Facts or Friends? Distinguishing Informational and Conversational Questions in Social Q&A Sites

F. Maxwell Harper, Daniel Moy, Joseph A. Konstan
GroupLens Research
University of Minnesota
{harper, dmoy, konstan}@cs.umn.edu

ABSTRACT
Tens of thousands of questions are asked and answered every day on social question and answer (Q&A) Web sites such as Yahoo Answers. While these sites generate an enormous volume of searchable data, the problem of determining which questions and answers are archival quality has grown. One major component of this problem is the prevalence of *conversational* questions, identified both by Q&A sites and academic literature as questions that are intended simply to start discussion. For example, a conversational question such as "do you believe in evolution?" might successfully engage users in discussion, but probably will not yield a useful web page for users searching for information about evolution. Using data from three popular Q&A sites, we confirm that humans can reliably distinguish between these conversational questions and other *informational* questions, and present evidence that conversational questions typically have much lower potential archival value than informational questions. Further, we explore the use of machine learning techniques to automatically classify questions as conversational or informational, learning in the process about categorical, linguistic, and social differences between different question types. Our algorithms approach human performance, attaining 89.7% classification accuracy in our experiments.

Author Keywords
Q&A, online community, machine learning.

ACM Classification Keywords
H.5.3 Group and Organization Interfaces: Web-based interaction.

INTRODUCTION
Social question and answer Web sites (Q&A sites) leverage the wisdom of crowds to address the limitations of search engines. These are sites that we can turn to when our search terms fail to turn up useful results, or when we seek personal advice and the opinions of others. Social Q&A sites work on a simple premise: that any user can pose a question, and in turn other users – potentially many of them – will provide an answer.

This simple premise has turned out to be very popular. For example, Yahoo Answers – the leading U.S. Q&A site – fields approximately 90,000 new questions every day. To put this number in perspective, Yahoo Answers generates as many new Web pages each month as are contained in the English language Wikipedia (2.5 million).

One of the reasons why social Q&A sites have been launched by major search engine companies is due to their ability to simultaneously expand their searchable corpus and engage users. An early example of success in this area is Knowledge iN, a social Q&A site that was central to the rise of South Korean internet portal Naver [4]. However, outside of Asia, social Q&A has failed to become a feasible alternative to search engines, or even a reliable source of high-quality searchable information. Operators of social Q&A sites speculate that one of (potentially many) reasons for this failure is the interference of conversational question asking – questions like "*what are you doing right now?*" that are unlikely to lead to the creation of archival-quality information.

In this paper, we seek to deepen our understanding of the differences between *conversational* questions and *informational* questions. Let us define these terms:

- *Informational questions* are asked with the intent of getting information that the asker hopes to learn or use via fact- or advice-oriented answers. An example: *What's the difference between Burma and Myanmar?*
- *Conversational questions* are asked with the intent of stimulating discussion. They may be aimed at getting opinions, or they may be acts of self-expression. An example: *Do you drink Coke or Pepsi?*

Research Questions
In this research, we explore the differences between conversational and informational Q&A using human coding, statistical analysis, and machine learning

CHI 2009, April 4–9, 2009, Boston, MA, USA.

algorithms. We organize this work using the following research questions:

RQ1: *Can humans reliably distinguish between conversational questions and informational questions?*

We first seek to validate our division of questions into two primary types. While other researchers have observed social question asking (e.g., [1], [2], [9]), we test the extent to which humans are able to agree when classifying random questions across several Q&A sites.

RQ2: *How do informational and conversational questions differ in terms of writing quality and archival value?*

Our second research question asks if conversational questions are perceived to have lower quality. We investigate two metrics in particular, one concerning the quality of the writing in the question, and the other concerning the long-term value of the question from the perspective of someone searching for information.

RQ3: *What are the structural differences between conversational questions and informational questions?*

In our third question, we explore the structure and nature of conversational and informational questions. We describe categorical, linguistic, and social differences between these two types of questions.

Research Challenge 1: *Can we build algorithms that reliably categorize questions as informational or conversational?*

We believe that developing automated techniques for separating informational and conversational content opens new possibilities for social media research. For instance, we can begin to develop information quality metrics that understand that not all user contributions are intended as archival-quality. Similarly, automated classification techniques open new possibilities for interaction designs, such as automated tagging of content and personalized content ranking systems that leverage user preferences for different content types.

RELATED WORK

Recently, social Q&A sites have become the focus of much research attention. A substantial fraction of this attention has been spent understanding the properties of these new forums for online social discourse. One study looked at the performance of social Q&A sites in comparison with more traditional online avenues for question asking such as online reference libraries, and found both benefits (more diverse opinions) and drawbacks (highly variable quality) [9]. Other researchers have reported on differences among categories in Q&A sites [1], user tendencies to specialize within topics in Q&A sites [7], and the effect of user experience and rewards on performance [19]. Collectively, these studies have revealed that Q&A sites are inconsistent in terms of question and answer quality, and that there is a large and poorly understood social component to these systems.

The apparently extreme variability in answer quality in sites such as Yahoo Answers has led to recent work that attempts to quantify quality, and to leverage computational techniques to improve the user experience. For example, researchers have investigated the use of machine learning techniques to predict askers' satisfaction with answers [15], build models of question and answer quality based on features derived from text and user relationships [2], predict the occurrence of "best answers" [1], and compute metrics of expected user quality [11]. These studies are encouraging: they all indicate that we can develop algorithms to infer the quality of Q&A discourse with some confidence. However, there are limitations to these studies. First, all this work has taken place on just one Q&A site – Yahoo Answers. While Yahoo is the largest social Q&A site in the United States, it is unclear the extent to which any of these methods generalize to the entire social Q&A space, rather than a single instance. Second, these studies treat all questions equally, but it seems likely that quality for a conversational task is a very different thing from quality for an informational task.

