August 2012

# Question Classification in the Cancer Domain

Adam Kurmally
*University of Wisconsin-Milwaukee*

# QUESTION CLASSIFICATION IN THE CANCER DOMAIN

by

Adam Y. Kurmally

A Thesis Submitted in

Partial Fulfillment of the

Requirements for the Degree of

Master of Science

Computer Science

at

The University of Wisconsin–Milwaukee

August 2012

ABSTRACT

QUESTION CLASSIFICATION IN THE CANCER DOMAIN

by

Adam Y. Kurmally

The University of Wisconsin-Milwaukee, 2012
Under the Supervision of Dr. Susan McRoy

We are investigating question classification for restricted domains with the broader goal of supporting mixed-initiative interaction on mobile phones. In this thesis, we present the development of a new domain-specific corpus of cancer-related questions, a new taxonomy of Expected Answer types, and our efforts toward training a classifier.

This work is the first of its kind in the cancer domain using a corpus consisting of real user questions gathered from cQA websites, and a taxonomy built from that corpus. Our goal is to create software to engage newly diagnosed prostate cancer patients in question-answering dialogs related to their treatment options. We are focusing our work on the interaction environment afforded by text and multimedia (SMS and MMS) messaging using mobile telephones, because of the prevalence of this technology and the growing popularity of text messaging, especially among underserved populations.

# TABLE OF CONTENTS

Page

# LIST OF FIGURES

# LIST OF TABLES

# Chapter 1: Introduction

This thesis describes the creation of a Question Classification framework for the cancer domain. Our research tests the hypothesis that a classifier trained on real users' questions can provide an effective means of identifying the information need expressed by a question. The end goal of this research is to support automated question answering through a publicly available service using SMS text messaging.

There are three main parts to the creation of our research. First, we mined cancer questions from Community-based Question Answering (cQA) websites [38], to collect a relevant set of patient questions. cQA websites build a virtual community where users can ask and answer questions [38]. The format can be unstructured (like a forum) where anyone can ask or answer a question [45]. They can also be structured such that only assigned experts may answer questions, but anyone may ask them [1, 6, 23, 41, 42]. We collected questions from both types of sites to build our corpus. A corpus is the set of documents comprising a data set in text mining systems [29]. In the context of this project, our corpus consists of all questions that we have mined form cQA websites, and coders have processed and classified; where each question is a separate document in the corpus.

Second, we iteratively built a taxonomy to divide the observed question corpus by Expected Answer Type (EAT) [24, 36, 37, 53]. The Expected Answer Type of a question is a category for the type of answer a person asking a question would expect to receive. Each category in our taxonomy constitutes a different EAT. The focus of developing a category taxonomy such as this is to create a set of EATs that allow each question to be placed in one and only one category.

Third, we created a set of test classifiers using our corpus as training and test data, and our taxonomy as our set of classifications. The performance of the classifiers was then compared to determine an optimal algorithm. We also explored simple dimensionality reduction techniques [14, 29] as well as automated spelling correction as methods to improve classifier performance.

## 1.1 Application Domain

Prostate cancer requires treatment that is personalized for each patient [9]. It has been shown that patients require a wide variety of information to make decisions about their treatment [4, 8–12, 18, 19]. Furthermore, different patients rate different pieces of information as important to their decision [18, 19]. The wide variety of information that individuals deem necessary for their treatment makes constructing a static general decision aid impossible without also including an overwhelming amount of information many patients will deem irrelevant [8, 18, 19].

Decision aids (both electronic and not) have been helpful to cancer patients [4, 5, 9, 12, 25], but the majority of these are static and cannot be tailored to individual patients' needs. The best solution is for clinicians to adequately answer all questions personally for each patient, whenever a patient needs information [4]. However, clinicians will not always be available to answer questions, and will not easily be able to deal with more than one patient at a time [25]. A computerized solution that is able to offer similar information would be available at any time, and would be able to handle large numbers of patients concurrently. Such a system could not diagnose or treat patients, but would be able to answer questions concerning basic facts and encourage patients to follow up with their physicians.

Question Answering (QA) agents have been used in numerous domains [13, 15, 24, 27, 38, 53] to answer factual questions in natural language. A dialogue-based question answering system could interact with patients in a similar way to a clinician. Such a system would give patients the information they seek in a familiar medium, while mitigating the high cost of a manual solution.

Mobile devices that possess Short Message Service (SMS) and Multimedia Messaging Service (MMS) capabilities are increasing in their ubiquity [3, 40]. These devices make two-way messaging dialogue between patients and clinicians possible, as well as between patients and software systems. Mobile phones with this ability have been used recently in the health care domain for health promotion and patient monitoring purposes [3]. Some studies have used two-way messaging via mobile phones in health

care; however these systems usually require a person to answer each message [3]. Another study has demonstrated two-way messaging with automated question answering for pregnant women [40]. However, this system did not support dialogue, and proved to be very brittle concerning the wording of questions. During testing it became so slow in responding to users that some gave up using the system [40]. The system we are proposing would be a fully functioning dialogue-based QA system capable of interacting with multiple users concurrently.

## 1.2 Question Classification

The primary Natural Language Processing (NLP) method for determining the semantic content of a question is parsing. There are a variety of parsing techniques that can be employed on various types of data, and these fall into two types: Deep and Shallow [29]. For Deep Parsing to be effective it requires a strong grammatical structure, which may not be present in SMS-based data [29]. Therefore, we must use a shallow parsing strategy to determine the semantic data in a text message. Part of a shallow parsing strategy is to use other text features (aside from grammar) to find the same semantic information [29].

A method to find semantic information from a document without analyzing its grammar directly is to use a classifier [16]. In a QA system, each document is a question; therefore a shallow parsing technique might also use a question classifier. A question classifier uses pattern classification techniques to determine the kind of answer the author is expecting [16, 29], which is called the Expected Answer Type (EAT) [24, 36, 37, 53]. The EAT along with the topic of the question can be used to replace the information that a deep parsing strategy is able to obtain.

The most common algorithms for QC, and text classification in general, use Supervised Machine Learning (SML) techniques [16, 29]. SML involves a pre-processing step called *training* [16]. Training requires example data sets that are already classified (usually by hand). These data sets are used to 'teach' an algorithm to distinguish between inputs that fit into the different categories [16]. The performance of a SML classifier is directly affected by the type and quality of the training data set [16,27,36].

Generally, a training data set that closely resembles the eventual input to the system results in a better classifier. Training data for use with a specific QA task should consist of questions that potential users of the system have. The data should match user questions in both content (spelling/grammar as well as information content) and EAT distribution. Therefore, the ideal training data set for a QA system for cancer questions should consist of real patient questions [36]. An SMS based QA system should have its classifier trained on short questions that are structurally and syntactically similar to SMS messages.

## 1.3 A New Question Corpus

The first contribution of this work is a new corpus of patient cancer questions [34, 35]. In the cancer question domain, there was no existing corpus of questions from cancer patients. This makes our corpus the first comprised of patient cancer questions. Our corpus is based on questions mined from cQA cites, and has not been corrected for spelling or grammar. The users that asked questions in our corpus are as close as possible to the target users of our proposed Cancer QA system. Therefore the construction of this corpus makes it as close to ideal as is currently possible.

There are only two ways to obtain a better corpus: creating a fully functioning QA system or conducing a Wizard-of-Oz test [29]. In order to create an automated QA system, a corpus would be required to train its question classifier. That would necessitate a corpus to "bootstrap" the QA system, which is essentially the purpose of our corpus. A Wizard-of-Oz system would be very costly in the medical domain, because a clinician would be required to answer most questions. Therefore, our approach is optimal given time and monetary constraints. Furthermore, it is our intent to create a QA system with a classifier based on this work which will gather all questions users ask it, so they can be used to create a more ideal corpus.

## 1.4 A New Expected Answer Type Category Taxonomy

The second contribution of this work is a new EAT Taxonomy for cancer questions. There are several existing question taxonomies [36], however, all of these were designed solely for factoid data in a general domain; hence they are unsuitable for a medical QA system. Furthermore, most of these taxonomies were not created based on questions posed by target users [13,27,36,53]. Instead, they were designed concurrently with artificial corpora that have correct spelling and grammar. Our taxonomy uses questions mined from those actually posed by target users, and reflects the EAT distribution they exhibit in the real world.

An effective EAT taxonomy must have two properties. The first is to fully cover all data in its domain. The second is to be reliable over varying human coders. This means each question must belong in exactly one category, and inter-coder agreement [20, 31] should be high. Our goals for this taxonomy are for each question in our domain to map to exactly one category and for a Krippendorff's Alpha score of 0.7 or higher. This Alpha score would indicate raters agree approximately 70% of the time above random chance. We achieved these metrics by developing our taxonomy iteratively, with the process using actual question data and inter-coder agreement scores based on the classification of partial data sets.

## 1.5 Roadmap

The following chapter will discuss general SML algorithms, their purpose, and how they work. The remainder of this thesis presents our approach to creating our corpus, EAT Taxonomy. We will present our initial design and all of our revisions to the taxonomy as well as its evaluation at each iteration. Next, we will show some experimental classifiers and their results when tested against our corpus. Finally, we will discuss the next steps in continuing this research to construct a functional QA dialog system for cancer questions.

# Chapter 2: Supervised Machine Learning and Metrics for Assessing the Reliability of the Coded Framing Set

The most common computational technique for classifying text is Supervised Machine Learning. It is a subset of the artificial intelligence field that allows machines to learn a function by example [16,29]. Examples constitute a set of problems (with answers) that are input to a machine learning algorithm. The algorithm then outputs a function that is able to perform a transformation from a problem to an answer [16,29].

In the case of this research, the function maps questions (text documents) onto an EAT category. The example data set is our corpus of hand-classified questions and their EAT classification. The output functions are the trained classifiers that can map cancer questions onto an EAT in our taxonomy.

## 2.1 Vector Space Model

A Vector Space Model is a mathematical model used to break different features of a datum apart. In our case, each vector will represent a single question, and each feature will represent a word in the question. A term in the vector corresponds to the number of occurrences of a specific word in the question. The number of terms in the vectors is the number of distinct words in our entire corpus, which is the dimensionality of the model.

## 2.2 Dimensionality Reduction

Dimensionality reduction techniques seek to improve the speed and accuracy of classifiers by reducing the number of features in a Vector Space Model [14, 16, 29]. This improves training and classification time since not as many features need to be processed. Reducing the number of features may improve the accuracy of a classification algorithm by removing irrelevant ones that may cause errors in the classifier. Improvements in classification results depend on a variety of other factors, including

the type of data, specific training data set, and the classification algorithm [14,16,29]. In text-based classification, this takes the form of removing irrelevant terms from the set of term-document vectors [16, 29].

## 2.3 Classification Algorithms

There are numerous classification algorithms that can be used to map a question document onto an EAT. The type of classification algorithm used in this work require a training set of example vectors that already have a class. The training step is a pre-processing task where the classification algorithm will use the training vectors to set certain parameters that the algorithm uses to map a single vector onto a single category. After the training step, the classifier is ready to classify real data. The input to the classifier is an input vector (in this case representing the terms in a question) and the output is a single class. Classification algorithms are generally tested with a separate known test set, or using an N-Fold cross validation [16, 29]. The three general algorithms tested in this work are Naive Bayes (NB) [28], Decision Trees [29], and Support Vector Machines (SVMs) [7].

A NB classifier uses Bayes theorem [16] to create a probabilistic classifier [28]. NB creates a probabilistic model using its training data that allows it to determine the probability an input vector belongs in each class. When a new item is input into the classifier it computes the probability that its vector belongs to each class and places it in the class with the highest probability. It is called "naive" because this model assumes that every variable in an input vector is independent of all others. In the QC task, this means it assumes the presence of a term in the vector does not depend on any other term, which is often false [16,28]. Still, NB classifiers have been used in the past for question classification with some success [27, 36, 48, 52].

Decision Trees do not use a vector space model to classify data. Instead, they model each feature that would be a term in a vector as a separate decision in a tree structure [29]. Decision Trees are a much more broad set of algorithms than Naive Bayes or SVMs; we focused on J48 Trees [46] as a representative algorithm, because it has been used successfully before for QC tasks [36, 48].

SVMs [29, 52] are a type of SML algorithm that classifies an input between categories by separating the categories in hyperspace. The divisions between each category are created during the training step. They consist of equations defining a hyper-plane that can be used to separate a defined hyperspace of possible data into regions representing each category [29]. In QC, SVMs use a Vector Space Model to create an n-dimensional hyperspace where n is the length of the vector. Training the SVM creates the hyperplanes used to discriminate between EAT categories.

## 2.4 Inter-Coder Agreement

Inter-coder agreement statistics define the amount of agreement among different coders classifying data. These statistics are used to determine how consistent a coding system is across different users. If inter-coder agreement is too low it indicates that there is a problem with the coding system, or coder training. Fleiss' Kappa [20] and Krippendorff's Alpha [31] are the two reliability statistics used in this work. Calculating multiple inter-coder agreement statistics helps ensure accuracy because the results of each statistic can be affected by different features of a data set or taxonomy, and those features are different for each statistic [47]. Checking that multiple inter-coder agreement statistics are close to each other ensures that this is not the case.

### 2.4.1 Fleiss' Kappa

Fleiss' Kappa [20] is an inter-coder agreement statistic for binary or nominal data. In the QC task, $\kappa$ uses the set of Expected Answer Types as nominal data in a flat hierarchy.

Fleiss' Kappa is related to Cohen's Kappa [20], with the advantage that it can account for more than two coders. It also requires that every question be classified by the same number of coders. This sometimes proved unfeasible for us given the large number of questions we were asking coders to classify, and the fact that we changed coders multiple times during our research. Therefore, $\kappa$ was only calculated during iterations where every question is classified by the same number of coders.

### 2.4.2 Krippendorff's Alpha

Krippendorff's Alpha [31] is an inter-coder agreement statistic similar to Kappa [20]. It can analyze binary, nominal, or ordinal data with more than two coders. It can also account for missing data, meaning every question does not have to be classified by the same number of coders for $\alpha$ to be useful. This makes Krippendorff's Alpha an ideal metric for use with distributed coders that only participate during part of a project, such as this work.

# Chapter 3: An Iterative, Data Driven Approach to Question Interpretation

This work addresses three questions. First, "How can we obtain real patients' cancer questions?" Second, "How can we develop a reliable coding scheme for cancer questions?" And third, "How well will different classifiers work after being trained on this data?"

This chapter describes how each of the three parts of this work were implemented. First, we created a web crawling framework to mine cancer questions from different cQA websites. Second, we used a sample of these questions to created an initial EAT Taxonomy. Third, we began an iterative process of manual question classification. In each iteration, we tested and revised the EAT Taxonomy to reflect a larger sample of the corpus. We accomplished this by building a custom website [33] to allow coders to work in a distributed fashion on any computer with a web browser and internet access. Finally, we created a set of classifiers using WEKA [22] to compare different SML algorithms [16]. We also used the WEKA API [22] to test dimensionality reduction and used Lucene [21] to test the effects of spelling correction on the classifiers.

## 3.1 Question Data Mining

The first step in this work was gathering questions for use in our corpus. Since there was no existing corpus of prostate cancer questions, we collected questions posted to public cQA websites [38]. We created a web-crawling application (called MedQuestionCrawler) to crawl specific cQA websites, parse their HTML, and save any questions found. The application is able to use simple keyword matching to determine relevant questions when crawling cQA sites not restricted to cancer. The questions were saved to XML files in a custom Document Object Model [49] created for medical questions. The saved data is then imported into a SQL database for processing and classification.

The cQA websites crawled for our corpus were: All Experts [1], The American Society of Clinical Oncology [42], The Cleveland Clinic [6], Med Help [23], Net

Wellness [41], and Your Cancer Questions [45]. Formats for the cQA websites include: live chats with medical professionals [6, 42], user answered forums [1, 45], and e-mailed or web form submitted questions that are answered publicly by a medical professional [23, 41]. Websites in which answers are posted by medical professionals have the answer text saved along with the question. No answers are recorded for websites where non-clinicians can answer questions.

Along with the question (and possibly answer) text, some meta-data is saved for each question. The saved data includes the original source URL, a short identifier specifying the source cQA site, and a numeric identifier that is assigned to track the question. Most cQA sites have no user data publicly available, and we do not save any user data that is not embedded in a question or answer. Questions are strictly anonymous, except where users have chosen to reveal personal information in their question.

We refer to the questions gathered in this manner as *Raw Questions*, meaning their question text is exactly as it appeared in its source. It is important to differentiate between Raw Questions and questions in our final classified corpus. We perform several transformations on mined questions in order to ensure our corpus of classified questions is more structurally similar to those users might ask via SMS.

## 3.2 Corpus/Taxonomy Creation

The second step in this work was to use the mined question data to build a reusable cancer question corpus and EAT Taxonomy for cancer questions. These two tasks were accomplished simultaneously through an iterative process. We used inter-coder agreement statistics [20, 31] measured from partial classifications of our corpus as feedback to inform our taxonomy construction. In this manner we could objectively measure the impact of each revision made to the taxonomy, with respect to how able coders were to reach a consensus on each category.

### 3.2.1 MedQuestionAdmin

The promotion and hand classification tasks described here necessitated a distributed application that would allow users to work independently, while aggregating their input in a single database. A web application called MedQuestionAdmin [33] was built to accomplish these tasks. MedQuestionAdmin allows registered users (coders) to perform different tasks that move mined questions from their initial state to *Classified Questions* we can use in our taxonomy. These tasks are *Promotion*, *Verification*, and *Classification*. MedQuestionAdmin also allows administrator users to review coders' work (it does not allow them to change any coder's work).

### 3.2.2 Question Promotion

Raw Questions are identified as relevant to cancer either by their source cQA website or by keyword matching by our web crawlers. However, that does not guarantee that each question is actually relevant to cancer in humans, and/or formatted like an SMS question. To this end, we process each question by hand in a step called *Question Promotion*.

During Question Promotion, coders have an opportunity to reject questions that are not relevant, and (under very limited circumstances) make changes to question text. The resultant questions from promotion are subject to another step in the process called *Question Verification*, where a different coder will either approve or reject each question. We refer to questions that are approved during the promotion step as *Promoted Questions*.

There are two reasons for coders to reject a question: irrelevance and length. Irrelevant questions may not be related to cancer at all, refer to cancer in animals other than humans, or state a complaint or problem without actually expressing an information need. These questions are not suitable to train a classifier for our purposes, since they are not related to the QA system we are targeting. Furthermore, we can ignore them as no serious user of an automated QA system would ask it these types of questions. Questions that are too long should be eliminated, because the SMS format forces users to ask very short questions. The SMS standard allows for

up to 160 characters per message, and the service we are currently using only allows up to 140 characters per message [40].

Some of our cQA sources for questions allow users to submit them in a forum or e-mail-like format, which means that they can be several thousand words long. Since we cannot use questions that far exceed our target character length we limit the maximum length of questions that can be promoted in MedQuestionAdmin [33]. However, we did not find enough existing cQA questions under 160 characters to build a training corpus. In order to balance the need for a large pool of questions with the maximum SMS length, the maximum question size in MedQuestionAdmin is 500 characters. This allows us to use a larger pool of questions to create our corpus and classifier while still limiting the question size to those that are more likely to resemble a question our target QA system would receive.

Sometimes questions will be formatted in such a way that coders will be able to edit them. The edits coders are instructed to make are minimal and fall into two categories. The first is if a question contains extraneous information that does not pertain to the information need expressed in the question. The second reason is if a Raw Question is *Compound*, meaning it contains more than one logical question.

Questions that contain extraneous information are often long, much greater than our 500 character limit. We are able to include more questions in our corpus if we edit out the irrelevant text so only the information need is present in our corpus. Since SMS character length is so short, these are more relevant to our target media as users will not be able to send large amounts of text to our system.

Compound Questions are different, in that their presence in our corpus will interfere with our inter-coder agreement. These documents can contain multiple EATs (due to multiple questions) that will make classification impossible, because coders will not necessarily be classifying the same question. Coders can split Compound Questions into multiple questions and promote them individually. In Compound Questions that contain pronouns across multiple questions, coders replace pronouns with the original noun phrases. Coders are instructed to only perform pronoun replacement on Compound Questions that have pronoun references across individual

questions. Coders are also instructed to not edit individual questions (non-compound) consisting of multiple sentences. Our rules for rejecting and reformatting questions during the promotion phase can be found in Appendix A.

### 3.2.3 Manual Classification

In order to use a corpus to train a Supervised Machine Learning classifier, items in the corpus must first be classified [16, 29]. We accomplished this by manually classifying the entire corpus of Verified Questions. Each promoted question is classified by as many coders as possible, but no less than two. This gives us the ability to compare coders' interpretation of the EAT taxonomy at each iteration in order to evaluate our taxonomy.

The manual classification step consisted of allowing coders to independently classify a subset of the promoted questions. In order to ensure their classifications are independent, coders were not able to see each others' work while classification was ongoing. Coders were also discouraged from discussing questions with each other, however they were able to discuss them with administrators who did not classify questions. The result of the manual classification step is a set of classified promoted questions that are suitable for evaluation. Our evaluations consisted of calculating inter-coder agreement statistics and manual examination of certain problematic questions as flagged by the statistics or coders.

### 3.2.4 Taxonomy Revision

Once inter-coder agreement statistics were calculated, we could begin evaluating them and revising the EAT taxonomy. In this step researchers collaborated with coders to discuss questions and categories with low inter-coder agreement scores. Researchers used this information to determine reason(s) why certain questions or categories in the taxonomy suffered lower agreement. The most common reasons for this were: misinterpretation of the category meaning on the part of coders, an ambiguity between two or more categories in the taxonomy, and a gap in the taxonomy where a question did not fit into any category.

If coders misinterpreted the meaning of a category, then we revised the taxonomy document shown in Appendix B to better explain the category. If an ambiguity existed between categories, then the categories were merged, clarified, or eliminated in order to make sure each question only fit into a single category. Similarly, if no category existed that clearly fit a question, then an existing one was expanded, or a new one was created to fill the gap. After all revisions were made to the taxonomy, we re-started the manual classification task using the new version of the taxonomy. In total, there were two pilot studies and six iterations in this work.

## 3.3 Hierarchical EATs

The taxonomy is a multi-level hierarchy that begins with three top level categories: Factual, PatientSpecific, and NonClinician. Factual questions are hypothetical questions that can be answered directly with medical facts by an automated QA system. This includes traditional factoid questions as well as more detailed descriptions such as "how" and "why" questions. PatientSpecific questions are also medical questions, except they are about a specific patient's condition or treatment. These questions can only be answered by a clinician that has examined the patient. NonClinician questions ask for information related to cancer that does not require medical information to answer. For example, questions about health insurance, legal issues, or emotional needs belong in the NonClinician category.

Using a hierarchical taxonomy as a series of classifiers has been shown to perform (at best) the same as a single level approach, except that it takes more time to run multiple classifiers in a production system [37]. Our taxonomy is divided into a hierarchy mainly to make manual classification easier for coders, by letting them narrow down each question to it's proper subcategory. We measure agreement and create classifiers with a flat set of classes, however we can divide the taxonomy into separate distributions based on its hierarchy. The different useful for testing agreement at different levels of the taxonomy

It is important for each valid distribution to cover the space of the entire taxonomy, so only a few distributions will be discussed. There are three main distributions that

are presented in this work for analysis. The *Level 1* distribution consists of all of the top-level categories in the taxonomy. Similarly, the *Level 2* distribution consists of all second-level categories. The *Terminal* distribution consists of all bottom-level categories (leaves in the taxonomy tree).

## 3.4 Final EAT Taxonomy

The final version of our taxonomy is described in terms of two trees. The first, shown in Figure 3.1, is a tree hierarchy where the Factual category is divided into three levels and the PatientSpecific and NonClinician categories are divided into two levels. The second, shown in Figure 3.2 is similar, except it is a two level hierarchy that is a transformation of the full hierarchy. We refer to this as a *Filtered Taxonomy*. Each category of the full taxonomy is described in detail (including examples) in Appendix B.

Our corpus proved to be too unbalanced with respect to the number of questions in each category in every distribution of the full taxonomy. In order to test classifiers more accurately, we created a 1 to 1 transformation that mapped this tree onto a smaller taxonomy where similar classes with low question counts were merged. Coders classified questions using the full taxonomy, while we used the filtered version to test classifiers. Having the corpus coded with the full taxonomy will be advantageous if our corpus becomes large enough that we can build a balanced training set with the full taxonomy.

