, not patients, would ask.

3) Mapping: Choudhury et al. [15] examined the intents of 197 survey respondents who use search engines to seek health information online. They found that the most common motivations of these users are, in decreasing order, identifying treatment options, diagnosing health conditions, understanding health conditions or procedures, and understanding medications. We see here that our formulated taxonomy classes from Section III-B2 match the most common user motivations. “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. The existence of this mapping validates our approach of deriving an intent taxonomy from an existing taxonomy of doctor’s questions.

4) Expansion: We make two observations with regards to health forums. First, we notice that some forum posts contain multiple medical inquiries corresponding to more than one intent. 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.

C. Summary

Our derivation in Section III-B yields five categories of intent, summarized with examples from HealthBoards below.

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

TABLE I

Top 10 generic questions by primary care doctors from [18].

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

2168-2194 (c) 2015 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
his throat and he says that he has some congestion. Does this sound typical of allergies?

**Adverse:** Can drugs or 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 talking about how certain nasal sprays has steroids in them which could be bad for you if you continue to take it regularly. I assume astelin, flonase, nasonex and other prescription nasal sprays are okay to take regularly?

**Combo:** Combination.

*Description:* 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; comments to garner 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.

### IV. 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 \( j = i \).

One caveat to this formulation is that it does not identify all intents from multiple intent posts, which is in practice quite common. Posts with multiple intents (i.e. Combo posts) will be considered to be correctly classified if one of its intents matches the predicted intent. Multiple intent identification will be left as future work.

### V. Methodology

Our classification method is based on the classic supervised learning framework. To do so, we design features to capture intents in forum posts. Our task is to apply these features to our dataset to construct a feature representation of each post, and then 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 intent. The intent with the highest score will be assigned to that post.

In our experiments, we use the popular Support Vector Machine (SVM) classifier, and assume that our choice of features will generalize well to other classifiers, leaving experimentation with different classifiers as future work.

### A. 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. In our experiments, we use the LIBSVM [12] implementation with a RBF kernel, and train the classifiers using a one-versus-all multiclass approach.

### B. A Hierarchical Classifier

We use a hierarchical classification setup that consists of a sequence of two cascading SVM classifiers using word and pattern features (see Section V-C). The first classifies posts that match at least one pattern feature (Pattern Classifier), while the second classifies all posts that do not match any pattern features using word features (Word Classifier).

### C. Features

As mentioned in Section V-B, we use two types of features in our hierarchical setup: words and patterns. This section describes each feature type in more detail.

1) **Word Features:** These features are based on the traditional bag-of-words model [33]. In this model, text is represented as a set of its words, disregarding grammar and word order but keeping multiplicity. For our purposes, we use standard unigram word features weighted with TF-IDF [34].

2) **Pattern Features:** These features are recurring sentence patterns found in forum posts, each indicating a specific intent. For example, the pattern “What could X be...” indicates strong Cause intent, but “What can X do...” indicates Manage intent. We carefully compile patterns that are most representative of Manage, Cause, and Adverse intents. We do not extract patterns from posts with Story intent because the posts have too much content variation. Each pattern has to be sufficiently discriminative for its respective intent, but also generic enough as to match content across many forum posts. As a result, we replace certain words inside our patterns with their stemmed roots, part-of-speech tags, and semantic group labels from the UMLS Metathesaurus, extracted using MetaMap [1].

### VI. Evaluation

#### A. Dataset

There is no existing dataset available for our classification task, so we create a new dataset consisting of 1,192 original forum posts from HealthBoards, evenly divided between four topics: allergies, breast cancer, depression, and heart disease. These topics give us a good mix of posts from both major and minor health disorders. Two computer science students with substantial agreement per Landis and Koch [27] (based on Cohen’s kappa [14] \( \kappa = 0.665 \)) and substantial agreement with two medical students (based on Fleiss’ kappa [20] \( \kappa = 0.67 \)) label the dataset using the five-class taxonomy as described in Section III-C. A third computer science student labels Combo posts with all intents from \{Manage, Cause, Adverse\}.

