Assessing Order Effects in Online Community-based Health Forums

Completed Research Paper

Reza Mousavi
Arizona State University
300 E. Lemon St.
Tempe, AZ 85287
mousavi@asu.edu

T. S. Raghu
Arizona State University
300 E. Lemon St.
Tempe, AZ 85287
raghu.santanam@asu.edu

Keith Frey
Dignity Health
3030 N Central Avenue
Phoenix, AZ 85012
keith.frey@dignityhealth.org

Abstract

Measuring the quality of health content in online health forums has been a challenging task. The majority of the existing measures are based on nonprofessional evaluations of forum users and may not be reliable. We employed machine learning techniques, text mining methods, and Big Data platforms to construct four measures of textual quality to automatically determine the similarity of a given answer to professional answers. We then used these measures to assess the quality of 66,888 answers posted on Yahoo! Answers Health section. All four measures of textual quality revealed a higher quality for asker-selected best answers indicating that askers, to some extent, have a proper judgment to select the best answers. We also studied the presence of order effects in online health forums. Our results suggest that the textual quality of the first answer positively influences the mean textual quality of the subsequent answers and negatively influences the quantity of the subsequent answers.

Keywords: Healthcare information systems, online communities, machine learning, text mining, information quality

Introduction

Seeking information on the web is now a nearly inevitable part of our lives. According to Pew's 2014 report, 87% of U.S. adults use the Internet. 72% of these users said they have looked at health information online within the past year. According to the same report, more than one third of U.S. adults say that at one time or another they have gone online specifically to try to figure out what medical condition they or someone else might have. The same study reports that 46% of online diagnosers' (i.e. users who seek health advice online) internet searches urged them to visit a medical professional. 38% of them could use the online information to take care of themselves at home and 11% of the online diagnosers fell in between the two options. Women were more likely to go online to find healthcare information. Furthermore, younger adults, those who have a college degree, and those from households who earn more than $75,000 per year were more inclined to seek health advice online (Pew Research Center 2014). Previous reports by Pew also reveal that online information seekers often take the recommendations seriously. 53% of

respondents talked to their clinicians about what they found online and 41% of respondents had their conditions confirmed by a clinician (Fox and Duggan 2013).

Given the abundance of health-related resources available online, online diagnosers have a variety of choices of where to seek and find information about health conditions. One of these options is a health-related community-based question answering (HCQA) platform, where users can ask their questions and seek advice from the other users of the platform. Availability, ease of use, low cost, and usefulness of HCQA platforms motivate users to ask health-related questions, as do the publicity and the social aspects of them.

One of the pioneers of such HCQA platforms is Yahoo! Answers, which was launched in June 2005. Yahoo! Answers is a question-and-answer platform that enables users to ask questions and post answers to other users' questions. Since its inception, Yahoo! Answers has been one of the most popular HCQA platforms of all time. Although the actual number of online diagnosers who use Yahoo! Answers to find relevant information about health conditions is not reported, it is estimated that Yahoo! Answers dominates the online Q&A platforms by gaining 74.05% of the overall market share (Jasra 2008). The following statement by an actual user of Yahoo! Answers provides a good explanation of its popularity among users seeking health advice:

"MOST people have the same problems with their health yr after yr after yr and are tired of going through this process. This is why they research their symptoms on the Web to see if MAYBE someone some where has the same thing and they can show the doctor so he might prescribe correctly or run tests to check and make sure it is right, then get the patient on the proper meds once and for all."[1]

The main issue in HCQA platforms is related to the quality of answers. Current HCQA platforms such as Yahoo! Answers rely on the judgments of the users themselves to provide a metric for quality. For instance, Yahoo! Answers enables the user who asked the question to choose the best answer of the answers posted and enables other users to "thumbs up" (approve) or "thumbs down" (disapprove) each answer. Since this approach still relies on subjective opinions of non-professional users, the quality of answers provided in HCQA platforms could largely deviate from that of healthcare professionals' opinions. For instance, a study by Oh et al. (2012) revealed that librarians and professional nurses rate the quality of answers in HCQA platforms significantly lower than non-professional users. As inappropriate health-related recommendations could have serious effects on advice–takers, reliable measures that scale the quality of recommendations are necessary.

Although the quality of answers is a multifaceted construct that may depend on a variety of elements such as the accuracy, completeness, and relevance of the advice and the credibility of the source, previous studies show that, to some extent, the textual features could be employed to determine the quality of online answers (Fallis and Frické 2002; Ghose and Ipeirotis 2011; Weimer et al. 2007). Therefore, by employing machine learning and text mining methods, we developed two measures for assessing the extent to which a given answer by a normal user resembles an answer given by a healthcare professional.

To evaluate the appropriateness of these metrics, we use them to measure and compare the quality of normal answers with the quality of best answers chosen by the askers. Furthermore, we employ these two metrics in addition to two previously developed metrics to study the effects of the quality of the first answer on the quality of the subsequent answers and the number of subsequent answers in Yahoo! Answers. We try to answer the following questions:

- Does the textual quality of the answer chosen as the best answer by the asker differ from that of a regular answer?
- Does the textual quality of the first answer influence the textual quality of the subsequent answers?
- Does the textual quality of the first answer influence the number of the subsequent answers?

The organization of this paper is as follows. The next section reviews the current literature relating to quality of answers in online forums. We present our data and variables in the next section. We describe the empirical model and present descriptive statistics along with the results of our analyses in the

[1] From: https://answers.yahoo.com/question/index?qid=20130513102251AAZ32kR

following section. We discuss our findings and conclude with limitations and potential extensions of this study in the final sections.

Literature Review

Evaluating the Quality of Answers

Initial attempts to evaluate the quality of answers in community-based question answering (CQA) platforms relied on human intelligence. The first mechanism for evaluating answers was introduced by CQA platforms themselves. In this approach, the CQA platform enables the asker to state whether the answers given were satisfactory. In most CQA platforms, the asker can select one answer as the best answer. As an extension to this approach, CQA platforms introduced voting mechanisms through which users other than the asker could evaluate the usefulness of the answers. In most CQA platforms, users can "thumbs up" (approve) or "thumbs down" (disapprove) the answers. Both of these approaches are part of a design paradigm called social computing, which is defined as a computational integration of social studies and human social dynamics together with the design and use of ICT technologies in a social context (Wang et al., 2007).

In subsequent attempts to evaluate the quality of the answers in CQA platforms, the social computing paradigm has applied more rigorous approaches. For instance, Shah & Pomerantz (2010) asked Amazon Mechanical Turk workers to rate the quality of a set of answers taken from Yahoo! Answers based on 13 different criteria. In this approach, the human intelligence is making the judgments about the quality of the answers. Along with social computing, researchers have also looked at the textual and non-textual features of the answers to evaluate their quality. Jeon et al. (2006) studies the effects of several factors including the length of the answers, the number of the answers in a given category, and the number of previous best answers by the answerer on the quality of the answer. Otterbacher (2009) studied consumer reviews on Amazon.com, deriving 22 measures quantifying their textual properties, authors' reputations, and product characteristics. Overall, this stream of research relies on social computing, textual clues, and non-textual clues such as answerers' reputation to evaluate the quality of the answers.

