Analyzing the Quality of Information Solicited from Targeted Strangers on Social Media

Jeffrey Nichols[*], Michelle X. Zhou[*], Huahai Yang[*], Jeon Hyung Kang[†], Xiaohua Sun[‡]

[*]IBM Research – Almaden
650 Harry Road
San Jose, CA 95120
{jwnichols,mzhou,hyang}@us.ibm.com

[†]USC ISI
4676 Admiralty Way
Marina del Rey, CA 90292
jeonhyuk@usc.edu

[‡]Tongji University
Shanghai, China
xsun@tongji.edu.cn

ABSTRACT

The emergence of social media creates a unique opportunity for developing a new class of crowd-powered information collection systems. Such systems actively identify potential users based on their public social media posts and solicit them directly for information. While studies have shown that users will respond to solicitations in a few domains, there is little analysis of the quality of information received. Here we explore the quality of information solicited from Twitter users in the domain of product reviews, specifically reviews for a popular tablet computer and L.A.-based food trucks. Our results show that the majority of responses to our questions (>70%) contained relevant information and often provided additional details (>37%) beyond the topic of the question. We compare the solicited Twitter reviews to other user-generated reviews from Amazon and Yelp, and found that the Twitter answers provided similar information when controlling for the questions asked. Our results also reveal limitations of this new information collection method, including its suitability in certain domains and potential technical barriers to its implementation. Our work provides strong evidence for the potential of this new class of information collection systems and design implications for their future use.

Author Keywords

Social Q&A; crowdsourcing; Twitter; product reviews

ACM Classification Keywords

H.5.2

General Terms

Algorithms; Experimentation; Human Factors

INTRODUCTION

Hundreds of millions of people express themselves every day on public social media, such as Twitter. This creates a unique opportunity for building a brand new class of crowd-powered information collection systems, which *actively* solicit information from the right people at the right time based on their public social media posts. For example, if a person just tweeted about getting a sandwich from a food truck, such a system can ask her to provide additional details about her experience. This approach offers several advantages over other crowd-powered information collection systems, such as social Q&A [3]. First, it can collect information about an event, such as a robbery, soon after that event occurred. Second, it can collect information from people who are most likely to share their "visceral reaction" to an event [4]. Third, it can also collect information from a range of people across a particular dimension of a population (e.g., liberal vs. conservative).

Although this approach and its feasibility have been demonstrated in a few domains [1, 12], there are still many unknowns about the approach that need to be explored. One unknown is the level of information *quality* obtainable through this collection method. While there is abundant research effort on studying the quality of crowd-sourced information, especially in the form of social Q&A systems [5, 8, 9, 10, 16, 17], there are differences in this new approach that may influence the outcomes:

- Answers are *actively* solicited from strangers, who have not opted-in *a priori* and most likely have no social or organizational ties to the question asker. This may causes strangers to be less likely to respond, but it may also cause the responses that are received to be more objective and balanced, because the information providers are not self-selected and may have fewer intrinsic motivations for providing information [4].
- Information exchange occurs mainly between the asker and answerer, without any moderation by a larger group. This removes the potential reputation and filtering benefits of typical Social Q&A sites, like Quora, to govern the quality of crowd-sourced information.
- Potential information providers are chosen based on their social media content, which may be misleading about their true ability to provide quality information.

To rigorously assess the quality of information collected using this approach, we have designed and conducted a set of experiments that focus on two aspects. First, we focus on analyzing the quality of crowd-sourced information collect-

CSCW '13, February 23–27, 2013, San Antonio, Texas, USA.

ed from targeted strangers on Twitter. Second, we focus on collecting reviews about "experienced goods"— products or services of which characteristics are difficult to observe in advance but can be learned after consumption [11]. By focusing in these directions, we contain the scope of our study and focus on collecting data that can rigorously and externally validated. A consequence of this choice is our experimental scenario does not explore the true potential of such information collection systems, which is to solicit time-sensitive information from a diverse population. Such information is unfortunately more subjective and difficult to validate externally.

We collect reviews about two "goods": a popular tablet computer and Los Angeles-area food trucks. We chose to focus on the product reviews domain for two reasons. First, products and services are a frequent topic of conversations among Twitter users, thus making it easier for us to find product owners or users of a service from whom we can solicit information. Second, other sources of product and service reviews are available, such as Amazon and Yelp, which can provide external validation for the reviews collected from Twitter. We chose these two particular goods because both inspired a reasonable volume of messages on Twitter but are fairly different from each other.

This paper presents our study design and results to answer three sets of questions:

1. Will strangers respond to our questions with relevant responses? What types of responses will we receive and what information will they contain? If we do not receive a relevant response, what might be the cause?
2. How does the quantity of information collected on Twitter compare with reviews on Amazon or Yelp when controlling for the specific questions asked?
3. How do people perceive the quality of information collected on Twitter versus that of Amazon or Yelp? Which reviews do people find more useful, objective, balanced, and trustworthy?

To address the first set of questions, we designed and conducted a live Q&A experiment on Twitter by manually identifying Twitter users who appeared to own the popular tablet computer or had recently visited one of the food trucks. Each of these users was asked to answer at least one question about the product or service they had used. The results demonstrate the feasibility of our approach: overall we received 369 responses (37.7% overall response rate) from strangers on Twitter, and 76.5% of the responses contained relevant answers.

