Debugging a Crowdsourced Task with Low Inter-Rater Agreement

Omar Alonso
Microsoft Corporation
omalonso@microsoft.com

Catherine C. Marshall
Texas A&M University
ccmarshall@cse.tamu.edu

Marc Najork*
Microsoft Corporation
najork@acm.org

ABSTRACT

In this paper, we describe the process we used to debug a crowdsourced labeling task with low inter-rater agreement. In the labeling task, the workers' subjective judgment was used to detect high-quality social media content—interesting tweets—with the ultimate aim of building a classifier that would automatically curate Twitter content. We describe the effects of varying the genre and recency of the dataset, of testing the reliability of the workers, and of recruiting workers from different crowdsourcing platforms. We also examined the effect of redesigning the work itself, both to make it easier and to potentially improve inter-rater agreement. As a result of the debugging process, we have developed a framework for diagnosing similar efforts and a technique to evaluate worker reliability. The technique for evaluating worker reliability, Human Intelligence Data-Driven Enquiries (HIDDENs), differs from other such schemes, in that it has the potential to produce useful secondary results and enhance performance on the main task. HIDDEN subtasks pivot around the same data as the main task, but ask workers questions with greater expected inter-rater agreement. Both the framework and the HIDDENs are currently in use in a production environment.

Categories and Subject Descriptors

H.1.2 [Models and Principles]*:* User/Machine Systems—*Human information processing*

General Terms

Design, Experimentation, Human Factors

Keywords

Crowdsourcing; labeling; inter-rater agreement; relevance judgment; debugging; Captchas; worker reliability

1. INTRODUCTION

Researchers and practitioners often rely on crowdsourcing as a quick and inexpensive means to produce labeled data sets. Labeled corpora may serve a range of functions: they may be used to train and evaluate machine learning algorithms; they may be used as reference datasets for implementing and evaluating information retrieval techniques; and they can aid in the curation of large information resources. In essence, labels are important metadata for systematically classifying and describing heterogeneous items. Thus care must be taken to ensure the crowdsourced labels are as high-quality as the metadata assigned by experts [1].

Assigning a label is usually considered a judgment task, where the judgments may be more or less subjective. A judgment on the objective end of the spectrum may suggest a single right answer, so we can rely on a single worker or a small number of workers to accurately label an item. If the judgment is more subjective, we might ask more workers the question, and use consensus or some other method of aggregation to determine a best-choice label.

We are focusing on labeling tasks in which the judgment is at the subjective end of the spectrum. These tasks can be thought of more like polling: although inter-rater agreement may not be high, the label choice should be replicable across groups of workers as long as the workers are diverse, reliable, and consistent in their judgments and the datasets can be adequately characterized to show their similarities and differences. In other words, if the same type of crowd labels the same type of data, the outcome will be predictable. Yet tasks with low inter-rater agreement are problematic to debug. If the judgments are more subjective, it may be difficult to tell whether the workers are indeed reliable (e.g. there is no gold set or consensus answer with which to vet their performance) or whether problems have been properly diagnosed and eliminated before the labeling process is scaled and put into a production environment.

In this paper, we'll address two research questions intrinsic to crowdsourced labeling tasks with low inter-rater agreement:

- Can we develop an effective way of debugging such a crowdsourced labeling task?
- How can we determine that the workers are performing this type of labeling task in good faith?

The example labeling application that has driven the investigation we describe in this paper is identifying high-quality content in a socially-produced feed. Specifically, we are asking workers to identify interesting tweets from a dataset that represents the contents of the high volume Twitter feed. Once a set of tweets can be reliably labeled as interesting (or not interesting), it can be used to train and evaluate a classifier that works on the entirety of the feed.

Naturally, what constitutes an interesting tweet is variable and relies on the judges' perceptions; we can think of it as a question at the subjective end of the range. Yet the ability to identify and label high-quality socially produced content has practical value.

JCDL'15, June 21–25, 2015, Knoxville, Tennessee, USA.

DOI: http://dx.doi.org/10.1145/2756406.2757741

* Author is now at Google.

For example, consider the use case of identifying and covering a news story via selection of interesting tweets. Predictive features have already been used to identify tweets at the other end of the value spectrum (uninteresting tweets) with similar applications in mind [2]. The next challenge is to extract sufficient signal to label interesting tweets. Alonso, Marshall, and Najork began that process by testing aspects of the data and workers, and by building a classifier whose performance may be improved [3].