DATA COLLECTION AND CODING METHODS

We picked three social Q&A sites to study that offer similar Q&A interfaces, but that differ in the volume of contribution and membership: Yahoo Answers, Answerbag, and Ask Metafilter.[*] These sites each offer an opportunity for users to ask questions on any topic for the community to answer. While there are other types of online forums for online question asking and answering, such as digital reference services and "ask an expert" sites, we do not consider these sites in this analysis, as they have a more restrictive Q&A process – relying on single "experts" to answer questions – and experience empirically different quality problems than social Q&A sites [9].

For each of the three sites, we collected information over a range of dates, including full text, user identifiers, category names and identifiers, and timestamps. See Table 1 for summary statistics of this dataset.

Yahoo Answers (answers.yahoo.com) is the largest Q&A site in the United States, claiming 74% of U.S. Q&A traffic [10]. We downloaded data using the Yahoo Answers web API for a period of seven weeks, resulting in a data set of over 1 million users, 4 million questions, and 24 million answers. Remarkably, the Yahoo Answers community asked an average of 88,122 questions per day over the period of our data collection. Notable features of the Yahoo

[*] These are three of the four most frequently visited Q&A sites in the U.S. as of March, 2008 [10]. We did not study the second-ranked site, WikiAnswers.com, because its interface is substantially different from the "typical" social Q&A site.

	Ask Metafilter	Answerbag	Yahoo Answers
# Days	808	180	49
# Users	11,060	51,357	1,575,633
# Questions	45,567	142,704	4,317,966
# Questions per Day	56	793	88,122
# Answers	657,353	806,426	24,661,775
# Answers per Question	14.43	5.65	5.71
% Questions Answered	99.7%	89.9%	88.2%

Table 1. Properties of the three datasets used in this work.

interface include questions that accept new answers for just 4 days (or at the asker's request, 8 days), browsing organized around categories, and a prominent system of rewarding users with "points" for answering questions.

Answerbag (www.answerbag.com) is a much smaller Q&A site, with an estimated 4.5% market share [10]. We downloaded data using the Answerbag web API, for a data set that spans 6 months of activity. We chose to collect data on questions asked in the first half of 2007, as Answerbag questions remain "open" indefinitely, and many questions continue to receive answers long after they are asked (in contrast with the short lifespan of questions at Yahoo). Answerbag distinguishes itself by sorting questions either by a user-rated "interestingness" metric, or by the last answer received. Also, Answerbag questions do not have a separate subject field, and are limited to just 255 characters. Finally, Answerbag answers may themselves be commented on – we exclude this data from our analysis to ensure consistency across the data sets.

Ask Metafilter (ask.metafilter.com) is the smallest of our Q&A sites, with an estimated 1.8% share of U.S. Q&A traffic [10], and an average of 56 questions per day over the course of our study. We collected over 2 years of data from Ask Metafilter using a custom scraper. Questions on this site may receive answers for a year after they are asked; we stopped scraping questions newer than May 2007 to ensure completeness. Ask Metafilter requires participants to pay $5 to join the community, which has led to a much lower volume of contributions. Also notable is the fact that nearly every question (99.7%) receives at least one answer.

Coding Methodology

Our research questions depend on human evaluations of question type and quality. We developed an online coding tool to better allow us to make use of available volunteers. To help ensure high-quality coding, the tool requires new users to complete a tutorial of instructions and quiz questions. The tutorial is designed to emphasize our own definitions of conversational and informational question asking, to ensure as much consistency across coders as possible. See Figure 1 for an example quiz question.

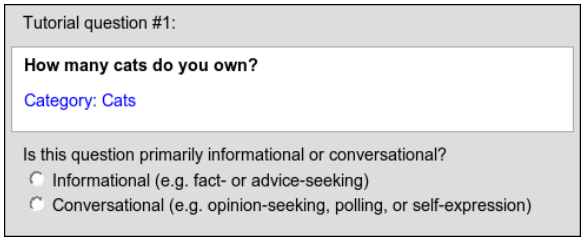

Figure 1. A sample quiz question from the coders' tutorial. This question is conversational, as the apparent *intent* of the question asker is to poll other users. Whether their answer is right or wrong, the tutorial shows an explanation that reinforces our definitions of the terms.

The online coding tool presents the text from all three Q&A sites in a common format, as shown in Figure 2. Other than the question text, the only information shown about each question is its category (which is site-specific), as some questions lack sufficient context without this information (e.g., the question "*How do I kick properly?*" does not make sense unless we understand it is part of the category "swimming"). We do not reveal the name of the Q&A site where the question was asked, and we do not provide a link to the original question or the answers – we wish to ensure that different questions are graded independent of any site-specific bias.

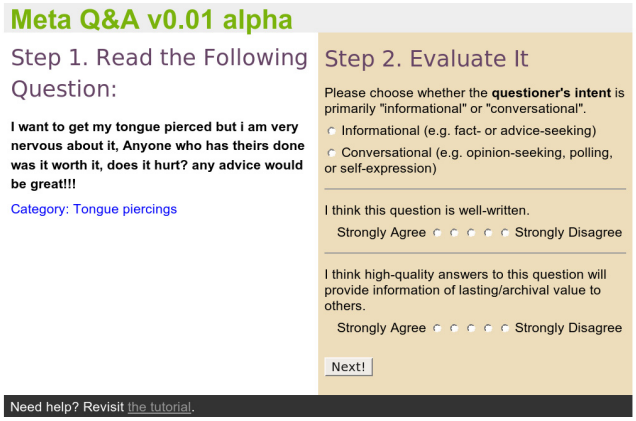

Figure 2. A screenshot from the online coding tool. Coders were asked (1) to determine if the question was asked primarily with informational or conversational intent, and (2) to assign subjective ratings of how well the question was written, and the potential archival value of the question.