## 3.5 Question Classification

Finally, we tested our final taxonomy and corpus with several SML classification algorithms using WEKA [22]. The purpose of our testing was to see how well classifiers could be trained with our data, and to see how different SML algorithms performed against each other. Our testing follows previous work [48] showing that while SVM classifiers do outperform others on contrived data [52], that may not be the case when they are confronted with real text data. Unfortunately we could not draw

- Factual
  - ClinicalDescription
    - Definition
    - Guideline
    - Purpose
    - ExplanationOfRelation
    - ExplanationOfProperty
  - Entity
    - Disease
    - Treatment
    - DiagnosticTechnique
    - Symptom
    - HealthEffect
    - RiskFactor
    - Prevention
    - NumericPropertyValue
  - ReferenceToInformation
    - ClinicalStudy
    - EducationalResource
- PatientSpecific
  - Diagnosis
  - HealthOutcome
  - MedicalRecommendation
  - Explanation
- NonClinician
  - NonClinicReference
  - NonClinicRecommendation
  - NonClinicDescription

Figure 3.1 Full Taxonomy

- Factual
  - Definition
  - Entity
  - EntityExplanation
  - NumericProsspertyValue
  - Reference
- PatientSpecific
  - PatientDiagnosis
  - PatientOutcome
  - PatientRecommendation
  - PatientExplanation
- NonClinician

Figure 3.2 Filtered Taxonomy

strong conclusions about the relative performance of the classification algorthms due to how unbalanced our corpus was.

# Chapter 4: Expected Answer Type Taxonomy Evolution

Our taxonomy evolved over the course of eight iterations. The first two iterations were pilot studies designed to test our procedure, custom classification software, web server, and training methods. The corpus used in these iterations did not contain enough questions to calculate meaningful inter-coder agreement statistics, therefore we did not calculate them. The following six iterations informed the majority of the changes to our taxonomy structure.

The main goal of our taxonomy is to identify EATs such that every question could be consistently placed into one and only one category by several independent coders. This means that no category would overlap another, and there were no gaps in the coverage of the set of all categories with respect to the EATs expressed by our corpus.

We had two sources of feedback for how well we achieved these goals: inter-coder agreement statistics and feedback from coders. Inter-coder agreement statistics were calculated based on coder classifications. Coders also identified some questions as impossible to classify because they did not fit into any category, or fit into more than one category. These questions were not counted in any agreement scores, but they did factor heavily into the evolution of our taxonomy by informing us of when and how to merge or create new categories.

Over the course of the experiment, there were several facts that came to light that made us change the procedure we followed and the tasks available to coders. First, some questions were promoted incorrectly and we would need the ability to reject these at any stage of the corpus construction (including after they were classified by one or more coders). Second, identifying why errors in Promotion and Classification were made as well as how to fix them could be more useful in identifying EATs than experimenting with revisions to the taxonomy. Third, the name of a category could alter what questions were placed into it, even if the definition did not change, so spending a lot of effort on EAT labels was a requirement. Last, interpreting the taxonomy and what constitutes the EAT of a question is a difficult task. Coders

needed a considerable amount of support in this area including examples, a (sometimes lengthy) training process, and the ability to ask experts questions in order to be able to promote or classify questions.

## 4.1 First Pilot Study

First, we performed a pilot study to test MedQuestionAdmin [33]. Coders were instructed to try each function of the application and provide feedback on its ease of use, speed, clarity, and any problems or errors. The corpus for this iteration consisted of 15 randomly chosen Raw Questions from our mined data. After promotion, there were 16 questions for coders to classify. This test occurred before the Verification step was added, so Promoted Questions were classified without review. Two paid coders participated in this iteration.

### 4.1.1 Bootstrap Taxonomy

Since this was the first test conducted, the taxonomy had to be constructed manually. We evaluated a random sample of 100 mined questions and tried to find patterns of EATs that could be used for classification. Next, we arranged these EATs into a two-level hierarchy shown in Figure 4.1.

### 4.1.2 First Pilot Study Results

We did not calculate any reliability statistics in this iteration because the small sample size would likely result in too much error to be useful [20, 31, 47]. To evaluate the classification results we looked at them as a group while discussing the performance of, and potential improvements to, MedQuestionAdmin [33]. This revealed almost no agreement among the two coders.

### 4.1.3 First Pilot Study Analysis

Discussions with the coders revealed that there was a lot of confusion about how to interpret the taxonomy documentation. Coders were classifying questions based on their interpretation of each category name and not the definition of the category.

- FactBased
  - ClinicalDescription
  - Entity
  - ReferenceToInformation
  - Complex
  - Statistic
- PeerOpinion
  - MentalHealth
  - Lifestyle
- PatientSpecific
  - Diagnosis
  - HealthOutcome
  - Recommendation
- NonMedical
  - Legal
  - Insurance

Figure 4.1 Bootstrap Taxonomy

This caused us to redesign the taxonomy document and add background information about the project, instructions on how to classify questions in general, and example questions for each category.

While redesigning the taxonomy document, we also realized that the Complex category was not an EAT. Its original intent was to cover questions that required an answer that was too long or complicated to be appropriate for an SMS message. That concept is important in a QA system, but does not constitute an EAT since it requires knowledge about a question's answer and questions belonging to any EAT may fit that description. We eliminated the Complex category to fix this problem, and decided that analyzing the complexity of an answer is best left in the answer processing portion of the QA system.

During the analysis phase of this test iteration we also evaluated more of our mined questions. In doing so, we found that there were very few questions in the Legal and Insurance categories. We decided that they did not warrant a top-level category of their own, so we merged PeerOpinion and NonMedical into a single NonMedical category. The resulting set of subcategories in NonMedical were defined based on subject matter rather than EAT. For example, the Legal category does not describe

a type of expected answer, but rather the subject matter of the question and answer. To fix this problem, we removed the Legal and Insurance categories and replaced them with categories similar to those in the FactBased and PatientSpecific groups.

## 4.2 Second Pilot Study

In light of the changes to the taxonomy document, we decided to perform a second test study. The corpus for this test consisted of 50 Raw Questions chosen at random from our mined data. After promotion there were 48 questions for coders to classify. The same two paid coders from the First Pilot Study participated in the second one.

### 4.2.1 Second Pilot Taxonomy

During the second pilot study, we used the revised taxonomy created at the end of the previous test iteration. We also changed some of the labels on categories for the sake of clarity. For example, PatientSpecific.Recommendation became PatientSpecific.MedicalRecommendation due to the addition of the category NonMedical.Recommendation. The taxonomy that coders used in the second test iteration is shown in Figure 4.2.

- Factual
  - ClinicalDescription
  - Entity
  - ReferenceToInformation
  - Statistic
- PatientSpecific
  - Diagnosis
  - HealthOutcome
  - MedicalRecommendation
- NonMedical
  - SocialReference
  - Recommendation
  - Description

Figure 4.2 Second Pilot Test Taxonomy

### 4.2.2 Second Pilot Study Results

We did not calculate inter-coder agreement statistics for this test because, like in the first pilot study, the number of questions was too low for them to be considered accurate [20, 31, 47]. Coders generally agreed on the first level of the taxonomy, especially in the Factual and PatientSpecific categories. However, there was still confusion about how to interpret some of the second level category meanings. This was especially true among subcategories of the Factual category.

### 4.2.3 Second Pilot Study Analysis

Despite our efforts to combat it, the main problem persisted from the First Pilot Study. Namely, coders were still unclear about the specific meaning of certain categories. Our proposed solution in the first pilot study was to change the classification instructions and add examples in the taxonomy document. Our solution to the problem in this study was to re-structure the EAT taxonomy hierarchy.

The corpus used in this test happened to skew heavily towards Factual questions, which revealed two things. First, many questions were classified as Factual.ClinicalDescription correctly, but they were so dissimilar that this information may not necessarily help a QA system. Second, many questions classified as Factual.ClinicalDescription or Factual.Entity were actually asking for a definition of medical terms.

We decided to add a third level to the taxonomy hierarchy under the Factual category. This allowed us to add more specific categories that would be more distinct to coders. It also allowed us to be more specific with respect to the EAT of Factual questions, which are the only type of question a medical QA system can directly answer. We did not add an additional level to the PatientSpecific and NonMedical categories, because coders were not as confused by their subcategories, and a QA system would not be directly answering these questions.

## 4.3 First Iteration

The next iteration we performed marked the first where inter-coder agreement statistics were applicable. The corpus consisted of 153 raw questions, which were reduced to 114 promoted questions. The set of raw questions contained all 65 questions used in both test iterations, and 88 additional randomly chosen questions. Four coders (two paid and two unpaid) categorized every question in the database. The two paid coders were the same as in the test iterations.

### 4.3.1 First Iteration Taxonomy

The taxonomy used in this iteration is shown in Figure 4.3. This is the first version of the taxonomy to contain a three level hierarchy in the Factual category. Some category names were changed from the previous version in the PatientSpecific and NonMedical subcategories for clarity.

### 4.3.2 First Iteration Results

The inter-coder agreement results for this iteration are shown in Table 4.1. There were 13 questions (11.4%) with an agreement of 0 (no two coders placed that question in the same category), and 12 questions (10.5%) with perfect agreement (all coders put them in the same category). As should be expected [37], the higher level category distributions showed higher agreement than the more specific ones.

We reviewed questions with low agreement manually to determine common features that could lead to corrections in the taxonomy. We also reviewed questions with perfect agreement to find parts of the taxonomy that were performing well.

There was high agreement in the Definition and Diagnosis categories. Questions involving relationships between entities or more specific properties of entities were among the lowest. Factual questions that were promoted with a significant amount of patient history also showed low agreement. In the Factual.Statistic category, coders expressed confusion between the DiseaseRisk and IncidenceRate categories. There were also a few Factual questions that did not appear to fit any existing category.

- Factual
  - ClinicalDescription
    - Definition
    - EntityRelation
    - ExplanationOfRisk
    - ExplanationOfCondition
    - ExplanationOfProcedure
  - Entity
    - Disease
    - Treatment
    - DiagnosticTechnique
    - Drug
    - Symptom
    - HealthEffect
    - RiskFactor
    - Prevention
  - ReferenceToInformation
    - ClinicalStudy
    - DescriptiveText
    - EducationalResource
- Factual (cont.)
  - Statistic
    - SurvivalRate
    - DiseaseRisk
    - IncidenceRate
- PatientSpecific
  - Diagnosis
  - HealthOutcome
  - MedicalRecommendation
- NonMedical
  - LifestyleReference
  - Recommendation
  - Description

Figure 4.3 First Iteration Taxonomy

While reviewing questions with low agreement, we found that coders promoted numerous Compound Questions. Having these questions in our classification corpus is problematic because coders will not be able to consistently assign a single classification to the question. What we observed is that coders will pick one of the questions in the Compound Question to categorize, but each coder may not pick the same one. Thus, no coder is classifying the question incorrectly, however they will not agree.

| EAT Distribution | $\kappa_1$ | $\alpha_1$ |
|---|---|---|
| Level 1 | 0.5308 | 0.5318 |
| Level 2 | 0.3611 | 0.3625 |
| Terminal | 0.3131 | 0.3146 |

Table 4.1 First Iteration Results

### 4.3.3 First Iteration Analysis

Questions with low-agreement asking for more detailed information about entities seemed to be grouped into one of two areas. Either they were asking for information about the relationship between two entities, or they were asking about a property of a single entity. The classifications of these questions were spread around the ExplanationOfRisk, ExplanationOfCondition, and ExplanationOfProcedure categories fairly evenly. Evidently, the ExplanationOf... categories were not distinct enough and these questions were overlapping. We decided to replace all three categories with ExplanationOfRelation and ExplanationOfProperty to cover these questions.

Some Factual questions with a significant amount of patient history require its presence to answer the question, but most of these questions do not. In this experiment, coders were classifying anything presenting patient information as PatientSpecific. We made coders aware verbally, and by updating the examples in the taxonomy document, that just because a question presents patient specific information does not mean the question is fits into the PatientSpecific category. The few Factual questions that did not fit in any category required three new ones to be added, two of which replaced existing categories that were either unclear or not distinct from others.

The presence of compound questions in the set of promoted questions required us to re-think our approach to promotion. The concept of a Verification Task was discussed, but we opted instead to instruct coders specifically to not promote Compound Questions. We did so verbally, and changed the promotion instructions to reflect the necessity of not promoting compound questions. We also decided we would add a verify step to the promotion process if we saw similar results after the next iteration of promotion.

## 4.4 Second Iteration

The corpus in this iteration consisted of 235 randomly selected raw questions. After promotion, there were 195 questions for the categorization task. There were five coders in this iteration, three paid and two unpaid. Two of the paid coders and one of the unpaid coders were the same as in the previous iteration. All five coders categorized every question in the corpus.

### 4.4.1 Second Iteration Taxonomy

Figure 4.4 is the version used in this iteration. A prefix was added to all Non-Medical categories for clarity, and the DescriptiveText category was removed. Other than these changes, this is the same version of the taxonomy as the revised version from the first iteration.

### 4.4.2 Second Iteration Results

The inter-coder agreement results are shown in Table 4.2. There were 4 questions (2%) with no agreement, and 28 (14.4%) questions with perfect agreement. While there was an improvement, the agreement was not much better than in the first iteration.

| EAT Distribution | $\kappa_1$ | $\boldsymbol{\kappa_2}$ | $\alpha_1$ | $\boldsymbol{\alpha_2}$ |
|---|---|---|---|---|
| Level 1 | 0.5308 | **0.5996** | 0.5318 | **0.6000** |
| Level 2 | 0.3611 | **0.3650** | 0.3625 | **0.3656** |
| Terminal | 0.3131 | **0.3747** | 0.3146 | **0.3753** |

Table 4.2 Second Iteration Results

Again, we analyzed questions with low agreement by hand to determine common characteristics that could be corrected. We found that a large number of promoted questions still contained Compound Questions that should have been split apart or not promoted. This was a much larger number than had been seen in the test iterations. The difference could be explained by two factors. First, the presence of new, less experienced coders promoting questions could have led to more mistakes being made.

- Factual
  - ClinicalDescription
    - Definition
    - Guideline
    - Purpose
    - ExplanationOfRelation
    - ExplanationOfProperty
  - Entity
    - Disease
    - Treatment
    - DiagnosticTechnique
    - Symptom
    - HealthEffect
    - RiskFactor
    - Prevention
  - ReferenceToInformation
    - ClinicalStudy
    - EducationalResource
- Factual (cont.)
  - Statistic
    - SurvivalRate
    - DiseaseRisk
    - AverageDuration
- PatientSpecific
  - Diagnosis
  - HealthOutcome
  - MedicalRecommendation
  - Explanation
- NonMedical
  - NonMedReference
  - NonMedRecommendation
  - NonMedDescription

Figure 4.4 Second Iteration Taxonomy

Second, questions in the corpus are chosen randomly, and this particular corpus may have coincidentally had a larger number of Compound Questions.

### 4.4.3 Second Iteration Analysis

The chief problem shown in this iteration was with promoting Compound Questions. After the first iteration, we attempted to correct this via written and verbal instruction on promotion. These attempts were, evidently, not enough to change the outcome. Therefore, we decided to add the Verification Task in the promotion process between Promotion and Classification.

## 4.5 Third Iteration

The third iteration was the first to include the Verification Task. The corpus consisted of 335 Raw Questions (235 Raw Questions from the previous iteration and 100 additional randomly chosen questions). All promoted questions from the previous iteration that contained Compound Questions were demoted to Raw Questions so coders could promote them again. After promotion, 254 questions were available for classification. 5 coders participated in this iteration, 3 paid and 2 unpaid. Of these coders, only 1 paid and 1 unpaid coder participated in the previous iteration. Not every question was classified by every coder, so a given question could have between 2 and 5 classifications.

### 4.5.1 Third Iteration Taxonomy

The purpose of this experiment was to detect whether or not the verification step kept Compound Questions out of the set of promoted questions. Therefore, the taxonomy for this iteration was exactly the same as in the previous iteration (shown in Figure 4.4).

### 4.5.2 Third Iteration Results

Since this iteration allowed a variable number of coders for each question, Fleiss' Kappa could not be used to calculate inter-coder agreement [20, 47]. The inter-coder agreement results are shown in Table 4.3. There were no questions with an agreement of 0, and 96 (37.8%) questions with perfect agreement. Again, we see improvement in the values, but they have not improved enough to allow us to use this taxonomy to train a classifier [15, 16, 31, 47, 53]. The improvement seen here can be attributed to the addition of the Verification Task, because no other changes were introduced to the experiment setup.

Hand evaluation of low agreement questions revealed almost no Compound Questions had been promoted in this iteration. The addition of the Verification step to promotion can therefore be considered successful. However, our estimates on how

| EAT Distribution | $\alpha_1$ | $\alpha_2$ | $\boldsymbol{\alpha_3}$ |
|---|---|---|---|
| Level 1 | 0.5318 | 0.6000 | **0.6855** |
| Level 2 | 0.3625 | 0.3656 | **0.4754** |
| Terminal | 0.3146 | 0.3753 | **0.4221** |

Table 4.3 Third Iteration Results

much the presence of compound questions impacted our inter-coder agreement scores were high.

Further evaluation of the low agreement questions revealed that there was no category for questions requesting a numeric value other than a statistic (AverageDuration). This led coders to categorize questions asking for numeric values incorrectly as either Factual.Statistic.AverageDuration or a subcateogry of ClinicalDescription.

During the classification phase of this iteration, some coders expressed a desire to be able to mark questions as not belonging to any category, or to belonging to multiple categories. Neither was possible in the version of MedQuestionAdmin [33] used for this iteration. Coders stated they classified questions fitting these descriptions as the "closest" category, or not at all. This could lead to classification errors, as coders were not always in agreement with each others' classifications.

### 4.5.3 Third Iteration Analysis

After examining questions that asked for a numeric value further, we decided that this warranted its own EAT. We also realized that the Factual.Statistic categories are essentially asking for a numeric value as well. To correct the gap in the taxonomy, we added a Factual.Entity.NumericPropertyValue category. This overlapped with the Statistic categories, so they were removed in favor of the more general NumericPropertyValue category. Answering these questions would require knowing the type of statistic the user asked for, but Named Entity Recognition or a similar technique would be better suited to obtaining that information [29].

In order to address the desire of coders to mark questions with meta information about the taxonomy (or the question promotion) we added *Wildcard Categories*. These are a different type of category that can be placed anywhere in the taxonomy

hierarchy to provide meta information about a question. We added three wildcard categories: OtherCategory, Ambiguous, and MultipleCategory.

OtherCategory encapsulates questions that do not belong to any category in the taxonomy. This gives coders the ability to flag questions that may indicate a gap in converage of the taxonomy.

Ambiguous questions are those that can be interpreted multiple ways, and are not able to be reliably classified. Compound Questions fit into the this category. Flagging questions as ambiguous gives us the ability to eliminate improperly promoted and verified questions during the classification task.

MultipleCategory encapsulates questions that fit equally into more than one category. This is different from Ambiguous in that the question will only have a single interpretation, yet still fit into more than one category. MultipleCategory gives us the ability to flag questions that may indicate categories which are not completely distinct.

## 4.6 Fourth Iteration

In the fourth iteration, we added many more questions to the database. This allowed coders to continue promoting, verifying, and classifying questions while our analysis was ongoing. However, it also meant that we finished this iteration before coders had actually seen all of the raw questions in the corpus.

The previous 335 questions and 3950 new ones were included in this iteration, for a total of 4,285 raw questions in the corpus. There were 1,767 promoted questions, and 558 of them were verified when inter-coder agreement statistics were calculated. Of the verified questions, 18 were placed into a Wildcard category by at least one coder. Wildcard questions were analyzed separately, and are not counted toward inter-coder agreement. Removing those 18 questions left 507 classified questions which were classified by 2 paid coders, one of which participated in the previous iteration.

### 4.6.1 Fourth Iteration Taxonomy

Figure 4.5 shows the taxonomy used in this iteration. It was the same as revised in the previous iteration, except that the NonMedical category was renamed to NonClinician for clarity. Similarly, the "NonMed" prefixed subcategories were also renamed to begin with "NonClinic". Coders explained that the name "NonMedical" indicated to them that questions in this category would have nothing to do with medicine (e.g. just for legal or insurance questions). Thus, they would often place "NonMed" questions in the Factual or PatientSpecific categories if, for example, the question was requesting a recommendation of a particular type of hospital. We decided that a more clear name was "NonClinician" to indicate that questions in this category could be answered by non-clinicians, but they still may have something to do with medical information.

### 4.6.2 Fourth Iteration Results

Each question was classified by the same number of coders, so we could use both Fleiss' Kappa and Krippendorff's Alpha in our analysis [20,31,47]. Table 4.4 shows the results for the fourth iteration. There were 231 (45.56%) questions with no agreement, and 276 (54.44%) with perfect agreement. These measures are vastly different from previous iterations because there were only two coders in this iteration, so every question either had no agreement or perfect agreement. Once again, there was an improvement in every EAT Distribution of the taxonomy.

| EAT Distribution | $\kappa_1$ | $\kappa_2$ | $\boldsymbol{\kappa_4}$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\boldsymbol{\alpha_4}$ |
|---|---|---|---|---|---|---|---|
| Level 1 | 0.5308 | 0.5996 | **0.7622** | 0.5318 | 0.6000 | 0.6855 | **0.7624** |
| Level 2 | 0.3611 | 0.3650 | **0.5461** | 0.3625 | 0.3656 | 0.4754 | **0.5466** |
| Terminal | 0.3131 | 0.3747 | **0.4921** | 0.3146 | 0.3753 | 0.4221 | **0.4926** |

Table 4.4 Fourth Iteration Results

There are two key features that these inter-coder agreement results illustrate. First, the Level 1 distribution passed the 0.7 threshold for both $\kappa$ and $\alpha$. This indicates that it is possible to train a functioning classifier with that data [47]. Second,

- Factual
  - ClinicalDescription
    - Definition
    - Guideline
    - Purpose
    - ExplanationOfRelation
    - ExplanationOfProperty
  - Entity
    - Disease
    - Treatment
    - DiagnosticTechnique
    - Symptom
    - HealthEffect
    - RiskFactor
    - Prevention
    - NumericPropertyValue
  - ReferenceToInformation
    - ClinicalStudy
    - EducationalResource
- PatientSpecific
  - Diagnosis
  - HealthOutcome
  - MedicalRecommendation
  - Explanation
- NonClinician
  - NonClinicReference
  - NonClinicRecommendation
  - NonClinicDescription
- OtherCategory
- Ambiguous
- MultipleCategory

Figure 4.5 Fourth Iteration Taxonomy

the values for $\kappa$ and $\alpha$ are very close to one another. The consistency of the different measures means that it is less likely that either statistic has been skewed by any (as yet unknown) natural feature of the data distribution [47].

We reviewed most questions in this iteration (not just those with low agreement), and all 18 wildcard questions. There were two areas where disagreement was most common. First, one or more coders did not place 43% of questions in the correct category according to the taxonomy in the Terminal Distribution. The second area showed that coders confused the Factual.Entity.Prevention and NonClinician.NonClinicRecommendation categories, which indicated a possible ambiguity between the two categories.

### 4.6.3 Fourth Iteration Analysis

In the case of questions where disagreement was caused by incorrect classification(s); it was clear (from other questions) that the coders understood the categories, but still classified questions incorrectly. Questions similar to those with errors were classified correctly by the same coders that made errors. It is possible that questions with errors were classified earlier, and the coders' understanding of the taxonomy had since improved. We cannot be sure of this, because coders can classify questions in any order and MedQuestionAdmin [33] does not record the time or order of any user action.

We expected some user error to be present, however we did not expect it to be so prevalent. Some questions with full agreement were also placed into an incorrect category by both coders. We determined this by having additional coders review certain questions independently. Our original goal was to create a training data set for building and testing our question classifier using automated methods based on the coders' data. However, the amount of errors encountered in this step led us to create a separate task for entering the classification for each question that is used to train our classifiers. This task allows an administrator to manually select the classification for each question while viewing all coder classifications (including Wildcard categories). We can check the separate classification data against the original coder data and manually investigate any discrepancies with multiple coders.

The confusion between Prevention and NonClinicRecommendation related exclusively to questions concerning diet and exercise. This can be traced back to an ambiguity in the EAT Taxonomy. Diet and exercise can qualify as preventative measures for cancer, but the taxonomy places these questions in the NonClinician parent category. Our solution was to move diet and exercise questions to the Factual parent category, which would place all ambiguous questions in the Factual.Entity.Prevention category. Diet and exercise questions pertaining to a specific patient would fit into the PatientSpecific categories.

Almost all of the 18 Wildcard questions in this iteration should not have been promoted or verified in the first place. They consisted of three types of questions. The first type were questions asking for a personal testimonial, which an automated system cannot provide. The second type were Compound Questions that should have been split in the Promotion phase. The third type were statements (rather than questions) that did not express any discernible information need. These results validate the decision to include Wildcard Categories in the taxonomy, even if they are not used in the EAT Distributions of any classifiers built with this corpus.

## 4.7 Fifth Iteration

Int the fifth iteration, we continued the approach taken in the fourth, and allowed coders to continue promoting, verifying, and classifying questions at will. In this iteration, there were 4487 promoted questions, of which 1503 were verified. Out of the 1503 verified questions, 1175 were classified by multiple coders. There were 42 questions placed in wildcard categories, leaving 1133 classified questions. The same two coders from the previous iteration classified all of these questions.

### 4.7.1 Fifth Iteration Taxonomy

At the conclusion of the fourth iteration we tested our hand-classified corpus as a training set for a classifier using WEKA [22]. In doing so, we discovered that questions were not evenly distributed across the set of terminal categories. For example, Factual.Entity.Disease and Factual.ReferenceToInformation.ClinicalStudy had a single question each, whereas there were 86 questions classified as Factual.ClinicalDescription.ExplanationOfRelation. This imbalance in the training data reduces the accuracy of SML algorithms, because they do not have a sufficient number of examples to properly build a model of the data [16].