To answer the second and third sets of questions, we designed and conducted a mixed-method study that assessed both quantity and quality of our Twitter answers against reviews collected from Amazon and Yelp, both objectively and subjectively. The comparisons show that the Twitter answers provided similar information when controlling for the questions asked. More importantly, the results reveal the advantages and disadvantages of our approach against existing crowd-powered systems, like Amazon and Yelp. Based on this finding, we discuss important design implications for building a new class of smart, crowd-powered information systems that leverage public social media.

RELATED WORK

Our work is related to two main research efforts: Social Q&A and studies on the answer quality of Social Q&A.

Social Q&A Systems

There are several different categories of Social Q&A systems. One type is web-based community Q&A (CQA) sites, such as Yahoo! Answers[1], Answers.com[2], StackOverflow[3], and Quora[4]. At these sites, askers post their questions and self-selected volunteers provide answers to the questions. Another type is instant messaging-based Q&A services, including Aardvark [7] and IM-an-Expert [17]. Such services route questions in real-time to users with matching, but often self-described, expertise. Compared to these efforts, our approach focuses on *actively* identifying and engaging suitable answerers based on estimates of their expertise gleaned from their public social media posts.

A third type of Q&A takes place on social networks, such as Twitter and Facebook, where users broadcast questions to their own social networks through their status update messages. These questions are likely to be seen only by users' Facebook friends or Twitter followers, many of whom may not know the answer, but this method may be preferable to some because users tend to trust the opinions of people they know more than those of strangers [10, 14]. To leverage a larger crowd, a few Q&A services on social networks also involve strangers in the answering process, such as TweetQA[5] and AskOnTwitter[6]. While these services allow users to find questions that have been answered in others' social networks, they do *not* allow the users to ask questions of strangers directly.

Closer to our work, there are systems that are targeting strangers to answer questions based on their social media content. For example, Moboq[7] utilizes Sina Weibo (a China-based service similar to Twitter) to identify and send location-based questions to strangers. Bulut et al. [1] crowd-sourced answers to location-based inquiries from users identified to be in a particular location either by their Twitter profiles or by their posts made automatically to Twitter from Foursquare. Nichols et al. explored asking questions of strangers on Twitter in two domains [12]: airport security wait time tracking and digital camera product

[1] http://answers.yahoo.com/
[2] http://answers.com/
[3] http://www.stackoverflow.com/
[4] http://www.quora.com/
[5] http://www.tweetqa.com/
[6] http://www.askontwitter.com/
[7] http://www.moboq.com/

reviews. Compared to these efforts, where the analysis of answer quality is limited, our focus in this paper is to rigorously examine answer quality under different conditions.

Answer Quality in Social Q&A

Harper et al. studied the quality of various Q&A services, and found that services that collected answers from *a group* outperformed services that routed questions to "expert" individuals [5]. Jeon et al. reanalyzed Harper et al.'s data specifically for services, such as Google Answers, where the asker could offer a reward for answering their question and found that while response time was affected by the reward, there was no effect on quality [9]. Hsieh et al. confirmed this result using a different data set from a different paid Q&A site, Mahalo Answers [8]. Our work builds on the notion that quality answers can be sourced for non-expert unpaid groups, and examines these findings in the context of answerers recruited from public social media.

Zhu et al. propose a set of 13 questions that can be used to evaluate the quality of an answer [18], and Shah et al. conducted an experiment using workers on Mechanical Turk to annotate a Q&A corpus from Yahoo! Answers with these 13 questions [16]. Shah et al. found that these answers strongly correlated with asker's ratings of the answer in the corpus. We also used a survey approach with Mechanical Turk workers to measure the quality of our Twitter responses. While our questions differ from those of Zhu et al., most of the same concepts, such as truthfulness, are present in our questions.

Shah et al. also experimented with automatically predicting quality ratings of answers using features such as answer length, number of answers, and properties of answerers' profiles [16]. Unfortunately, many of these features are sparse or unavailable when analyzing strangers on public social media, and thus we could not make use of these models in our work.

There are also two studies that have looked at answer quality of Q&A on social networks. Paul et al. conducted an in-depth study of questions asked on Twitter, examining the response rate and relevance of the responses [13]. Panovich et al. studied tie strength and the quality of responses received from friends to questions posted as Facebook status updates [14]. The results show that answers from people with stronger ties provide slightly more informational answers than those from people with weaker ties. Unlike our approach that utilizes strangers to answer questions, both of these studies focus on Q&A scenarios that revolve around users' social networks. Their results may serve as interesting point of comparison however.

DATA COLLECTION

For the purpose of our experiments, we collected data for two products: one for a popular tablet computer at the time we asked questions (Samsung Galaxy 10.1) and another for food trucks based in the Los Angeles area. For each product, we needed to determine both the set of questions to ask and our method for finding the potential answerers using the public Twitter stream.