The coding tool first asks users to determine if a question is asked with primarily informational or conversational intent. It then asks users to rate the question on two dimensions using Likert scales (5=strongly agree, 1=strongly disagree):

- *WRITING QUALITY*: I think this question is well-written.
- *ARCHIVAL VALUE*: I think high-quality answers to this question will provide information of lasting/archival value to others.

We ask coders to evaluate *ARCHIVAL VALUE* assuming that the question receives high-quality answers, rather than showing the actual answers, because our goal is to

understand questions (not answers) and because we want consistent judgment across sites without the bias of different answer quality.

The coding tool randomly samples questions from our data set, balanced across the three test sites. To ensure some consistency in our coding, each question was independently coded by at least two volunteers. In the case that the first two volunteers disagree about the question type, we solicit two additional volunteers to code the question. For analysis that depends on question type, we do not consider questions where the four voters cannot come to a majority classification. See Figure 3 for an overview of the process.

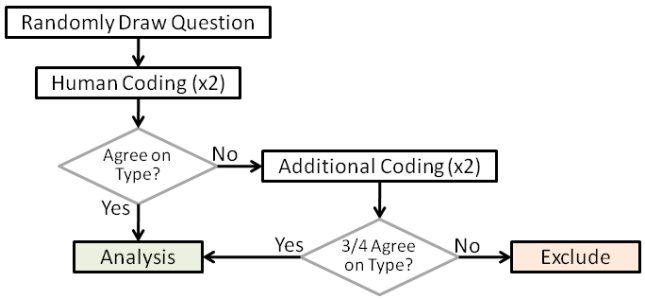

Figure 3. A flowchart showing how a randomly drawn question was classified by two or more coders.

RESULTS OF HUMAN CODING

In all, 30 people participated as coders, submitting a total of 1,106 evaluations across a set of 490 questions. In this section, we present descriptive statistics about this dataset, in the process addressing our first two research questions.

Human Coder Agreement on Question Type

427/490 questions (87.1%) received agreement on question type from the first two coders. Of the remaining 63 questions that were coded by four people, 23 (4.7%) received a "split decision", where two coders voted conversational and two voted informational. Thus, for our machine learning sections below, we consider the set of 467 questions that received majority agreement among coders.

Human Coder Agreement on Quality Metrics

The *WRITING QUALITY* and *ARCHIVAL VALUE* metrics reflect coders' responses to 5 point Likert scale questions. Coders agreed within one point 74.4% of the time on the *WRITING QUALITY* metric (mean disagreement = 1.02), and 70.4% of the time on the *ARCHIVAL VALUE* metric (mean disagreement = 1.08). The distribution of coder disagreements does not change significantly across the different sites in our study.

Site Characteristics

Conversational questions are common in Q&A sites: we found that overall, 32.4% of the questions in our sample were conversational. However, it is clear from these data that different sites have different levels of conversational content ($p < 0.01$, $\chi^2 = 117.4$), ranging from Answerbag where 57% of the questions are coded conversational to Ask Metafilter where just 5% of the questions are coded conversational. All three pairwise comparisons are statistically significant (Answerbag (57%) > Yahoo (36%) > Metafilter (5%); $p<0.01$ for all pairwise tests). See Table 2 for more details.

	Total	Inf.	Conv.	Disagreement
Yahoo	163	93 (57%)	58 (36%)	12 (7%)
Answerbag	161	64 (40%)	92 (57%)	5 (3%)
Metafilter	166	151 (91%)	9 (5%)	6 (4%)
Overall	490	308 (63%)	159 (32%)	23 (5%)

Table 2. Number of questions by coding result. "Disagreement" represents the case where four coders failed to reach a majority classification.

In Figure 4, we compare the three Q&A sites in terms of coders' responses to our *WRITING QUALITY* and *ARCHIVAL VALUE* questions. Ask Metafilter was rated, on average, to have the highest *WRITING QUALITY* scores (means: Answerbag=3.2, Metafilter=3.9, Yahoo=2.7). Using a Tukey-Kramer HSD test [12], we find that all pairwise comparisons indicate statistically significant differences ($\alpha=0.05$). Ask Metafilter was also rated, on average, to have the highest *ARCHIVAL VALUE* scores (means: Answerbag=2.3, Metafilter=3.69, Yahoo=2.4). An HSD test shows that the difference between Ask Metafilter and the other two sites is significant, but that the difference between Answerbag and Yahoo is not ($\alpha=0.05$). More notably, the most common assessment for Yahoo and Answerbag is that the questions will not yield archival value (1 or 2 on a 5-point scale), while 1 and 2 are the least common assessments for Metafilter.

Archival Value and Writing Quality by Question Type

On average, conversational questions received lower scores on both our *WRITING QUALITY* metric (means: Conversational: 2.9, Informational: 3.4) and our *ARCHIVAL VALUE* metric (means: Conversational: 1.7, Informational: 3.3). Figure 5 shows the full results. Notably, we find that fewer than 1% of conversational questions received *ARCHIVAL VALUE* scores of 4 (agree) or above.