Since our data consists solely of real world questions, we cannot easily change the category distribution. The only way to balance the categories is to classify enough questions that we can select the same number from each category, or change the

taxonomy such that it is more balanced. Our coders were classifying questions as fast as they could, so we opted for the second solution.

We collapsed related categories with low question counts together at the same level, or into their parent category (if possible). This allowed us to apply a transform to existing classifications and arrive at a classification in the new taxonomy hierarchy. Coders continued to categorize questions using the previous version of the taxonomy, since it gives a higher resolution for the EAT of each question.

We collapsed all of the NonClinician subcategories into a single category since those questions were the most rare in our corpus. We also collapsed the three-level hierarchy for Factual questions into two levels. All of the ReferenceToInformation subcategories were collapsed into the Reference category. All of the Entity subcategories (except for NumericPropertyValue) were collapsed together as well. The ExplanationOfRelation and ExplanationOfProperty categories were merged into a single EntityExplanation category. The resulting taxonomy is shown in Figure 4.6.

- Factual
  - Definition
  - Entity
  - EntityExplanation
  - NumericPropertyValue
  - Reference
- PatientSpecific
  - PatientDiagnosis
  - PatientOutcome
  - PatientRecommendation
  - PatientExplanation
- NonClinician
- OtherCategory
- Ambiguous
- MultipleCategory

Figure 4.6 Fifth Iteration Taxonomy

### 4.7.2 Fifth Iteration Results

The agreement results for the fifth iteration are shown in Tables 4.5 and 4.6. There were 435 (38.3%) questions with no agreement, and 698 (61.61%) questions

with perfect agreement. The taxonomy revision for the fifth iteration contains only two levels, so the Level 2 and Terminal distributions are now the same.

| $\kappa$ EAT Distribution | $\kappa_1$ | $\kappa_2$ | $\kappa_4$ | $\boldsymbol{\kappa_5}$ |
|---|---|---|---|---|
| Level 1 | 0.5308 | 0.5996 | 0.7622 | **0.6955** |
| Terminal | 0.3131 | 0.3747 | 0.4921 | **0.5404** |

Table 4.5 Fifth Iteration $\kappa$ Results

| $\alpha$ EAT Distribution | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\boldsymbol{\alpha_5}$ |
|---|---|---|---|---|---|
| Level 1 | 0.5318 | 0.6000 | 0.6855 | 0.7624 | **0.6956** |
| Terminal | 0.3146 | 0.3753 | 0.4221 | 0.4926 | **0.5406** |

Table 4.6 Fifth Iteration $\alpha$ Results

The two key features in the inter-coder agreement results is that $\kappa$ and $\alpha$ for the Level 1 distribution went down, while agreement for the Terminal distribution went up. With the exception of a label change, the Level 1 distribution has not changed since the second iteration. Therefore, changes to the taxonomy in this version cannot be the cause of the drop in Level 1 agreement.

The Level 1 distribution did not changed, and the Terminal distribution was collapsed to compensate for an uneven category distribution in the corpus. We expected Level 1 agreement to remain the same, and Terminal agreement to increase as would be consistent with the changes we made to the taxonomy. Instead, Level 1 agreement decreased and Terminal agreement increased, leading us to believe that the lower agreement seen in Level 1 was not due to changes in the taxonomy. We suspected that the decrease in Level 1 agreement was related to coder errors when classifying questions.

### 4.7.3 Fifth Iteration Analysis

We evaluated the agreement results by manually inspecting coder classifications. We found that a large portion of what are likely the most recently classified questions were classified incorrectly according to the taxonomy by one or more coders. MedQuestionAdmin [33] does not timestamp classifications, nor keep track of any

classification order; so we cannot be certain that coder errors increased over time. However, most errors seemed to occur on questions with larger IDs, which are generally classified later; as coders are shown questions ordered by ID for every task in MedQuestionAdmin [33]. The fact that inter-coder agreement improved in the Level 1 distribution continuously up to the fifth iteration, and there were no changes to that level of the taxonomy since the second iteration also points to an increase in errors.

During this iteration, both coders were also working more sporadically than previously and had taken long breaks from working on the project. This led us to believe that they may simply have forgotten some parts of the taxonomy and may be trying to classify questions from memory. Further discussions with coders revealed that they did perform most classifications from memory, and only consulted the taxonomy document when they were unsure. Our solution to this problem was to instruct the coders to review the taxonomy document, and then review their most recent classifications to try and correct any errors.

## 4.8 Sixth Iteration

The purpose of the sixth and final iteration was to see if we could remedy the classification errors seen in the fifth iteration. There were 7692 raw questions in our database in this iteration. 1503 questions were promoted, verified, and classified by the same two paid coders as the previous iteration. There were 48 questions placed in Wildcard categories, leaving a total of 1455 classified questions in our corpus.

### 4.8.1 Sixth Iteration Taxonomy

Since the point of this iteration was to correct errors made by coders during the previous iteration, the taxonomy remained the same as in the previous iteration.

### 4.8.2 Sixth Iteration Results

The results from this iteration are shown in Tables 4.7 and 4.8. The $\kappa$ and $\alpha$ values are largely the same as in the previous iteration. There were 567 (38.97%) questions with no agreement and 888 (61.03%) questions with perfect agreement.

| $\kappa$ EAT Distribution | $\kappa_1$ | $\kappa_2$ | $\kappa_4$ | $\kappa_5$ | $\boldsymbol{\kappa_6}$ |
|---|---|---|---|---|---|
| Level 1 | 0.5308 | 0.5996 | 0.7622 | 0.6955 | **0.6969** |
| Terminal | 0.3131 | 0.3747 | 0.4921 | 0.5404 | **0.5288** |

Table 4.7 Sixth Iteration $\kappa$ Results

| $\alpha$ EAT Distribution | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ | $\boldsymbol{\alpha_6}$ |
|---|---|---|---|---|---|---|
| Level 1 | 0.5318 | 0.6000 | 0.6855 | 0.7624 | 0.6956 | **0.6970** |
| Terminal | 0.3146 | 0.3753 | 0.4221 | 0.4926 | 0.5406 | **0.5290** |

Table 4.8 Sixth Iteration $\alpha$ Results

### 4.8.3 Sixth Iteration Analysis

Again, we manually inspected question classifications. Analyzing new classifications as well as those that were previously classified revealed the same types of errors as seen in the previous iteration. Some older questions appear to have been corrected by coders, but others also appear to have been changed from the correct classification to an incorrect one. The number of classification errors mean that we cannot reliability use this data to make changes to the taxonomy, since the questions have not been classified according to the taxonomy.

Coder agreement declined in the fifth iteration, which was also the same iteration where we did not place an artificial limit on the number of questions they were instructed to process. In every previous iteration there were a set number of questions for coders to process, and they knew when they would be done with them. But, in every subsequent iteration, there were even more questions added to the database also with no set goal for coders. It is possible that the errors we have seen are due to coders' rushing through the classification process in order to finish more questions.

To combat coders' rushing through questions, we proposed changes to the GUI in MedQuestionAdmin [33] to slow them down and force them to consider more of the taxonomy when classifying questions. Modifying the control structure for selecting a class so only the current level of the taxonomy could be seen at any one time would prevent coders from visually considering all 22 terminal classes at once. It would break up the process of classification and, hopefully, coders would not skip to the first class they though of, but rather consider only which class fit on a level-by-level basis. This change would also free space on the web page for text describing each category, which could also help correct errors as coders have previously stated they do not like looking through the taxonomy document while classifying.

These UI changes would take time to implement, and not help with existing classifications. We tried to have coders review completed questions in the previous iteration, with mixed results. Removing classifications from iterations 5 and 6 and having coders re-do them with a new UI could potentially solve the problem, but would take weeks or months to complete. At this point in our study, we decided that time would be better spent evaluating classification algorithms rather than revising the EAT taxonomy with our current data. A coder can take up to 3 months to train, and we did not have the time to recruit new people to continue classifying questions.

# Chapter 5: Classifier Analysis

Although we have not built a large enough corpus to train a production classifier (previous research used corpora with as many as 5,500 questions [52]) on the Terminal distribution, we began using WEKA [22] to build and test classifiers based on our taxonomy and corpus. Our testing involved multiple classification algorithms that have been used previously for text classification [13, 16, 29, 36, 48] as well as other techniques that can compliment this type of question classification [14, 21, 29].

A total of 1904 Raw Questions were processed for our testing, with 757 (39.7%) rejected by coders. The 1147 remaining Raw Questions yielded 1503 Promoted Questions. All 1503 Promoted Questions were classified by two coders, with 1279 placed in the final classified corpus. The remaining 224 (15%) questions were thrown out by coders (during classification) or administrators (post classification). These questions were promoted improperly either by being edited or split incorrectly, or were irrelevant to our target QA task from the beginning. There were no relevant questions that coders were unable to place in a Level 1 category.

Unfortunately, due to the uneven distribution of questions across categories, and the relatively low number of questions in our corpus, we were unable to draw strong conclusions from our testing. We did successfully build a framework for collecting questions, creating a corpus, creating an EAT taxonomy, and testing different classification techniques with our corpus and taxonomy. Our framework is also suitable for our own ongoing research in medical and cancer question answering, as well as for QA tasks in other domains using a medium other than SMS.

## 5.1 Corpora

We applied several transformations to our corpus to test dimensionality reduction techniques with real user data. Our main goal was to see how many terms could be trimmed from our corpus, and how that would affect classifiers trained on the resulting corpora. We attempted the Latent Semantic Indexing (LSI) [14,16] attribute selection in WEKA [22], however it proved to require too much memory to be practical. We

implemented two Local Relevancy LSI techniques similar to a a Ladder-Weighted LSI [14]. We also tested automated spelling correction as a means to reduce the dimensionality of our corpus.

We first investigated a flat threshold Local Relevancy LSI technique for the incidence rate of each term. A term was trimmed from the corpus if it appeared in more than a certain percentage of questions in every category. We designed this test to try and remove terms that occurred frequently across all categories. We tested the range of 5-25% as the threshold in increments of 5% (there were no terms that appeared in 30% or more of questions every category) to see how different threshold values affected the term count of the corpus and classifier results.

The second Local Relevancy LSI technique we tried used a range of incidence rates. A term was trimmed from the corpus if its incidence rate across all categories fell within a specified percentage of the mean incidence rate. This eliminated terms that had a similar rate of occurrence across all categories, not just a high rate. We tested ranges of 30-50% of the median in increments of 5% to see how different ranges affected the corpus term count and classifier results.

Since our corpus is not corrected for spelling or grammar, classifiers will interpret different spellings (some valid and some not) of the same word as different terms in our corpus. It would reduce the dimensionality of our corpus, as well as increase the similarity of questions' term-doc vectors to have a corpus with correct spelling. However, we cannot hand correct any spelling since we are targeting an automated QA system. Instead, we experimented with automated spelling correction on our corpus.

We tested spelling correction with Lucene's [21] built-in spell-checking functionality. This spell checker takes a dictionary text file as input for the correct spelling of words. Our corpus deals with cancer, therefore our dictionary required medical terminology as well as standard English. We combined the ISpell [32] standard American English dictionary as compiled by WordList [2] and the Consumer Health Vocabulary [51] dictionaries into a single input file.

We chose Lucene [21] spelling correction because it was the most convenient to incorporate into our software and was the simplest to use. It did not require any pre-parsed [29] or otherwise formatted data, just a flat list of words. This was important because we were combining multiple dictionaries using custom software and it would be much more difficult and time consuming to edit them to match a more complicated input file format.

Table 5.1 contains the number of unique terms (term count) in each transformation of our corpus. Corpora with the threshold transform applied are labeled as $T_{xx}$, were $xx$ is the threshold percent for trimming a term from the corpus. Similarly, Corpora with the range transform applied applied are labeled as $R_{xx}$, where $xx$ is the maximum percent deviance from the mean that incidence rates can occur in for a term to be trimmed from the corpus. The unmodified corpus (no transformation applied) is denoted by a U, and the corpus with corrected spelling is denoted by SP.

| Corpus | Term Count |
|---|---|
| U | 4356 |
| SP | 4164 |
| $T_5$ | 4290 |
| $T_{10}$ | 4323 |
| $T_{15}$ | 4333 |
| $T_{20}$ | 4339 |
| $T_{25}$ | 4344 |
| $R_{30}$ | 4308 |
| $R_{35}$ | 4284 |
| $R_{40}$ | 4260 |
| $R_{45}$ | 4228 |
| $R_{50}$ | 4202 |

Table 5.1 Corpora Unique Term Counts

Table 5.1 shows that the dimensionality reduction techniques did not impact the term count of the corpus by more than 3.5 %. In most cases, the term reduction was much less. Matrices made of term-doc vectors are usually very sparse [29], and our corpus is no exception. Additionally, our corpus has a much shorter document length that probably served to make its matrix even more sparse. This could explain why the threshold and range based term trimming did not remove very many terms.

Our corpora consisted of 1279 questions total, all of which were promoted, verified, and classified by multiple coders. This is less than the 1455 questions in the last iteration of our research because some questions were thrown out as not being relevant to the project after they were classified. These 176 questions should not have been promoted, or have been classified in Wildcard categories', but they were missed by coders.

## 5.2 Classification Algorithms

There are numerous classification algorithms that have been used for QA tasks [13, 16, 27, 29, 35, 48, 52]. Previous research suggests that Support Vector Machines [7] are superior to other techniques [52], however more recent research shows that this conclusion may be based on incomplete information [48]. Since there is some disagreement as to the most advantageous algorithm to use, we tested multiple algorithms and implementations using the default implementations in WEKA [22]. The algorithms we tested were Naive Bayes (NB) [28], Multinomial Naive Bayes (MNB) [39], J48 Trees [46], and a Sequential Minimal Optimization (SMO) implementation of Support Vector Machines (SVM) [44].

The "standard" NB and the multinomial algorithms differ in the type of data they use. MNB uses a term-document frequency vector (e.g. the number of times a term occurs in the document). In contrast, a traditional approach uses a binary value to indicate whether or not a term is present in the document.

The SMO implementation uses a different algorithm to train a SVM classifier [29, 44]. SMO breaks the complex task of training the SVM into numerous smaller, simpler problems and is able to train a classifier much faster than previous implementations [44]. The resultant classifier is nearly identical to one that would be created by a more standard implementation such as LibSVM [17].

We used a linear kernel for SMO (the weka.classifiers.functions.SMO Polynomial kernel option, with an exponent of 1) for the entirety of our testing. We tried several others including a radial kernel and a quadratic kernel, however they did not achieve results anywhere close to the linear kernel. Our findings with respect to the type

of kernel in an SMO implementation match the majority of other research in using SVMs to classify questions [26, 48, 52].

All classifiers were tested with a 10-fold cross validation [16, 22]. The cross validation implementation in WEKA [22] is randomized, so the folds will be different every time a test is run, which leads to slightly different accuracy statistics. It also means that across each corpus and algorithm test the same folds were likely not run twice, so some variation is to be expected, and may not indicate any difference in the performance of different classifiers.

## 5.3 Level 1 Classifiers

The final Level 1 Distribution agreement was close to 0.7, which is considered to be the minimum value for a reliable coding scheme [47]. This makes the Level 1 Distribution the most likely to create a successful classifier. We expected this distribution to have the highest accuracy of all distributions' classification results. The distribution of questions across each category is shown in Table 5.2.

| Category | Question Count |
|---|---|
| Factual (F) | 561 |
| PatientSpecific (PS) | 613 |
| NonClinician (NC) | 105 |

Table 5.2 Level 1 Corpus Question Distribution

The distribution of questions across the three categories in the Level 1 Distribution illustrates how unbalanced our corpus turned out to be. This definitely impacted the training and performance of the classifiers negatively, although how much and in what way is difficult to tell. The confusion matrices for NB and SMO for the unmodified corpus are shown in Table 5.3. The remaining matrices can be found in Appendix C. They all illustrate some of these effects. Factual and PatientSpecific questions show a much better rate of true positives than NonClinician questions, which is at least partly attributable to the fact that there are 5-6 times as many examples used to train the classifiers.

| NB | F | PS | NC |
|---|---|---|---|
| F | 415 | 116 | 30 |
| PS | 129 | 428 | 56 |
| NC | 37 | 28 | 40 |

| SMO | F | PS | NC |
|---|---|---|---|
| F | 453 | 106 | 2 |
| PS | 122 | 479 | 12 |
| NC | 34 | 37 | 34 |

Horizontal Axis = Classified As
Vertical Axis = Actual Category

Table 5.3 Unmodified Level 1 Corpus Confusion Matrices

The questions included in our corpus represent a random sample of the cQA questions we collected, however it is unknown whether this is a representative sample or not. We have promoted, verified, and classified only 1503 questions out of 5062 that have been promoted, and 7692 in our current database. However, there are over 50.000 questions that we have collected, but not yet added to our database (for testing), and have not been reviewed. Therefore it is not possible for us to make any conclusions about the overall distribution of cancer questions when we have classified only 3 % of the ones we have collected. The percentage of correctly classified questions for all configurations of the Level 1 Distribution classifiers are shown in Table 5.4.

| Level 1 Corpus | NB | MNB | J48 Tree | SMO |
|---|---|---|---|---|
| U | 69.0 % | 73.4 % | 69.1 % | 75.5 % |
| SC | 69.4 % | 73.9 % | 66.4 % | 73.6 % |
| $T_5$ | 68.8 % | 70.4 % | 66.6 % | 71.6 % |
| $T_{10}$ | 70.0 % | 72.0 % | 69.3 % | 72.6 % |
| $T_{15}$ | 69.2 % | 71.6 % | 68.5 % | 72.3 % |
| $T_{20}$ | 69.5 % | 73.3 % | 67.5 % | 73.9 % |
| $T_{25}$ | 70.1 % | 72.9 % | 68.2 % | 73.6 % |
| $R_{30}$ | 69.4 % | 72.2 % | 70.0 % | 75.3 % |
| $R_{35}$ | 69.8 % | 72.5 % | 68.6 % | 74.4 % |
| $R_{40}$ | 70.0 % | 72.4 % | 70.2 % | 72.8 % |
| $R_{45}$ | 69.4 % | 72.5 % | 68.2 % | 73.7 % |
| $R_{50}$ | 70.2 % | 72.0 % | 70.7 % | 71.1 % |

Table 5.4 Level 1 Classifier Percent Correct

It remains unclear from Table 5.4 which algorithm (if any) performs the best for this type of QC task. SMO and Multinomial Naive Bayes outperformed J48 and Naive Bayes with our corpus. However, we cannot draw any conclusions for a general

QC task given the unbalanced category distribution of our corpus. Also, while our corpus is composed of user questions in the closest format that we could achieve to SMS questions, they are not actual SMS questions. The method users have to ask questions (e.g. a phone versus a computer) may impact the type of questions they ask, which could make the natural distribution of questions an SMS system would receive different from the distribution in our corpus.

While not as accurate as other classifiers [48, 50, 52], these results are similar to what other QC research comparing different algorithms has found [48, 52]. This serves to further illustrate the need for building a large corpus of cancer questions in order to develop better QC and QA techniques in the cancer domain. Fortunately, our research group has already begun collecting questions for a pilot study involving cancer patients actually using SMS devices to ask questions.

Results in Table 5.4 suggest that dimensionality reduction and spelling correction techniques had very little impact on actual question classification. Table 5.1 showed that our dimensionality reduction did not trim many terms, so it follows that it would also not affect classifiers a great deal. Our application that performs those tasks outputs all trimmed/corrected terms to log files, so we could inspect the transformations they made on the corpus.

We examined these results by hand to learn why they had so little impact on our classifiers. There were 584 instances where a term was changed by our spelling correction. The term count of the corpus was altered the most by the spelling correction, yet there did not seem to be any real change in the classifier results. Closer inspection of the corrected terms revealed two main problems.

The first problem was that, while we were correcting spelling for single words, mistakes in typing do not necessarily adhere to single misspelled words. There were numerous instances in our corpus where users' mistakes carried across multiple terms. For example, typing "qualitycare" instead of "quality care", or "th etreatment" instead of "the treatment". In these cases, our spelling correction would make a guess

as to the intended term, but it would always be wrong. More advanced spelling correction could mitigate this problem, as there are other techniques available to correct these types of errors [29].

The second problem was that correcting misspelled medical terms is not as simple as other English words, because terms with radically different meanings can be very close to one another lexicographically. Upon examining the log file from our spell checker we found many instances where the corrected term was actually wrong. In these cases, the misspellings were actually lexicographically closer to other terms in our dictionary rather than the correct one. For example, a user misspelled "chemotherapy" as "quimotherapy", and it was corrected to "biotherapy". This error removed a feature that could have placed the question in a correct category, and added a feature that did not exist in the question to begin with. These types of errors yield unpredictable results in a classifier as they distort the information content of the questions they affect.

While the first problem with our spelling correction could be solved using existing techniques, we did not find a possible solution to the second problem. Some specialized medical spell checking programs do exist [30, 43], however these all require a clinician to discriminate between the results. This makes them no better than the approach we already tried.

Detailed classification results for the Level 1 distribution are shown in 5.5 for the SMO classifier. The full results for every classification algorithm are shown in Appendix C. As expected, the categories with a larger number of questions in our corpus (Factual and PatientSpecific) had much better results than the one with comparably fewer questions (NonClinician). Precision and Recall were both consistently high in the Factual and PatientSpecific categories across different corpus modifications, whereas NonClinician had slightly worse Precision and terrible Recall. It is possible that NonClinician results would increase to the same level as other categories if more NonClinician questions were present in our corpus.

| SMO Results | | U | SC | $T_5$ | $T_{25}$ | $R_{30}$ | $R_{50}$ |
|---|---|---|---|---|---|---|---|
| Factual | P | 0.744 | 0.727 | 0.696 | 0.721 | 0.739 | 0.685 |
| | R | 0.807 | 0.774 | 0.804 | 0.797 | 0.797 | 0.786 |
| | F1 | 0.774 | 0.75 | 0.746 | 0.757 | 0.767 | 0.732 |
| PatientSpecific | P | 0.77 | 0.753 | 0.751 | 0.757 | 0.769 | 0.742 |
| | R | 0.781 | 0.78 | 0.718 | 0.752 | 0.791 | 0.716 |
| | F1 | 0.776 | 0.766 | 0.734 | 0.755 | 0.78 | 0.729 |
| NonClinician | P | 0.708 | 0.617 | 0.289 | 0.66 | 0.721 | 0.698 |
| | R | 0.324 | 0.276 | 0.39 | 0.314 | 0.295 | 0.286 |
| | F1 | 0.444 | 0.382 | 0.332 | 0.426 | 0.419 | 0.405 |
| Weighted Average | P | 0.754 | 0.73 | 0.701 | 0.733 | 0.752 | 0.713 |
| | R | 0.755 | 0.736 | 0.688 | 0.736 | 0.753 | 0.711 |
| | F1 | 0.748 | 0.727 | 0.694 | 0.729 | 0.744 | 0.704 |

P = Precision
R = Recall
F1 = F-1 Score

Table 5.5 SMO Level 1 Classifier Results

## 5.4 Terminal Classifiers

Inter-coder agreement for the Terminal distribution was not as high as in the Level 1 distribution, so we expected classifiers trained on this taxonomy distribution to perform worse than those in the Level 1 distribution. Our hypothesis proved to be correct across all classifiers. Like the Level 1 distribution, questions were distributed across categories unevenly, which impacted our classifiers. The distribution of questions across the categories is shown in Table 5.6. The maximum ratio of questions in the Level 1 distribution is 5.84, whereas in the Terminal distribution it is 13.23, so the Terminal distribution is more unbalanced than Level 1. This feature of our corpus likely made the classifiers trained on the Terminal distribution perform even worse.