The next wave of studies that evaluated the quality of the answers in CQA platforms deviated from the social computing paradigm and employed machine-learning techniques to assess the answers given in CQA platforms. This recent set of studies has used text mining methods to automatically detect the answers that could be candidates for the best answers chosen by the askers. Agichtein et al. (2008) used textual features, user information, and usage statistics to evaluate the quality of the answers posted on Yahoo! Answers. Liu et al. (2010) employed Latent Dirichlet Allocation model to predict the best answers for new questions in CQA platforms. Hu (2013) took the problem of detecting a high-quality answer as a classification task and employed a multimodal deep belief nets-based approach to perform the classification. This approach includes two steps. In the first step, a specially-designed deep network is given to learn the unified representation using both textual and non-textual information. In the second step, the outputs of the previous step are used as inputs for a classifier to make prediction. In another study by Yao et al. (2015), the quality of the answers is tied to the quality of the questions. Yao and colleagues developed algorithms that use the votes on the questions and answers as indicators of their quality.

The Effects of First Answer's Quality

Psychology literature has long been studying the effects of item orders on the quality of survey responses (Hogarth and Einhorn 1992; Krosnick and Alwin 1987; McFarland 1981). Such effects are called "order effects" and have been reportedly observed during survey administrations. For instance, in a 2008 experiment by Pew Research center, when people were asked, "All in all, are you satisfied or dissatisfied with the way things are going in this country today?" immediately after having been asked, "Do you approve or disapprove of the way George W. Bush is handling his job as president?" 88% said they were dissatisfied, compared to only 78% without the context of the prior question.[2]

[2] From: http://www.people-press.org/methodology/questionnaire-design/question-order

Another form of order effects is observed in response alternatives (Holbrook et al. 2007). According to Lavrakas (2008), "[R]esponse order effect occurs when the distribution of responses to a closed-ended survey question is influenced by the order in which the response options are offered to respondents" (P. 95). In general, there are two types of response order effects: primacy and recency (Krosnick and Alwin 1987). Primacy refers to the phenomenon that occurs when the placement of an answer as the first answer increases its likelihood of being selected. Recency refers to the case where placing an answer as the last alternative increases its likelihood to be selected. Particularly when response options are presented orally, the recency effect is dominant, as respondents cannot think much about the first option they heard and therefore weigh the last alternative more (Holbrook et al. 2007). Research shows that when choices are presented visually, the primacy effect is to be expected. There are two major reasons for the primacy effect. First, an alternative that is presented as the first option may establish a cognitive framework or standard for comparison that guides the interpretation of later items. Second, an alternative that is presented first is subjected to a deeper cognitive processing and therefore might occupy the mind of the respondent even when s/he is reading the subsequent alternatives. In other words, the first items "suffer less competition for time and space in immediate memory from other items" (Schwarz et al. 1992, p. 188). Due to the latter effect, the respondent is less stimulated by the subsequent answers and may select them less frequently (Klayman and Ha 1987; Krosnick and Alwin 1987; Tversky and Kahneman 1981).

Another evidence supporting the importance of the first item in a list of items is provided by Galesic et al. (2008). Using eye-tracking technology to record the eye movements of online survey respondents, Galesic and colleagues observed what respondents did and did not look at while they answered questions. Their results suggest that respondents do in fact spend more time looking at the first few options in a list than those at the end of the list. This study provides evidence for the previous claims that were made about the primacy effect.

Based on the studies reviewed above, we propose that the first answers in HCQA platforms would serve as reference points for the subsequent answers. That is, the providers of the subsequent answers would rely partly on the features of the first answer. Therefore, we argue that the quality of the first answer could influence the quality of the subsequent answers.

H1: Ceteris paribus, the textual quality of the first answer has a positive influence on the average textual quality of the subsequent answers for a given question in an HCQA platform.

Another significant yet indirect effect of high-quality first answers would be on the number of answers posted on HCQA platforms. Since HCQA platforms have different policies regarding closing a question thread or keeping it open, and since an HCQA platform's policy may influence the number of answers users are able to post, we first need to discuss Yahoo! Answers' policies with regard to closing a question thread.

In Yahoo! Answers, once a question is posted, other users can answer it. When the person who asked the question is satisfied with one of the answers, s/he can close the question thread by selecting one of the answers as the best answer. The question and all of its answers will be available on Yahoo! Answers for future use although no one can provide a new answer after the asker selects the best answer. In this setting, the respondents are not aware of the best answer at the time of posting their answers, as the asker has not yet selected the best answer. Therefore, the respondents see all answers in chronological order.

We argue that when an HCQA platform imposes a closed-form policy, where further answers cannot be posted after the best answer is determined, the number of answers received by each question shrinks as the quality of the first answer increases. The reason is that a high-quality first answer could either be a candidate for the best answer, or, according to our first hypothesis, it could increase the chance of higher-quality answers to be posted subsequently. In both ways, the asker will receive high-quality answers early on and may settle the thread by selecting one of those initial answers as the best answer. Once the best answer is selected, the thread is closed and no more answers can be posted. Therefore, the total number of subsequent answers decreases when the first answer has a high quality. Therefore:

H2: Ceteris paribus, the textual quality of the first answer has a negative influence on the quantity of the subsequent answers for a given question in an HCQA platform with closed-form policy.

Data & Variables

Yahoo! Answers

A total of 17,824 questions and 66,888 answers (3.75 answers per question on average) were collected from Yahoo Answers' Health section. The questions were posted starting from 7/7/2005 and ending to 9/2/2006 and the answers were posted starting from 7/8/2005 and ending to 9/9/2006. This data set includes information about the questions (the content of the question, the date and time the question was posted, the topic of the question, and user id of the asker) and the answers (the content of the answer, the date and time the answer was posted, the user id of the respondent, and whether the answer was chosen as the best answer by the asker). We only collected question sets that had a best answer. The questions are categorized in 21 health-related topics ranging from women's health to injuries and diseases. Table 1 reports the frequencies of the answers for each of the health topics.

Table 1. The Frequencies of Answers for Each Health Topic

Health Topic	Frequency	Percent	Cumulative
Allergies	431	0.64	0.64
Alternative Medicine	2,038	3.05	3.69
Cancer	738	1.1	4.79
Dental	2,328	3.48	8.28
Diabetes	522	0.78	9.06
Diet & Fitness	9,592	14.34	23.4
First Aid	138	0.21	23.6
General Health	10,121	15.13	38.73
Heart Diseases	602	0.9	39.63
Infectious Diseases	878	1.31	40.95
Injuries	191	0.29	41.23
Men's Health	4,225	6.32	47.55
Mental Health	9,205	13.76	61.31
Other – Diseases	3,550	5.31	66.62
Other - General Health Care	2,529	3.78	70.4
Other – Health	10,168	15.2	85.6
Pain & Pain management	47	0.07	85.67
Respiratory Diseases	376	0.56	86.23
STDs (Sexually Transmitted Diseases)	1,077	1.61	87.84
Skin Conditions	974	1.46	89.3
Women's Health	7,158	10.7	100

According to Table 1, "Other Health", "General Health", and "Diet & Fitness" were the most answered topics; "Pain and Pain Management", "First Aid", and "Allergies" were the least answered topics.