To isolate the effects of question type from the site in which it is asked, we build a regression model that includes Q&A site, question type, and the interaction of these two variables. After controlling for site, we still find that conversational question type is a significant negative predictor for both *WRITING QUALITY* ($p<0.01$, $F=10.7$) and *ARCHIVAL VALUE* ($p<0.01$, $F=125.7$).

Discussion

RQ1. Humans can reliably distinguish between conversational and informational questions in most cases. The first two coders agreed 87.1% of the time, while only 4.7% of questions failed to receive a majority classification.

However, the 12.9% of questions where the first two coders disagreed indicate that there is a class of questions that contain elements of both conversational and informational questions. Two example questions that received split

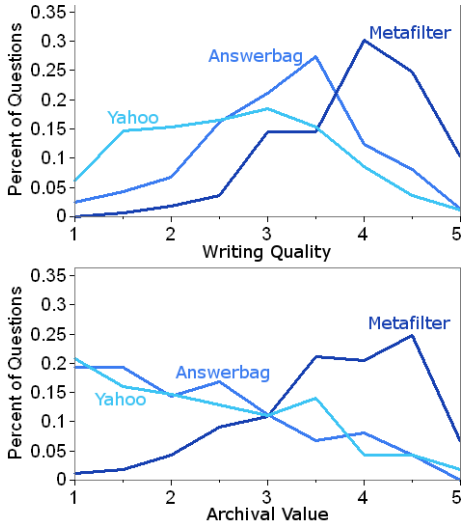

Figure 4. Distribution of aggregate scores assigned to questions (rounded to the nearest 0.5), split by Q&A site. Higher values are better.

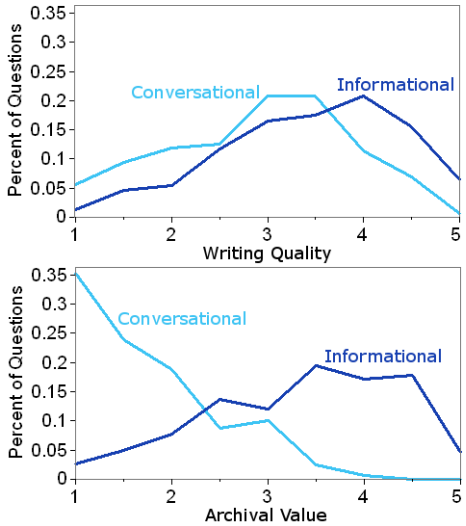

Figure 5. Distribution of aggregate scores assigned to questions (rounded to the nearest 0.5), split by question type. Higher values are better.

decisions are "*how many people are on answer bag*" and "*Is it me or do ipods vibrate?*" In each case, it is hard to determine whether the primary intent of the question asker is to learn information or to start a conversation. Because of these ambiguities, we cannot expect either humans or computers to achieve 100% accuracy in classifying questions.

RQ2. Conversational questions are associated with lower writing quality and lower potential archival value than informational questions. This effect is robust even after controlling for the site in which a question is asked. Few conversational questions appear to have the potential to provide long-term informational value, even if we assume the given answers are high quality.

STRUCTURAL DIFFERENCES AND CLASSIFIERS

We now turn our attention to the task of using machine learning algorithms to detect question type at the time a question is asked. These models help us to understand the structural properties of questions that are predictive of relationship-oriented or information-oriented intent. Specifically, we address RQ3 and begin to address Research Challenge 1 by learning about the categorical, linguistic, and social differences that are indicators of question type.

Machine Learning Methods and Metrics

In general terms, our task is to use the data available at the time a question is asked online to predict the conversational or informational intent of a question asker. To accomplish this task, we employ supervised machine learning algorithms (see [16] for more information).

We use three primary metrics in reporting the performance of our machine learning classifiers: sensitivity, specificity, and area under the ROC curve (AUC). Because the output of our classification algorithm does not have a clear division between "positive" and "negative" values, we arbitrarily consider a positive to be a conversational question. For example, a "true-positive" is correctly predicting a conversational question, and a "false-negative" is incorrectly predicting an informational question. Our metrics may be interpreted as:

- **Sensitivity** - proportion of conversational questions that are correctly classified
- **Specificity** - proportion of informational questions that are correctly classified
- **AUC** - (area under ROC curve) a single scalar value representing the overall performance of the classifier.

For more information about these metrics, see [5].

There is an important reason why we choose to use sensitivity and specificity (more common in the medical literature) over precision and recall (more common in the information retrieval literature): our data have different proportions of positives and negatives across the three Q&A sites. In general, precision and recall suffer from what is known as "class skew", where apparent classification performance is impacted by the overall ratio of positives to negatives in the data set [5]. Thus, if we were to use precision and recall metrics, we could not fairly compare classifier performance across sites with different conversational/informational ratios: performance would appear worse for sites with few conversational questions.

Unless otherwise noted, we employ 5-fold cross validation to evaluate performance. Our data set excludes those "split decision" questions that received 2 votes each for informational and conversational by the coders. "Overall" performance of classifiers across all three sites represents a mathematical combination of the individual performance statistics rather than a fourth (i.e., unified) model. Finally, because the three site-specific ROC curves represent

predictions over different datasets, in our overall performance we simply report the mean of AUC scores, rather than first averaging the individual ROC curves. Unless otherwise noted, we use the Weka data mining software package to build these predictive models [22].