The confusion matrix for the SMO classifier with the Unmodified corpus in the Terminal distribution is shown in Table 5.7. Matrices for other algorithms and corpus modifications are shown in Appendix C. Much like in the Level 1 distribution results, it is clear that the categories with more examples in the training data performed better. Training classifiers with a corpus of questions more evenly distributed across

| Category | Question Count |
|---|---|
| Definition (DEF) | 26 |
| NumericPropertyValue (NPV) | 56 |
| PatientDiagnosis (PD) | 127 |
| Entity (ENT) | 106 |
| Reference (REF) | 26 |
| PatientExplanation (PE) | 192 |
| NonClinician (NC) | 105 |
| PatientOutcome (POUT) | 91 |
| EntityExplanation (ENTE) | 344 |
| PatientRecommendation (PREC) | 205 |

Table 5.6 Terminal Corpus Question Distribution

the categories may give better results if we changed nothing else about our classifiers or taxonomy.

| SMO | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 6 | 1 | 3 | 2 | 0 | 4 | 0 | 1 | 9 | 0 |
| NPV | 1 | 31 | 3 | 0 | 0 | 6 | 0 | 2 | 12 | 1 |
| PD | 5 | 1 | 61 | 5 | 1 | 28 | 2 | 3 | 16 | 5 |
| ENT | 1 | 1 | 1 | 47 | 1 | 2 | 0 | 3 | 39 | 11 |
| REF | 0 | 0 | 1 | 4 | 3 | 2 | 4 | 0 | 9 | 3 |
| PE | 3 | 4 | 23 | 3 | 2 | 94 | 0 | 15 | 22 | 26 |
| NC | 0 | 3 | 1 | 7 | 2 | 16 | 32 | 4 | 17 | 23 |
| POUT | 0 | 4 | 4 | 3 | 0 | 29 | 1 | 28 | 10 | 12 |
| ENTE | 2 | 5 | 10 | 23 | 1 | 25 | 6 | 3 | 252 | 17 |
| PREC | 2 | 1 | 8 | 24 | 2 | 33 | 14 | 7 | 18 | 96 |

Horizontal Axis = Classified As
Vertical Axis = Actual Category

Table 5.7 Unmodified Terminal Corpus SMO Confusion Matrix

The accuracy results for all configurations of the Terminal distribution are shown in Table 5.8. The results are similar to those in the Level 1 distribution, albeit lower, as was expected based on our inter-coder agreement results. Also similar to the Level 1 distribution is the fact that SMO and Multinomial Naive Bayes outperformed J48 and Naive Bayes, with the performance being close enough not to be able to clearly tell one algorithm as clearly better.

| Terminal Corpus | NB | MNB | J48 Tree | SMO |
|---|---|---|---|---|
| U | 40.1 % | 45.7 % | 38.9 % | 50.7 % |
| SC | 40.2 % | 46.0 % | 38.4 % | 51.9 % |
| $T_5$ | 41.1 % | 45.8 % | 39.6 % | 48.1 % |
| $T_{10}$ | 41.0 % | 45.6 % | 37.9 % | 50.2 % |
| $T_{15}$ | 40.1 % | 44.4 % | 39.0 % | 50.7 % |
| $T_{20}$ | 40.6 % | 44.7 % | 39.9 % | 50.5 % |
| $T_{25}$ | 40.6 % | 45.5 % | 39.4 % | 50.9 % |
| $R_{30}$ | 40.8 % | 46.0 % | 38.4 % | 51.2 % |
| $R_{35}$ | 40.3 % | 46.0 % | 39.0 % | 50.5 % |
| $R_{40}$ | 39.7 % | 45.1 % | 39.7 % | 52.6 % |
| $R_{45}$ | 41.0 % | 45.1 % | 40.6 % | 51.2 % |
| $R_{50}$ | 41.3 % | 45.7 % | 40.2 % | 52.1 % |

Table 5.8 Filtered Terminal Classifier Accuracy

Detailed accuracy results for the SMO algorithm are shown in Table 5.4. Similar tables for the other algorithms are shown in Appendix C. Again, the categories with a larger number of examples in our corpus like EntityExplanation, and PatientRecommendation showed significantly better results than those that had very few examples.

| SMO Results | | U | SC | $T_5$ | $T_{25}$ | $R_{30}$ | $R_{50}$ |
|---|---|---|---|---|---|---|---|
| Definition | P | 0.3 | 0.333 | 0.136 | 0.278 | 0.308 | 0.286 |
| | R | 0.231 | 0.269 | 0.115 | 0.192 | 0.154 | 0.154 |
| | F1 | 0.261 | 0.298 | 0.125 | 0.227 | 0.205 | 0.2 |
| NumericPropertyValue | P | 0.608 | 0.593 | 0.53 | 0.526 | 0.6 | 0.576 |
| | R | 0.554 | 0.571 | 0.625 | 0.536 | 0.536 | 0.607 |
| | F1 | 0.579 | 0.582 | 0.574 | 0.531 | 0.566 | 0.591 |
| PatientDiagnosis | P | 0.53 | 0.533 | 0.523 | 0.5 | 0.488 | 0.542 |
| | R | 0.48 | 0.512 | 0.449 | 0.488 | 0.496 | 0.512 |
| | F1 | 0.504 | 0.522 | 0.483 | 0.494 | 0.492 | 0.526 |
| Entity | P | 0.398 | 0.437 | 0.382 | 0.386 | 0.433 | 0.383 |
| | R | 0.443 | 0.491 | 0.472 | 0.462 | 0.491 | 0.462 |
| | F1 | 0.42 | 0.462 | 0.422 | 0.421 | 0.46 | 0.419 |
| Reference | P | 0.25 | 0.429 | 0.2 | 0.3 | 0.25 | 0.333 |
| | R | 0.115 | 0.115 | 0.077 | 0.115 | 0.115 | 0.115 |
| | F1 | 0.158 | 0.182 | 0.111 | 0.167 | 0.158 | 0.171 |
| PatientExplanation | P | 0.393 | 0.379 | 0.355 | 0.375 | 0.397 | 0.396 |
| | R | 0.49 | 0.464 | 0.37 | 0.453 | 0.484 | 0.484 |
| | F1 | 0.436 | 0.417 | 0.362 | 0.41 | 0.437 | 0.436 |
| NonClinician | P | 0.542 | 0.603 | 0.509 | 0.636 | 0.564 | 0.661 |
| | R | 0.305 | 0.362 | 0.257 | 0.333 | 0.295 | 0.352 |
| | F1 | 0.39 | 0.452 | 0.342 | 0.437 | 0.388 | 0.46 |
| PatientOutcome | P | 0.424 | 0.412 | 0.419 | 0.493 | 0.439 | 0.453 |
| | R | 0.308 | 0.308 | 0.286 | 0.363 | 0.319 | 0.319 |
| | F1 | 0.357 | 0.352 | 0.34 | 0.418 | 0.369 | 0.374 |
| EntityExplanation | P | 0.624 | 0.647 | 0.587 | 0.622 | 0.619 | 0.623 |
| | R | 0.733 | 0.747 | 0.735 | 0.735 | 0.747 | 0.735 |
| | F1 | 0.674 | 0.694 | 0.653 | 0.674 | 0.677 | 0.675 |
| PatientRecommendation | P | 0.495 | 0.479 | 0.469 | 0.514 | 0.5 | 0.529 |
| | R | 0.468 | 0.449 | 0.444 | 0.454 | 0.449 | 0.483 |
| | F1 | 0.481 | 0.463 | 0.456 | 0.482 | 0.473 | 0.505 |
| Weighted Average | P | 0.505 | 0.518 | 0.472 | 0.51 | 0.506 | 0.522 |
| | R | 0.509 | 0.519 | 0.481 | 0.509 | 0.512 | 0.521 |
| | F1 | 0.5 | 0.511 | 0.469 | 0.501 | 0.501 | 0.513 |

SMO Filtered Terminal Classifier Results

# Chapter 6: Next Steps

We have developed a new cancer question corpus and EAT taxonomy, as well as the tools to iteratively revise them. However, there is still much work left to perform in order to arrive at a question classifier suitable for use in a production QC system.

The biggest impediment to creating a classifier is the relatively small size of our corpus. Training a successful classifier requires many more hand-classified questions than we have in our database. We have only processed around 3.5 % of our available questions, and there are other cQA websites (and newer questions available on those we have already mined) from which to gather questions. We also have a very unbalanced training corpus, which we suspect has affected the performance of classifiers created using it. Collecting more questions may give us enough to create a training corpus with an even distribution of questions across all categories.

Although our inter-coder agreement is high enough for a classifier in the Level 1 distribution, it is not in the Terminal distribution. Classifying more questions may give insight into further revisions to our EAT taxonomy that would improve inter-coder agreement. Successfully classifying thousands of questions would require more than two coders working part time, though.

In our experience, an individual coder can take between 3 and 12 weeks to train in all tasks depending on skill level and how much time they have available. Furthermore, coders can (and did) abandon the project at any time during or after their training, which created a lot of wasted effort in training them. Finding and training coders with enough free time to work on the project was the biggest challenge we faced throughout the duration of this research.

Our attempts at dimensionality reduction did not prove fruitful, however there are other techniques that may work better. We tried simple LSI, but it proved too complex for our hardware to handle in Java. A custom LRLW-LSI [14] approach (which we did not have time to implement) could remove more terms. Also, using newer, more powerful hardware could ameliorate this problem.

We also ran into a problem with medical terminology and our spell checker. We tried a well-known approach, and did not achieve promising results. What is needed is an approach more specific to our problem domain.

An ideal corpus for training a classifier in a SMS-based QA system would be actual SMS questions. Our corpus is built from processed cQA questions because we did not have access to any SMS questions. We are hopeful that the two will be very similar, but it is impossible to tell unless we have SMS questions from real target users. We are currently working on a pilot study similar to a previous one with pregnant women [40] that will provide us with prostate cancer questions asked via SMS. The questions we have gathered thus far are shown in Figure 6.1, and they are very similar to the cQA questions in our corpus in grammar and lexical choice. However, this is not a large enough sample to show definitively that SMS questions are similar to cQA questions.

The questions in Figure 6.1 could be used in our classification procedure with our software to create a new corpus using our existing EAT taxonomy to bootstrap the process. This would create a representative corpus for both the question data, and category distribution of the questions in the taxonomy. However, this pilot study only provides a small number of questions at a time, as a nurse is currently required to answer all incoming questions. A larger study requires a machine-aided approach to encourage people to ask questions while the system collects question data.

One way to use our taxonomy, corpus, and classifiers in a machine-aided approach while gathering more SMS question data would be to use the Level 1 distribution as a EAT taxonomy by itself. The Level 1 classifiers can discriminate between questions we could attempt to answer (Factual) and the questions we would have to forward to a person (PatientSpecific and Non-Clinician). This would let us focus more effort on the QA task with real users, while simultaneously allowing us to gather many more real SMS questions than would be possible of a person had to answer all incoming questions.

Although we created a corpus and taxonomy for training a classifier for question answering, this is just the first step in the creation of a functioning system. We plan

1. *Why is the prostate only found in men?*
2. *How does the prostate gland make fluid?*
3. *Why is prostate cancer more common in African Americans?*
4. *Why is prostate cancer so common and how is the cancer treated?*
5. *What is the #1 cause of cancer related deaths?*
6. *I recently quit smoking. How long does it take for my lungs to clear?*
7. *How long can you live with prostate cancer?*
8. *Would someone have a lot of pain if they developed prostate cancer?*
9. *Where can someone get tested for free?*
10. *How long does a screening take?*
11. *Whats the youngest someone should be screened for prostate cancer?*
12. *Are there a lot of free screenings out there for people?*
13. *Is radiation very painful?*
14. *What does PSA stand for?*
15. *If the prostate is enlarged or infected does that mean there is cancer there?*
16. *Is the patient put under during a biopsy?*
17. *Is there an actual cure for prostate cancer?*
18. *Does drinking water help you to lose weight? Other than filling you up so you might not eat as much.*
19. *What is a liver count?*
20. *I like to run, is there anything wrong with running every day?*
21. *Is there any problem with donating plasma twice a week every week?*
22. *Are there any symptoms to know that u might have prostate cancer?*
23. *Why is prostate cancer so common in black men?*
24. *What would be the perfect blood pressure for a man age 41?*
25. *How often shovld i get checked*
26. *After getting treatment can it reoccur*
27. *Does it affect sexual performance*
28. *Can u have ur prostate removed*
29. *Could i still lead a normal life with prostate cancer*
30. *Are there long term side effects from the treatments*
31. *What causes prostate cancer*
32. *Do women have a prostate gland*

Figure 6.1 Collected SMS Questions

on continuing this work by processing and classifying more of the cQA questions we already have as well as adding real SMS questions gathered in our pilot study. Having more classified questions will help us be able to determine whether or not we need to continue revising our EAT taxonomy, and how classifiers perform on our data. This

will, in turn, inform us on the true performance of different classification algorithms and pattern recognition techniques to be able to more accurately determine the EAT of an incoming cancer question.

# LIST OF REFERENCES


[1] About.com. All experts. http://www.allexperts.com/.

[2] Kevin Atkinson. Ispell english word lists. http://wordlist.sourceforge.net/.

[3] Holly Blake. Innovation in practice: mobile phone technology in patient care. *British journal of community nursing*, 13(4):160, 162–5, April 2008.

[4] Phyllis Butow, Rhonda Devine, Michael Boyer, Susan Pendlebury, Michael Jackson, and Martin H N Tattersall. Cancer consultation preparation package: changing patients but not physicians is not enough. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 22(21):4401–9, 2004.

[5] Donald J Cegala. Patient Communication Skills Training, 2000.

[6] The Cleveland Clinic. Cleveland clinic live chats. http://my.clevelandclinic.org/multimedia/transcripts/default.aspx.

[7] Corinna Cortes and Vladimir Vapnik. Support-vector networks. In *Machine Learning*, pages 273–297, 1995.

[8] B Joyce Davison, Lesley F Degner, and Thomas R Morgan. Information and decision-making preferences of men with prostate cancer. *Oncology Nursing Forum*, 22(9):1401–1408, 1995.

[9] B Joyce Davison, S Larry Goldenberg, Kristin P Wiens, and Martin E Gleave. Comparing a Generic and Individualized Information Decision Support Intervention for Men Newly Diagnosed with Localized Prostate Cancer. *Cancer Nursing*, 30(5):E7–15, 2007.

[10] B Joyce Davison, Peter Kirk, Lesley F Degner, and Thomas H Hassard. Information and patient participation in screening for prostate cancer. *Patient education and counseling*, 37(3):255–63, July 1999.

[11] B Joyce Davison, Patricia A Parker, and S L Goldenberg. Patients preferences for communicating a prostate cancer diagnosis and participating in medical decision-making. *BJU International*, pages 47–51, 2004.

[12] B Joyce Davison, Alan So, S Larry Goldenberg, Jonathan Berkowitz, and Martin E Gleave. Measurement of factors influencing the participation of patients with prostate cancer in clinical trials: a Canadian perspective. *BJU international*, 101(8):982–7, 2008.

[13] Dina Demner-Fushman and Jimmy Lin. Answering Clinical Questions with Knowledge-Based and Statistical Techniques. *Computational Linguistics*, 33(1):63–103, March 2007.

[14] Wang Ding, Shanqing Yu, Wei Wei, and Qianfeng Wang. LRLW-LSI: An Improved Latent Semantic Indexing (LSI) Text Classifier. *Lecture Notes in Computer Science*, 5009, 2008.

[15] Stephen Dubien. *Question Answering Using Document Tagging and Question Classification*. PhD thesis, University of Lethbridge, 2005.

[16] Richard O Duda, Peter E Hart, and David G Stork. *Pattern Classification*. John Wiley & Sons, second edition, 2001.

[17] Yasser EL-Manzalawy and Vasant Honavar. *WLSVM: Integrating LibSVM into Weka Environment*, 2005.

[18] Deb Feldman-Stewart, Sarah Brennenstuhl, and Michael D. Brundage. The information needed by Canadian early-stage prostate cancer patients for decision-making: stable over a decade. *Patient education and counseling*, 73(3):437–42, 2008.

[19] Deb Feldman-Stewart, Michael D. Brundage, Charles Hayter, Patti Groome, Curtis Nickel, Heather Downes, and William J. Mackillop. What Questions Do Patients with Curable Prostate Cancer Want Answered? *Medical Decision Making*, 20(7):7–19, 2000.

[20] J.L. Fleiss and Others. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[21] The Apache Software Foundation. Lucene. `http://lucene.apache.org/`.

[22] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The weka data mining software: An update. *SIGKDD Explorations*, 11, 2009.

[23] Med Help. Med help forums. `http://www.medhelp.org/`.

[24] Eduard Hovy, Laurie Gerber, Ulf Hermjakob, Chin-Yew Lin, and Deepak Ravichandran. Toward semantics-based answer pinpointing. *Proceedings of the first international conference on Human language technology research - HLT '01*, pages 1–7, 2001.

[25] Emily Hsu. Patient's Orders: How Information Technology and the Media are Changing the Patient-Doctor Relationship. *The Next Generation: An Introduction to Medicine*, 3(7):5–7, 2007.

[26] Yan Huang. Support vector machines for text categorization based on latent semantic indexing. Technical report, 2001.

[27] Abraham Ittycheriah, Martin Franz, Wei-Jing Zhu, Adwait Ratnaparkhi, and Richard J Mammone. IBMs statistical question answering systemTREC-10. In *Proc. of TREC-9*, volume 10, 2000.

[28] George John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pages 338–345. Morgan Kaufmann, 1995.

[29] Daniel Jurafsky and James H Martin. *Speech and Language Processing*. Prentice Hall, second edition, 2009.

[30] A. S. Ivan Kpiachem, Christoph Champ, Y.A. Hamed, Madalina Marin, and Sam Vaknin. Xterm medical dictionary. `http://www.medical-dictionary.ro/index.html`.

[31] Klaus Krippendorff. Computing krippendorffs alpha-reliability, 2007.

[32] Geoff Kuenning. International ispell. `http://www.lasr.cs.ucla.edu/geoff/ispell.html/`.

[33] Adam Kurmally. MedQuestionAdmin. `http://owl.cs.uwm.edu:8008/MedQuestionAdmin`.

[34] Adam Kurmally. MedQuestion Category Taxonomy. 2010.

[35] Adam Kurmally, Barbara Di Eugenio, Charles E Kahn, and Susan McRoy. Building a Corpus and Developing a Question Classifier to Support Messaging-Based Question Answering. In *International Health Informatics Symposium*, 2010.

[36] Xin Li and Dan Roth. Learning question classifiers. *Proceedings of the 19th international conference on Computational linguistics*, 2002.

[37] Xin Li and Dan Roth. Learning question classifiers: the role of semantic information. *Natural Language Engineering*, 12(03):229, December 2005.

[38] Yuanjie Liu, Shasha Li, Yunbo Cao, Chin-Yew Lin, Dingyi Han, and Yong Yu. Understanding and summarizing answers in community-based question answering services. *Proceedings of the 22nd International Conference on Computational Linguistics - COLING '08*, (August):497–504, 2008.

[39] Andrew McCallum and Kamal Nigam. A comparison of event models for naive bayes text classification, 1998.

[40] Susan McRoy, Vishnuvardhan Vaidhayanathan, Amy May, and Hayeon Song. An Open Architecture for Messaging-Based Consumer-Health Question-Answering. In *International Health Informatics Symposium*, 2012.

[41] University of Cincinnati. Net wellness questions. `http://www.netwellness.org`.

[42] American Society of Clinical Oncology. Asco expert chats. http://connection.asco.org/.

[43] National Library of Medicine. Medlineplus. http://www.nlm.nih.gov/medlineplus/.

[44] John C Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, Microsoft Research, 1998.

[45] Your Cancer Questions. Your cancer questions. http://yourcancerquestions.com/.

[46] J. Ross Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1993.

[47] Steven E Stemler. A Comparison of Consensus , Consistency , and Measurement Approaches to Estimating Interrater Reliability. *Practical Assessment, Research & Evaluation*, 9(4), 2004.

[48] H. Sundblad. A Re-examination of Question Classification. In *Proceedings of the Nordic Conference on Computational Linguistics*, pages 394–397, 2007.

[49] W3C. Document Object Model. http://www.w3.org/DOM/ Page last accessed 13 July 2011, 2005.

[50] Hong Yu, Carl Sable, and Hai Ran Zhu. Classifying medical questions based on an evidence taxonomy. In *Workshop on Question Answering in Restricted Domains. 20th National Conference on Artificial Intelligence (AAAI-05)*, pages 27–35, 2005.

[51] Qing T Zeng, Tony Tse, Jon Crowell, Guy Divita, Laura Roth, and Allen C Browne. Identifying consumer-friendly display (cfd) names for health concepts. *AMIA Annual Symposium proceedings AMIA Symposium AMIA Symposium*, 2005:859–863.

[52] Dell Zhang and Wee Sun Lee. Question Classification Using Support Vector Machines. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval - SIGIR '03*, page 26, New York, New York, USA, 2003. ACM Press.

[53] Zhiping Zheng. AnswerBus question answering system. In *Proceedings of the second international conference on Human Language Technology Research*, page 404. Morgan Kaufmann Publishers Inc., 2002.

# Appendix A: Question Promotion Taxonomy



This document covers question classifications that will either not be included in the final question classification training set; or will be included with hand modifications. It explains how to code these questions and what to do with each classification once they are coded.

## A.1 Unusable

The first set of questions are to be eliminated from the pool of candidate questions for various reasons. They should be marked as Unusable questions in the database.

### A.1.1 Not a Question/Makes No Sense

These "questions" are not actually asking anything discern-able. They may be part of a title for a message board posting, or the user may have submitted the question before completing typing.

#### A.1.1.1 Example 1

*Breast Cancer?*

Obviously this is not a complete question, so it cannot be included in a question classification corpus.

#### A.1.1.2 Example 2

*Can you get prostate cancer from ************?*

This is a general form question that was probably the title of a posting to a message board. If it contained a specific entity as a potential cause for cancer then it would be a valid question for this corpus. However since it is not specified this question is unusable.

#### A.1.1.3 Example 3

*How is it possible that Darth Vader dies of pancreatic cancer in Clear and Present Danger?*

This is from an Internet joke that some jerk actually submitted to a website for actual cancer patients. Obviously, this is not a usable question.

### A.1.2 Not About Cancer

These questions were picked up by one of the web crawlers because they appear on a web page that mentions cancer, but they are not questions about cancer.

#### A.1.2.1 Example 1

*Hi, I am hoping someone could ease my mind with my current situation. My husband and I decided on an abortion on April 2, 2008 - as difficult as that decision was, at that time it was the best for our situation. Having taken all of the necessary precautions to prevent that situation again...it seems our contraceptive failed and I am pregnant. Is it safe to repeat that procedure and to be honest, I am hesitant about scheduling the procedure for fear of being judged or criticized for this.*

*Your advice is greatly appreciated. Thank you Sara*

This question is about abortion, not cancer. Therefore it is unusable in this corpus.

#### A.1.2.2 Example 2

*I am a 23 year old female having some chronic symptoms. I have been having intermittent fevers and night sweats for 2 years now. The episodes generally last 2-3 months and come every 3-4 months. I have had a bone marrow biopsy which turned out negative. I have one chronically enlarged lymph node on the left side of my neck. I was on antibiotics for over four months for this. I have been having chronic abnormal pap smears for the past 5 years with a diagnosis of HPV. The pap smears have been coming back with CIN 1-2. I have had numerous colposcopies and one LEEP. I have had no weight loss. I have occasional abdominal pain, but recently had a HIDA scan done which showed sludge in my gallbladder. Doctors also say I have IBS which would be reason for the occasional abdominal pain accompanied by diarrhea and constipation. My CRPs and ESRs have been fluctuating from normal to high for the past 2 years. CMV and EBV were positive. My cortisol levels have been tested and range from too low, to too high, to normal. I am constantly fatigued...seeming to get worse as time goes on. I have been diagnosed with Raynauds phenomenon...so my hands and feet are constantly cold and clammy... I don't know what to think, but all these small symptoms are starting to become bothersome, not to mention worrisome. Any ideas on what it might be?*

This question discusses a litany of medical problems, but does not mention cancer as a possible cause or ask for information about cancer.

### A.1.3 Do My Homework for Me

Some of the questions ask for a certain amount of text or a general survey of a topic, but it is possible that they are questions from actual patients or their loved ones. Questions that fit into this category will literally ask for answers to their school work. The reason these questions are to be eliminated is they are not necessarily

semantically related to the type of question a patient would ask; so they are unusable in this corpus.

#### A.1.3.1 Example 1

*I am a 9th grade student doing a research paper on smoking. i have found what diseases and cancers are related to tobacco use and what harms smoking does to What is the update on fighting smoking related diseases and cancers? Is there a cancer fighting breakthrough coming up? Any information would be greatly appreciated.*

*Thank you, Danyelle Harp*

Here the user specifies that they are doing a research paper and that is why they are only looking for general information on lung cancer. Since we cannot be certain that this question fits in semantically with those a patient would ask it must be labeled unusable.

#### A.1.3.2 Example 2

*HELLO I M WORKING ON MY FINAL BIOLOGY PROJECT ON BREAST CANCER AND I REALLY WANT TO KNOW WHAT GENE IS IT LOCATED ON AND WHAT RACE ETHNICITY DOES IT AFFECTS THANK YOU*

The user states that they are working on a school project in this question. Note that the questions are very general and that a patient may not necessarily need to know this information in light of more immediate concerns such as treatment options, prognosis, etc. The fact that the topic is fundamentally different makes this question unusable in this corpus.

### A.1.4 Out of Sync

Some of the questions are read from live web chat transcripts, and sometimes the parser gets out of sync with the conversation and marks an answer as a question. It then mark the next question as an answer. These questions should be marked as unusable since the appropriate context cannot be determined after the text has been written to the database.

#### A.1.4.1 Example 1

*Any patient who goes on long-term hormone therapy can become hormone resistant. While it used to be thought that the earlier you start hormones, the earlier you became hormone resistant, that may not be the case, and it certainly is not a reason to not use hormones when required.If your cancer relapses, and you require hormone therapy, the average response duration to hormones for patients with a climbing PSA after radical prostatectomy probably exceeds 10 years.*

This question is actually an answer to another question. Since the original order and context cannot be determined once questions are entered into the database, these questions should be marked unusable. Some of these may be hard to discern at first

glance, however they will almost certainly have an "answer" associated with them that is really a question.