---

3See http://thejason.co/intents/patterns.txt, README.txt.

4http://www.nlm.nih.gov/research/umls
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2015.2416252, IEEE Journal of Biomedical and Health Informatics

Two experiments are conducted with their descriptions below. Note that for all experiments, Combo posts are excluded for training, but every forum post in the dataset is used for testing. A Combo post is considered to be correctly classified if its predicted class label matches at least one of its labels. Otherwise, we pick the first label in the order of Manage, Cause, and Adverse and consider the post to be misclassified for that particular class.

1) Five-Fold Cross-Validation: In this experiment, we evaluate and compare five-fold cross-validation performance of the hierarchical classifier in Section V-B with that of a baseline Word Classifier.

2) Four-Fold Forum Cross-Validation: Here, each fold in cross-validation consists of only forum posts from one topic. This experiment therefore 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 be used in a real-life scenario.

### TABLE II

**Baseline Word Classifier CV Results.**

<table>
<thead>
<tr>
<th>Intent</th>
<th>Five-Fold CV</th>
<th>Four-Fold Forum CV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>Manage</td>
<td>58.25</td>
<td>62.85</td>
</tr>
<tr>
<td>Cause</td>
<td>61.92</td>
<td>59.75</td>
</tr>
<tr>
<td>Adverse</td>
<td>39.47</td>
<td>29.41</td>
</tr>
<tr>
<td>Story</td>
<td>39.54</td>
<td>40.74</td>
</tr>
<tr>
<td>Overall</td>
<td>53.44</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE III

**Hierarchical Classifier CV Results.** *indicates statistical significance at \( \alpha = 0.05 \) against baseline results in Table II.

<table>
<thead>
<tr>
<th>Intent</th>
<th>Five-Fold CV</th>
<th>Four-Fold Forum CV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>Manage</td>
<td>58.71</td>
<td>65.26</td>
</tr>
<tr>
<td>Cause</td>
<td>61.72</td>
<td>66.67</td>
</tr>
<tr>
<td>Adverse</td>
<td>60.81</td>
<td>48.39</td>
</tr>
<tr>
<td>Story</td>
<td>47.28</td>
<td>38.05</td>
</tr>
<tr>
<td>Overall</td>
<td>57.63</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE IV

**Pattern Classifier CV Performance, Isolated from the Hierarchical Classifier.**

<table>
<thead>
<tr>
<th>Intent</th>
<th>Five-Fold CV</th>
<th>Four-Fold Forum CV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>P</td>
<td>R</td>
</tr>
<tr>
<td>Manage</td>
<td>75.51</td>
<td>36.72</td>
</tr>
<tr>
<td>Cause</td>
<td>73.53</td>
<td>43.86</td>
</tr>
<tr>
<td>Adverse</td>
<td>80.85</td>
<td>40.86</td>
</tr>
<tr>
<td>3-Class</td>
<td>75.05</td>
<td>40.34</td>
</tr>
</tbody>
</table>

### C. Results

We make several observations pertaining to the results:

**Pattern Classifier achieves high precision but low recall.** 75.05% for five-fold CV, and 72.55% for four-fold forum CV (see Table IV), which means the classifier can accurately predict intents of posts that match at least one pattern feature. To increase recall we need to expand our pattern feature set.

**Hierarchical classifier improves over baseline.** 57.63% from 53.44% in five-fold CV, and 55.87% from 50.59% in four-fold forum CV. These results are both statistically significant per McNemar’s test [32] at 0.05-level.

**Four-fold forum CV compares with five-fold CV.** 55.87% with 57.63%. This shows our method generalizes well to posts from forums not represented in the training set.

Both baseline and hierarchical classifiers struggle to identify *Story* intent. 40.13% and 42.16% for five-fold CV, and 39.10% and 41.61% for four-fold forum CV (F1 scores). This is expected as we did not extract pattern features for posts with *Story* intent.