Baseline

We provide a baseline model as a frame of reference for interpreting our results: a 0-R algorithm that always predicts the most commonly occurring class. For example, in Yahoo Answers, 62% of the (non-excluded) questions were coded as informational, so the 0-R algorithm will always predict informational. See Table 3 for results. Note that our baseline outperforms random guessing, which would converge to an overall performance of sensitivity=0.5 and specificity=0.5.

	Sensitivity	Specificity	AUC
Yahoo	0.00	1.00	0.50
Answerbag	1.00	0.00	0.50
Metafilter	0.00	1.00	0.50
Overall	0.58	0.79	0.50

Table 3. Performance of the 0-R baseline classifier.

Predicting Type Using Category Data

Social Q&A sites typically organize questions with categories or tags. Prior work on Yahoo Answers showed that some categories resemble "expertise sharing forums", while others resemble "discussion forums" [1]. In this section, we test the accuracy with which it is possible to classify a question as conversational or informational with just knowledge of that question's category.

All three sites in our study use categories, but they use them in different ways. At the time of our data collection, Metafilter had a flat set of 20 categories, while Answerbag had over 5,700 hierarchical categories and Yahoo had over 1,600 hierarchical categories. The problem with building a classifier across so many categories is that it is difficult to collect a training set that is populated with sufficient data for any given category. However, none of the three sites had more than 26 "top-level categories". Thus, to improve the coverage of our training data, we develop two features:

- "Top-level category" (TLC), a feature that maps each question to its most general classifier in the category hierarchy. For example, in Yahoo, the "Polls & Surveys" category is part of the TLC "Entertainment & Music".
- "Low-level category" (LLC), a feature that maps each question to its most specific category. However, we only populate this feature when we've seen a low-level category at least 3 times across our dataset. For example, we only have one coded example from Answerbag's "Kittens" category, and so leave that question's LLC unspecified in the feature set.

We implement this classifier with a Bayesian network algorithm. This classifier is able to improve on the baseline classifier (sensitivity=0.77, specificity=0.72, AUC=0.78). Compared to the 0-R baseline, the category-based classifier improves sensitivity 18%, but worsens specificity by 4%; see Table 4. In particular, we note that this classifier appears to work quite well on the Metafilter dataset – one where there are only 9 conversational instances – achieving AUC of 0.82.

	Sensitivity	Specificity	AUC
Yahoo	0.66	0.82	0.81
Answerbag	0.82	0.41	0.71
Metafilter	0.56	0.95	0.82
Overall	0.77	0.72	0.78

Table 4. Performance of the category-based classifier.

To better understand the relationship between category and question type, we may look at examples from across the three sites. Table 5 shows the three most popular categories in each Q&A site in our coded dataset. We may see from these examples that some categories provide a strong signal concerning question type. For example, questions in Answerbag's "Outside the bag" category are unlikely to be informational. However, we find few TLCs that are unambiguously predictive.

The addition of low-level categories to the set of features does slightly improve performance as compared with a classifier trained only on TLCs. These categories appear to be especially important in the case of Yahoo, where some categories provide a very strong signal concerning the presence of conversational questions. For instance, Yahoo's Polls & Surveys category was rated 100% conversational (14/14), Singles & Dating was rated 70% conversational (7/10), and Religion & Spirituality was rated 100% conversational (6/6).

Yahoo	Entertainment & Music (20/26) Health (6/22) Science & Mathematics (2/12)
Answerbag	Outside the bag (19/20) Relationship advice (12/15) Entertainment (9/14)
Metafilter	computers & internet (0/34) media & arts (3/19) travel & transportation (1/16)

Table 5. Top 3 Top-Level Categories (TLCs) in the coded dataset and the fraction of conversational questions. Few TLCs provide an unambiguous signal regarding question type.

Predicting Type Using Text Classification

One of the most effective tools in distinguishing between legitimate email and spam is the use of text categorization techniques (e.g., [18]). One of the insights of this line of work is that the words used in a document can be used to categorize that document. For example, researchers have used text classification to algorithmically categorize Usenet posts based on the presence of requests or personal

introductions [3]. In this section, we apply this approach to our prediction problem, in the process learning whether conversational questions contain different language from informational questions.

The text classification technique that we use depends on a "bag of words" as the input feature set. To generate this feature set, we parse the text of the question to generate a list of lower-case words and bigrams. To improve the classifier's accuracy, we only kept the 500 most-used words or bigrams in each type of question. We did not discard stopwords as they turned out to improve the performance of the classifier. We used Weka's sequential minimum optimization (SMO) algorithm.

The text classifier does outperform the baseline classifier (sensitivity=0.70, specificity=0.85, AUC=0.62) but, surprisingly, does not improve on the category-based classifier. However, there is reason for optimism, as this is the best-performing classifier so far on the Answerbag data set. See Table 6 for details.

	Sensitivity	Specificity	AUC
Yahoo	0.48	0.71	0.60
Answerbag	0.79	0.70	0.75
Metafilter	0.00	1.00	0.50
Overall	0.70	0.85	0.62

Table 6. Performance of the text-based classifier.

We now evaluate individual tokens, based on their impact on classification performance as measured by information gain, a metric that represents how cleanly the presence of an attribute splits the data into the desired categories [13]. In this study, information gain can range from 0 (no information) to .93 (perfect information).

Question words. Questions in English often contain "interrogative words" such as "who", "what", "where", "when", "why", and "how". 75.4% of the questions in our coded dataset contain one or more of these words, the most common being the word "what", used in 40.3% of questions. While several of these words ("who", "what, "when") appear to be used in roughly equal proportion across conversational and informational questions, the words "how" and "where" are used much more frequently in informational questions, while the word "why" is used much more frequently in conversational questions. See Table 7 for details.