Textual Quality Measure 1: Readability (FKGL)

To evaluate the readability of each question and answer, we employed Flesch-Kincaid Grade Level (FKGL) method.[3] This metric assesses the readability of an English text. FKGL was first used by the United States Army for measuring the difficulty of technical manuals in 1978. This grade is also widely used in educational institutes and presents a score as a U.S. grade level. For instance, a score of 6 means a sixth grader would be able to read and comprehend the text and a score of 9 means that a ninth grader would be able to read and comprehend the text. FKGL is calculated based on the number of words, number of sentences, and number of syllables in each document. It is worth noting that FKGL is widely used in

[3] The score were calculated by using PHP Text Statistics. The source code and documentations are available at: https://github.com/DaveChild/Text-Statistics.

assessing the readability of health related documents (Cherla et al. 2013; Eltorai et al. 2014). Table 2 reports the summary statistics for readability scores for the entire sample.

Table 2. The Readability Scores for Questions & Answers

Variable	Observations	Mean	Std. Dev.	Min	Max
Question_FKGL	17,824	2.703	3.513	0	12
Answer_FKGL (all answers)	66,888	5.302	3.483	0	12
Answer_FKGL (regular answers)	49,064	4.868	3.421	0	12
Answer_FKGL (best answers)	17,824	6.493	3.73	0	12

For both questions and answers, FKGL ranges from zero (very easy to read) to 12 (the maximum plateau of secondary education). According to table 2, questions have lower FKGL scores when compared to the answers. The average FKGL score for the entire answers is 5.302 which is approximately 0.5 points higher than the regular answer only sample. Best answers (the answers chosen by the asker as the best answer) have the highest FKGL score on average. According to these findings a 7th grader would be able to easily read the best answers provided in Yahoo! Answers.[4]

Textual Quality Measure 2: External Links

Another feature that could convey a high quality answer, particularly in health related topics, is the existence of links to external online resources in the answer. The existence of external links can partly capture the presence of tangible information in the answer (Weimer et al. 2007). The existence of links to other online resources signals that the responder either had a knowledge about the topic prior to answering the question or took time and effort to search for clues while answering the question. Either way, the existence of the links may show that the responder attempted to provide a good answer. Therefore, we employed the presence of external links in the answers as another indicator of textual quality. An answer that includes at least one external link will take the value of one on this measure and zero otherwise. We found that out of 66,888 answers in our sample, 10,974 (16.40%) answers contained at least one external link.

Textual Quality Measure 3: Predicted Professional

To be able to examine the quality of the content of the answers provided in Yahoo! Answers, we took a novel approach. In this approach, we determined which answers were similar to answers provided by a health care professional (doctors and specialists). Since the occupations or affiliations of the respondents is not known in the Yahoo! Answers platform, we borrowed from supervised machine learning and text mining literatures to build a classifier that could predict the probability of an answer being similar to an answer provided by a healthcare professional. To construct such classifier, we needed to have two samples: 1- a collection of documents that includes health-related texts only provided by non-professionals. And 2- a collection of documents that includes health-related texts only provided by healthcare professionals. With these two collections, we could train a classifier that would compare the content of an unknown text with each of the two collections and determine if the unknown text is more similar to professional collection or non-professional collection.

The preparation of the two collections (from now on we call this the training data) is indeed the main challenge of this supervised machine learning method. To prepare the training data, researchers traditionally hire independent raters who rate a subsample of data and provide the desired training data. However, in the context of health-related text, finding raters who can professionally evaluate and rate the content of medical advices could be costly and time-consuming. To overcome this challenge, we prepared the training data by collecting answers that were posted on another health Q&A platform. We extracted 1,874 health related Q&A samples that were posted starting from 11/20/2011 and ending to 3/1/2012

[4] It is worth noting that we initially employed two other readability measures (Gunning Fog Index and Flesch Reading Ease) along with FKGL. Since the results with either of these two measures did not significantly differ from the results derived by FKGL after accounting for scale differences, we dropped these two readability measures from this manuscript.

from askthedoctor.com website. Askthedoctor.com is a free Q&A platform through which users can submit their health related questions. The website claims that a board certified Doctor will respond to the question without charging any fees. The website only has one unique answer for each answered question and covers almost any health related topic. We argue that the answers provided by the Doctors can be used as the health professional collection. To create a sample of non-professional answers, we first extracted all of the Yahoo! Answers health answers that were posted from 11/20/2011 to 3/1/2012. Then we removed all of the answers except those answers that were provided by a user who has posted at least one health related question in Yahoo! Answers. The logic is that while health professionals may answer some questions in Yahoo! Answers, it is highly unlikely for them to ask health related questions in this community question answering platform. Thus, we assume that the users who have posted a question in Yahoo! Answers are not healthcare professionals and therefore we can treat their answers as non-professional answers. To balance the number of non-professional answers with the number of professional answers, we randomly selected a subset of 1,874 answers from the remaining answers and used them as non-professional collection.

As mentioned earlier, we used a text mining approach to train a classifier that would automatically classify an unknown document in professional and non-professional classes. We employed Rapidminer V5 for processing the training data, building the model, and applying the model on Yahoo! Answers data. In the first step, we tokenized the texts in the training data. We also transformed all of the tokens to lower cases and dropped all of the tokens with less than 3 or more than 60 characters. We further removed English stop words such as "are", "was", etc. and then applied Porter's stemming function to transform the tokens to their stems. These pre-processing steps would result in a better training data for building a classifier. We used a Naïve Bayes classifier with Laplace Correction criterion.[5]

To evaluate the overall performance of this classifier we used cross-validation method. In cross-validation approach, the training data is sliced into k folds (partitions). The classifier then uses k-1 partitions for building the model and the remaining partition to evaluate the performance. In the next iteration, another partition will be selected for testing and the remaining partitions will be used for building the classifier. The software repeats this process until each partition is used once for testing. In our example, we used 10 folds. The software reports the mean and the standard deviation of the accuracies achieved in each iteration. The Naïve Bayes classifier had a mean accuracy of 89.63% and a 0.86% standard deviation in our case.[6] Table 3 reports the confusion matrix for this classifier. "Pred. Non-Pro" in Table 3 stands for the texts that were predicted to be dissimilar to professional texts. "True Non-Pro" stands for the texts posted by non-professionals. "Pred. Pro" stands for texts that were predicted to be similar to professional texts, and "True Pro" are the actual texts posted by healthcare professionals.

According to Table 3, the class recall and precision for both classes are above 88% which indicates a high quality classification.

Table 3. Confusion Matrix for Naïve Bayes Classifier

	True Non-Pro	True Pro	Class Precision
Pred. Non-Pro	1661	176	90.42%
Pred. Pro	213	1698	88.85%
Class Recall	88.63%	90.61%	

Figure 1 provides an illustration for the steps taken in processing data and building models.

[5] We also performed the classification task by employing support vector machine (SVM), ANN, random forest, logistic regression, and KNN. Among all of the models, Naïve Bayes had the highest accuracy, class recall, and class precision and therefore was selected for the subsequent step.