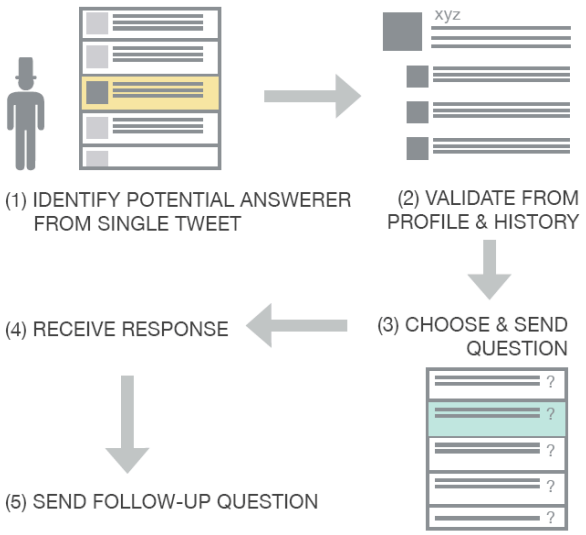

Figure 1. Question asking process flow used in our studies.

Twitter Data Collection

To collect responses from strangers on Twitter, we developed an experimental system where a human operator monitors the Twitter stream, identifies suitable strangers on Twitter, and then sends questions to them. The same human operator asked all of the questions for the studies described in this paper.

Experimental System Overview

Figure 1 shows the main question-asking flow used for collecting reviews for both products. The first step for the human operator was to identify potential answerers by watching a filtered feed of publicly posted Twitter status updates that was updated in real time. Filtering was primarily done by keywords related to the scenario, though spam filters and other scenario-specific heuristics were applied to eliminate as many irrelevant tweets as possible. The tweets of users who had previously been asked a question were also filtered to prevent the operator from accidentally engaging with the same user more than once. When the human operator identified a potential answerer, he would examine the user's profile and recent tweets, verify that user was a good candidate, and then send the user a question. Questions were sent as @replies to one of the answerers' recent relevant tweets, thus giving the answerers some context for why that question was directed to them.

The operator would also monitor Twitter for responses to questions. Up to two questions were asked of each answerer, and if the answerer responded to the first question then a second question would be sent. The operator could send custom responses to users who asked follow-up questions back to the question-asking account. Note that not

Table 1. Questions asked for the Tablet. The 1st question was prefaced with "Trying to learn about tablets...sounds like you have Galaxy Tab 10.1." The 2nd question was prefaced with "Thanks!". The "Rnd" column indicates in which round that question was used.

Q#	Rnd	Question Text
1a	*1*	How fast is it? Does it ever noticeably lag?
1b	*2*	How fast is it?
2a	*1*	How does it feel? Does it seem solid?
2b	*2*	How does it feel?
3a	*1*	How is the display? Readable in sun or from angle?
3b	*2*	How is the display?
4a	*1*	How is the camera quality? (both front and back)
4b	*2*	How is the camera quality?
5a	*1*	How are the speakers? Is the sound quality acceptable?
5b	*2*	How are the speakers?
6a	*1*	How is the battery life? How long does it take to charge?
6b	*2*	How is the battery life?
7a	*1*	Where do you use yours most? (home, work...)
7b	*2*	Where do you use yours most?
8a	*1*	Who would you recommend it for? (techie, novice...)
8b	*2*	Who would you recommend it for?
9	*2*	How easy is it to carry around?
10	*2*	Are you finding all of the apps that you want?
11	*2*	Is it easy to personalize the device for your use?
12	*2*	What do you use it for the most?
13	*2*	How easy is it to connect to other devices?

every user was sent a follow-up question, particularly in cases where they responded that they did not own the tablet or had not visited the food truck, or if their response came in long after the initial question was asked.

For both products, a set of questions was chosen in advance. In general, the next question to ask was chosen randomly by the operator with a weight towards questions that had so far received the fewest responses. A *question-tracking* view in the dashboard let the operator track how many times each question had been answered. Our goal was to collect roughly the same number of responses to each question. All tweets received and sent by our system were collected in a database for later analysis.

Collecting Tablet Reviews

We chose a consumer electronics product for our study because there is a great deal of discussion around such products on Twitter. Tablets seemed like a particular good choice because they were popular products at the time, and the particular device that we chose (Samsung Galaxy Tab 10.1) was among the most popular in the tablet category at the time. One negative to choosing this particular device, which we did not anticipate until after we began, is that Samsung has a line of Galaxy-branded products that also include phones. This made the task of identifying potential answerers more challenging, and may be in part responsible for the resulting lower response rates compared to previous work [12] and the food truck scenario (see Table 3).

We could have chosen the Apple iPad instead of the Galaxy Tab, but at the time spam about the iPad was quite common, which might have had two negative effects. First, it would have made it more difficult to find appropriate answerers due to the high volume of spam tweets about iPads. Second, users might have been more likely to disregard our question as a spam because they might have already received many spam messages about iPads.

We conducted two separate rounds of questioning for the Galaxy Tab, and the questions we chose are shown in Table 3. For the first round, we chose questions based on reading expert tablet reviews from Cnet.com and Engadget.com, and our own intuitive notions of what might be important to a purchaser of a tablet. The questions we wrote were all composed as a general question followed by some form of clarification, because we felt such questions would be easier for users to answer. After sending these questions to a number of users, we decided to add a second round of questions for two reasons. First, we wanted to ensure that our questions covered all of the major types of information commonly discussed in Amazon reviews to facilitate the content comparison between the Twitter and Amazon reviews. For this purpose, we analyzed the top-50 Amazon reviews for the Galaxy Tab 10.1 to identify the top-10 major aspects of information revealed about the tablet. Based on our findings, we added questions #9-13. Second, we removed the clarification phrases from our 1st round of questions to explore whether and how answer behavior would change if more open-ended versions were used.