#### A.1.4.2 Example 2

*It would be virtually impossible to remove all sources of sugar from the diet. The most important thing you can do is get adequate dietary consultation though your cancer center or hospital.*

Again this is much more a statement than a question. It may be easier to group these with the "Not a Question" category since they do not ask for anything, however the cause is important to understand so we can potentially quantify how often this happens. If a lot of questions are being marked unusable because of conversation synchronization problems it may be worthwhile to change the web crawlers so they do not make these errors.

### A.1.5 Non-Human

Some of the questions found concern cancer in non-human animals. These questions and answers are not appropriate for a human cancer related corpus for two reasons. The first is that often the animals have cancers that are not very common in humans but occur frequently in their species. The second reason is that the discussion often centers around whether or not the animal should be put down. This is, of course, inappropriate for humans with cancer.

#### A.1.5.1 Example 1

*Dear Dr. Krempels,*

*I had to put my bunny Tia down today, now I am wondering if I made the right decision. I am very knowledgable about health and biology as I am going for a PhD in science, however I just need to set my mind at ease.*

*She was about 2 years old and prefectly healthy. Yesterday she stopped eating, drinking, urinating, and passing pellets. The next morning, she was bleeding from her vulva and passing these brown/red clot like stuff. I took her to the vet immediately, he thought that it was sludge or stones. He did a physical and said that he felt a very hard mass in her bladder and wanted to do x-rays. The x-ray showed that she had a mass in her bladder, that was not stones or sludge, he said he knew this as it did not show crystals???? He said it was a tumor. He said I could do an ultrasound, bloodwork, and biopsy, which if it did show cancer there is not much we can do. He said surgery would be very tough on her even if it was not malignant and chemotherapy can be fatal to rabbits. He then said to take her home give antibiotics and anti-inflammatory drugs or euthanize. She was in so much pain and could not sit or pass urine, so I decided to let her go.*

*Im wondering now if this was the correct diagnosis and if it was really just sludge and we cant see it on x-ray.*

*Thanks, Candice*

This question concerns cancer in rabbits, not humans; so it is not applicable to a corpus concerning cancer in humans. Any questions about non-human animals should be marked unusable.

### A.1.6 Group Address

Some of the questions are retrieved from message-board type websites that may contain postings asking if their peers have experienced or heard about a treatment/condition etc. While the subject of referrals to peer information is relevant, the language used in these questions is semantically different enough from a request for a reference to peer information or advice that they are not usable.

#### A.1.6.1 Example 1

*Anyone had difficulty getting pregnant after thyroid cancer?*

A reference to information or a statistic about pregnancy after thyroid cancer do apply to a QA system concerning cancer. However, the wording in this question is not applicable to a QA system since it is personalized toward the answering entity. A machine will not be asked it it has had trouble getting pregnant, rather users would ask it about the statistics of pregnancy after thyroid cancer, or the effects of thyroid cancer on pregnancy.

#### A.1.6.2 Example 2

*Have you had ovarian cancer and what was your experience?*

This question is phrased in a way that would be inappropriate for a machine to answer, so it is not applicable to this corpus.

## A.2 Edit

The second set of questions cannot be promoted in the database in their raw format primarily because they are too long to apply to an SMS or MMS-based system. Some of these questions may be under the length limit, however they must be split up for logical reasons. These include containing multiple questions, or text that is irrelevant to the question. It is very important that in editing questions the expected answer type is preserved. If it is not, then the promoted question will not be a valid user question since it is constructed by the promoter. For example, if the question asks for a personalized health recommendation it should not be changed such that it asks about a general health process. This would change the category of the question from PatientSpecific to Factual.

Questions that are to be edited will often exhibit traits from multiple edit subcategories. It is perfectly acceptable to apply procedures from multiple subcategories to the same document.

### A.2.1 Multiple Questions

These questions may or may not be under the 500 character length limit for promotion, however they do contain multiple questions that can be used in the database. The problem is that they will contain more than one question and are unsutable for use in a system with short be split up. It may be necessary to repeat some information in every split question created in order for them to all make sense as individual questions. Many of the websites that the web crawler searches have a much different format than this corpus, which can lead to an e-mail like document or a conversation getting lumped into a single question. These should be broken up for the purposes of this corpus, so each document expresses a single question with a single information need.

#### A.2.1.1 Example 1

*Is there new research or treatments in the works for Barrett's Esophogus? Am I correct in my understanding that currently there can be no change to that condition with medication? Only control of symptoms?*

This question could be logically broken into 2 or 3 questions depending on how it is done. Even though the question is under the maximum character limit, it should be broken into multiple questions. Ideally each document in the corpus should contain a single question expressing a single information need.

#### A.2.1.2 Example 2

*Good morning Doctor - Regarding Sarcoidosis and long-term monitoring. Does Pulmonary Sarcoidosis lead to Lymphoma or one of its other forms? Outside of the regular checkups with our General Practitioner, is there anything we should be looking out for?*

This question can obviously be broken in to two questions.

#### A.2.1.3 Example 3

*Hello :-) My husband has AML(acute myelogenous leukemia)in remission for 6 years now ( a third remission). He took place in a clinical trial of a new drug which stripped the nerves in his leg, ear and hand, resulting in painful neuropathy, especially in his leg. We have done several procedures, including acupuncture, radio frequency neuralgia, epidural nerve blocks, and various medications My question is two-fold: 1. What remedies are available that are new that are not in the list I mentioned? 2. They have suggested a neurostimulator, to be implanted in his spine, what are the chances of the leukemia cells mutating from the electrical impulses being generated? His doctors do not have the answers to the above questions, since no one with leukemia has had stats taken. really appreciate your time*

In this question the user has numbered the different questions. The patient history presented at the beginning of the document also applies to both questions, therefore at least some of it will have to be preserved when the question is split.

### A.2.2 Too Much History

Sometimes long documents actually contain a simple short applicable question along with a lot of extraneous information that is not necessary to know in order to answer the question. Often these will contain a lot of situational history dealing with the patient and their life situation, which does not directly address the question. If the extraneous information can be removed while the question is intact, then the information should be removed or re-worked in order to use the question.

#### A.2.2.1 Example 1

*i am a 25 yr. old female who just weeks ago was diganosed with a chronic pulmonary bullae of the right lung. i am a smoker yet the doctor told me it had been there for a while. i have no insurance so it is very hard for me to see a specialist and i was wondering if you could provide me some information on what will happen next. i have slowed on the cigerettes. is this the first stages of emphysmia or cancer? i have a 5 yr. old and this news and not knowing what is likely to happen next has both stressed me out and made me have slight trouble with deppression. i would be very grateful if you could give me a little insight on what i have and how it will affect me later. thank you for your time.*

This question could probably be broken into two separate questions. One about the next steps following the diagnosis, another about how to deal with her mental issues regarding her health. However, not all of the back history applies to both questions. The fact that she has been ill for some time does not really apply to her mental health, which would really be affected by her knowledge that she is ill, for example. This question can be broken up and edited such that each split question contains only information related to that question.

#### A.2.2.2 Example 2

*Hi: First thank you for this wonderful site. It is very helpful and a great way to find out information about MG. Now, the question: I was diagnosed in 1981 with MG, had a thymectomy, was on Mestinon and went into remission in 1990. Last year (2001), I started having symptoms again. I am currently on 240mg of Mestinon a day, 150mg of Imuran and will be having my first IVIG next week. Yes, the question, can "bad" stress cause relapse. My dad was diagnosed with Metastatic Prostate Cancer which just blew me away. A few months after finding that out, symptoms started reocurring. I guess I'm looking for a reason for relapse after 10 years of remission. Could there be any other causes for relapse? Once again thank you for your help and this forum. Regards*

This question contains lots of superfluous greetings and thanks because the user was asking a person. Users will not address a machine this way, especially when they are interfacing with it via text message. Therefore some of this information may be removed in order to make the question shorter and more like one that will be received via text message.

#### A.2.2.3 Example 3

*Hello, I thank you in advance for your time and effort in amsweing my question and reading my story. 2 years ago my boyfriend's mother was diagnosed with cancer. She had a large mass under her arm on the side of her breast. Her's was the aggressive form and spread quickly. A month and a half later after chemo and radiation she passed. Now the mother of his child has been diagnosed with stage 3b cancer. There is a large mass in her lung and the doctors have said it has spread to her hip and thigh. They have her on oxycodone for now as she starts chemo and radiation this week. She is in extreme pain and has to now use a cane or walker to get around, she needs help dressing herself and can't get to the restroom by herself. I.ve never heard of cancer spreading to the hip or thigh. My question is, if this is where it has spread, does that mean to the tissue surrounding the hip and thigh or in the bones of the hip and thigh?*

The question the user is actually asking is about the second patient mentioned, and has nothing to do with the first patient. This information should be removed as it is superfluous and would probably not be given in a media like a text message.

#### A.2.2.4 Example 4

*I am a long time former smoker. I have had a feeling in my right ear, for 5 months, that is like just before an ear infection. On a few occasions it feels like fluid in there. It now sometimes leaves my right side and travels to my left side which isnt as severe. I have a small spot in my throat that started about a month ago that feels more like a real sore and is aggravated by moving my head side to side. It is also real sore when pressed on. My ear is sometimes painful upon movement of my head. Ity seems like these two things are connected. This happens on both sides. I dont know if I have swollen glands or not as it is hard to tell. I think I can feel something but it is much larger than pea size, more like a couple of inches. What I feel is not painful but hard. I definately feel lumps under my chin. It seems like I also have lump on the right side next to my throat under the jaw. I have been taking a decongestant every 3 hours for about 6 weeks. Sometimes this will take all the symptoms away. Sometimes it takes a couple of doses to calm things down. I have slight clear drainage from the nostril on the side my ear is having problems. Clearing of throat is much better since taking decongestant. Not any of the other acute sinus or allergy symptoms. Would a decongestant help if these were head and neck cancer symptoms? I know I'm not making things better with poking and pushing all the time. I do have an ENT appointment but couldnt get in for awhile. Any help would be appreciated*

This question presents a very detailed list of symptoms, however the actual question asked only relates to a possible relationship between decongestant and head/neck cancer symptoms. Some symptom information is necessary in this question so it is classified as user-specific and not a general question, however most of the symptom list is irrelevant to the actual question.

# Appendix B: MedQuestion Expected Answer Type Taxonomy



## B.1 Introduction

This document describes a classification taxonomy for cancer questions. The purpose of classifying these questions is to act as a precursor to a Question Answering system that will automatically answer them. Each question will express some type of information need related to cancer (in humans) which may or may not be medical. Each question placed into a category by its Expected Answer Type. That is, the type of answer the person who asked the question expects.

### B.1.1 Background

The purpose of this question classification taxonomy is to act as a guideline for the hand classification of cancer-related medical questions. The intent is to form a database of classified medical questions that can be used to train supervised machine learning algorithms to determine the Expected Answer Type of a question. This trained classification algorithm may them be used by a Question Answering system in order to determine the Expected Answer Type of questions that are posed to it by human users. This is a useful step in answering user questions.

Hand classification can be performed by users using the web application located at: `http://owl.cs.uwm.edu:8009/MedQuestionAdmin/`

The web application will allow raters to review, edit, and classify natural language medical questions asked by real users. This data will be used to create a question set suitable for use as a corpus of cancer question documents in Question Answering and Question Classification tasks.

This corpus is being created to accomplish two things. First it will the be first corpus of natural language questions to be limited to the cancer domain. Second, it will be the first corpus to contain only medical questions asked by the target user population. The intent is that this corpus will not only be used for the development of the Question Answering system it is being created for; but also as a training and classification corpus for further research.

### B.1.2 Definitions

The following terminology is useful for understanding the context of the classification task that this document defines. The information presented is by no means an exhaustive explanation of any of the concepts or fields in question. Further research is encouraged for question raters reading this document.

#### B.1.2.1 Question Answering (QA)

The Question Answering task is create a system that automatically answers natural language questions posed by a human being. The sub task of Question Answering concerning this project is to answer factual questions about cancer. Proper medical treatment requires that a professional examine a patient before giving a diagnosis or treatment recommendation. Therefore, at no point should a QA system attempt to give a medical diagnosis or recommendation to a patient. To this end, before a question can be answered, the system must determine if it is within its scope to answer. This task is accomplished via a question classifier. Questions that ask for a diagnosis or treatment recommendation can be filtered out before the system spends resources attempting to answer them.

#### B.1.2.2 Question Classification (QC)

The Question Classification task is to place a natural language question into a category by its Expected Answer Type (EAT). The purpose of this document is to define a set of EAT classifications for the QC task. This will primarily help a QA system determine whether a question is within its scope to attempt to answer. However, the EAT of a question can also be used when searching for the answer to a question and translating it to natural language.

Question classification is usually accomplished via a supervised classifier. This means the classification algorithm must be trained using example data that is already classified. Generally the example data is classified by hand. This document defines a taxonomy for use by raters that will hand classify questions. The end result is a set of training data suitable for use by a supervised machine learning classifier.

#### B.1.2.3 Expected Answer Type (EAT)

The Expected Answer Type of a question is a label for the type of answer the questioner expects to receive. Each category in this taxonomy is an EAT that questions can exhibit. The entire space of cancer questions should be contained by the taxonomy. If a question does not seem to fit in any category then this indicates that the taxonomy may not be complete and may need to be revised.

In the case of a real world example, as occurs when creating classification training data, the EAT is based on the text of the question and not the answer. This is because it is possible that the information need of a question is fulfilled by an answer that does not match the question's EAT. However, a classifier must still be able to categorize the EAT of each question correctly. And since this taxonomy is aimed at creating a data set for the classification task; all questions must be categorized with the most appropriate EAT possible.

#### B.1.2.4 Supervised Machine Learning

Supervised machine learning algorithms are the most most commonly used algorithm for solving the QC task. They are a set of algorithms that require training data in order to 'learn' their function. These algorithms consist of a basic mathematical formula that requires parameters to 'calibrate' it to a particular type of data. The training data is used to calculate these parameters which can then be used by the classification algorithm on new data. In the case of the QC task, the data takes the form of natural language questions that have been labeled with a category. Usually in supervised machine learning tasks initial data sets are hand classified, and this project is no exception.

The machine learning algorithms most prevalent in text and question classification include Bayesian Networks, Support Vector Machines, and Decision Trees (J48 in particular). It is postulated that Support Vector Machines perform better than other techniques for the QC task, however they have never been tried on a large data set containing text errors (misspellings, type-os, SMS/IM type spellings and abbreviations etc). One of the goals of this research is to determine which supervised learning algorithm performs the best under these circumstances.

## B.2 Question Taxonomy

The taxonomy divides questions into 3 levels of categories in a tree hierarchy. Categories can have an unlimited number of child categories, but only a single parent category. There are 3 top-level categories that divide questions into a single category (Factual) that consists of questions that can be answered by a QA system, and 2 categories that require too much other information to be answered by a machine. Level 2 categories further refine these in order to allow a QA system to tailor its answer to the needs of a user question. This is important even in questions that cannot be answered by a machine, as the system will have to respond to the question coherently to maintain a dialogue. Level 3 categories are slightly different in that they only apply to the child categories of the Factual category. Since a QA system is able to answer these questions, more detail is needed about their EAT when searching for and formulating an answer.

### B.2.1 Factual

This level 1 category contains questions a Knowledge-Based QA system can either answer; or provide a reference containing the answer. Questions fitting into the Factual category are hypothetical in nature and do not address specific patients. Questions that ask for a diagnosis, treatment recommendation, or health outcome for a real patient are classified in the PatientSpecific category.

#### B.2.1.1 ClinicalDescription

These questions will ask for a description of medical terms and/or processes. This includes definitions and medical data descriptions that do not ask for external references. It also includes possible health outcomes given non-specific or hypothetical situations. The answers to these questions describe generic medical cases that may or may not apply to the user directly. They do not diagnose or give advice on specific user cases; those questions are classified in the PatientSpecific category.

It is possible for a ClinicalDescription question to not have an answer, or not be possible to answer in the space of an SMS or MMS message. For the QA task these questions must be separated from questions that are answerable, however the QC task is ill suited to do this. These questions do not differ in Expected Answer Type, but rather in the length or complexity of their actual answer. Since this information is not available when Question Classification takes place separating these questions is left to another part of the QA task.

**Definition** A Definition question will ask for the definition of, or factual information about, a single medical entity or process. Questions in the ExplanationOf... categories will indicate that the questioner understands the basic meaning, but requires clarification or more specific information. Questions in the Definition category will commonly be phrased as "What is..." or "Is X..." and directly ask for basic information.

**Example 1** *What is stereotactic radiotherapy? I'm facing early stage lung cancer.*

The presence of patient history in this question seems to indicate it could be placed in the PatientSpecific category. However, the patient history portion of the question is irrelevant to the information content of the answer. The user is asking for a definition of the entity "stereotactic radiotherapy", which will not change regardless of any personal information provided.

**Example 2** *is polycythemia vera a type of cancer...*

This question asks for a very basic definition of polycythemia vera. It falls under the Definition category because it is basic information included in the definition of the term, even though the question actually only asks for a small part of the definition.

**Guideline** Questions in the Guideline category ask for the specification or explanation of a guideline or set of guidelines for a procedure. Most of these questions will ask for general guidelines on when someone should be tested for cancer. However, this category could also apply to questions about treatments as well. A specific procedure does not have to be named for a question to be placed in the Guideline category.

**Example 1** *What are the current recommendations for mammograms? I heard there were some changes.*

This question asks for the "current recommendations" for a testing procedure. It is not asking for a personal recommended course of action like a PatientSpecific.MedicalRecommendation question would. Instead, it is asking for a generic set of recommendations (e.g. guidelines) for the procedure.

**Purpose** Questions in the Purpose category ask for the specification or clarification of the general purpose of a procedure. Usually this will be a treatment, but this category can also apply to questions about testing and diagnosis. Purpose questions cannot have a specific patient in mind; for example, they cannot ask "Why am I being treated with X...". These types of questions fit in the PatientSpecific category.

**Example 1** *Why is it more important to be screened for prostate cancer as you get older?*

This question asks about the importance (e.g. purpose) of the recommended cancer screening process for a generic patient.

**ExplanationOfRelation** Questions in the ExplanationOfRelation category will ask for a description of the relationship between two or more entities. The answer to which may or may not require an embedded definition of one or more entities. Any question that asks for an explanation of how multiple entities are related fits in the ExplanationOfRelation category. This is true regardless of the actual relationship between the entities, even if there is no relationship.

**Example 1** *When a prostate cancer Gleason score indicates an aggressive cancer does that mean there is a greater chance that the malignant cells have seeped (?) out of the prostate into other parts of the body?*

This question asks to relate a definition entity (Gleason score) to a process entity (spread of malignant cells).

**Example 2** *Is there a connection between hot dog consumption and leukemia in children?*

This question is asking about a causal relationship between hot dog consumption and leukemia. It falls into the ExplanationOfRelation category because the question asks for the relationship between a possible risk entity and a disease entity.

**Example 3** *Does alcohol consumption affect the incidence of breast cancer?*

Similar to the previous example, this question asks about a causal relationship between a disease and the consumption of a drug.

**ExplanationOfProperty** Questions placed in the ExplanationOfProperty category will ask for more detailed information about a specific part of a single entity. If a question concerns more than one entity, and their relationship; then it falls in the ExplanationOfRelation category.

The ExplanationOfProperty category does not cover questions that ask for a definition of an entity. Definitions are general and describe the entity as a whole. ExplanationOfProperty questions will ask for further information about some specific part of an entity (e.g. the steps to a treatment, risk associated with a behavior or condition, a measure or metric of the entity etc.), and not a general definition. It will be implicit in the question that the person asking it understands the basic definition of the entity in question.

**Example 1** *Is it likely that someone who is 18 years old to have prostate cancer if they are experiencing some slight symptoms?*

This question is asking for an explanation of the risk of cancer given that a patient is displaying some of the symptoms at a young age.

**Example 2** *In most cases when BC is dx, is there usually one lump involved?*

The entity in question here is breast cancer (BC). The property the user requests more information about is

**Example 3** *Should all polyps in the intestines and rectum be removed?*

The entity in question is intestinal polyps, and the property is whether they should be removed.

**Example 4** *What is the procedure for a prostate biopsy?*

This question asks for an explanation of the 'procedure' property of a prostate biopsy entity.

#### B.2.1.2 Entity

An entity in this case is a disease, treatment, diagnostic technique, medication, or similar 'thing' relating to cancer. Questions in the Entity category will ask for the identification of one of these items. An entity can also be a set. Therefore, a set of symptoms for a disease or drugs that all treat a certain condition are also entities.

The difference between the Entity and ClinicalDescription categories is that Entity questions ask to identify a specific noun or set of nouns and ClinicalDescription questions ask to describe a process or relationship. Entity questions will give a definition or description of the entity and ask for a name or set of names. Questions that contain the name of their subject cannot be Entity questions.

**Disease** Questions in this category will provide details of health conditions and ask to identify a disease that can cause them.

**Example 1** *Other than STD's are there any cervical infections with symptoms of painful intercourse, bleeding during/after intercourse (red blood and more than a little), odor from the vaginal area, severly swollen cervix?*

This question is asking to identify a set of disease entities with the listed symptoms.

**Treatment** Questions in this category will specify symptoms or a disease and ask for the identification of a treatment. They may also identify a treatment and ask for possible alternative treatments.

**Example 1** *Overall, what is the best treatment for cancer, and why?*
This question is very broad, and probably not answerable. In fact, the response to this question by a QA system will likely not be an answer along the lines of what the questioner intended. However, it is apparent that the intent of the question is to find the name of a treatment entity.

**Example 2** *Are there other options, besides a mastectomy?*
The 'options' that this question asks for would be alternative treatment entities to a mastectomy.

**Example 3** *If breast cancer metastasize to bones which causes unbearable pain.. is there any medicine which can ammeliorate the pain as well as cure the condition???*
This question specifically asks for a drug. However, since the drug is used to treat cancer this question belongs in the Treatment category.

**DiagnosticTechnique** Questions in this category will ask for the specification of a particular diagnostic technique or test. Most questions in this category will ask for a list of the diagnostic techniques for a particular type of cancer. However, some will ask for alternatives to techniques of which the questioner is aware.

**Example 1** *Is there a test for Non Hodgkins Lymphoma?*
This question asks for the identification of a diagnostic technique for Non-Hodgkin's Lymphoma.

**Example 2** *Is there a way to determine if a large fibroid is cancerous, short of a hysterectomy?*
In requesting "a way to determine" something the questioner is looking for a diagnostic technique or test.

**Symptom** Questions in the Symptom category will typically ask for a set of symptoms for a health condition. This may be a type of cancer, or it may be symptoms of a condition associated with cancer. For example, a question asking for the symptoms of radiation poisoning from an improper treatment dosage would be placed in the Symptom category.

**Example 1** *I want to know the symptoms of throat and neck cancer?*
This question asks to identify the set entity for the description "symptoms of throat and neck cancer". This is a case where the entity being described is a set.

**HealthEffect** Questions in this category will usually ask for the health effects of a disease or condition. They may also ask for potential or known health effects of a cancer risk factor such as smoking. Typically these are effects due to the treatment of cancer, but they can also be effects of the cancer itself. This is different from the Symptom category in that it will ask about possible future conditions given a diagnosis. Symptom questions will ask about current conditions that may point to a diagnosis that has yet to be made.

**Example 1** *What are the known long-term effects of carboplatin (Paraplatin)/docetaxel (Taxotere)?*

This question is asking for the identification of the entity for the description "long-term effects of carboplatin". Since the user wants more long term effects of the chemotherapy drug carboplatin this question belongs in the HealthEffect category.

**Example 2** *What are the dangers of a cyst on the Pituitary gland?*

The use of the work 'dangers' in this question (as opposed to 'symptoms') indicates that the questioner is looking for information on more long term conditions.

**RiskFactor** Questions in this category will ask for the identification of risk factors for cancer. The type of cancer may or may not be specified in the question. Usually these questions will ask for a set of entities as opposed to a single entity.

**Example 1** *What are the Risk Factors for Breast Cancer?*

This question asks for a set of risk factors for breast cancer.

**Prevention** Questions in this category will ask for the identification of methods to prevent or lower the risk of cancer. This can include lifestyle changes such as diet and exercise information as well as medical information. Prevention questions may or may not specify a type of cancer.

**Example 1** *What ways are there to prevent Leukemia?*

This question asks to identify some preventative measures for leukemia.

**Example 2** *Are there any vaccines under study to prevent a recurrence of lung cancer?*

Even though the type of preventative measure (vaccine) is in the question, this is classified as Prevention. The actual answer to this question will identify a specific vaccine (or say one does not exist), which qualifies as a preventative measure entity as well.

**NumericPropertyValue** The NumericPropertyValue category covers questions that ask for any kind of numeric value of a property of an entity. This will usually be a min, max, average, or estimated value. Questions asking for the dosage of a drug, statistic, length of time, size, quantity, etc. belong in this category.

**Example 1** *What is the sucess rate of the seed thearpy?*

This question is asking for a statistical rate, which is a numeric value. Therefore it is a NumericPropertyValue question.