[6] It is worth noting that the cross-validation approach is an efficient method for using the training data for both building the model and validating it. Particularly, since the training data is not large and there are only 1,874 records in the training data, the traditional approaches such as using 70% of the records for building the model and 30% of the records for evaluating the model may not be efficient.

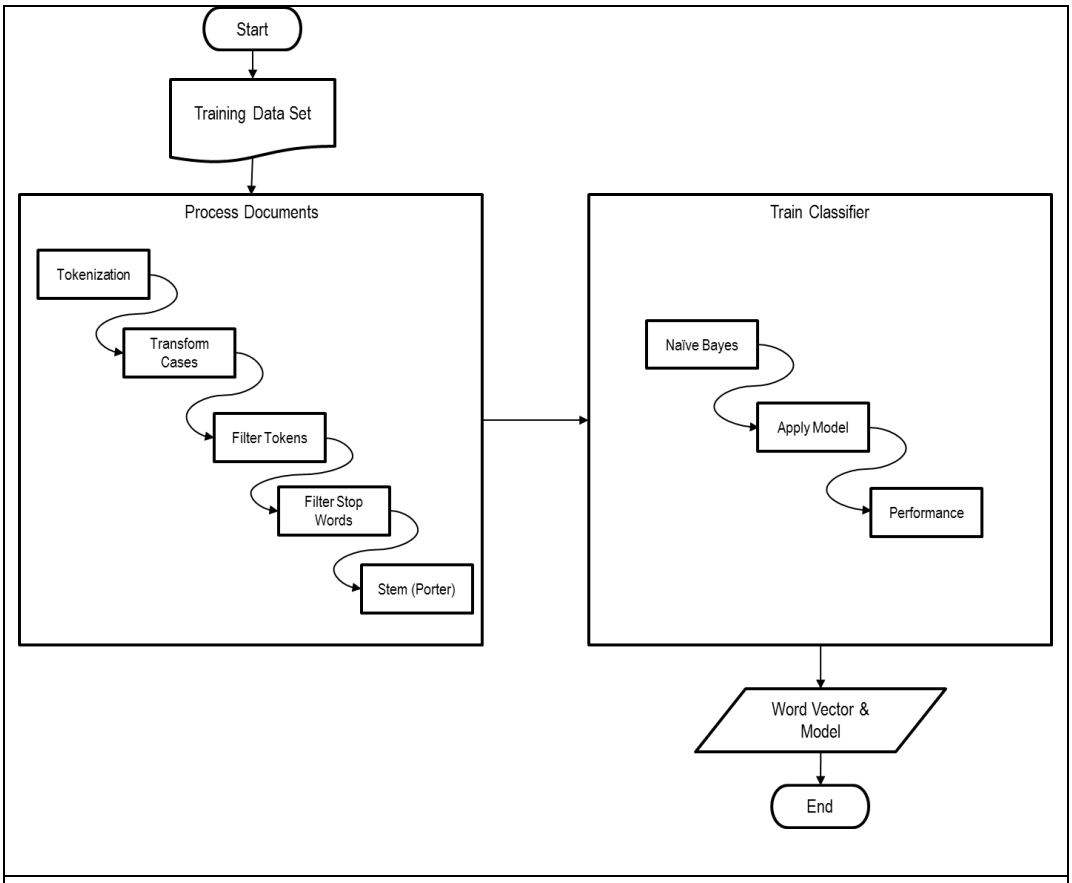

Figure 1. Processing Training Data and Building the Classifier

The final step in the text mining process is to apply the model to the unknown data (in our case the answers in Yahoo! Data). The goal is to determine if a given answer in Yahoo! Answers platform belongs to the professional class or the non-professional class. To apply the model, we performed the exact same pre-processing operations (tokenization, transformation, filtration, and stemming) on the Yahoo! Answers Data. The output of applying the model is a binary class label (professional/ non-professional) for each answer. Out of 66,888 answers, 4,174 answers were predicted to be in the professional class and the rest were predicted to be in the non-professional class. As will be discussed later, we used this binary variable as one of the measures for the textual quality. That is, if the answer belongs to the professional class, it is similar to an answer provided be a healthcare professional and therefore has a high quality. The answer will be considered a low quality answer when it belongs to the non-professional class.

Textual Quality Measure 4: Medical Terms Density

It is fair to assume that a medical text generated by a healthcare professional may have a higher ratio of count of medical terms to count of total terms. Therefore, an answer that resembles a professional answer is likely to have higher medical term frequency to total term frequency ratio. To count the number of medical terms used in each question and answer in Yahoo! Answers data set, we collected 27,455 medical terms from the Medical Subject Headings (MeSH) Thesaurus. This way we can externally evaluate the validity of the classification task by measuring the density of medical terms in the answers. It is worth noting that the "Medical Subject Headings (MeSH®) thesaurus is a controlled vocabulary produced by the National Library of Medicine (NLM) and used for indexing, cataloging, and searching for biomedical and health-related information and documents. 2015 [version of] MeSH includes the subject descriptors appearing in MEDLINE®/PubMed®, the NLM catalog database, and other NLM databases (U.S.

National Library of Medicine 2015)." To find the medical terms matches in Yahoo! Answers data, we needed to check the presence of each medical term in each answer. Since there are 27,455 medical terms and 66,888 answers in our data set, such task could be computationally intensive. Therefore, we decided to employ Apache Spark™ which is a fast and general engine for large-scale data processing.[7] We used Scala programming language to hire a single-node spark cluster to determine the scores for each answer and each question separately. Table 4 reports the summary statistics for the density of the medical terms in Yahoo! Q&A data set.

Table 4. The Density of Medical Terms in Questions and Answers

Variable	Observations	Mean	Std. Dev.	Min	Max
Medical Terms Density in Questions	17,824	0.002	0.015	0	0.4
Medical Terms Density in Answers (all answers)	66,888	0.006	0.040	0	1
Medical Terms Density in Answers (best answers)	17,824	0.007	0.036	0	1
Medical Terms Density in Answers (regular answers)	49,064	0.005	0.041	0	1
Medical Terms Density in Predicted Pro. Answers	4,174	0.008	0.019	0	0.4
Medical Terms Density in Predicted non-Pro. Answers	62,714	0.006	0.042	0	1

According to Table 4, the mean medical terms density for the answers is three times higher than that of the questions. The average medical terms density for the best answers is 33% higher than the average medical terms density of other answers (t-test significant at 0.001). Interestingly, the answers that were predicted to be professional answers had the mean of 0.008, which is 0.002 higher than those that were predicted to be non-professional. In other words, those answers that were predicted to be professional had 29% higher medical term density than did the non-professional answers on average (t-test significant at 0.001). This finding confirms that the answers that were predicted to be professional answers indeed differed from those that were predicted to be non-professional in terms of the use of medical terms.

We employ the density of the medical terms as another measure for quality. A higher medical terms density score signals a higher quality and a lower medical terms density score signals a low quality answer.