For the first round of questioning, the initial question sent to a user was chosen from questions #1-6, and the follow-up question was either #7 or #8. We chose this approach because we felt the latter questions were more open-ended and might be too difficult to ask initially. For the second round of questioning, we removed this restriction and the operator randomly selected questions as described earlier.

All questions for this scenario were sent from the Twitter account **@tabletsqa**. The first round of questions took place on September 11-14, 2011, and the second round took place between October 5, 2011 and January 5, 2012.

Collecting Food Truck Reviews

We chose food truck reviews for our second scenario because of the increasing popularity of gourmet food trucks and the use of Twitter by many food truck owners to broadcast their trucks' locations and converse with their customers. Many users when talking about a food truck on Twitter will use its Twitter handle instead of its proper name (e.g., "grlldcheesetruk" instead of "Grilled Cheese Truck"), which makes filtering for conversations about specific trucks much easier. Los Angeles has a large vibrant food truck community, and we thus focused on trucks specifically in this area. We collected a list of 90 active L.A.-based trucks' Twitter handles by browsing web pages and Twitter lists dedicated to the topic.

In order to maximize the number of answers that could be collected, we did not filter the list of 90 trucks. Due to the

Table 2. Questions asked for the Food Trucks. The 1st question was prefaced with "Interested in <food truck handle>...sounds like you've eaten there." The 2nd one had no preface.

Q#	Question Text
1	What do you prefer to order?
2	Are there vegetarian options?
3	Is the menu large or small (for a food truck)?
4	How is the service?
5	Does the price match the amount of food you get?
6	Is it clean?
7	Does it move a lot, or is it often in the same places?
8	How far would you travel to get food from this truck?

popularity of the trucks however, most of our answers are for three: grlldcheesetruk, GrillEmAllTruck, and kogibbq.

The questions asked in this scenario are shown in Table 2. These questions were chosen based on our experience and intuition of what might be interesting to a prospective customer of a food truck. Unlike for the tablet questions, we conducted only one round of question asking and we also chose not to include a second clarification question for any of our questions. Initially our phrasing included the food truck Twitter handle using the @ symbol, which notifies the food truck owner that we mentioned their account. Due to the number of questions being sent, we later removed the @ symbol to avoid creating too many mentions that might spam the truck owners. Use of the @ symbol in our tweets did lead to responses from several food truck owners in addition to the potential answerers we were targeting. Questions were sent using the weighted random method from the Twitter account **@foodtruckqa**. Questions were asked during the period of September 16-28, 2011.

Data Collection from Amazon and Yelp

To compare the quality of the product reviews collected from targeted strangers on Twitter to a baseline, we collected a set of user-generated reviews on Amazon and Yelp.

From Amazon, we first collected the top-50 rated most useful reviews of the Samsung Galaxy Tab 10.1. By carefully examining these reviews, we found that the top 10 reviews covered 8.7 key features of the tablet (e.g., display, performance, and software) on average, while the next 10 reviews covered only 3.9 features already in the top 10 reviews. The reviews rated below the top 20 contained even less information. We thus decided to use the top 10 reviews as our comparison data set for the tablet scenario.

From Yelp, we collected the first page of reviews in Yelp-sort order (a ranking that combines usefulness, recency, and other factors), for the three most popular food trucks mentioned previously: *Kogi BBQ*, *The Grilled Cheese Truck*, and *Grill Em All*. This collection contained 40 reviews for each truck, which we examined using the same approach as for the Amazon reviews. Based on our analysis, we selected the top 10 reviews of each truck for comparison.

CONTENT ANALYSIS METHODOLOGY

We examined the content of collected reviews and analyzed their quality in two methods: hand coding the reviews for an objective content analysis, and a survey conducted on Mechanical Turk for a subjective analysis.

Review Coding

To help understand the user responses and to compare the content of the reviews collected on Twitter with those from Amazon and Yelp, we hand-coded all reviews from all sources to answer two main questions:

1. How did the Twitter users respond to our questions? This question gets at more generic aspects of the responses, such as whether users provided a relevant answer, or whether they asked us a question back, etc.
2. What types of information do the reviews contain, and how do the Twitter responses compare to those collected from Amazon and Yelp in terms of their content? This question requires more scenario-specific coding to match the different information types conveyed in the different types of reviews.

We coded the reviews using an open coding approach, guided by our research questions (see the Introduction). For the tablet reviews, four coders coded each collected tweet and each Amazon review independently and then discussed their coding and reconciled the differences. Similarly, for the food truck reviews, two coders first coded each collected tweet and each Yelp review independently, and then reconciled their differences through discussion.

Through this process, our coding scheme evolved to contain two types of codes to address our first two sets of research questions. Twitter-specific codes about users' response behavior are shared across the Twitter reviews for both products, aiming at answering the first set of questions. For example, one Twitter-specific code indicates whether a response is relevant to the question. To answer the second set of questions, domain-specific codes were created to describe particular types of information contained within a review. For example, one tablet-specific code indicates that a user mentioned that the display was bright, and one food truck-specific code indicates the cleanliness of the truck. As a result, our scheme included 10 Twitter-specific codes, 86 informational codes about tablets, and 29 about food trucks. The informational codes were grouped into a hierarchy with 9 top-level topics for each product type, which were used to construct the survey described below.