**Example 2** *How many women in the U.S would you say die per day from Breast Cancer?*

This question asks for a quantity value, which places it in the NumericPropertyValue category.

**Example 3** *If there is a promising new treatment, how long does it take for the general patient population to benefit?*

This question asks for a time value.

**Example 4** *My dads bloodwork shows PSA numbers at 18, They were 4 three years ago. He just turned 81. He already went thru the false negative biopsy stuff, they are obviously higher now. The doctor says, "the numbers are quite high, would you like to try antibiotics first?". He cannot urinate when he thinks he needs to. Just a little urine but frequent feelings of having to go. How high have some numbers have gone without having cancer associated with it?*

This question includes a lot of patient specific information that may make it look like a PatientSpecific question. However, the actual question at the end is not patient specific. The max PSA values patients have had without cancer does not depend on this specific patient. Also of note is that the patient specific information is necessary due to the phrasing of the question (it refers back to the patient history to identify PSA as the value in question). It is Factual.Entity.NumericPropertyValue because the question is really asking for the maximum observed value for PSA, where the patient did not have cancer.

#### B.2.1.3 ReferenceToInformation

If a question requests a reference to external information such as medical cases, studies, or papers it is placed in this category. Usually the answer will involve some sort of link (hyperlink to a website, phone number of a clinic etc) to the information. For example, a question such as "has there ever been a study on X?" could be answered with "Such a study cannot be found".

EAT categorization uses the expressed intent of the question to determine answer type. Therefore, the ReferenceToInformation category includes only questions where the user specifically requests external information. There are questions where a reference to information will likely be required to answer the question, but the user has not requested it. These questions are likewise classified based on the intent of the user as expressed by the question.

**ClinicalStudy** ClinicalStudy questions ask for a reference to a clinical study or trial, or an article describing one for the layperson. These questions may also ask for

a reference to the results of a clinical study. Primarily these will be concerning cancer drugs or treatments, however the ClinicalStudy category is not limited to these types of studies.

**Example 1** *The other day on the news, I heard that the skin cancer rate in Ohio is down. Could you give a short report on that, or tell me where I could find the study?*

This question states an assertion the user heard on the news, and asks for either a direct answer (short report) or a reference to the original study.

**Example 2** *Have there been any reports or studies published about the long-term side-effects of birth control pills?*

This implicitly asks for a reference to reports or studies about the effects of birth control pills.

**DescriptiveText** Questions in this category ask for a reference to a resource that is not a clinical study. Most DescriptiveText questions will ask for a research summary or a link to a website dedicated to a specific type of cancer. The requested external reference could also include a book, person, or physical location.

**Example 1** *Hi! I was wondering if you could give me about a page or two on lung cancer and the effects it has on the lungs. I would really appreciate it. Thank you,*

This question contains a specific request for information on the effects of cancer on the lungs. Due to the imposed length requirement ("a page or two") it is infer-able that the user is requesting indirect information.

**Example 2** *Is there a resource for learning more about chromosomal abnormalities and NHL or other diseases linked to translocations, deletions and additions?*

This question asked for a learning resource. Implicit in the question is a request for a reference to text describing chromosomal abnormalities; hence the question is asking for a reference to descriptive text.

### B.2.2 PatientSpecific

All questions that ask about specific patient cases are placed in this category. These questions will refer to a particular patient's condition or care in a way that requires a medical professional to examine the patient before the question can be answered. Questions in the PatientSpecific category cannot be answered by any QA system, since they have to do with the diagnosis and treatment of patients.

#### B.2.2.1 Diagnosis

Diagnosis questions request a personal diagnosis about a current health condition. These questions will generally give a set of symptoms or medical background of a person (which may or may not be the questioner of the question) and ask if they have a disease. A specific type of cancer may or may not be specified. It is also possible for Diagnosis questions to ask about multiple types of cancer. Diagnosis questions usually ask something of the form "I am experiencing conditions X, Y, and Z. Could this mean I have disease A right now?".

**Example 1** *I have been having pain under my left rib, my rib is not sore to touch, its under my rib. I have had it for about 4 months now. I have had an ultrsound of pancres(sp),spleen and liver. All came back ok. I mostly get the pain at night, It will wake me up. Now I'm getting worse. The pain stays longer and I cough more and now my chest hurts and sometimes it is harder to breath. Do you think it could be cancer? I do smoke 1 pack a day.*

The person that asked this question describes the patient's current condition and symptoms. Then they ask if their current condition could indicate cancer and give more history. This fits the "here is my condition, what is wrong with me" formula that most Diagnosis questions fall into.

**Example 2** *I had a ct scan which says I have " nondule 1.5cmm right middle lobe non calcified. I am having a pet scan tomorrow. I am 49. Never smoked in my life but parents did. What are the chances it is malignant?*

The user in this case specifies themselves as the patient. The requested diagnosis is to find out if the user's nodule is malignant or not. This question differs from a Factual question in that it identifies a patient and their condition. The user is requesting information about their condition, not that of the "average" patient.

**Example 3** *My husband has been diagnosed with a prostate infection. Does this mean he could have cancer?*

This is another example where the identification of a specific patient places the question in the PatientSpecific.Diagnosis category. A generic version of this question like "Does having a prostate infection mean you could have cancer?" would be placed in a subcategory of Factual.ClinicalDescription.

**Example 4** *I have one cyst that is over 5 inches and one over 4. Does this imply the presence of ovarian cancer?*

This question is a good example of a PatientSpecific.Diagnosis question. The user clearly identifies themselves as the patient and asks what their condition is based on their symptoms.

#### B.2.2.2 HealthOutcome

The HealthOutcome category consists of questions that ask about the possibility of a future health condition given the current condition of a patient. Common topics include likelihood of disease recurrence, quality of life, and the possibility of developing cancer later in life. Generally HealthOutcome questions take a form similar to "I have condition X, what could happen to me later?".

HealthOutcome questions differ from Diagnosis questions in that they ask about patient condition in the future; as opposed to the patient's current condition.

**Example 1** *My mom has been diagnosed stage four colon cancer. What is her survial rate? She is on 7 pills of chemo per day.*

This question identifies a specific patient and asks for the likelihood of her survival. This is not a Factual question because it concerns a specific patient. A generic question such as "What is the survival rate for a person with stage four colon cancer on 7 pills of chemo per day?" would be placed in the Factual.Entity.NumericPropertyValue category.

**Example 2** *I am 26 years old and was told I have moderate dysplasia.If it gose untreated what can happen and how long will it take?*

This question specifies the user as the patient and provides some background and condition information. It is classified as a HealthOutcome question because the user is not requesting information on a disease or treatment, but rather on the eventual outcome of their current condition.

**Example 3** *I finished radiation for throat cancer about 3 months ago. Will the thick saliva that came because of this ever go away?*

In this question the user identifies themselves as the patient and gives some history. The question is placed in the HealthOutcome category because the question is asking for an end result as opposed to a diagnosis or recommendation for treatment.

**Example 4** *my mother is in the first stage of breast cancer i was wondering with it being in first stage what are the chances of reacurring cancer? she is seventy-nine years of age. thanks*

This question is similar to a Diagnosis question since it asks about the chances of the patient having a recurring cancer. The patient background indicates that they have breast cancer currently, so this is actually an outcome of the current treatment; not a new diagnosis given current symptoms. Therefore this question is classified in the HealthOutcome category.

**Example 5** *Does having a granuloma mean that if I continue to smoke it can turn into lung cancer or emphesyma?*

This question specifies the user as the patient and asks about the health outcome associated with smoking. This is considered a longer term outcome since the questioner does not indicate that they are suffering from any symptoms of lung cancer or emphysema.

#### B.2.2.3 MedicalRecommendation

The MedicalRecommendation category consists of questions that ask for help in making a decision associated with the diagnosis or treatment of cancer. These questions may be similar to Factual.ClinicalDescription questions, however Factual questions will talk about treatment options in terms of risks vs. benefits, clinical precedent, or measured statistics. PatientSpecific.MedicalRecommendation questions essentially ask "what should I do?". The implication being that there is a patient that will act on the recommendation as if it came from a doctor that has examined them.

**Example 1** *My dad 84, has dementia and bladder cancer. Should he have the surgery to remove bladder - concerned about quality of life?*

This question directly asks whether the questioner's father should have surgery to treat his cancer (or not). Presumably the answer will weigh into the decision as to whether the patient will undergo surgery.

**Example 2** *My husband is concerned about the hereditary nature of prostate cancer - his father and paternal grandfather have been diagnosed with it. How soon should he start thinking about screenings? He is 31 years old now.*

This question asks for a recommendation about cancer screening. The questioner wants to know when the patient should start getting tested for prostate cancer given his family history.

**Example 3** *I recently saw something on the news about African Americans and their risk for colon cancer. What are some things that I need to do, as an African American, to prevent this disease?*

This question is not about cancer risks or recommendations for the "general" African American population. Instead it is about what, specifically, the questioner should do about the cancer risks posed to the African American population.

**Example 4** *My right breast suddenly has an inverted nipple. When I stand up and bend over it will usually come out. Should I be concerned? I'm getting a mamogram in 5 days. Should I see a doctor before I get the mamogram?*

This question asks for a recommendation related to cancer screening, given patient symptoms.

#### B.2.2.4 Explanation

Questions in the PatientSpecific.Explanation category will ask for an explanation of either a patient's current condition or treatment. Since this is a subcategory of PatientSpecific, Explanation questions reference a specific patient case and not a general explanation.

**Example 1** *My 61 year old mom (non smoker)went for a chest xray yesterday after having difficulty breathing and a cough for 2 weeks. She was told there was a 9mm tumor in the upper portion of her left lung. Is there any way at this point to know if it is more likely a cancer from somewhere else that spread to lung since she is non smoker??*

This question asks for an explanation of how a non-smoker could be diagnosed with lung cancer. Note that the question cannot be a Diagnosis question since the diagnosis is give in the question.

### B.2.3 NonClinician

Questions in the NonClinician category will be related to cancer, but concerned with insurance/financing of health care, legal issues related to care, emotional health for patients and their loved ones, or hospital policies/specialties rather than medical/clinical data. This is in contrast to the Factual and PatientSpecific categories which deal only with diagnosis and treatment of cancer. Factual and PatientSpecific questions require medical knowledge to answer, NonClinician questions do not.

Many NonClinician questions will be similar to Factual or PatientSpecific questions in how they are phrased, however the subject matter will differ. NonClinician questions may identify a specific patient and give some patient history (similar to PatientSpecific questions), or they may ask for factual information (like Factual questions). It is also possible for NonClinician questions to not have specific answers, or have multiple possible answers.

#### B.2.3.1 NonClinicReference

These questions ask for a reference to information like the Factual.ReferenceToInformation category, except the reference will be to non-clinical data. Instead of referring to medical studies and clinical evidence as in Factual.ReferenceToInformation questions; these questions concern patient counseling, legal issues, financial issues, and other non-clinical items related to cancer. The contact information of a physician or hospital is a reference, so questions asking for a doctor or hospital to treat a certain disease will fit in NonClinicReference.

**Example 1** *I am a cancer survivor and have started dating again, but feel like I don't have much in common with the people I meet. Are there any dating services or online networks specifically geared for cancer survivors?*

This question asks for a reference to a dating resource. That falls under the umbrella of counseling, which places the question in the NonClinicReference category.

**Example 2** *Do you know of any books/resources to help explain cancer to a young child?*

This question asks for resources that explain how to explain cancer in terms a young child can understand. The question does not specify whether it is for children dealing with cancer in their family, or children who have cancer. That does not matter though, because either way it concerns counseling.

**Example 3** *Do you have any suggestions on where I could go for faith-based cancer support groups?*

This question asks for a reference to support groups, which are a form of counseling.

**Example 4** *Friends have said I need to get into a clinical trial right away. Where can I sign up for one?*

This question implicitly asks for a reference to information on clinical trials. While participation in a clinical trial is a medical decision and will likely come with many medical decisions; this question is non medical. The question is classified as NonClinician.NonClinicReference because it is asking for information on how/where to apply for a clinical trial. It is not asking for medical information about any clinical trial.

**Example 5** *My wife was notified that she has lung cancer. She was a smoker of Marlboro Light for over 30 years. Is there a lawfirm that represents this type of case?*

This question asks about a legal issue, which means it's NonClinician. Since the user would like a reference to a law firm that means it's NonClinicReference.

#### B.2.3.2 NonClinicRecommendation

The NonClinicRecommendation category is very similar to the PatientSpecific.MedicalRecommendation category. Questions in this category will also ask for a suggested course of action; however the questions will not ask for treatment or diagnosis recommendations. Instead, questions will ask for suggested actions regarding mental health and counseling, insurance/financing health care, legal issues, etc.

**Example 1** *My brother smokes, and I've nagged him for years to stop. (Most people in our family have died from one type of cancer or another). What are the most effective ways for a person to quit smoking?*

This question asks for a recommendation on effective ways to get someone to quit smoking. While smoking is a risk factor for cancer, the question does is not concerned with the medical implications or details of the relationship between smoking and cancer. It is only asking for techniques to stop smoking

**Example 2** *How do I tell my children i have stage 4 ovarian cancer?*

This question concerns counseling and mental health (even though it's not of the cancer patient's mental health). Therefore, it is NonClinician. It is also asking for a recommendation, which places it in the NonClinicRecommendation category.

**Example 3** *A coworker will be returning after four months of cancer treatments. We are all concerned about saying things that would upset him or acting in a way that would make him uncomfortable. Any suggestions?*

This question asks for a recommendation on how to interact with a cancer patient. This would fall under counseling/mental health, making the question NonClinician.

#### B.2.3.3 NonClinicDescription

The NonClinicDescription category fits questions requesting the description or identification of an entity or process that is non-clinical. This includes describing health care coverage, legal issues, and other processes. It also includes personal testimony about the life effects of cancer on patients and their loved ones. The types of questions classified as NonClinician.NonClinicDescription will be much more varied than those contained in Factual.ClinicalDescription. It is also possible for these questions to contain a patient identification and patient history as well.

**Example 1** *How difficult is traveling with lung cancer and emphazema?*

This question is asking for a description of the difficulty traveling with certain conditions. Likely this question would require a reference to answer, however since it does not ask for one it is placed in the NonClinicDescription category.

**Example 2** *About how much does it cost to go through genetic testing and counseling?*

This question asks for a description of the cost structure of genetic testing and counseling. Since it concerns both cost and counseling this is a NonClinician question. The request for a description makes it a NonClinicDescription question.

**Example 3** *I am military, my fiancee was just diagnosed with cervical cancer. If we marry will my benefits cover her?*

This question is asking for a description of the policy associated with health care for military personnel. Questions about health care policies and benefits belong in the NonClinician category. Since it is asking for a description this question is also in the NonClinicDescription category.

### B.2.4 Wildcard Categories

The wildcard categories are a set of meta-categories that do not directly describe the EAT of a question, but rather how the question fits the taxonomy design. They are called wildcards because they may be placed anywhere in the category hierarchy. Each

of these categories indicates that the question doesn't fit properly in the taxonomy at that level. It is possible that there was something wrong with the promotion of the question which causes it to be unclassifiable or there is a problem in the taxonomy.

#### B.2.4.1 OtherCategory

OtherCategory questions do not fit in any category at the given taxonomy level. These questions will be completely outside of the scope of any category at this level, but be within the scope of it's parent category. Classifying a question as OtherCategory indicates that there may be a hole in the taxonomy where not every EAT is covered.

#### B.2.4.2 Ambiguous

Ambiguous questions cannot be classified because they can be interpreted multiple ways. Ideally these questions should not have been promoted, however this category allows them to be formally flagged for review. Ambiguous questions will have the ability to be in two or more different categories depending on how a word or phrase is interpreted by the reader. Depending on the interpretation the question will then be in one and only one category.

#### B.2.4.3 MultipleCategory

MultipleCategory questions fit into two or more categories at the given level. The question must fit into multiple categories at the same time, not just be able to be interpreted multiple ways by the reader (which would make it Ambiguous). MultipleCategory questions indicate that there may be an overlap between categories in the taxonomy, where the EATs are not distinct enough.

## B.3 Revision Notes

### B.3.1 Taxonomy Draft rev08

- Move all questions concerning exercise and diet from NonClinician to Factual

### B.3.2 Taxonomy Draft rev07

- Added wildcard categories OtherCategory, Ambiguous, and MultipleCategory categories
- Changed NonMedical to NonClinician along with subcategories. Still encompass the same set of questions.
- Replaced Factual.Statistic category and all its subcategories with Factual.Entity.NumericPropertyValue

### B.3.3 Taxonomy Draft rev06

- Added Factual.ClinicalDescription.Guideline category
- Added Factual.ClinicalDescription.Purpose category
- Added Factual.Statistic.AverageDuration category
- Added PatientSpecific.Explanation category
- Added Ambiguous category
- Merged Factual.ClinicalDescription.ExplanationOfCondition, ExplanationOfRisk, and ExplanationOfProcedure into ExplanationofProperty (covers detailed explanations of single entities)
- Renamed Factual.ClinicalDescription.EntityRelation to ExplanationOfRelation (covers detailed explanations of multiple entities)
- Merged Factual.ReferenceToInformation.DescriptiveText and EducationalResource into EducationalResource
- Merged Factual.Statistic.DiseaseRate and Factual.Statistic.IncidenceRate into DiseaseRisk category
- Added a 'NonMed' prefix to all NonMedical subcategories
- Removed Factual.Entity.Drug category (redundant with Treatment)
- Revised examples

### B.3.4 Taxonomy Draft rev05

- Added Subsubcategories for the Factual category
- Revised the Introduction section & definitions

### B.3.5 Taxonomy Draft rev04

- The FactBased category has been renamed to the Factual category. The category remains the same, but the name was changed for clarity.
- The Advice category has been renamed to the NonMedical category. The category remains the same, but the name was changed for clarity.
- The Advice.ExternalReference category has been renamed to the NonMedical.SocialReference category. The category remains the same, but the name was changed for clarity.

- The FactBased.Complex category has been absorbed into the Factual.ClinicalDescription category. It is still important to distinguish the two questions, but that will happen at a later stage in the QA process.

# Appendix C: Classifier Data

The raw classifier data is very similar, and is included here only for completeness. Versions of the corpus that have had terms trimmed in threshold or range based modifications are especially similar, therefore only the minimum and maximum of the respective values are shown. This means $T_5$, $T_{25}$, $R_{30}$, and $R_{50}$ are shown; while $T_{10}$, $T_{15}$, $T_{20}$, $T_{15}$, $R_{35}$, $R_{40}$, and $R_{45}$ are not. The omitted values are very similar to the results show, although they do not always fall between them.

The following terms will be used throughout the remainder of this Appendix:

**L1** Level 1 Taxonomy Distribution

**FT** Filtered Terminal Taxonomy Distribution

**U** Unmodified corpus version

**SC** Spelling Corrected corpus version

$\boldsymbol{T_5}$ 5% Threshold corpus version

$\boldsymbol{T_{25}}$ 25% Threshold corpus version

$\boldsymbol{R_{30}}$ 30% Range corpus version

$\boldsymbol{R_{50}}$ 50% Range corpus version

**NB** Naive Bayes classification algorithm

**MNB** Multinomial Naive Bayes classification algorithm

**SMO** Sequential Minimal Optimization implementation of a Support Vector Machine classification algorithm

**J48** J48 Decision Tree classification algorithm

**P** Precision

**R** Recall

**F1** F-1 Score

**F** Factual category

**PS** PatientSpecific category

**NC** NonClinician category

**DEF** Definition category

**NPV** NumericPropertyValue category

**PD** PatientDiagnosis category

**ENT** Entity category

**REF** Reference category

**PE** PatientExplanation category

**NC** NonClinician category

**POUT** PatientOutcome category

**ENTE** EntityExplanation category

**PREC** PatientRecommendation category

Only the Level 1 and Filtered Terminal taxonomy are shown. The Level 2 taxonomy distribution is very similar to the Filtered Terminal taxonomy, so it would be moot to discuss them separately. The full Terminal Taxonomy results are not very accurate, but some categories had very low representation in the training data, with some categories having less than 10 instances). Therefore, no conclusions can be drawn for this distribution until more questions are classified and added to our corpus.

## C.1 Level 1 Distribution

The following sections contain data for the Level 1 distribution.

### C.1.1 Confusion Matrixes

The confusion matrixes for all algorithms and selected corpus variants are shown below for the Level 1 distribution. The horizontal axis is the category each question was classified as, and the vertical axis is the actual category each question belonged to for all algorithms.

| NB | F | PS | NC |
|---|---|---|---|
| F | 415 | 116 | 30 |
| PS | 129 | 428 | 56 |
| NC | 37 | 28 | 40 |

U L1 Corpus NB Confusion Matrix

| MNB | F | PS | NC |
|---|---|---|---|
| F | 392 | 164 | 5 |
| PS | 76 | 537 | 0 |
| NC | 40 | 55 | 10 |

U L1 Corpus MNB Confusion Matrix

| SMO | F | PS | NC |
|---|---|---|---|
| F | 453 | 106 | 2 |
| PS | 122 | 479 | 12 |
| NC | 34 | 37 | 34 |

U L1 Corpus SMO Confusion Matrix

| J48 | F | PS | NC |
|---|---|---|---|
| F | 421 | 131 | 9 |
| PS | 142 | 454 | 17 |
| NC | 43 | 53 | 9 |

U L1 Corpus J48 Confusion Matrix

| NB | F | PS | NC |
|---|---|---|---|
| F | 414 | 118 | 29 |
| PS | 130 | 430 | 53 |
| NC | 37 | 24 | 44 |

SC L1 Corpus NB Confusion Matrix

| MNB | F | PS | NC |
|---|---|---|---|
| F | 398 | 161 | 2 |
| PS | 78 | 535 | 0 |
| NC | 39 | 54 | 12 |

SC L1 Corpus MNB Confusion Matrix

| SMO | F | PS | NC |
|---|---|---|---|
| F | 434 | 120 | 7 |
| PS | 124 | 478 | 11 |
| NC | 39 | 37 | 29 |

SC L1 Corpus SMO Confusion Matrix

| J48 | F | PS | NC |
|---|---|---|---|
| F | 415 | 125 | 21 |
| PS | 156 | 428 | 29 |
| NC | 44 | 54 | 7 |

SC L1 Corpus J48 Confusion Matrix

| NB | F | PS | NC |
|---|---|---|---|
| F | 402 | 125 | 34 |
| PS | 109 | 437 | 67 |
| NC | 28 | 36 | 41 |

$T_5$ L1 Corpus NB Confusion Matrix

| MNB | F | PS | NC |
|---|---|---|---|
| F | 398 | 161 | 2 |
| PS | 78 | 535 | 0 |
| NC | 39 | 54 | 12 |

$T_5$ L1 Corpus MNB Confusion Matrix

| SMO | F | PS | NC |
|---|---|---|---|
| F | 363 | 190 | 8 |
| PS | 103 | 505 | 5 |
| NC | 30 | 42 | 33 |

$T_5$ L1 Corpus SMO Confusion Matrix

| J48 | F | PS | NC |
|---|---|---|---|
| F | 409 | 148 | 4 |
| PS | 163 | 438 | 12 |
| NC | 53 | 47 | 5 |

$T_5$ L1 Corpus J48 Confusion Matrix

| NB | F | PS | NC |
|---|---|---|---|
| F | 412 | 118 | 31 |
| PS | 117 | 440 | 56 |
| NC | 34 | 26 | 45 |

$T_{25}$ L1 Corpus NB Confusion Matrix

| MNB | F | PS | NC |
|---|---|---|---|
| F | 391 | 164 | 6 |
| PS | 92 | 520 | 1 |
| NC | 33 | 50 | 22 |

$T_{25}$ L1 Corpus MNB Confusion Matrix

| SMO | F | PS | NC |
|---|---|---|---|
| F | 447 | 110 | 4 |
| PS | 139 | 461 | 13 |
| NC | 34 | 38 | 33 |

$T_{25}$ L1 Corpus SMO Confusion Matrix

| J48 | F | PS | NC |
|---|---|---|---|
| F | 401 | 154 | 6 |
| PS | 133 | 456 | 24 |
| NC | 40 | 49 | 16 |

$T_{25}$ L1 Corpus J48 Confusion Matrix

| NB | F | PS | NC |
|---|---|---|---|
| F | 417 | 110 | 34 |
| PS | 122 | 427 | 64 |
| NC | 30 | 31 | 44 |

$R_{30}$ L1 Corpus NB Confusion Matrix

| MNB | F | PS | NC |
|---|---|---|---|
| F | 365 | 188 | 8 |
| PS | 76 | 537 | 0 |
| NC | 28 | 56 | 21 |

$R_{30}$ L1 Corpus MNB Confusion Matrix

| SMO | F | PS | NC |
|---|---|---|---|
| F | 447 | 111 | 3 |
| PS | 119 | 485 | 9 |
| NC | 39 | 35 | 31 |

$R_{30}$ L1 Corpus SMO Confusion Matrix

| J48 | F | PS | NC |
|---|---|---|---|
| F | 429 | 123 | 9 |
| PS | 124 | 458 | 31 |
| NC | 39 | 57 | 9 |

$R_{30}$ L1 Corpus J48 Confusion Matrix

| NB | F | PS | NC |
|---|---|---|---|
| F | 420 | 110 | 31 |
| PS | 122 | 424 | 67 |
| NC | 30 | 21 | 54 |

$R_{50}$ L1 Corpus NB Confusion Matrix

| MNB | F | PS | NC |
|---|---|---|---|
| F | 365 | 182 | 14 |
| PS | 90 | 513 | 10 |
| NC | 19 | 43 | 43 |

$R_{50}$ L1 Corpus MNB Confusion Matrix

| SMO | F | PS | NC |
|---|---|---|---|
| F | 441 | 118 | 2 |
| PS | 163 | 439 | 11 |
| NC | 40 | 35 | 30 |

$R_{50}$ L1 Corpus SMO Confusion Matrix

| J48 | F | PS | NC |
|---|---|---|---|
| F | 424 | 132 | 5 |
| PS | 125 | 472 | 16 |
| NC | 49 | 48 | 8 |

$R_{50}$ L1 Corpus J48 Confusion Matrix

### C.1.2 Detailed Accuracy

Detailed accuracy results for the classifiers (Precision, Recall, and F-Measure) are shown below for each algorithm/category. The weighted average (across the categories) is also shown for each algorithm.