	% inf.	% conv.	inf. gain
where	**15.5%**	1.9%	0.033
how	**29.8%**	11.3%	0.029
why	5.6%	**17.0%**	0.012
what	30.2%	34.0%	0.000
who	13.9%	11.3%	0.000
when	17.0%	13.2%	0.000

Table 7. Six common interrogative words, and the percentage of questions that contain one or more instances of these words.

I vs. you. Conversational questions are more often directed at readers by using the word "you", while informational questions are more often focused on the asker by using the word "I". The word "I" is the highest-ranked token in our dataset based on information gain. See Table 8 for details.

	% inf.	% conv.	inf. gain
I	**68.6%**	27.4%	0.124
you	25.8%	**54.7%**	0.050

Table 8. A higher percentage of questions with informational intent contain the word "I", while a higher percentage of questions with conversational intent contain the word "you".

Strong Predictors. To find textual features that possess a strong signal concerning question type, we sort all features by their information gain. In Table 9, we show the top features after filtering out common stopwords (they tended to be predictive of informational questions). Again, we find that questions directed at the reader (i.e., using the word "you") are conversational: phrases such as "do you", "would you" and "is your" have high information gain and are predictive of conversational intent. On the other hand, informational questions use words that reflect the need for the readers' assistance, such as "can", "is there", and "help".

	% inf.	% conv.	inf. gain
can	**35.1%**	9.4%	0.069
is there	**10.4%**	0.6%	0.032
help	**19.8%**	5.7%	0.029
do I	**12.3%**	1.9%	0.028
do you	4.9%	**22.0%**	0.047
would you	1.3%	**8.8%**	0.023
you think	1.3%	**8.2%**	0.021
is your	0.0%	**3.7%**	0.020

Table 9. Tokens that are strong predictors of conversational or informational intent, sorted by information gain.

Predicting Type Using Social Network Metrics

Social network methods have emerged to help us visualize and quantify the differences among users in online communities. These methods are based on one key insight: to understand a user's role in a social system, we might look at who that user interacts with, and how often [6]. In this way, each user can be represented by a social network "signature" [21] – a mathematical representation of that user's *ego network*: the graph defined by a user, that user's neighbors, and the ties between these users [8]. Researchers have used these signatures to better understand canonical user roles in online communities [21], including *answer people*, who answer many people's questions, and *discussion people*, who interact often with other discussion people. While these roles were developed in the context of Usenet, a more traditional online discussion forum, they are adaptable (and useful) in the Q&A domain [1].

In this section, we use social network signatures to predict question type. This analysis builds on features that describe each user's history of question asking and answering. We

treat the Q&A dataset (described in Table 1) as a directed graph, where vertices represent users and directed edges represent the act of one user answering another user's question (see Figure 6 for an example). We allow multiple edges between pairs of vertices. For example, if user A has answered two of user B's questions, there are two edges directed from A to B. We model the three datasets collected from separate Q&A as three separate social networks. Because we wish to build models that can make predictions at the time a question is asked, we filter these structures by timestamp, ensuring that we have an accurate snapshot of a user's interactions up to a particular time.

To make use of these social networks in our machine learning framework, we construct the following features:

- *NUM_NEIGHBORS*: The number of neighbors to the question asker. (How many other users has the question asker interacted with?)
- *PCT_ANSWERS*: The question asker's number of outbound edges as a percentage of all edges connected to that user. (What fraction of the question asker's contributions have been answers?)
- *CLUST_COEFFICIENT*: The clustering coefficient [20] of the question asker's ego network (How inter-connected are the question asker's neighbors?)

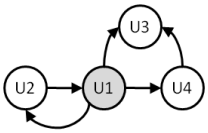

Figure 6. An example ego network for user U1. U1 has answered questions by U2, U3, and U4, while U2 has answered a question by U1. U1's metrics: *NUM_NEIGHBORS*=3, *PCT_ANSWERS*=0.75, *CLUST_COEFFICIENT*=0.17.

The social network-based classifier, implemented with a Bayesian network algorithm, is able to correctly classify 69% of conversational questions and 89% of informational questions, overall. While the model appears successful across both Yahoo and Answerbag, performance is low at Metafilter (AUC=0.64). See Table 10 for details.

	Sensitivity	Specificity	AUC
Yahoo	0.71	0.87	0.81
Answerbag	0.87	0.61	0.72
Metafilter	0.00	1.00	0.64
Overall	0.69	0.89	0.72

Table 10. Performance of the social network-based classifier.

Analyzing a social network built from Q&A discourse reveals strong differences between users who ask questions with a conversational intent and users who ask questions with informational intent (see Figure 7 for an overview). The first metric, *NUM_NEIGHBORS*, shows that users who ask conversational questions tend to have many more neighbors than users who ask informational questions (means: Conversational=757, Informational= 252). This effect is significant across the three sites ($p<0.01$), as well as within each of the three sites ($p<0.01$ for all three).

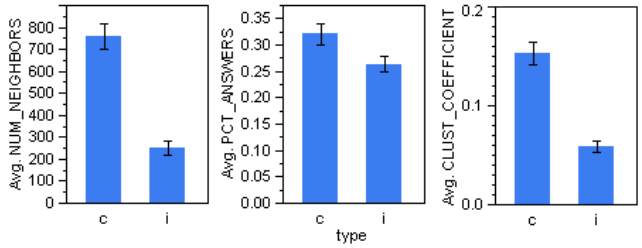

Figure 7. Differences in the three social network metrics across question type (c=conversational, i=informational), shown with standard error bars.