Mechanical Turk Studies

To assess users' perceived quality of the reviews collected from different sources to answer our third set of research questions, we designed and conducted a survey study using Amazon Mechanical Turk. Four different surveys were constructed for each combination of review types (Twitter and Amazon/Yelp reviews) and products (tablet and food truck). The first page of each collected basic demographic

information, such as age, and information about the turker's experience with the product under survey.

The turker was then asked to read the complete set of reviews corresponding to the particular survey. For example, all Twitter reviews about the 3 LA-area food trucks, or the top 10 Amazon reviews about the Samsung Galaxy Tab 10.1. The reviews were displayed in a separate browser window so the turker could refer to them throughout the survey. Each survey consisted of multiple pages, each corresponding to the top-level topic identified in our collected Twitter reviews for that product. For each topic, the turker was asked to use the reviews that they just read to come up with at least 5 keywords to describe the topic. This exercise not only was intended to test the turkers' comprehension of the reviews, but it also forced the turkers to carefully read all the reviews before answering any survey questions.

After completing each topic page, the turker were asked 5 5-point Likert scale questions about the reviews:

- *Usefulness*: how well have the reviews provided overall and relevant ideas about the product that may affect your buying decision?
- *Objectiveness*: how well do the reviews provide unbiased information about the product?
- *Trustworthiness*: how much do you trust the reviews?
- *Balance*: how well do the reviews cover multiple aspects, including pros and cons?
- *Readability*: how easy is it to understand the reviews?

After completing the survey, the turker was given a code to enter into the Mechanical Turk web site to receive payment.

To ensure the quality of the survey, we allowed only turkers located in United States and with greater than 96% approval ratings to participate. A browser cookie was checked to ensure that a turker filled out only one of the four surveys. Each turker response was individually screened to ensure the quality of answers. The four surveys were posted until we received 36 valid responses to each. A total of 19 work assignments were rejected, primarily due to the submission of incorrect completion codes. Some surveys were rejected for copying/pasting or entering random text in the responses. Each approved response was paid $2.00. Among the respondents, 57% were male; 49% were between age 20 and 30, 22% between age 30 and 40, and 21% above 40.

TWITTER RESPONSE RESULTS

To address our first set of research questions, we examined the response rate and time of the Twitter users in our experiment and compare them with previous work. We then analyzed the response quality based on our coding.

Response Rate and Time

In previous work [12], it was found that response rates and times differ for the first question asked and any follow-up questions, so here we break out results by the order in which a user saw a question. Table 3 shows the response rates and times, broken down by product and question order. Figure 2 shows the response time behavior for both tablet and food truck.

Table 3. Summary of response statistics for both products.

	Questions	Responses	Response Rate (%)	Median Response Time (min)	% Responses in 30 Min	95% Responses Received In (hrs)
Tablet						
Q1	633	183	29%	180.5	36%	128
Q2	117	75	64%	23.3	52%	39
Food Truck						
Q1	171	70	41%	34.7	46%	128
Q2	57	41	72%	32.1	49%	21

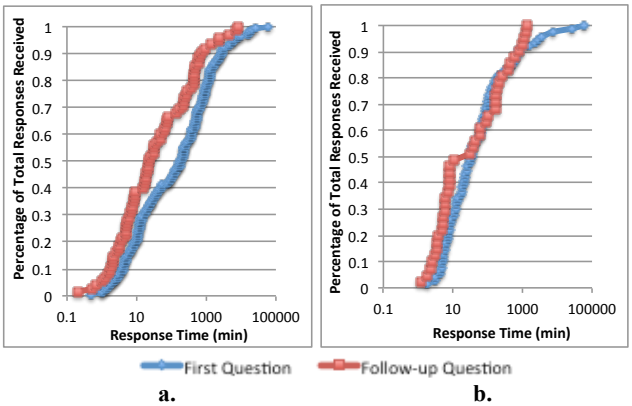

Figure 2. Response times for the (a) Tablet scenario, and (b) Food Truck scenario.

The response rates and times for food truck are comparable to previous results that obtained an average response rate of 42% and received 44% of responses in 30 minutes [12]. The first question in the tablet scenario has both a noticeably lower response rate and slower response times. The lower response rates may be due to the greater difficulty in identifying users who owned a Galaxy Tab 10.1 as compared to other products in the Galaxy line, as questions sent to owners of the wrong product were less likely to receive a response. It is not clear what would lead to the longer response times. We examined the difference between the two rounds of Q&A for the tablet to see if a change between the two might account for these differences. However, response rates were nearly identical for both rounds (34.9% for both) despite more than twice as many questions being asked in the second round (209 questions were sent in the first round vs. 541 in the second round).

Table 4. Summary of answer coding statistics.

	Response Count	Relevant Answer	Wrong Answer But Useful Info	Multi- Message Response	Average Info per Response	Off-topic Info per Response
Tablet	258	71%	19%	3%	1.82	0.48
Food Truck	111	82%	6%	6%	1.69	0.46

Table 5. Summary of reasons for low quality answers.

	# Irrelevant Responses	No Experience	Didn't know or understand	Thinks we're a bot
Tablet	75	63%	11%	7%
Food Truck	20	25%	30%	0%

Response Quality

Table 4 presents our analysis of the relevance of the Twitter responses that we received, as determined from hand coding. The majority of responses for both products provided answers to the questions asked. Even when the specific question was not answered, the answerer often provided useful additional information.