Method & Results

Empirical Model

To study the effects of first answer's quality on the subsequent answers quality and quantity, we employed the following empirical model:

$$y_i = \beta_0 + \beta_1 x_i + \sum_{j=1}^{3} \gamma_{ji} z_{ji} + \sum_{k=1}^{21} \varphi_k p_k + \sum_{s=1}^{14} \tau_s \theta_s + \varepsilon_i$$

where y_i is the arithmetic mean of the textual quality of the subsequent answers for question *i*. The textual quality of the answers are measured by four metrics: 1- readability score (FKGL), 2- existence of external links (External Links) 3- similarity to professional answers (Predicted Professional), and 4- density of the medical terms (Medical Terms Density). x_i is the textual quality of the first answer for question i, z_{ji} is a vector of question *i*'s attributes including readability of the question, the number of words in the question, and density of the medical terms in the question, p_k are the dummies for 21 health topics, and θ_s are dummies for each month of the study starting from 7/8/2005 and ending to 9/9/2006. p_k and θ_s would allow us to control for topic-specific and time-specific effects.

We also used the same equation to study the influence of first answer's quality on the total number of subsequent answers. Only this time (Model 5 in Table 8) y_i is the total number of subsequent answers

[7] Please visit https://spark.apache.org/ for more information.

posted after the first answer for question *i*. Since there should be a subsequent answer after the first answer for our model to be valid, we limited our analysis to those Q&As where there were more than one answer posted for the question.

It is worth noting that the dependent variables in Models 2 and 4 are the arithmetic means of binary variables and therefore are bounded to range from 0 to 1. Furthermore in Model 3, Medical Term Density is based on a proportion (i.e. density is calculated by dividing the number of medical terms by the number of words) and is only allowed to range from 0 to 1. Therefore, for these three regressands we used zero-and-one inflated beta regression (ZOIB) specification in Models 2 and 3. As the OLS specification assumes a normal distribution, the linear specification may not work in this setting. According to Kieschnick and McCullough (2003), parametric regression models based on beta distribution are recommended for these data. The ZOIB model consists of three separate regression models: 1- A logistic regression model for whether or not the proportion equals 0, 2- a logistic regression model for whether or not the proportion equals 1, and 3- a beta regression model for the proportions between 0 and 1 (Buis 2010). Since we do not have a specific hypothesis about the upper bound and the lower bound, we only reported the coefficients for the proportions between 0 and 1 (Table 6 Models 2, 3, and 4).

In Model 5, we employed negative binomial regression specification as the dependent variable in this model is a count measure (number of subsequent answers) and follows a negative binomial distribution.

Best Answers VS Other Answers

The first question in this study is whether the answers chosen by the users as the best answers would differ from other answers in terms of textual quality. According to Table 5, best answers had a mean FKGL score 29% higher than regular answers. The second measure of textual quality (External Links) is also significantly different between regular answers and best answers. The third measure of textual quality (Predicted Professional) also reveals significant difference between best answers and regular answers. The average score for best answers is 0.139 and the average score for regular answers is 0.038. Furthermore, best answers have higher medical term density than do regular answers. Two sample t-tests were performed to examine the significance of the differences between regular answers and best answers in terms of textual quality. According to Table 5, best answers have higher textual quality than the regular answers for all measures.

Table 5. Comparison of The Best Answers And The Regular Answers

Variables	Regular Answer (mean)	Best Answer (mean)	Regular Answers VS Best Answers	
			t-value	Sig.
Answer FKGL	4.868	6.493	54.567	<0.001
External Link	0.118	0.289	53.942	<0.001
Predicted Professional	0.038	0.139	46.072	<0.001
Medical Term Density	0.005	0.007	5.033	<0.001

Testing H1: Order Effects on Quality

The second question in our study is to evaluate the impact of the quality of the first answer in the subsequent answers. To study this relationship, we first regressed the mean readability scores (FKGL) of subsequent answers against the readability scores of the first answers after controlling for the readability scores of the questions, question topics, and time trends. Table 6 reports the regression results. As mentioned earlier, we narrowed the sample to those questions that had at least two answers. Imposing this condition returned 9,047 sets of Q&As.

According to Model 1, the readability of the first answer and the readability of the question are positively associated with the mean readability of the subsequent answers. Question's word count is however, negatively associated with the mean readability of the subsequent answers. The density of medical terms in the question is not significant in this model.

According to Model 2, the existence of external links in the first answer is positively associated with the existence of external links in the subsequent answers. To interpret the magnitude of this effect, marginal effects at means (MEMS) are reported. With binary regressor (External Link in First Answer), marginal effects measure the discrete change, i.e. how do predicted probabilities change as the binary regressor changes from 0 to 1? The marginal effect for External Link in First Answer is 0.056. This means that if the first answer contains at least one external link, the Mean number of answers with external links in them increases by 5.6%. Given that only 16.40% of the answers contained an external link, the effect of External Link in First Answer on Mean Number of Subsequent Answers with External Links is a sizable effect. Question's FKGL and medical terms density are also positively associated with the existence of external links in the subsequent answers. Question's word count is however, negatively associated with the existence of external links in the subsequent answers.

Table 6. Effects of First Answer's Quality on Subsequent Answers' Quality

Variable	Model 1 DV=Mean FKGL for Subsequent Answers	Model 2 DV= Mean Number of Subsequent Answers with External Links		Model 3 DV= Mean Medical Terms Density in Subsequent Answers		Model 4 DV= Mean Predicted Professional Answers in Subsequent Answers	
First Answer's FKGL	**0.082*** (0.008)**						
External Link in First Answer		**0.501*** (0.032)**	**0.056*** (0.004)**				
First Answer's Medical Terms Density				**0.719* (0.311)**	**0.004** (0.001)**		
Predicted Professional First Answer						**0.559*** (0.122)**	**0.014*** (0.003)**
Controls:							
Question's FKGL	**0.091*** (0.009)**	**0.081*** (0.003)**	**0.009*** (<0.001)**	**0.054*** (0.004)**	**0.003*** (<0.001)**	**0.093*** (0.006)**	**0.002*** (<0.001)**
Question's Word Count (log)	**-0.141*** (0.039)**	**-0.128*** (0.014)**	**-0.014*** (0.002)**	**-0.121*** (0.018)**	**-0.001*** (<0.001)**	**-0.103*** (0.027)**	**-0.002*** (<0.001)**
Question's Medical Terms Density	2.561 (2.049)	**1.588* (0.682)**	**0.177* (0.076)**	**1.876*** (0.601)**	**0.010*** (0.003)**	0.112 (2.234)	0.003 (0.053)
Observations	9,047	9,047	9,047	9,047	9,047	9,047	9,047
Adj. R-Sq	0.086						
Wald χ^2		1308.83		404.34		463.82	
Specification	OLS	ZOIB		ZOIB		ZOIB	

Note 1: The health topics are controlled for in all models.
Note 2: We employed time dummies to account for time-fixed effects.
Note 3: Huber/White/sandwich estimator of variance was used in all models. The robust standard errors are reported in parentheses.

According to Model 3, the density of the medical terms in the first answer is positively associated with the mean medical terms density in the subsequent answers. The readability of the question (Question's FKGL) and Question's Medical Term Density are both significant and positive in this model. However, the count of words in the question is significant and negative. Marginal effects for continuous variables

measure the instantaneous rate of change. The marginal effect of First Answer's Medical Terms Density equals 0.004. Given that the average Medical Terms Density for all of the answers in our data set is 0.006 (table 4), First Answer's Medical Terms Density has a sizable effect on Mean Medical Terms Density in Subsequent Answers. More than the first answer, Question's Medical Terms Density influences Mean Medical Terms Density in Subsequent Answers as the marginal effects estimate for this variable is 0.010. Since the answers may contain keywords from the questions, it makes sense that Question's Medical Terms Density has a higher impact on Mean Medical Terms Density in Subsequent Answers than does First Answer's Medical terms density. For instance, if the asker uses medical terms such as the name of an allergy or a specific disease, the answers may contain those terms too.