Conversational question askers tend to have higher *PCT_ANSWERS* scores than informational question askers (mean *PCT_ANSWERS*: Conversational=0.32 vs. Informational=0.26, $p=0.02$). This effect is robust across Yahoo ($p=0.02$) and Answerbag ($p<0.01$), though Metafilter exhibits the reverse effect ($p<0.01$) where informational askers have a higher *PCT_ANSWERS* score than conversational askers.

Finally, looking at *CLUST_COEFFICIENT* reveals that users asking conversational questions have more densely interconnected ego networks than users asking informational questions (mean *CLUST_COEFFICIENT*: conversational=0.15 vs. informational=0.06, $p<0.01$). This effect is robust across Yahoo ($p<0.01$) and Answerbag ($p<0.01$), but not significant in Metafilter ($p=0.35$).

Discussion

RQ3. There are several structural properties of conversational and informational questions that help us to understand the differences between these two question types. Though site-specific, some categories are strongly predictive of question type, such as the top-level category "computers & internet" in Metafilter (predictive of informational) and the low-level category "Polls & Surveys" in Yahoo Answers (predictive of conversational). Certain words are also strongly predictive of question type. For example, the word "you" is a strong indicator that a question is conversational. Finally, users that ask conversational questions tend to have larger, more tightly interconnected ego networks.

Research Challenge 1. All three of the models presented in this section outperform our baseline model. However, each individual model exhibits strengths and weaknesses. Moving forward, we attempt to address these limitations through the use of ensemble methods.

AN ENSEMBLE FOR PREDICTING QUESTION TYPE

Intuitively, it would seem that our different feature sets complement one another, rather than simply presenting several different ways of arriving at the same predictions. In general, we assert a belief that in online content analysis, the use of multiple complementary feature sets is superior to the use of a single feature set.

Classifier Diversity

The key idea of ensemble learning methods is that we may build many classifiers over the same data, then combine their outputs in such a way that we capitalize on each classifier's strengths. In our case, we may intuitively note that a text classifier may perform better in the presence of more text, or that a category classifier will more accurately classify questions in some categories than in others. Thus, to assess the potential of our classifiers to collaborate, we first assess whether they are making errors on the same questions.

To measure our potential for improvement through ensemble methods, we may turn to diversity metrics [17]. These metrics quantitatively measure the tendency for a pair of classifiers to make the same mistakes. In this analysis, we choose Yule's Q Statistic [23]. This metric varies from -1 to 1; classifiers that tend to categorize the same instances correctly take more positive values, while classifiers that tend to categorize different instances incorrectly take more negative values. Thus, lower values indicate greater diversity [14].

We find that the category-based and social network-based classifiers appear to pick up on much of the same signal (Q=0.72), perhaps showing that the same "types" of users tend to post in the same categories. On the other hand, we see better diversity scores when comparing the output of text-based and category-based classifiers (Q=0.31) as well as the social network-based and text-based classifiers (Q=0.58).

Algorithm Details and Results

We construct an ensemble classifier by running a meta-classifier on the output of three individual classifiers: a confidence score between 0 and 1 that a question is conversational. Intuitively, we might believe that our meta-classifier picks up on signals concerning which classifier is strongest in each site and what different confidence scores mean in each classifier. We used JMP's neural network algorithm for this classifier, and report the results of 5-fold cross validation on reordered data.

The ensemble method shows strong improvement over any of the individual classifiers for each site. This classifier returns AUC values > 0.9 for each site, correctly classifying 79% of conversational content and 95% of informational content (see Table 11), for an overall accuracy of 89.7%.

	Sensitivity	Specificity	AUC
Yahoo	0.78	0.95	0.95
Answerbag	0.85	0.84	0.91
Metafilter	0.33	1.00	0.91
Overall	0.79	0.95	0.92

Table 11. Performance of the ensemble classifier.

All three individual classifiers mutually agreed on a classification 61.9% of the time – in these cases, the ensemble achieved 93.8% accuracy. In the remaining 38.1% of the instances where there was disagreement, the ensemble achieved just 83.2% accuracy. For example, the question "*What operating system do you prefer? Windows, Linux, Mac etc."* was correctly classified as conversational, despite a wrong prediction by the category-based classifier (the question is in Answerbag's Operating systems category). However, the question "*Which 1 do u need more??? Money or Love???*" was incorrectly classified as informational, as only the category-based classifier was correct (the question is in Yahoo's Polls & Surveys category). There is future work in using multiple learners as a means for generating confidence scores in the resulting classification.

Discussion

Research Challenge 1. Our classifier achieves 89.7% classification accuracy across the three Q&A sites, close to the rate at which the first two human coders agreed on a classification (91.4%). Unsurprisingly, the classifier was more accurate on these questions where the first two coders agreed than on questions where four coders were required (92.0% vs. 65.9% classification accuracy). Also, recall that coders were unable to achieve consensus on 5% of the questions; we do not expect our algorithm to achieve classification where humans cannot. Based on these results, we are optimistic about the potential for algorithms that are at least as reliable as humans for distinguishing conversational and informational questions.

SUMMARY DISCUSSION AND DESIGN IMPLICATIONS

In this paper, we investigated the phenomenon of conversational question asking in social Q&A sites, finding that few questions of this type appear to have potential archival value. We explored the use of machine learning techniques to automatically distinguish conversational and informational questions, finding that an ensemble of techniques is remarkably effective, and in the process learning about categorical, linguistic, and social differences between the question types.