Table 5 shows a breakdown of the responses that failed to answer the question. In the tablet scenario, the majority of irrelevant responses (63%) came from users who had no experience with the device we asked about. This verifies that identifying users of the Galaxy Tab 10.1 was difficult for our human operator, which likely led to a decrease in response quality. Two other causes can be attributed to the irrelevant responses: 1) some users did not understand our questions (11%), and 2) some believed that a bot sent the questions (7%). However, these two cases combined occurred in only 5% of the overall responses (19 out of 369).

Table 6 shows some actual responses that we received to our questions. We examined the number of information pieces contained in each response, and found for both products that users on average provided more than one piece of information per response. While the majority of this extra information seems relevant to the response, such as providing evidence for a claim (Table 6a), users would occasionally provide other information that was not directly asked for, such as about a different feature that was not mentioned (Table 6). In about 40% of responses for the tablet and 87% for the food truck, the users gave a summary statement about the product (e.g., "great device!").

COMPARING QUANTITIES OF INFORMATION

To objectively compare the *quantities of information* contained in the two sets of reviews for each product, we applied Shannon's information theory [15] to our human coding of the amount and type of information contained in each set of reviews (Twitter, Amazon, and Yelp).

Table 6. Questions and answers sent as part of our experiments with Twitter user names anonymized.

a.	**foodtruckqa:** "Is it clean?"
	user1: "yes. They have a health grade of A."
b.	**tabletsqa:** "How fast is it?"
	user2: "Its fast. More flexibility than ipad. Allows SIM card."
c.	**foodtruckqa:** "Does the price match the food you get?"
	user3: "yupp the best tots and hot dogs on the planet"
d.	**foodtruckqa:** "What do you prefer to order?"
	user4: "Chow fun, chimales are my faves, but it's all good. In alley off 6th at Main tonight."

Table 7. Information entropy computed based on the hand-coded content including only information points occurred in both of a scenario's review sets.

	Tablet		Food Truck	
	Amazon	Twitter	Yelp	Twitter
Information (bits)	4.09	3.73	3.27	3.02

Specifically, we computed Shannon entropy $H(X)$ to measure the average amount of information in *bits* contained in each set of reviews (e.g., Twitter tablet reviews) [15]. Here X is an *information point*, a random variable with values ranging over all the features of a data set (e.g., the display of the tablet or the location of the food truck). $H(X)$ was calculated using a shrinkage entropy estimator [6]. We computed the entropy for the four review data sets (Table 7). To make a fair comparison, here we limit the entropy analysis to only information points that exist in both sets.

For both products, the Twitter reviews contained slightly less information than their web-based counterparts. To better understand the differences in their information quantity, we also calculated the relative entropy (KL divergence) between each two sets [2], which suggests the corresponding review sets contain similar amounts of information for their common information points.

MECHANICAL TURK COMPARISON RESULTS

In Mechanical Turk studies, we examined how people perceive the quality of data collected from different sources.

Tablet Results

On average, turkers finished reading and answering questions for the Amazon reviews in 26.5 minutes, Twitter reviews in 25.8 minutes. A t test showed no significant difference between the two.

Turker Background

Among our subjects, 68% did not own any tablet; 58% did not know anything about the tablet under review, and only 5% had a lot of experience with the tablet. However, 68% of the readers did have plans to buy a tablet at some point in

Table 8. Turker's subjective rating of review quality for Tablet reviews

	Amazon	Twitter	Mann-Whitney	p
Usefulness	3.19	2.64	868.5	0.006
Objectiveness	2.94	2.53	814.5	0.042
Trustworthiness	2.94	2.39	861.0	0.008
Balance	3.00	2.11	936.0	0.001
Readability	2.92	2.61	741.5	0.270

the future. On a 5-point Likert scale, the turkers were asked how often they search and consult online reviews before a purchase (0 being "never" and 4 being "always"). The median answer was 3 "often", with mean=3.32, sd=0.78.

Subjective Ratings of Review Quality

Table 8 summarizes the results of the five final questions in our survey, where turkers were asked to rate the reviews in five aspects on a 5-point likert scale (1 being most negative and 5 being most positive). Mean ratings and Mann-Whitney U statistics of comparison are shown. Since the turkers rated Amazon reviews consistently higher than Twitter reviews in all five aspects, we re-examined both sets of reviews. We found that Amazon reviews provided more details (e.g., one's experience with the tablet) and context (e.g., comparing the Galaxy tablet to iPad) about the tablet. The comments from the turkers confirmed these findings. Amazon reviews are also better written compared to the informal text in Twitter responses, which may be another factor affecting the ratings (e.g., trustworthiness).

As mentioned by Gilbert and Karahalios [4], there are professionals who write elaborate reviews to gain social capital on sites like Amazon. We examined the profiles of the 10 reviewers of the top-10 reviews used in our study. Seven out of ten reviewers had written at least ten reviews (23 reviews per person on average) on Amazon on various topics (e.g., electronics and computers). Five of them were also highly ranked among the millions of reviewers on Amazon. In contrast, we selected Twitter answerers merely based on their mention of "Galaxy tablet" in their tweets. Our ongoing work is performing an in-depth analysis of Twitter users based on their social media posts and social behavior to infer their intrinsic traits, including personality and motivations, which will form the basis for us to select those who are *willing*, *able*, and *ready* to answer our questions. However, this topic is beyond the scope of this paper. This also suggests that building a better profile of social media users (e.g., knowing a person who has extensive experience with multiple devices) and more careful selection of answerers might help improve the quality of responses.