According to Model 4, Predicted Professional First Answer is significant and positive. To interpret the magnitude of this effect, marginal effects at means are reported. The marginal effect for Predicted Professional First Answer is 0.014. This means that if the first answer is a professional answer, the Mean Predicted Professional Answers in Subsequent Answers increases by 1.4%. Given that out of 66,888 answers only 4,174 (6.2%) answers were predicted to be in professional class the effect of Predicted Professional First Answer on Mean Predicted Professional Answers in Subsequent Answers is a sizable effect. The readability of the question (Question's FKGL) has a positive effect on Mean Predicted Professional Answers in Subsequent Answers. Question's Word Count is also negative and significant, yet it has a minimal effect on the dependent variable.

To test for heteroskedasticity, we exploited modified Wald test for heteroskedasticity in regression model. Since we identified the presence of heteroskedasticity, we employed Eicker-Huber-White robust standard errors in all models. We also estimated variance inflation factor (VIF) for all models and did not detect a VIF larger than 2. To account for the heterogeneous popularity of health topics in Yahoo! Answers, we included topic-specific dummies for the topics. To control for changes in all users' propensity to shift the quality of their answers over time, we included dummies for each month from July 2005 to September 2006.

Overall, the results reported in Table 6 support the presence of order effects in HCQA platforms for all four textual quality measures. This finding confirms the first hypothesis that the textual quality of the first answer has a positive influence on the average textual quality of the subsequent answers for a given question in a HCQA platform.

Testing H2: Order Effects on Quantity

To measure the influence of the textual quality of the first answer on the quantity of the subsequent answers, we regressed the quantity against the readability score for the first answer (First Answer's FKGL), a binary variable that takes the value of one if the first answer contains any external link and zero otherwise (External Link in First Answer), the density of the medical terms in the first answer (First Answer's Medical Terms Density), and a binary variable that takes the value of one if the first answer was predicted to be similar to professional answers by our classifier and zero otherwise (Predicted Professional First Answer). Table 7 reports the results of the regressions.

According to the results of the binomial regression models, the readability of the first answer and existence of external links in the first answer have negative associations with the number of answers posted. However, the use of medical terms in the first answer does not influence the total number of subsequent answers. An important finding in Table 7 is the negative relationship between Predicted Professional First Answer and the number of answers posted. To interpret the magnitude of this effect, we estimated the marginal effects of this regressor by keeping all of the other regressors at their means. The marginal effect estimate revealed that if the first answer is similar to a professional answer, there will be almost 1 fewer subsequent answer than in the case where the first answer is not similar to a professional answer. Another interesting finding according to Model 5 is the negative effect of Question's FKGL on the total number of subsequent answers.

To test for heteroskedasticity, we exploited modified Wald test for heteroskedasticity in regression model. Similar to Models 1 through 4, we identified the presence of heteroskedasticity in Model 5. Therefore, we employed Eicker-Huber-White robust standard errors. We also estimated variance inflation factor (VIF) for Model 5 and did not detect a VIF larger than 2. Similar to the previous models, we included topic-

specific dummies for the topics to account for the heterogeneous popularity of health topics in Yahoo! Answers. To control for changes in all users' propensity to shift the quantity of their answers over time, we included dummies for each month from July 2005 to September 2006.

Table 7. Effects of First Answer's Quality on Subsequent Answers' Quantity

Variables:	Model 5: DV= Number of Subsequent Answers			
First Answer's FKGL	**-0.011*** (0.002)**			
External Link in First Answer		**-0.586*** (0.042)**		
First Answer's Medical Terms Density			0.206 (0.170)	
Predicted Professional First Answer				**-0.314*** (0.027)**
Controls:				
Question's FKGL	**-0.029*** (0.002)**	**-0.178*** (0.003)**	**-0.031*** (0.002)**	**-0.030*** (0.002)**
Question's Word Count (log)	0.007 (0.010)	0.173*** (0.014)	0.010 (0.010)	0.008 (0.010)
Question's Medical Terms Density	-0.690 (0.358)	0.689 (0.654)	-0.769 (0.366)	-0.629 (0.360)
Observations	9,047	9,047	9,047	9,047
Pseudo R-Sq.	0.294	0.079	0.293	0.297
Wald χ^2	2192.91	4401.95	2336.18	2329.84
Specification	Negative Binomial			

Note 1: The health topics are controlled for in all models.
Note 2: We employed time dummies to account for time-fixed effects.
Note 3: Huber/White/sandwich estimator of variance was used in all models. The robust standard errors are reported in parentheses.

Overall, the results reported in Table 7 partly confirm the second hypothesis that *the textual quality of the first answer has a negative influence on the quantity of the subsequent answers for a given question in a HCQA platform with closed-form policy*. The only insignificant quality measure was First Answer's Medical Terms Density.

Discussion

Measuring the Textual Quality of Answers

The quality of health-related content on online websites has been an important topic in recent years. Perhaps the pioneering study in this domain was conducted by Culver et al. (1997). They performed a study in which they examined the medical information provided in a health-related online discussion group, in terms of the professional status of the individuals providing information, the consistency of the information with standard medical practice, and the nature of the evidence cited in support of specific claims or recommendations. Their findings suggested that medical information available on Internet discussion groups may come from non-professionals, may be unconventional, may be based on limited evidence, and/or may be inappropriate. In a more recent study, Chung et al. (2012) assessed the accuracy of the infant sleep safety recommendations provided on online websites. Chung and colleagues found that 43.5% of the 1300 reviewed websites provided accurate information about infant sleep safety, while almost 28% of the websites provided inaccurate information. Fairly similar results were reported by Scullard et al. (2010). In this study, Scullard and his colleagues assessed the reliability and accuracy of

medical advices that are returned by Google search engine. After analyzing the content of 500 websites that provided advice about children's health, Scullard and colleagues concluded that 49% of the websites failed to answer the question, 11% provided incorrect information, and 39% provided accurate and reliable advice. Government websites turned to be the most accurate sources of information for health-related questions. In a similar study, Quinn et al. (2012) examined the accuracy of information about breast cancer by analyzing the content of 500 websites. They found that 42% of the websites provided information that was inapplicable to the question asked. However, the accuracy of the applicable suggestions hinged around 80%.

Perhaps the most relevant study is a 2012 study by Oh et al. (2012). They assessed the quality of online health answers in Yahoo! Answers by relying on the ratings of three groups of participants: librarians, nurses, and users of Yahoo! Answers. Their findings indicate that there was a significant difference between those two expert groups (librarians and nurses) and users. Librarians and nurses rated the quality of answers significantly lower on most of the evaluation criteria than did the users (Oh et al., 2012). Regarding the accuracy of the answers provided, nurses and librarians gave an overall rating of 2.6 to the answers, while the users gave a 3.6 rating.