Not included in this paper are several classification methods that did not work well. For instance, we built a classifier from quantitative features extracted from the text (e.g., question length, and Flesch-Kincaid grade level) that barely outperformed our baseline classifier: apparently conversational and informational questions "look" similar! Also, we tried a suite of additional features to supplement the social network model, none of which improved performance. However, there is potentially more signal to be mined from the text of a question than we have realized. Natural language-based methods could extract additional features to supplement our naïve "bag of words" feature set, though NLP programmers beware: the language used in Q&A sites often barely resembles English! (LOL)

We acknowledge that looking at questions in isolation is not necessarily the best way to classify a Q&A thread. For example, the question "*Why is the sky blue?*" might appear

to be informational in intent, until you realize that this question has been asked over 2,000 times in Yahoo Answers, and often receives no replies with serious answers to the question. Although we addressed the classification problem at the time a question is asked, it is certainly possible to classify a question hours or days after it is asked, utilizing information about the answerers and the text of the answers.

However, classifying questions at the time they are asked can allow for effective automated tagging, and for the development of systems that direct questions to the appropriate place. Such classification could lead to the development of interfaces that support both social users (who drive traffic numbers and advertising revenue) and informational users (who generate searchable content). Q&A sites could adjust user reward mechanisms based on question type, offer different "best answer" interfaces to enhance the fun of participating in conversational Q&A, or create search tools that allow users to perform informational search (emphasizing keywords) or conversational search (emphasizing timeliness).

Social Q&A is certainly not the only online domain struggling to better understand how to simultaneously attract committed users and encourage high quality contributions. Wikis, discussion forums, blogs, and other forms of social media are an increasing fraction of the searchable Web; across all of these types of sites, it is important to understand the implications of users' roles, motivations, and intentions. In this work, we use a combination of human evaluation and machine learning techniques to explore Q&A questions, discovering that conversational question asking is an important indicator of information quality. We hope that these methods will prove useful to others in the research community as they pursue related questions across different domains to better understand how to identify and harvest archival-quality user-contributed content.

ACKNOWLEDGEMENTS

We gratefully acknowledge the help of Marc Smith, Shilad Sen, Reid Priedhorsky, Sam Blake, and our friends who helped to code questions. This work was supported by the National Science Foundation, under grants IIS 03-24851 and IIS 08-12148.

REFERENCES

1. Adamic, L., Zhang, J., Bakshy, E., Ackerman, M. Knowledge Sharing and Yahoo Answers: Everyone Knows Something. In *Proc. WWW* (2008).
2. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G. Finding High-Quality Content in Social Media. In *Proc. WSDM* (2008).
3. Burke, M., Joyce, E., Kim, T., Anand, V., Kraut, R. Introductions and Requests: Rhetorical Strategies that Elicit Community Response. In *Proc. Communities and Technologies* (2007).
4. Chae, M., Lee, B. Transforming an Online Portal Site into a Playground for Netizen. *Journal of Internet Commerce*, 4 (2005).
5. Fawcett, T. ROC Graphs: Notes and Practical Considerations for Researchers. *HP Labs Tech Report HPL-2003-4* (2004).
6. Fisher, D., Smith, M., Welser, H. T. You Are Who You Talk To: Detecting Roles in Usenet Newsgroups. In *Proc. HICSS* (2006).
7. Gyöngyi, Z., Koutrika, G., Pedersen, J., Garcia-Molina, H. Questioning Yahoo! Answers. In First Workshop on Question Answering on the Web (2008).
8. Hanneman, R., Riddle, C. *Introduction to Social Network Methods*. Riverside, CA: University of California, Riverside (2005).
9. Harper, F., Raban, D., Rafaeli, S., Konstan, J. Predictors of Answer Quality in Online Q&A Sites, In *Proc. CHI* (2008).
10. Hitwise. U.S. Visits to Question and Answer Websites Increased 118 Percent Year-over-Year (2008). http://www.webcitation.org/5alK5xpWh
11. Jurczyk, P., Agichtein, E. HITS on Question Answer Portals: Exploration of Link Analysis for Author Ranking. In *Proc. SIGIR* (2007).
12. Kramer, C. Extension of Multiple Range Tests to Group Means with Unequal Numbers of Replications. *Biometrics*, 12 (1956).
13. Kullback, S., Leibler, R. On Information and Sufficiency. *Annals of Mathematical Statistics*, 22 (1951).
14. Kuncheva, L., Whitaker, C. Measures of Diversity in Classifier Ensembles. *Machine Learning*, 51:2 (2000).
15. Liu, Y., Bian, J., Agichtein, E. Predicting Information Seeker Satisfaction in Community Question Answering. In *Proc SIGIR* (2008).
16. Mitchell, T. *Machine Learning* (1997).
17. Polikar, R. Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6 (2006).
18. Sahami, M., Dumais, S., Heckerman, D., Horvitz, E. A Bayesian Approach to Filtering Junk E-mail. In *Learning for Text Categorization* (1998).
19. Shah, C., Oh, J., Oh, S. Exploring Characteristics and Effects of User Participation in Online Social Q&A Sites. *First Monday*, 13 (2008).
20. Watts, D., Strogatz, S. Collective Dynamics of 'Small-World' Networks. *Nature*, 393 (1998).
21. Welser, H. T., Gleave, E., Fisher, D., Smith, M. Visualizing the Signatures of Social Roles in Online Discussion Groups. *The Journal of Social Structure*, 8 (2007).
22. Witten, I., Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques* (2005).
23. Yule, G. On the Association of Attributes in Statistics, In *Proc. Royal Society of London*, 66 (1900).