Comparison of Perceived Content

To understand turkers' comprehension of targeted reviews, we asked them to use a list of keywords to describe each of the nine topics covered by the reviews. For each topic, we then compared the two sets of turker-entered keywords based on their reading of the Twitter and Amazon reviews.

Depending on the topic, the overlap between the two sets of keywords differed, which may imply the amount of content perceived by the turkers differed. For example, the *build* topic of the tablet is rather broad, covering multiple facets, such as look, feel, and weight of the device. In this case, the turkers used more keywords to describe the build of the tablet after reading the Amazon reviews. In contrast, our Twitter questions solicited information about the build from only three facets (Table 1, questions 2a, 2b, and 9). Compared to the build topic, the two sets of keywords generated for the display topic overlapped greatly. This suggests that our 2 questions about displays were sufficient to cover what was said in the Amazon reviews.

While the overlap between the two turker-generated keyword sets varied by topics, we see a clear trend: the more specific the topic is, the greater the overlap is between the two keyword sets. As discussed more later, this implies that our Twitter-based information solicitation is more suitable for clearly defined topics compared to broad topics.

Table 9. Turker's subjective rating of review quality for Food Truck reviews

	Yelp	Twitter	Mann-Whitney	p
Usefulness	2.86	2.56	734.0	0.309
Objectiveness	2.17	2.08	672.0	0.783
Trustworthiness	2.58	2.14	800.5	0.071
Balance	2.47	1.72	921.0	0.002
Readability	2.89	2.11	896.0	0.004

Food Truck Results

On average, turkers finished reading and answering questions about the Yelp reviews in 19.9 minutes and the Twitter reviews in 16.8 minutes. A t test finds no significant difference, with $t = 1.73$, $p = 0.08$.

Subjective Rating of Review Quality

Table 9 summarizes the five aspects of review quality as judged by the turkers. As can be seen, the Twitter responses are not significantly different from the Yelp reviews in term of perceived usefulness and objectiveness. The Yelp reviews were perceived slightly more trustworthy compared to the Twitter responses, but it is not statistically significant. The Yelp reviews are perceived to be significantly more balanced and readable than Twitter responses.

Note that the perceived difference between the two sets of reviews for the tablet is larger than that of the food trucks. Again, this may imply the suitability of domains for this method of information collection. In the tablet case, we

asked more generic questions (e.g., how is the display) versus more personalized experience in the food truck case (e.g., what they prefer to order).

Comparison of Perceived Content

We compared the two turker-generated keyword sets for each of the nine topics about the food trucks based on Twitter and Yelp. The results were similar to that of the tablet.

DISCUSSION

The results of our experiments show both the advantages and limitations of this new information collection approach.

Advantages

We have seen that greater than 70% of the responses to our questions contained relevant information, and many answers contained additional information beyond what was asked. We also saw that many of the cases in which we received irrelevant answers were due to targeting strangers who did not have the background or experience to answer the question. This is encouraging because it suggests we may be able to further improve quality by better choosing strangers to target or focusing on scenarios where the challenge of targeting is less difficult.

Our entropy-based analysis of information quantity suggests that the Twitter method produces similar information to other methods for the questions asked. This suggests that our approach could be more valuable if we also involve a crowd in question selection. For example, our system can collect reader-driven product reviews by allowing review readers to select the questions. This would differ greatly from the current state of the art where reviews are primarily writer-driven. Our food truck reviews provide some evidence that this could work already, as certain features of food trucks, like *cleanliness*, were not frequently discussed in the Yelp reviews that we looked at.

Limitations

A key negative result is that the Mechanical Turk users who took our surveys found the Twitter-based reviews universally lacking in balance, a feature we hypothesized would be a strength of our approach. Comments from turkers who took our survey suggest that concrete examples of positive or negative experiences would be desirable to include in the reviews, as well as more background about the users who are providing the reviews. This might be addressable by structuring our question asking to draw out scenarios and background information, which we did not attempt here.

Design Implications

Our findings suggest several important design considerations for building such a system.

Domain Suitability

A unique advantage of our approach is its *agility*: we solicit information from the right people at the right time on social media. This implies that our method could be better used to collect information that existing systems cannot offer. Given the dynamic nature of social media, our approach is particularly suitable for collecting *time-sensitive*, *context-specific* information (e.g., the current wait time at a popular restaurant). Due to the constraints of social media interaction (e.g., length of a tweet), our experimental results also indicate that our work is more effective for collecting information along a narrow dimension (e.g., the display vs. the build of a tablet) associated with an *easy-to-identify* entity (a specific food truck vs. a tablet in varied sizes). The latter criterion also helps reduce the difficulty in identifying the right users. In short, our approach could be applied to answer questions that are not easily possible to answer with existing systems (e.g., the crowd mood at an event) or to supplement existing systems, like Amazon and Yelp, to collect missing (e.g., cleanliness) information.