In a similar study by Bowler et al. (2013), the researchers investigated answers to 81 informational questions about eating disorders posted in Yahoo! Answers. Through a content analysis, they found that users do not always respond to eating disorder questions with credible, factual information, even if the need for it is expressed in the question. The findings suggest that people who post questions in Yahoo! Answers use it as a social and emotional scaffold rather than an informational source, even if their questions are couched in terms that suggest they are seeking information. Furthermore, a large portion of the people who answer such questions understand this to be the purpose and rarely provide answers drawn from evidence-based medicine or reliable, credible sources for health information.

According to these studies, textual quality metrics that are not reliant on the subjective judgments of non-professional users are useful in helping online diagnosers to find reliable resources. To respond to this demand, we borrowed from the machine learning and text mining literatures to build an accurate classifier that measures the similarity between health advice given on Yahoo! Answers and health advice given by professionals. According to our results, only a little above 6% of all of the answers provided in Yahoo! Answers were similar to professional answers.

We used the density of the medical terms in the answers as another indicator of textual quality. For each answer, we counted the number of times any of the 27,455 medical terms from the Medical Subject Headings (MeSH) Thesaurus were mentioned in the answers. Our findings suggest that oftentimes 0.6% of the words used in Yahoo! Answers are medical terms. Moreover, we measured the readability of the answers by using the Flesch-Kincaid Grade Level (FKGL) method. According to our results, an average 5th grader would be able to read the majority of health-related content posted on Yahoo! Answers. Last but not least, we found that only a little above 16% of the answers in Yahoo! Answers contained an external link to online resources.

Four measures of textual quality were used to compare the textual quality of best answers to that of regular answers. For all four measures, the best answers chosen by the askers were of higher quality than the other answers.

Order Effects

According to the survey design literature, the order of the questions in a survey or the order of the alternatives in response to survey questions may influence the quality of responses. The two types of order effects are identified as primacy effects and recency effects. The primacy effect is often present where a visual format is used to present the questions or alternatives. The recency effect is usually present when a verbal format is used for surveying. Since HCQA platforms present the questions and answers in visual form, it is expected that the primacy effect would be present. That is, the first answer in the list serves as an anchor for the subsequent answers. Therefore, we argue that a high-quality first answer may attract high-quality subsequent answers.

To study this effect, we measured the textual quality of the first answers and the mean textual quality of the subsequent answers by using the four measures. Our results confirmed the presence of the primacy effect, as all four measures of textual quality indicated that a high-quality first answer will attract high-

quality subsequent answers, even after controlling for question-specific, topic-specific, and time-specific factors. There are two explanations for the primacy effect. One, an alternative that is presented as the first option may establish a cognitive framework or standard for comparison that guides the interpretation of later items. Two, an alternative that is presented first is subjected to a deeper cognitive processing and therefore might occupy the respondent's mind even when s/he is reading the subsequent alternatives. In other words, the first alternatives "suffer less competition for time and space in immediate memory from other items" (Schwarz et al. 1992, p. 188). Due to this latter effect, the respondent is less stimulated by the subsequent answers and may select them less frequently (Klayman and Ha 1987; Krosnick and Alwin 1987; Tversky and Kahneman 1981).

We further tested the effect of the first answer's textual quality on the total number of subsequent answers posted in response to the question. We proposed that in two ways the number of subsequent answers would diminish as a consequence of a high-quality first answer. First, if the first answer is a high-quality answer, the asker could select it as the best answer and close the thread. According to Table 5, the quality of the best answers is significantly higher than the quality of the regular answers for all four measures of textual quality. This finding indicates that askers' evaluations of the quality of a given answer are aligned with the four measures of textual quality. Therefore, the asker is likely capable of choosing the same answer that we predicted as a high-quality answer as the best answer. If this answer is the first answer, then it will be selected earlier than when the best answer is posted later on and therefore the thread will be closed and will not receive further answers. Second, as discussed in H1, high-quality first answers may attract high-quality subsequent answers. Therefore, even if the asker cannot detect the first high-quality answer as the best answer, there is a high chance that s/he will select one of the high-quality subsequent answers and close the thread. Either way, the textual quality of the first answer could influence the quantity of the subsequent answers. Our results confirmed this proposition, except where the textual quality was measured by medical terms density.

Conclusion & Implications

Measuring the quality of health content in health-related community-based question answering (HCQA) platforms has been a challenging task that has attracted many scholars from a variety of disciplines. Previous studies employed a variety of cues to evaluate the quality of the answers. However, the majority of these factors were based on non-professional evaluations of HCQA platform users. In rare studies where professional judgment was used to evaluate the quality of answers in HCQA platforms, the procedure was not scalable to be employed for automatic evaluation of answers' quality. To respond to this gap, we borrowed from machine learning and text mining literature to construct a classifier to automatically detect the similarity of a given answer to a professional answer. We also employed Apache Spark platform to develop another measure of textual quality, the density of the medical terms, which is defined by the ratio of the count of medical terms in text to the total count of terms (words). We also used a traditional readability score and the presence of external links as other indicators of textual quality.

The four measures of textual quality were first employed to compare the quality of best answers chosen by the askers with the quality of the regular answers. All four measures of textual quality revealed a higher quality for best answers, indicating that askers tend to have proper judgment to select the best answers. We also studied the presence of order effects in HCQA platforms. Our results suggest that the textual quality of the first answer positively influences the mean textual quality of subsequent answers and negatively influences the quantity of subsequent answers.

The textual quality measures introduced in this study can be employed by HCQA platforms to automatically evaluate the quality of the answers and the questions right after they are posted. HCQA platforms then could filter out questions with a quality lower than a certain threshold and could present the predicted highest-quality answer as the first answer for the questions that survived the previous step. This would improve the quality of the subsequent answers, according to our findings. This practice could diminish the number of answers received, particularly if the HCQA platforms impose a closed-form policy (which closes the thread after the best answer is selected by the asker). Presenting the best answer as the best answer would typically lead to higher-quality answers and therefore may attract more online diagnosers to consume the resources provided in HCQA platforms, yet would decrease the number of answers posted. Therefore, the overall effect of this practice on user traffic is not clear.

With respect to the utility of online diagnosers, presentation of the highest-quality answer as the best answer would be desirable. Online diagnosers would have access to high-quality answers without spending time reading numerous answers by browsing the pages up and down. That is, for them, there are only a few high-quality answers to be read. This way, the online diagnosers would find a better answer in a shorter period of time.

Acknowledgement

This research was supported in part through the Robert D. and Patricia E. Kern Center for the Science of Healthcare Delivery.

References

Agichtein, E., Castillo, C., Donato, D., Gionis, A., and Mishne, G. 2008. "Finding High-Quality Content in Social Media," in *Proceedings of the International Conference on Web Search and Web Data Mining - WSDM '08*, New York, New York, USA: ACM Press, February 11, p. 183.

Bowler, L., Mattern, E., Jeng, W., Oh, J. S., and He, D. 2013. "I Know What You Are Going Through: Answers to Informational Questions about Eating Disorders in Yahoo! Answers: A Qualitative Study," *76th Association for Information Science and Technology Annual Meeting (ASIS&T 2013)*, Montreal, Quebec, Canada, 1-5 November 2013, pp. 1-9.