**Pattern Classifier makes mistakes on posts that match features across different intents by accident.** For example, a post could contain a sentence like "the doctor says (disorder) could be caused by..." but at the end ask "does anyone know if (treatment) helps?". It is clear that the post intent is *Manage*, but the classifier might be confused by the matched feature in the first sentence that appears in many *Cause* posts. We find that most of the classification errors occur due to this phenomenon.

### VII. MedHelp Distribution

We can use our classification method to analyze the distribution of intents for any collection of original health forum posts. To demonstrate this, we crawl 61,225 posts from MedHelp from the same topics as those in our HealthBoards dataset (i.e. allergy, breast cancer, depression, and heart disease).

Recall that since *Story* intents are hard to identify (see Section VI-C), we instead focus on identifying the other three intents (*Manage, Cause, and Adverse*). We decided to use a Pattern Classifier to classify the MedHelp dataset because it gives impressive precision on posts it classifies, albeit with low recall. Furthermore, since patterns are less forum dependent than words, we believe that a Pattern Classifier generalizes better to forum posts across different topics.

Figure 1 shows the distribution of classified intents for the MedHelp dataset as separated by topics. From these statistics, we can make several observations:

**Cause makes up a majority in three of the four topics.** This may be due to the propensity of forum users to use health forum information to make preliminary diagnosis of their conditions before consulting a medical professional.

**Manage makes up a majority in the topic of depression.** This leads us to believe that depressed patients are mostly concerned with finding ways to mitigate their symptoms.

**Depression contains the greatest proportion of Adverse posts.** We believe that this is due to many medications listing depression as a side effect. This in turn causes users of these medications to ask about them on health forums.

**Allergy forum contains a smaller ratio of Cause to Manage posts.** Since allergies are relatively minor ailments, patients are more interested in asking about treatment options than obtaining an accurate diagnosis.
VIII. 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 is 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 demonstrated that a classifier trained on novel pattern features is capable of identifying intents of forum posts with high precision. Third, we showed, with statistical significance, that a hierarchical classifier that uses both pattern and word features outperforms a one that uses only word features. Finally, we find 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 did not conduct a study of health forum user intents due to limited resources, which would have better justified the intents we decided to include in our taxonomy. Second, we need to work on expanding our pattern feature set in order to improve classification performance. Third, our current pattern classifier does not handle classification of posts with Story intent, so future work is needed to identify such posts. Fourth, we would ideally want a larger annotated dataset for more accurate evaluation and a set of annotated forum posts from MedHelp to obtain further validation that our classifiers do in fact generalize well to posts from other data sources. Finally, we are currently unable to extract all intents for posts with Combo intent and further work is required to identify all of the intents in a multi-intent forum post.

IX. 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 gratitude to Jump Trading and the Siebel Foundation for their generous scholarships that supported this research.

REFERENCES

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBMI.2015.2416252, IEEE Journal of Biomedical and Health Informatics


Thomas Zhang obtained his bachelor’s and master’s degrees in Computer Science at the University of Illinois in Urbana-Champaign in 2014, where he was a member of the Text Information Management and Analysis Research Group. His research interests focus on the development of smart systems and applications using text mining, natural language processing, and machine learning methods.

Jason H.D. Cho obtained his bachelor’s and master’s, and is currently pursuing his Ph.D. in Computer Science at the University of Illinois in Urbana-Champaign. His research interests lie in applications of natural language processing and text mining on health domains. In particular, he is interested in obtaining deeper understanding of what patients and clinicians write, and what general medical population are interested in.

Chengxiang Zhai is a Professor of Computer Science at the University of Illinois at Urbana-Champaign, and an ACM Distinguished Scientist. He received a Ph.D. in Computer Science from Nanjing University in 1990, and a Ph.D. in Language and Information Technologies from Carnegie Mellon in 2002. He worked at Clarivoyance Corp. as a Research Scientist and a Senior Research Scientist from 1997-2000. His research interests are in information retrieval, text mining, machine learning, biomedical informatics, and intelligent education systems.