Selection of Target Answerers

Looking forward, if our method is to become useful, then it will likely be necessary to automate or greatly streamline the answerer selection process. Improved methods of user profiling are needed to identify users that are *willing*, *able*, and *ready* to provide the requested information. Such methods should also identify a diverse crowd from which balanced information can be collected (e.g., positive and negative reviews of a product). Not only will these considerations help address the current deficiencies in our approach (e.g., the perceived a lack of balance in our Twitter reviews), but they can also help improve the quality of solicited answers. For example, we could ask a person who has gone to several food trucks to provide a comparison or ask people with different backgrounds about their opinion of a product.

Questioning Method

The question method directly affects the quality of responses received. In our tablet scenario, for example, we asked "*Where do you use yours most?*" One person responded, "*elaborate 'where'. As in apps or where I usually use my tab.*" A more targeted question such as, "*What apps do you use the most?*" may solicit more useful responses regarding the usage of the tablet. We could also consider structuring questions to elicit more balanced and useful responses. For example, in the food truck scenario, when asking "*what do you prefer to order*," adding "*tell us also what you dislike*" may help solicit a more balanced answer. In this same domain, we would obtain more useful rating about the price-to-value ratio of a food truck, if we ask users how often they frequent the food truck or whether they have visited other trucks.

Since our approach engages with strangers on social media, a challenging research question is how to maximize the information gain while minimizing the cost. For example, a multi-step conversation may allow the system to acquire more information and accommodate platform restrictions (e.g., number of characters allowed per message), but users may dislike engaging in a long conversation with a stranger.

Another key ingredient toward automation is to decide how many times a question should be sent. The amount of information obtained may not grow linearly as the number of responses grows. Monitoring the information entropy of the responses (i.e., the amount of information gained) as they are collected may help with making this decision, and carefully crafting or dynamically adjusting questions may even allow the system to optimize the rate of information gain.

CONCLUSION

In this paper, we explored the quality of crowd-sourced information that is actively solicited from strangers based on their public social media status updates. We found that users answered questions at rates similar to those found previously, the answers contained information relevant to the question over 70% of the time, and over 37% of answers provided additional details beyond the specific question asked. The information collected was also similar to that of other sources when controlling for the set of questions that were asked. While our work demonstrates the potential of this new type of information collection systems, our finding that users preferred traditional reviews suggests that challenges still remain in selecting questions and answerers, or displaying the content in a more acceptable fashion, in order to produce a better review experience.

ACKNOWLEDGEMENTS

Research was sponsored by the U.S. Defense Advanced Research Projects Agency (DARPA) under the Social Media in Strategic Communication (SMISC) program, Agreement Number W911NF-12-C-0028. The views and conclusions contained in this document are those of the author(s) and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Defense Advanced Research Projects Agency or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

REFERENCES

1. Bulut, M.F., Yilmaz , Y.S., and Demirbas, M. "Crowdsourcing Location-based Queries," in *PerCol 2011*.
2. Cover. T., and Thomas, J., (2006) *Elements of Information Theory*. Wiley.
3. Gazan, R. Social Q&A. *J. of the Amer. Soc. for Info Science & Technology,* 62(12): 2301-2312.
4. Gilbert, E. and Karahalios, K. "Understanding déjà reviewers", in *CSCW 2010*: 225-228.
5. Harper, F.M., Raban, D., Rafaeli, S., and Konstan, J.A. "Predictors of answer quality in online Q&A sites," *CHI'08*: 865-874.
6. Hausser, J. and Strimmer, K. (2009) "Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks," In *J. Mach. Lear. Res.* 10: 1469-1484.
7. Horowitz, D. and Kamvar, S.D. "The anatomy of a large-scale social search engine," *WWW'10*: 431–440.
8. Hsieh, G., Kraut, R.E., and Hudson, S.E. "Why pay?: exploring how financial incentives are used for question & answer," *CHI'10*: 305-314.
9. Jeon, G.Y., Kim, Y-M., and Chen, Y. "Re-examining price as a predictor of answer quality in an online q&a site," *CHI'10*: 325-328.
10. Morris, M.R., Teevan, J., and Panovich, K. "What do people ask their social networks, and why?: A survey of status message q&a behavior," *CHI '10*:1739-1748.
11. Nelson, P. Information and Consumer Behavior, 78(2) *Journal of Political Economy* 311-329 (1970).
12. Nichols, J. and Kang, J.H. "Asking Questions of Targeted Strangers on Social Media," *CSCW'12*: 999-1002.
13. Paul, S.A., Hong, L., and Chi, E.H. "Is Twitter a Good Place for Asking Questions? A Characterization Study," *ICWSM'11 Posters*.
14. Panovich, K., Miller, R.C., and Karger, D. "Tie Strength in Question and Answer on Social Network Sites," *CSCW'12*: 1057-1066.
15. Shannon, C. E. "A Mathematical Theory of Communication". *Bell Sys.Tech. J.* 27 (3): 379–423, 1948.
16. Shah, C. and Pomerantz, J. "Eval. and predicting answer quality in community Q&A," *SIGIR'10*: 411-418.
17. Richardson, M. and White, R.W. "Supporting Synchronous Social Q&A Throughout the Question Lifecycle," *WWW'11*: 755-764.
18. Zhu, Z., Bernhard, and D., Gurevych, I. (2009) "A Multi-dimensional Model for Assessing the Quality of Answers in Social Q&A Sites," Tech. Rep. TUD-CS-2009-0158, Technische Universität Darmstadt.
19. Zhuang, L., Jing, F., and Zhu, X. "Movie review mining and summarization," *CIKM'06*: 43-50.