Buis, M. 2010. Analyzing Proportions http://www.stata.com/, Berlin-Mitte: http://www.stata.com/ (available at http://www.stata.com/meeting/germany10/germany10_buis.pdf).

Cherla, D. V, Sanghvi, S., Choudhry, O. J., Jyung, R. W., Eloy, J. A., and Liu, J. K. 2013. "Readability Assessment of Internet-based Patient Education Materials Related to Acoustic Neuromas," *Otology & Neurotology* (34:7), pp. 1349–1354.

Chung, M., Oden, R. P., Joyner, B. L., Sims, A., and Moon, R. Y. 2012. "Safe Infant Sleep Recommendations on the Internet: Let's Google It.," *The Journal of pediatrics* (161:6), pp. 1080–1084.

Culver, J. D., Gerr, F., and Frumkin, H. 1997. "Medical Information on the Internet," *Journal of General Internal Medicine* (12:8), pp. 466–470.

Eltorai, A. E. M., Ghanian, S., Adams, C. A., Born, C. T., and Daniels, A. H. 2014. "Readability of Patient Education Materials on the American Association for Surgery of Trauma Website," *Arch Trauma Res.* (3:2), p. e18161.

Fallis, D., and Frické, M. 2002. "Indicators of Accuracy of Consumer Health Information on the Internet: A Study of Indicators Relating to Information for Managing Fever in Children in the Home," Journal of the American Medical Informatics Association: JAMIA (9:1), pp. 73–79.

Fox, S., and Duggan, M. 2013. "Health Online 2013," Washington, D.C., p. 55 (available at http://www.pewinternet.org/files/old-media//Files/Reports/PIP_HealthOnline.pdf).

Galesic, M., Tourangeau, R., Couper, M. P., and Conrad, F. G. 2008. "Eye-Tracking Data: New Insights on Response Order Effects and Other Cognitive Shortcuts in Survey Responding," *Public Opinion Quarterly* (72:5), pp. 892–913.

Ghose, A., and Ipeirotis, P. G. 2011. "Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics," *IEEE Transactions on Knowledge and Data Engineering* (23:10), IEEE, pp. 1498–1512.

Hogarth, R. M., and Einhorn, H. J. 1992. "Order Effects in Belief Updating: The Belief-adjustment Model," *Cognitive Psychology* (24:1), pp. 1–55.

Holbrook, A. L., Krosnick, J. A., Moore, D., and Tourangeau, R. 2007. "Response Order Effects in Dichotomous Categorical Questions Presented Orally: The Impact of Question and Respondent Attributes," *Public Opinion Quarterly* (71:3), pp. 325–348.

Hu, H., Liu, B., Wang, B., Liu, M., and Wang, X. 2013. "Multimodal DBN for Predicting High-Quality Answers in cQA portals," in *Annual Meeting of the Association of Computational Linguistics*, Sofia, Bulgaria, August 4-9 2013, pp. 843–847.

Jasra, M. 2008. "Yahoo Answers Down but Dominating Q&A Websites," web Analytics World (available at http://www.webanalyticsworld.net/2008/03/yahoo-answers-dominates-q-websites.html).

Jeon, J., Croft, W. B., Lee, J. H., and Park, S. 2006. "A Framework to Predict the Quality of Answers with Non-Textual Features," in *Proceedings of the 29th annual international ACM SIGIR conference on*

Research and development in information retrieval - SIGIR '06, New York, New York, USA: ACM Press, August 6, pp. 228–235.

Kieschnick, R., and McCullough, B. D. 2003. "Regression Analysis of Variates Observed On (0, 1): Percentages, Proportions and Fractions," *Statistical Modelling* (3:3), pp. 193–213.

Klayman, J., and Ha, Y.-W. 1987. "Confirmation, Disconfirmation, and Information in Hypothesis Testing," *Psychological Review* (94:2), pp. 211–228.

Krosnick, J. A., and Alwin, D. F. 1987. "An Evaluation of a Cognitive Theory of Response-Order Effects in Survey Measurement," *Public Opinion Quarterly* (51:2), pp. 201-219.

Lavrakas, P. 2008. "Encyclopedia of Survey Research Methods," *SAGE Publications, Inc*, p. 1072.

Liu, M., Liu, Y., and Yang, Q. 2010. "Predicting Best Answerers for New Questions in Community Question Answering," *Springer-Verlag*, pp. 127–138.

McFarland, S. G. 1981. "Effects of Question Order on Survey Responses," *Public Opinion Quarterly* (45:2), pp. 208-215.

Oh, S., Yi, Y. J., and Worrall, A. 2012. "Quality of Health Answers in Social Q&A," *Proceedings of the American Society for Information Science and Technology* (49:1), pp. 1–6.

Otterbacher, J. 2009. "'Helpfulness' in Online Communities," in *Proceedings of the 27th international conference on Human factors in computing systems* - CHI 09, New York, New York, USA: ACM Press, April 4, p. 955-964.

Pew Research Center. 2014. "Health Fact Sheet," (available at http://www.pewinternet.org/fact-sheets/health-fact-sheet/).

Quinn, E. M., Corrigan, M. A., McHugh, S. M., Murphy, D., O'Mullane, J., Hill, A. D. K., and Redmond, H. P. 2012. "Breast Cancer Information On The Internet: Analysis of Accessibility and Accuracy," *Breast* (Edinburgh, Scotland) (21:4), pp. 514–517.

Schwarz, N., Hippler, H.-J., and Noelle-Neumann, E. 1992. "A Cognitive Model of Response-Order Effects in Survey Measurement," in Context Effects in Social and Psychological ResearchN. Schwarz and S. Sudman (eds.), New York, NY: Springer New York, p. 353.).

Scullard, P., Peacock, C., and Davies, P. 2010. "Googling Children's Health: Reliability of Medical Advice on the Internet," *Archives of Disease in Childhood* (95:8), pp. 580–592.

Shah, C., and Pomerantz, J. 2010. "Evaluating and Predicting Answer Quality in Community QA," in *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR '10*, New York, New York, USA: ACM Press, July 19, pp. 411–418.

Tversky, A., and Kahneman, D. 1981. "The Framing of Decisions and the Psychology of Choice," *Science* (211:4481), pp. 453–458.

U.S. National Library of Medicine. 2015. "Medical Subject Headings," (available at http://www.nlm.nih.gov/mesh/intro_preface.html#pref_rem).

Wang, F.-Y., Carley, K. M., Zeng, D., and Mao, W. 2007. "Social Computing: From Social Informatics to Social Intelligence," *IEEE Intelligent Systems* (22:2), pp. 79–83.

Weimer, M., Gurevych, I., and Mühlhäuser, M. 2007. "Automatically Assessing the Post Quality in Online Discussions on Software," in ACL '07 Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions, Stroudsburg, PA: Association for Computational Linguistics, June 25, pp. 125–128.

Yao, Y., Tong, H., Xie, T., Akoglu, L., Xu, F., and Lu, J. 2015. "Detecting High-Quality Posts in Community Question Answering Sites," *Information Sciences* (302), Forthcoming, pp. 70–82.