| NB Results | | U | SC | $T_5$ | $T_{25}$ | $R_{30}$ | $R_{50}$ |
|---|---|---|---|---|---|---|---|
| Factual | P | 0.714 | 0.713 | 0.746 | 0.732 | 0.733 | 0.734 |
| | R | 0.74 | 0.738 | 0.717 | 0.734 | 0.743 | 0.749 |
| | F1 | 0.727 | 0.725 | 0.731 | 0.733 | 0.738 | 0.741 |
| PatientSpecific | P | 0.748 | 0.752 | 0.731 | 0.753 | 0.752 | 0.764 |
| | R | 0.698 | 0.701 | 0.713 | 0.718 | 0.697 | 0.692 |
| | F1 | 0.722 | 0.726 | 0.722 | 0.735 | 0.723 | 0.726 |
| NonClinician | P | 0.317 | 0.349 | 0.289 | 0.341 | 0.31 | 0.355 |
| | R | 0.381 | 0.419 | 0.39 | 0.429 | 0.419 | 0.514 |
| | F1 | 0.346 | 0.381 | 0.332 | 0.38 | 0.356 | 0.42 |
| Weighted Average | P | 0.698 | 0.702 | 0.701 | 0.71 | 0.707 | 0.717 |
| | R | 0.69 | 0.694 | 0.688 | 0.701 | 0.694 | 0.702 |
| | F1 | 0.693 | 0.697 | 0.694 | 0.705 | 0.7 | 0.708 |

NB L1 Classifier Results

| MNB Results | | U | SC | $T_5$ | $T_{25}$ | $R_{30}$ | $R_{50}$ |
|---|---|---|---|---|---|---|---|
| Factual | P | 0.772 | 0.773 | 0.732 | 0.758 | 0.778 | 0.77 |
| | R | 0.699 | 0.709 | 0.647 | 0.697 | 0.651 | 0.651 |
| | F1 | 0.733 | 0.74 | 0.687 | 0.726 | 0.709 | 0.705 |
| PatientSpecific | P | 0.71 | 0.713 | 0.685 | 0.708 | 0.688 | 0.695 |
| | R | 0.876 | 0.873 | 0.824 | 0.848 | 0.876 | 0.837 |
| | F1 | 0.785 | 0.785 | 0.748 | 0.772 | 0.77 | 0.759 |
| NonClinician | P | 0.667 | 0.857 | 0.717 | 0.759 | 0.724 | 0.642 |
| | R | 0.095 | 0.114 | 0.314 | 0.21 | 0.2 | 0.41 |
| | F1 | 0.167 | 0.202 | 0.437 | 0.328 | 0.313 | 0.5 |
| Weighted Average | P | 0.734 | 0.751 | 0.708 | 0.734 | 0.73 | 0.724 |
| | R | 0.734 | 0.739 | 0.704 | 0.729 | 0.722 | 0.72 |
| | F1 | 0.711 | 0.717 | 0.696 | 0.715 | 0.706 | 0.714 |

MNB L1 Classifier Results

| SMO Results | | U | SC | $T_5$ | $T_{25}$ | $R_{30}$ | $R_{50}$ |
|---|---|---|---|---|---|---|---|
| Factual | P | 0.744 | 0.727 | 0.696 | 0.721 | 0.739 | 0.685 |
| | R | 0.807 | 0.774 | 0.804 | 0.797 | 0.797 | 0.786 |
| | F1 | 0.774 | 0.75 | 0.746 | 0.757 | 0.767 | 0.732 |
| PatientSpecific | P | 0.77 | 0.753 | 0.751 | 0.757 | 0.769 | 0.742 |
| | R | 0.781 | 0.78 | 0.718 | 0.752 | 0.791 | 0.716 |
| | F1 | 0.776 | 0.766 | 0.734 | 0.755 | 0.78 | 0.729 |
| NonClinician | P | 0.708 | 0.617 | 0.289 | 0.66 | 0.721 | 0.698 |
| | R | 0.324 | 0.276 | 0.39 | 0.314 | 0.295 | 0.286 |
| | F1 | 0.444 | 0.382 | 0.332 | 0.426 | 0.419 | 0.405 |
| Weighted Average | P | 0.754 | 0.73 | 0.701 | 0.733 | 0.752 | 0.713 |
| | R | 0.755 | 0.736 | 0.688 | 0.736 | 0.753 | 0.711 |
| | F1 | 0.748 | 0.727 | 0.694 | 0.729 | 0.744 | 0.704 |

SMO L1 Classifier Results

| SMO Results | | U | SC | $T_5$ | $T_{25}$ | $R_{30}$ | $R_{50}$ |
|---|---|---|---|---|---|---|---|
| Factual | P | 0.695 | 0.675 | 0.654 | 0.699 | 0.725 | 0.709 |
| | R | 0.75 | 0.74 | 0.729 | 0.715 | 0.765 | 0.756 |
| | F1 | 0.722 | 0.706 | 0.69 | 0.707 | 0.744 | 0.732 |
| PatientSpecific | P | 0.712 | 0.705 | 0.692 | 0.692 | 0.718 | 0.724 |
| | R | 0.741 | 0.698 | 0.715 | 0.744 | 0.747 | 0.77 |
| | F1 | 0.726 | 0.702 | 0.703 | 0.717 | 0.732 | 0.746 |
| NonClinician | P | 0.257 | 0.123 | 0.238 | 0.348 | 0.184 | 0.276 |
| | R | 0.086 | 0.067 | 0.048 | 0.152 | 0.086 | 0.076 |
| | F1 | 0.129 | 0.086 | 0.079 | 0.212 | 0.117 | 0.119 |
| Weighted Average | P | 0.667 | 0.644 | 0.638 | 0.667 | 0.677 | 0.681 |
| | R | 0.691 | 0.665 | 0.666 | 0.683 | 0.701 | 0.707 |
| | F1 | 0.675 | 0.653 | 0.646 | 0.671 | 0.687 | 0.688 |

J48 L1 Classifier Results

## C.2 Filtered Terminal Distribution

The following sections contain data for the Filtered Terminal distribution.

### C.2.1 Confusion Matrixes

The confusion matrixes for all algorithms and selected corpus variants are shown below for the Filtered Terminal distribution. The horizontal axis is the category each question was classified as, and the vertical axis is the actual category each question belonged to for all algorithms.

| NB | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 5 | 0 | 1 | 3 | 0 | 4 | 0 | 1 | 12 | 0 |
| NPV | 1 | 7 | 1 | 3 | 0 | 9 | 2 | 4 | 25 | 4 |
| PD | 3 | 3 | 53 | 0 | 1 | 32 | 2 | 7 | 16 | 10 |
| ENT | 1 | 4 | 2 | 29 | 2 | 4 | 6 | 6 | 43 | 9 |
| REF | 0 | 2 | 0 | 2 | 3 | 5 | 4 | 1 | 6 | 3 |
| PE | 7 | 4 | 29 | 1 | 6 | 61 | 3 | 23 | 18 | 40 |
| NC | 1 | 1 | 5 | 5 | 4 | 4 | 36 | 7 | 28 | 14 |
| POUT | 1 | 5 | 10 | 3 | 1 | 21 | 4 | 26 | 11 | 9 |
| ENTE | 9 | 7 | 12 | 21 | 8 | 22 | 6 | 12 | 228 | 19 |
| PREC | 4 | 4 | 17 | 12 | 4 | 36 | 20 | 14 | 29 | 65 |

U FT Corpus NB Confusion Matrix

| MNB | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 0 | 0 | 1 | 0 | 0 | 8 | 1 | 0 | 11 | 5 |
| NPV | 0 | 0 | 1 | 0 | 0 | 14 | 1 | 4 | 28 | 8 |
| PD | 0 | 0 | 64 | 0 | 0 | 49 | 0 | 1 | 3 | 10 |
| ENT | 0 | 0 | 4 | 11 | 0 | 8 | 1 | 2 | 56 | 24 |
| REF | 0 | 0 | 1 | 0 | 0 | 2 | 3 | 1 | 12 | 7 |
| PE | 0 | 0 | 15 | 0 | 0 | 124 | 0 | 10 | 14 | 29 |
| NC | 0 | 0 | 5 | 5 | 0 | 9 | 21 | 1 | 17 | 47 |
| POUT | 0 | 0 | 6 | 0 | 0 | 51 | 0 | 8 | 6 | 20 |
| ENTE | 0 | 0 | 13 | 6 | 0 | 52 | 2 | 1 | 236 | 34 |
| PREC | 0 | 0 | 13 | 1 | 0 | 51 | 2 | 6 | 12 | 120 |

U FT Corpus MNB Confusion Matrix

| SMO | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 6 | 1 | 3 | 2 | 0 | 4 | 0 | 1 | 9 | 0 |
| NPV | 1 | 31 | 3 | 0 | 0 | 6 | 0 | 2 | 12 | 1 |
| PD | 5 | 1 | 61 | 5 | 1 | 28 | 2 | 3 | 16 | 5 |
| ENT | 1 | 1 | 1 | 47 | 1 | 2 | 0 | 3 | 39 | 11 |
| REF | 0 | 0 | 1 | 4 | 3 | 2 | 4 | 0 | 9 | 3 |
| PE | 3 | 4 | 23 | 3 | 2 | 94 | 0 | 15 | 22 | 26 |
| NC | 0 | 3 | 1 | 7 | 2 | 16 | 32 | 4 | 17 | 23 |
| POUT | 0 | 4 | 4 | 3 | 0 | 29 | 1 | 28 | 10 | 12 |
| ENTE | 2 | 5 | 10 | 23 | 1 | 25 | 6 | 3 | 252 | 17 |
| PREC | 2 | 1 | 8 | 24 | 2 | 33 | 14 | 7 | 18 | 96 |

U FT Corpus SMO Confusion Matrix

| J48 | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 2 | 1 | 4 | 1 | 0 | 3 | 2 | 1 | 9 | 3 |
| NPV | 3 | 23 | 6 | 1 | 1 | 5 | 0 | 6 | 8 | 3 |
| PD | 2 | 0 | 42 | 8 | 1 | 31 | 4 | 3 | 16 | 20 |
| ENT | 0 | 0 | 6 | 22 | 0 | 14 | 6 | 2 | 42 | 14 |
| REF | 0 | 0 | 1 | 0 | 10 | 4 | 4 | 0 | 5 | 2 |
| PE | 6 | 5 | 34 | 7 | 1 | 69 | 11 | 9 | 23 | 27 |
| NC | 2 | 3 | 8 | 5 | 4 | 11 | 27 | 3 | 22 | 20 |
| POUT | 1 | 8 | 9 | 6 | 2 | 23 | 2 | 13 | 16 | 11 |
| ENTE | 6 | 6 | 18 | 22 | 6 | 38 | 9 | 6 | 206 | 27 |
| PREC | 3 | 5 | 17 | 17 | 3 | 31 | 18 | 5 | 23 | 83 |

U FT Corpus J48 Confusion Matrix

| NB | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 4 | 0 | 0 | 1 | 0 | 5 | 0 | 3 | 13 | 0 |
| NPV | 1 | 11 | 0 | 1 | 0 | 8 | 2 | 3 | 26 | 4 |
| PD | 2 | 1 | 54 | 0 | 2 | 33 | 5 | 6 | 16 | 8 |
| ENT | 0 | 3 | 0 | 32 | 2 | 5 | 5 | 6 | 40 | 13 |
| REF | 0 | 2 | 1 | 1 | 4 | 2 | 6 | 1 | 5 | 4 |
| PE | 9 | 4 | 37 | 4 | 3 | 59 | 4 | 26 | 16 | 30 |
| NC | 2 | 1 | 5 | 6 | 5 | 4 | 35 | 5 | 27 | 15 |
| POUT | 1 | 5 | 10 | 4 | 1 | 18 | 4 | 28 | 8 | 12 |
| ENTE | 11 | 7 | 11 | 29 | 9 | 22 | 7 | 10 | 221 | 17 |
| PREC | 5 | 3 | 14 | 12 | 7 | 33 | 17 | 19 | 29 | 66 |

SC FT Corpus NB Confusion Matrix

| MNB | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 0 | 0 | 0 | 0 | 0 | 11 | 0 | 0 | 9 | 6 |
| NPV | 0 | 1 | 1 | 0 | 0 | 15 | 1 | 4 | 28 | 6 |
| PD | 0 | 0 | 65 | 0 | 0 | 48 | 0 | 1 | 3 | 10 |
| ENT | 0 | 0 | 4 | 11 | 0 | 9 | 2 | 2 | 57 | 21 |
| REF | 0 | 0 | 1 | 0 | 0 | 3 | 2 | 1 | 13 | 6 |
| PE | 0 | 0 | 13 | 0 | 0 | 129 | 0 | 10 | 14 | 26 |
| NC | 0 | 0 | 4 | 3 | 0 | 8 | 24 | 1 | 20 | 45 |
| POUT | 0 | 0 | 7 | 0 | 0 | 51 | 0 | 9 | 8 | 16 |
| ENTE | 0 | 0 | 13 | 6 | 0 | 49 | 1 | 4 | 236 | 35 |
| PREC | 0 | 0 | 13 | 2 | 0 | 55 | 2 | 4 | 16 | 113 |

SC FT Corpus MNB Confusion Matrix

| SMO | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 7 | 1 | 1 | 2 | 0 | 5 | 0 | 1 | 9 | 0 |
| NPV | 1 | 32 | 2 | 0 | 0 | 3 | 2 | 4 | 8 | 4 |
| PD | 3 | 2 | 65 | 3 | 1 | 31 | 0 | 1 | 11 | 10 |
| ENT | 1 | 2 | 1 | 52 | 0 | 3 | 0 | 5 | 32 | 10 |
| REF | 0 | 0 | 1 | 5 | 3 | 5 | 2 | 0 | 8 | 2 |
| PE | 4 | 4 | 26 | 2 | 1 | 89 | 1 | 14 | 25 | 26 |
| NC | 0 | 2 | 2 | 8 | 0 | 11 | 38 | 3 | 18 | 23 |
| POUT | 1 | 6 | 2 | 3 | 0 | 30 | 2 | 28 | 10 | 9 |
| ENTE | 2 | 4 | 10 | 24 | 1 | 22 | 6 | 2 | 257 | 16 |
| PREC | 2 | 1 | 12 | 20 | 1 | 36 | 12 | 10 | 19 | 92 |

SC FT Corpus SMO Confusion Matrix

| J48 | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 5 | 1 | 2 | 3 | 0 | 2 | 0 | 1 | 10 | 2 |
| NPV | 2 | 19 | 5 | 3 | 0 | 7 | 2 | 7 | 9 | 2 |
| PD | 3 | 0 | 37 | 4 | 2 | 31 | 7 | 6 | 22 | 15 |
| ENT | 1 | 2 | 3 | 26 | 1 | 7 | 5 | 9 | 41 | 11 |
| REF | 0 | 0 | 3 | 4 | 6 | 3 | 1 | 1 | 6 | 2 |
| PE | 3 | 6 | 37 | 7 | 0 | 65 | 10 | 10 | 26 | 28 |
| NC | 0 | 0 | 6 | 3 | 4 | 16 | 29 | 6 | 22 | 19 |
| POUT | 0 | 9 | 9 | 5 | 0 | 25 | 5 | 13 | 8 | 17 |
| ENTE | 8 | 7 | 15 | 27 | 7 | 30 | 8 | 9 | 211 | 22 |
| PREC | 5 | 7 | 24 | 8 | 2 | 34 | 19 | 8 | 18 | 80 |

SC FT Corpus J48 Confusion Matrix

| NB | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 4 | 0 | 1 | 2 | 0 | 4 | 1 | 2 | 11 | 1 |
| NPV | 1 | 14 | 1 | 2 | 1 | 9 | 2 | 3 | 19 | 4 |
| PD | 3 | 1 | 47 | 0 | 1 | 37 | 4 | 7 | 16 | 11 |
| ENT | 0 | 3 | 2 | 30 | 2 | 9 | 4 | 4 | 41 | 11 |
| REF | 0 | 2 | 0 | 3 | 4 | 4 | 4 | 1 | 5 | 3 |
| PE | 9 | 5 | 29 | 3 | 3 | 61 | 7 | 24 | 14 | 37 |
| NC | 2 | 2 | 6 | 3 | 4 | 3 | 35 | 7 | 26 | 17 |
| POUT | 2 | 4 | 9 | 3 | 1 | 27 | 3 | 24 | 13 | 5 |
| ENTE | 10 | 4 | 13 | 28 | 7 | 26 | 9 | 10 | 218 | 19 |
| PREC | 3 | 3 | 16 | 8 | 6 | 31 | 20 | 15 | 26 | 77 |

$T_5$ FT Corpus NB Confusion Matrix

| MNB | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 0 | 0 | 0 | 0 | 0 | 12 | 1 | 0 | 10 | 3 |
| NPV | 0 | 13 | 0 | 0 | 0 | 13 | 0 | 4 | 18 | 8 |
| PD | 0 | 0 | 64 | 0 | 0 | 45 | 0 | 2 | 5 | 11 |
| ENT | 0 | 0 | 3 | 15 | 0 | 10 | 2 | 3 | 50 | 23 |
| REF | 0 | 0 | 1 | 1 | 0 | 1 | 4 | 1 | 11 | 7 |
| PE | 1 | 0 | 18 | 0 | 1 | 115 | 0 | 9 | 18 | 30 |
| NC | 0 | 1 | 4 | 3 | 0 | 10 | 36 | 3 | 15 | 33 |
| POUT | 0 | 1 | 4 | 0 | 0 | 49 | 1 | 12 | 7 | 17 |
| ENTE | 0 | 3 | 16 | 10 | 1 | 56 | 1 | 8 | 217 | 32 |
| PREC | 0 | 0 | 13 | 3 | 0 | 51 | 1 | 6 | 13 | 118 |

$T_5$ FT Corpus MNB Confusion Matrix

| SMO | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 3 | 1 | 1 | 1 | 0 | 7 | 0 | 0 | 12 | 1 |
| NPV | 1 | 30 | 3 | 1 | 0 | 4 | 0 | 1 | 13 | 3 |
| PD | 2 | 2 | 54 | 2 | 1 | 29 | 1 | 2 | 21 | 13 |
| ENT | 1 | 1 | 2 | 52 | 1 | 3 | 0 | 3 | 32 | 11 |
| REF | 0 | 0 | 1 | 7 | 2 | 4 | 1 | 0 | 10 | 1 |
| PE | 6 | 5 | 19 | 7 | 2 | 79 | 2 | 14 | 31 | 27 |
| NC | 0 | 2 | 6 | 9 | 2 | 11 | 29 | 3 | 26 | 17 |
| POUT | 2 | 7 | 4 | 7 | 0 | 29 | 2 | 22 | 13 | 5 |
| ENTE | 3 | 5 | 3 | 25 | 1 | 22 | 6 | 1 | 265 | 13 |
| PREC | 3 | 5 | 10 | 24 | 1 | 33 | 11 | 8 | 22 | 88 |

$T_5$ FT Corpus SMO Confusion Matrix

| J48 | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 5 | 1 | 0 | 0 | 0 | 1 | 0 | 2 | 12 | 5 |
| NPV | 0 | 23 | 3 | 2 | 0 | 9 | 0 | 3 | 15 | 1 |
| PD | 3 | 5 | 36 | 5 | 3 | 25 | 6 | 8 | 22 | 14 |
| ENT | 0 | 1 | 1 | 20 | 1 | 19 | 4 | 2 | 46 | 12 |
| REF | 0 | 0 | 1 | 3 | 7 | 1 | 5 | 1 | 7 | 1 |
| PE | 1 | 6 | 29 | 6 | 0 | 67 | 12 | 15 | 30 | 26 |
| NC | 1 | 5 | 9 | 5 | 4 | 5 | 27 | 3 | 25 | 21 |
| POUT | 2 | 7 | 11 | 12 | 0 | 17 | 5 | 16 | 16 | 5 |
| ENTE | 5 | 6 | 14 | 30 | 7 | 20 | 8 | 6 | 227 | 21 |
| PREC | 2 | 6 | 17 | 21 | 3 | 27 | 21 | 4 | 26 | 78 |

$T_5$ FT Corpus J48 Confusion Matrix

| NB | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 4 | 1 | 1 | 2 | 0 | 4 | 0 | 3 | 11 | 0 |
| NPV | 1 | 10 | 1 | 0 | 0 | 10 | 2 | 3 | 26 | 3 |
| PD | 3 | 2 | 50 | 0 | 2 | 35 | 3 | 8 | 15 | 9 |
| ENT | 3 | 4 | 1 | 29 | 2 | 5 | 4 | 4 | 41 | 13 |
| REF | 0 | 2 | 1 | 1 | 4 | 2 | 6 | 1 | 5 | 4 |
| PE | 10 | 2 | 29 | 2 | 5 | 67 | 6 | 26 | 17 | 28 |
| NC | 1 | 1 | 5 | 2 | 6 | 4 | 37 | 7 | 29 | 13 |
| POUT | 3 | 3 | 9 | 3 | 1 | 26 | 3 | 24 | 9 | 10 |
| ENTE | 5 | 6 | 10 | 31 | 10 | 22 | 9 | 13 | 219 | 19 |
| PREC | 4 | 3 | 14 | 6 | 6 | 37 | 15 | 13 | 37 | 70 |

$T_{25}$ FT Corpus NB Confusion Matrix

| MNB | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 0 | 0 | 1 | 0 | 0 | 10 | 1 | 0 | 10 | 4 |
| NPV | 0 | 1 | 1 | 1 | 0 | 14 | 4 | 3 | 22 | 10 |
| PD | 0 | 0 | 67 | 0 | 0 | 48 | 0 | 0 | 3 | 9 |
| ENT | 0 | 0 | 4 | 18 | 0 | 9 | 2 | 1 | 49 | 23 |
| REF | 0 | 0 | 1 | 1 | 0 | 2 | 2 | 1 | 11 | 8 |
| PE | 0 | 0 | 15 | 0 | 0 | 122 | 0 | 8 | 16 | 31 |
| NC | 0 | 0 | 3 | 3 | 0 | 9 | 29 | 1 | 16 | 44 |
| POUT | 0 | 0 | 4 | 1 | 0 | 53 | 0 | 9 | 7 | 17 |
| ENTE | 0 | 2 | 16 | 12 | 1 | 55 | 2 | 3 | 215 | 38 |
| PREC | 0 | 0 | 13 | 2 | 0 | 55 | 2 | 4 | 13 | 116 |

$T_{25}$ FT Corpus MNB Confusion Matrix

| SMO | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 6 | 0 | 2 | 2 | 0 | 3 | 0 | 1 | 11 | 1 |
| NPV | 0 | 32 | 3 | 0 | 0 | 8 | 0 | 3 | 9 | 1 |
| PD | 3 | 1 | 64 | 1 | 1 | 30 | 2 | 4 | 15 | 6 |
| ENT | 1 | 1 | 0 | 49 | 1 | 2 | 0 | 5 | 37 | 10 |
| REF | 0 | 0 | 1 | 4 | 3 | 3 | 2 | 0 | 10 | 3 |
| PE | 5 | 3 | 28 | 2 | 2 | 84 | 2 | 15 | 25 | 26 |
| NC | 0 | 4 | 4 | 4 | 1 | 11 | 31 | 3 | 26 | 21 |
| POUT | 1 | 6 | 4 | 3 | 0 | 27 | 2 | 27 | 10 | 11 |
| ENTE | 2 | 4 | 9 | 29 | 1 | 23 | 7 | 1 | 252 | 16 |
| PREC | 2 | 3 | 11 | 24 | 1 | 30 | 8 | 9 | 19 | 98 |

$T_{25}$ FT Corpus SMO Confusion Matrix

| J48 | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 3 | 1 | 0 | 5 | 1 | 2 | 0 | 0 | 10 | 4 |
| NPV | 1 | 23 | 7 | 4 | 0 | 6 | 0 | 3 | 10 | 2 |
| PD | 2 | 3 | 39 | 4 | 1 | 26 | 6 | 6 | 22 | 18 |
| ENT | 2 | 3 | 5 | 29 | 0 | 13 | 3 | 3 | 34 | 14 |
| REF | 0 | 0 | 2 | 0 | 8 | 6 | 4 | 0 | 5 | 1 |
| PE | 2 | 9 | 21 | 6 | 2 | 71 | 6 | 12 | 29 | 34 |
| NC | 0 | 0 | 8 | 9 | 8 | 11 | 20 | 6 | 18 | 25 |
| POUT | 2 | 7 | 14 | 5 | 1 | 22 | 6 | 13 | 17 | 4 |
| ENTE | 9 | 5 | 8 | 27 | 7 | 31 | 4 | 7 | 217 | 29 |
| PREC | 2 | 3 | 20 | 18 | 3 | 22 | 14 | 8 | 34 | 81 |

$T_{25}$ FT Corpus J48 Confusion Matrix

| NB | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 5 | 1 | 1 | 2 | 0 | 4 | 0 | 2 | 11 | 0 |
| NPV | 1 | 9 | 1 | 2 | 0 | 9 | 2 | 3 | 26 | 3 |
| PD | 2 | 2 | 53 | 0 | 1 | 30 | 5 | 9 | 17 | 8 |
| ENT | 1 | 3 | 1 | 32 | 2 | 6 | 5 | 5 | 39 | 12 |
| REF | 0 | 2 | 2 | 0 | 1 | 2 | 8 | 1 | 6 | 4 |
| PE | 10 | 6 | 30 | 4 | 2 | 61 | 6 | 25 | 18 | 30 |
| NC | 2 | 2 | 6 | 4 | 6 | 4 | 33 | 5 | 27 | 16 |
| POUT | 2 | 3 | 9 | 3 | 1 | 24 | 3 | 26 | 12 | 8 |
| ENTE | 10 | 7 | 9 | 27 | 8 | 22 | 7 | 14 | 220 | 20 |
| PREC | 3 | 3 | 16 | 10 | 5 | 34 | 18 | 16 | 29 | 71 |

$R_{30}$ FT Corpus NB Confusion Matrix

| MNB | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 0 | 11 | 5 |
| NPV | 0 | 0 | 1 | 0 | 0 | 14 | 2 | 3 | 29 | 7 |
| PD | 0 | 0 | 68 | 0 | 0 | 46 | 0 | 0 | 3 | 10 |
| ENT | 0 | 0 | 3 | 15 | 0 | 9 | 1 | 2 | 52 | 24 |
| REF | 0 | 0 | 1 | 0 | 0 | 2 | 2 | 1 | 13 | 7 |
| PE | 0 | 0 | 15 | 0 | 0 | 127 | 1 | 7 | 13 | 29 |
| NC | 0 | 0 | 5 | 4 | 0 | 8 | 21 | 1 | 24 | 42 |
| POUT | 0 | 0 | 5 | 0 | 0 | 53 | 0 | 9 | 7 | 17 |
| ENTE | 0 | 0 | 13 | 7 | 0 | 52 | 2 | 4 | 232 | 34 |
| PREC | 0 | 0 | 15 | 1 | 0 | 54 | 0 | 5 | 15 | 115 |

$R_{30}$ FT Corpus MNB Confusion Matrix

| SMO | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 5 | 1 | 2 | 1 | 0 | 5 | 0 | 1 | 9 | 2 |
| NPV | 1 | 32 | 3 | 0 | 0 | 5 | 0 | 4 | 10 | 1 |
| PD | 3 | 1 | 63 | 3 | 1 | 33 | 1 | 1 | 13 | 8 |
| ENT | 1 | 1 | 1 | 50 | 1 | 2 | 1 | 5 | 35 | 9 |
| REF | 0 | 0 | 1 | 3 | 3 | 4 | 2 | 0 | 11 | 2 |
| PE | 4 | 4 | 24 | 3 | 2 | 85 | 2 | 16 | 28 | 24 |
| NC | 0 | 2 | 2 | 13 | 1 | 10 | 32 | 5 | 17 | 23 |
| POUT | 0 | 3 | 3 | 3 | 0 | 31 | 2 | 28 | 11 | 10 |
| ENTE | 2 | 4 | 13 | 23 | 1 | 24 | 4 | 5 | 255 | 13 |
| PREC | 3 | 3 | 10 | 27 | 2 | 29 | 15 | 10 | 19 | 87 |

$R_{30}$ FT Corpus SMO Confusion Matrix

| J48 | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 4 | 1 | 2 | 3 | 0 | 4 | 0 | 2 | 8 | 2 |
| NPV | 0 | 19 | 3 | 1 | 0 | 7 | 2 | 6 | 16 | 2 |
| PD | 4 | 3 | 42 | 6 | 1 | 27 | 4 | 7 | 14 | 19 |
| ENT | 2 | 5 | 3 | 28 | 0 | 9 | 5 | 6 | 38 | 10 |
| REF | 0 | 1 | 0 | 2 | 8 | 2 | 5 | 2 | 5 | 1 |
| PE | 3 | 10 | 31 | 3 | 1 | 60 | 7 | 19 | 33 | 25 |
| NC | 2 | 2 | 8 | 6 | 5 | 12 | 30 | 5 | 20 | 15 |
| POUT | 3 | 8 | 10 | 6 | 0 | 24 | 3 | 14 | 18 | 5 |
| ENTE | 4 | 8 | 12 | 30 | 8 | 31 | 9 | 7 | 207 | 28 |
| PREC | 1 | 8 | 18 | 19 | 2 | 31 | 21 | 6 | 20 | 79 |

$R_{30}$ FT Corpus J48 Confusion Matrix

| NB | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 6 | 0 | 0 | 2 | 0 | 5 | 0 | 2 | 11 | 0 |
| NPV | 1 | 9 | 0 | 1 | 0 | 9 | 1 | 4 | 26 | 5 |
| PD | 1 | 1 | 53 | 0 | 2 | 32 | 3 | 9 | 15 | 11 |
| ENT | 2 | 5 | 3 | 30 | 2 | 6 | 5 | 6 | 40 | 7 |
| REF | 0 | 2 | 2 | 2 | 4 | 2 | 5 | 0 | 5 | 4 |
| PE | 8 | 6 | 28 | 4 | 4 | 64 | 6 | 27 | 17 | 28 |
| NC | 2 | 1 | 5 | 4 | 3 | 6 | 36 | 7 | 27 | 14 |
| POUT | 2 | 1 | 6 | 4 | 0 | 26 | 4 | 27 | 10 | 11 |
| ENTE | 10 | 3 | 10 | 31 | 7 | 24 | 6 | 12 | 220 | 21 |
| PREC | 3 | 2 | 15 | 9 | 6 | 36 | 17 | 17 | 32 | 68 |

$R_{50}$ FT Corpus NB Confusion Matrix

| MNB | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 0 | 0 | 1 | 0 | 0 | 12 | 1 | 0 | 9 | 3 |
| NPV | 0 | 1 | 1 | 2 | 0 | 16 | 3 | 3 | 22 | 8 |
| PD | 0 | 0 | 60 | 1 | 0 | 51 | 0 | 0 | 3 | 12 |
| ENT | 0 | 0 | 3 | 19 | 0 | 11 | 2 | 2 | 48 | 21 |
| REF | 0 | 0 | 1 | 0 | 0 | 2 | 1 | 0 | 15 | 7 |
| PE | 0 | 0 | 15 | 0 | 0 | 125 | 0 | 7 | 16 | 29 |
| NC | 0 | 0 | 3 v 4 | 0 | 7 | 32 | 1 | 14 | 44 | |
| POUT | 0 | 0 | 7 | 1 | 0 | 52 | 0 | 10 | 5 | 16 |
| ENTE | 0 | 1 | 15 | 10 | 0 | 61 | 3 | 5 | 215 | 34 |
| PREC | 0 | 0 | 14 | 2 | 0 | 50 | 2 | 4 | 16 | 117 |

$R_{50}$ FT Corpus MNB Confusion Matrix

| SMO | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 4 | 1 | 3 | 2 | 0 | 5 | 0 | 0 | 11 | 0 |
| NPV | 1 | 30 | 3 | 0 | 0 | 6 | 0 | 6 | 8 | 2 |
| PD | 2 | 3 | 65 | 4 | 1 | 29 | 0 | 2 | 14 | 7 |
| ENT | 1 | 1 | 1 | 52 | 1 | 3 | 0 | 4 | 35 | 8 |
| REF | 0 | 0 | 1 | 5 | 3 | 4 | 2 | 0 | 8 | 3 |
| PE | 3 | 5 | 26 | 5 | 1 | 90 | 1 | 13 | 25 | 23 |
| NC | 0 | 4 | 6 | 10 | 0 | 9 | 37 | 2 | 14 | 23 |
| POUT | 1 | 6 | 5 | 3 | 0 | 31 | 1 | 25 | 10 | 9 |
| ENTE | 2 | 3 | 9 | 33 | 1 | 24 | 5 | 1 | 251 | 15 |
| PREC | 2 | 4 | 11 | 25 | 3 | 29 | 10 | 6 | 27 | 88 |

$R_{50}$ FT Corpus SMO Confusion Matrix

| J48 | DEF | NPV | PD | ENT | REF | PE | NC | POUT | ENTE | PREC |
|---|---|---|---|---|---|---|---|---|---|---|
| DEF | 5 | 3 | 1 | 1 | 0 | 3 | 0 | 0 | 10 | 3 |
| NPV | 2 | 24 | 5 | 4 | 0 | 2 | 1 | 3 | 14 | 1 |
| PD | 3 | 3 | 49 | 4 | 1 | 31 | 5 | 6 | 12 | 13 |
| ENT | 5 | 6 | 3 | 24 | 2 | 8 | 3 | 4 | 42 | 9 |
| REF | 0 | 0 | 2 | 3 | 5 | 1 | 4 | 0 | 10 | 1 |
| PE | 3 | 9 | 29 | 8 | 4 | 66 | 3 | 8 | 34 | 28 |
| NC | 1 | 3 | 13 | 6 | 5 | 13 | 21 | 2 | 19 | 22 |
| POUT | 4 | 8 | 6 | 9 | 1 | 21 | 7 | 14 | 12 | 9 |
| ENTE | 3 | 8 | 15 | 18 | 6 | 33 | 6 | 6 | 231 | 18 |
| PREC | 2 | 5 | 28 | 22 | 4 | 18 | 21 | 4 | 26 | 75 |

$R_{50}$ FT Corpus J48 Confusion Matrix

### C.2.2 Detailed Accuracy

Detailed accuracy results for the classifiers (Precision, Recall, and F-Measure) are shown below for each algorithm/category. The weighted average (across the categories) is also shown for each algorithm.

| NB Results | | U | SC | $T_5$ | $T_{25}$ | $R_{30}$ | $R_{50}$ |
|---|---|---|---|---|---|---|---|
| Definition | P | 0.156 | 0.114 | 0.111 | 0.111 | 0.086 | 0.114 |
| | R | 0.192 | 0.154 | 0.154 | 0.154 | 0.115 | 0.154 |
| | F1 | 0.172 | 0.131 | 0.129 | 0.129 | 0.098 | 0.131 |
| NumericPropertyValue | P | 0.189 | 0.297 | 0.359 | 0.273 | 0.317 | 0.265 |
| | R | 0.125 | 0.196 | 0.25 | 0.161 | 0.232 | 0.161 |
| | F1 | 0.151 | 0.237 | 0.295 | 0.202 | 0.268 | 0.2 |
| PatientDiagnosis | P | 0.408 | 0.409 | 0.392 | 0.438 | 0.368 | 0.437 |
| | R | 0.417 | 0.425 | 0.37 | 0.441 | 0.362 | 0.433 |
| | F1 | 0.412 | 0.417 | 0.381 | 0.439 | 0.365 | 0.435 |
| Entity | P | 0.367 | 0.356 | 0.395 | 0.37 | 0.357 | 0.33 |
| | R | 0.274 | 0.302 | 0.302 | 0.283 | 0.283 | 0.274 |
| | F1 | 0.314 | 0.327 | 0.342 | 0.321 | 0.316 | 0.299 |
| Reference | P | 0.103 | 0.121 | 0.138 | 0.143 | 0.097 | 0.121 |
| | R | 0.115 | 0.154 | 0.154 | 0.154 | 0.115 | 0.154 |
| | F1 | 0.109 | 0.136 | 0.145 | 0.148 | 0.105 | 0.136 |
| PatientExplanation | P | 0.308 | 0.312 | 0.311 | 0.308 | 0.34 | 0.338 |
| | R | 0.318 | 0.307 | 0.354 | 0.333 | 0.354 | 0.37 |
| | F1 | 0.313 | 0.31 | 0.331 | 0.32 | 0.347 | 0.353 |
| NonClinician | P | 0.434 | 0.412 | 0.452 | 0.407 | 0.446 | 0.447 |
| | R | 0.343 | 0.333 | 0.362 | 0.333 | 0.352 | 0.362 |
| | F1 | 0.383 | 0.368 | 0.402 | 0.366 | 0.394 | 0.4 |
| PatientOutcome | P | 0.257 | 0.262 | 0.234 | 0.272 | 0.243 | 0.278 |
| | R | 0.286 | 0.308 | 0.242 | 0.308 | 0.275 | 0.297 |
| | F1 | 0.271 | 0.283 | 0.238 | 0.289 | 0.258 | 0.287 |
| EntityExplanation | P | 0.548 | 0.551 | 0.556 | 0.55 | 0.561 | 0.55 |
| | R | 0.663 | 0.642 | 0.654 | 0.642 | 0.657 | 0.645 |
| | F1 | 0.6 | 0.593 | 0.601 | 0.592 | 0.605 | 0.594 |
| PatientRecommendation | P | 0.376 | 0.391 | 0.421 | 0.393 | 0.41 | 0.416 |
| | R | 0.317 | 0.322 | 0.351 | 0.332 | 0.346 | 0.337 |
| | F1 | 0.344 | 0.353 | 0.383 | 0.36 | 0.376 | 0.372 |
| Weighted Average | P | 0.393 | 0.398 | 0.41 | 0.401 | 0.406 | 0.409 |
| | R | 0.401 | 0.402 | 0.412 | 0.406 | 0.408 | 0.413 |
| | F1 | 0.394 | 0.398 | 0.408 | 0.401 | 0.404 | 0.408 |

NB FT Classifier Results

| MNB Results | | U | SC | $T_5$ | $T_{25}$ | $R_{30}$ | $R_{50}$ |
|---|---|---|---|---|---|---|---|
| Definition | P | 0 | 0 | 0 | 0 | 0 | 0 |
| | R | 0 | 0 | 0 | 0 | 0 | 0 |
| | F1 | 0 | 0 | 0 | 0 | 0 | 0 |
| NumericPropertyValue | P | 0 | 1 | 0.75 | 0.75 | 0 | 1 |
| | R | 0 | 0.018 | 0.214 | 0.054 | 0 | 0.018 |
| | F1 | 0 | 0.035 | 0.333 | 0.1 | 0 | 0.035 |
| PatientDiagnosis | P | 0.52 | 0.537 | 0.548 | 0.525 | 0.521 | 0.524 |
| | R | 0.504 | 0.512 | 0.535 | 0.504 | 0.496 | 0.512 |
| | F1 | 0.512 | 0.524 | 0.542 | 0.514 | 0.508 | 0.518 |
| Entity | P | 0.478 | 0.5 | 0.395 | 0.472 | 0.5 | 0.447 |
| | R | 0.104 | 0.104 | 0.142 | 0.16 | 0.123 | 0.198 |
| | F1 | 0.171 | 0.172 | 0.208 | 0.239 | 0.197 | 0.275 |
| Reference | P | 0 | 0 | 0 | 0 | 0 | 0 |
| | R | 0 | 0 | 0 | 0 | 0 | 0 |
| | F1 | 0 | 0 | 0 | 0 | 0 | 0 |
| PatientExplanation | P | 0.337 | 0.341 | 0.323 | 0.33 | 0.335 | 0.335 |
| | R | 0.646 | 0.672 | 0.589 | 0.661 | 0.656 | 0.651 |
| | F1 | 0.443 | 0.453 | 0.417 | 0.44 | 0.444 | 0.442 |
| NonClinician | P | 0.677 | 0.75 | 0.698 | 0.689 | 0.767 | 0.674 |
| | R | 0.2 | 0.229 | 0.352 | 0.295 | 0.219 | 0.276 |
| | F1 | 0.309 | 0.35 | 0.468 | 0.413 | 0.341 | 0.392 |
| PatientOutcome | P | 0.235 | 0.25 | 0.286 | 0.243 | 0.25 | 0.297 |
| | R | 0.088 | 0.099 | 0.154 | 0.099 | 0.066 | 0.121 |
| | F1 | 0.128 | 0.142 | 0.2 | 0.141 | 0.104 | 0.172 |
| EntityExplanation | P | 0.597 | 0.584 | 0.582 | 0.61 | 0.575 | 0.607 |
| | R | 0.686 | 0.686 | 0.631 | 0.642 | 0.683 | 0.634 |
| | F1 | 0.639 | 0.631 | 0.605 | 0.626 | 0.624 | 0.62 |
| PatientRecommendation | P | 0.395 | 0.398 | 0.403 | 0.38 | 0.418 | 0.388 |
| | R | 0.585 | 0.551 | 0.537 | 0.532 | 0.595 | 0.556 |
| | F1 | 0.472 | 0.462 | 0.46 | 0.443 | 0.491 | 0.457 |
| Weighted Average | P | 0.439 | 0.49 | 0.468 | 0.473 | 0.446 | 0.486 |
| | R | 0.457 | 0.46 | 0.459 | 0.455 | 0.46 | 0.457 |
| | F1 | 0.414 | 0.419 | 0.438 | 0.425 | 0.416 | 0.427 |

MNB FT Classifier Results

| SMO Results | | U | SC | $T_5$ | $T_{25}$ | $R_{30}$ | $R_{50}$ |
|---|---|---|---|---|---|---|---|
| Definition | P | 0.3 | 0.333 | 0.136 | 0.278 | 0.308 | 0.286 |
| | R | 0.231 | 0.269 | 0.115 | 0.192 | 0.154 | 0.154 |
| | F1 | 0.261 | 0.298 | 0.125 | 0.227 | 0.205 | 0.2 |
| NumericPropertyValue | P | 0.608 | 0.593 | 0.53 | 0.526 | 0.6 | 0.576 |
| | R | 0.554 | 0.571 | 0.625 | 0.536 | 0.536 | 0.607 |
| | F1 | 0.579 | 0.582 | 0.574 | 0.531 | 0.566 | 0.591 |
| PatientDiagnosis | P | 0.53 | 0.533 | 0.523 | 0.5 | 0.488 | 0.542 |
| | R | 0.48 | 0.512 | 0.449 | 0.488 | 0.496 | 0.512 |
| | F1 | 0.504 | 0.522 | 0.483 | 0.494 | 0.492 | 0.526 |
| Entity | P | 0.398 | 0.437 | 0.382 | 0.386 | 0.433 | 0.383 |
| | R | 0.443 | 0.491 | 0.472 | 0.462 | 0.491 | 0.462 |
| | F1 | 0.42 | 0.462 | 0.422 | 0.421 | 0.46 | 0.419 |
| Reference | P | 0.25 | 0.429 | 0.2 | 0.3 | 0.25 | 0.333 |
| | R | 0.115 | 0.115 | 0.077 | 0.115 | 0.115 | 0.115 |
| | F1 | 0.158 | 0.182 | 0.111 | 0.167 | 0.158 | 0.171 |
| PatientExplanation | P | 0.393 | 0.379 | 0.355 | 0.375 | 0.397 | 0.396 |
| | R | 0.49 | 0.464 | 0.37 | 0.453 | 0.484 | 0.484 |
| | F1 | 0.436 | 0.417 | 0.362 | 0.41 | 0.437 | 0.436 |
| NonClinician | P | 0.542 | 0.603 | 0.509 | 0.636 | 0.564 | 0.661 |
| | R | 0.305 | 0.362 | 0.257 | 0.333 | 0.295 | 0.352 |
| | F1 | 0.39 | 0.452 | 0.342 | 0.437 | 0.388 | 0.46 |
| PatientOutcome | P | 0.424 | 0.412 | 0.419 | 0.493 | 0.439 | 0.453 |
| | R | 0.308 | 0.308 | 0.286 | 0.363 | 0.319 | 0.319 |
| | F1 | 0.357 | 0.352 | 0.34 | 0.418 | 0.369 | 0.374 |
| EntityExplanation | P | 0.624 | 0.647 | 0.587 | 0.622 | 0.619 | 0.623 |
| | R | 0.733 | 0.747 | 0.735 | 0.735 | 0.747 | 0.735 |
| | F1 | 0.674 | 0.694 | 0.653 | 0.674 | 0.677 | 0.675 |
| PatientRecommendation | P | 0.495 | 0.479 | 0.469 | 0.514 | 0.5 | 0.529 |
| | R | 0.468 | 0.449 | 0.444 | 0.454 | 0.449 | 0.483 |
| | F1 | 0.481 | 0.463 | 0.456 | 0.482 | 0.473 | 0.505 |
| Weighted Average | P | 0.505 | 0.518 | 0.472 | 0.51 | 0.506 | 0.522 |
| | R | 0.509 | 0.519 | 0.481 | 0.509 | 0.512 | 0.521 |
| | F1 | 0.5 | 0.511 | 0.469 | 0.501 | 0.501 | 0.513 |

SMO FT Classifier Results

| J48 Results | | U | SC | $T_5$ | $T_{25}$ | $R_{30}$ | $R_{50}$ |
|---|---|---|---|---|---|---|---|
| Definition | P | 0.08 | 0.185 | 0.263 | 0.13 | 0.174 | 0.179 |
| | R | 0.077 | 0.192 | 0.192 | 0.115 | 0.154 | 0.192 |
| | F1 | 0.078 | 0.189 | 0.222 | 0.122 | 0.163 | 0.185 |
| NumericPropertyValue | P | 0.451 | 0.373 | 0.383 | 0.426 | 0.292 | 0.348 |
| | R | 0.411 | 0.339 | 0.411 | 0.411 | 0.339 | 0.429 |
| | F1 | 0.43 | 0.355 | 0.397 | 0.418 | 0.314 | 0.384 |
| PatientDiagnosis | P | 0.29 | 0.262 | 0.298 | 0.315 | 0.326 | 0.325 |
| | R | 0.331 | 0.291 | 0.283 | 0.307 | 0.331 | 0.386 |
| | F1 | 0.309 | 0.276 | 0.29 | 0.311 | 0.328 | 0.353 |
| Entity | P | 0.247 | 0.289 | 0.192 | 0.271 | 0.269 | 0.242 |
| | R | 0.208 | 0.245 | 0.189 | 0.274 | 0.264 | 0.226 |
| | F1 | 0.226 | 0.265 | 0.19 | 0.272 | 0.267 | 0.234 |
| Reference | P | 0.357 | 0.273 | 0.28 | 0.258 | 0.32 | 0.179 |
| | R | 0.385 | 0.231 | 0.269 | 0.308 | 0.308 | 0.192 |
| | F1 | 0.37 | 0.25 | 0.275 | 0.281 | 0.314 | 0.185 |
| PatientExplanation | P | 0.301 | 0.295 | 0.351 | 0.338 | 0.29 | 0.337 |
| | R | 0.359 | 0.339 | 0.349 | 0.37 | 0.313 | 0.344 |
| | F1 | 0.328 | 0.316 | 0.35 | 0.353 | 0.301 | 0.34 |
| NonClinician | P | 0.325 | 0.337 | 0.307 | 0.317 | 0.349 | 0.296 |
| | R | 0.257 | 0.276 | 0.257 | 0.19 | 0.286 | 0.2 |
| | F1 | 0.287 | 0.304 | 0.28 | 0.238 | 0.314 | 0.239 |
| PatientOutcome | P | 0.271 | 0.186 | 0.267 | 0.224 | 0.189 | 0.298 |
| | R | 0.143 | 0.143 | 0.176 | 0.143 | 0.154 | 0.154 |
| | F1 | 0.161 | 0.352 | 0.212 | 0.174 | 0.17 | 0.203 |
| EntityExplanation | P | 0.566 | 0.647 | 0.533 | 0.548 | 0.546 | 0.563 |
| | R | 0.613 | 0.747 | 0.66 | 0.631 | 0.602 | 0.672 |
| | F1 | 0.589 | 0.694 | 0.59 | 0.586 | 0.573 | 0.613 |
| PatientRecommendation | P | 0.404 | 0.479 | 0.424 | 0.382 | 0.425 | 0.419 |
| | R | 0.39 | 0.449 | 0.38 | 0.395 | 0.385 | 0.366 |
| | F1 | 0.397 | 0.463 | 0.401 | 0.388 | 0.404 | 0.391 |
| Weighted Average | P | 0.378 | 0.518 | 0.382 | 0.382 | 0.378 | 0.39 |
| | R | 0.384 | 0.519 | 0.396 | 0.394 | 0.384 | 0.402 |
| | F1 | 0.38 | 0.511 | 0.386 | 0.385 | 0.38 | 0.392 |

J48 FT Classifier Results