Figure 1 shows our two-phase model. In the figure, information is represented as boxes and functions as labeled arcs. In Phase 1, the focus of this paper, the dataset we start with consists of a representative portion of the data that will eventually be processed via Phase 2. One or more judges assigns a label (usually from a constrained set of possibilities) to each item from the dataset. Since judge characteristics (e.g. expertise, interests) and item characteristics (e.g. topic, provenance) may influence label production, the degree to which the judgments vary depends on both. This variability (i.e. differences in judges' assessment of which label is appropriate for an item) may be addressed by developing an aggregation method that resolves the differences. For example, some aggregation methods rely on worker consensus. Others might use the labels assigned by the most reliable workers. In Phase 2 of the process, a portion of the labeled items are reserved to form a *holdout set*. The rest, the labeled training set, are used by a machine learning algorithm to produce a model that predicts labels based on item characteristics. The holdout set can thus be compared with the predictions to evaluate how well the model captures the judges' assessments.

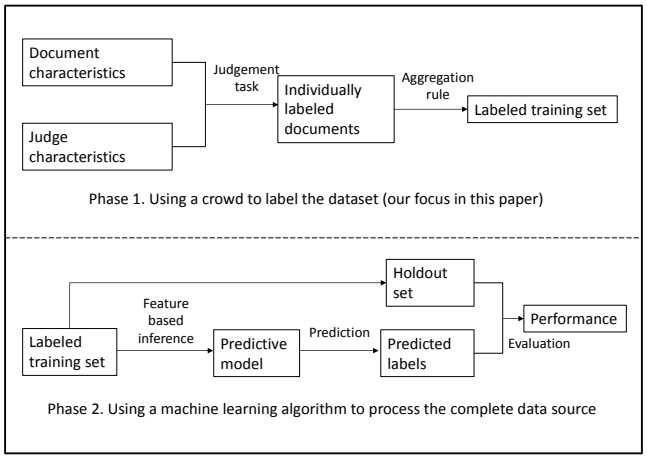

Figure 1. A two-phase model of labeling documents and assessing the performance of a classifier that uses the labels.

This paper will first consider related work. Then we will present a series of investigations designed to debug our labeling task. The debugging process accounts for the workers, the datasets they are labeling, and the design of the template they use to perform the work. We will also describe a promising technique to vet the workers' reliability and improve their task performance. We will conclude by discussing some avenues for future work, especially the evaluation of the techniques we developed during the course of the research presented in this paper.

2. RELATED WORK

Three types of related work informed our effort: (1) using crowdsourcing to evaluate content quality, including such efforts in the TREC community; (2) improving the crowdsourcing process by applying techniques for crowdsourcing at scale and for validating worker reliability; and (3) developing techniques to address low inter-rater agreement. We discuss each in turn.

Evaluating content quality. This work aims to identify high-quality content in very short documents, especially tweets and comments. André, Bernstein, and Luther [4] rely on self-selected volunteers to rate tweets from accounts that they follow to characterize worthwhile tweets. We are similarly trying to identify interesting tweets; however, we are ultimately concerned with identifying predictive features so that the process can be scaled to evaluate tweets in near real-time. Momeni et al. use a crowdsourcing approach to label a set of useful comments against which to build a classifier [14]. Although TREC ranking algorithms estimate a tweet's relevance to a query, some of the features identified by Metzler and Cai [13] are similar to interestingness features used by Alonso, Marshall, and Najork [3], work that is the basis for ours. Alonso et al. have taken a related subtractive approach by identifying tweets that are not interesting [2]. Like Lin, Etzioni, and Fogarty [11], we began by looking for consensus on what is interesting, although our crowd has a more diverse view on what constitutes interestingness; we are building on these results to understand if we can harness this diversity and work with the reduced level of inter-rater agreement in labeling.

Crowdsourcing techniques. Crowdsourcing at scale has been the focus of recent workshops and conversations [8]. Because we are planning to use our technique in production, we have paid particular attention to work that considers experimentation as the first step to scaling up [3]. Worker reliability has been a focus of von Ahn, Blum, and Langford's Captcha research [17]. Captcha-like techniques were introduced to crowdsourced user studies by Kittur, Chi, and Suh [9]. Because user studies tend to involve more effort per Human Intelligence Task (HIT) than labeling tasks, we have taken a variant approach to ensuring worker reliability; this approach is described in Section 4.2.

Addressing low inter-rater agreement. Much of the related work uses inter-rater agreement to label content. What is an acceptable minimum signal upon which to base a binary classifier? Alonso, Marshall, and Najork were only able to achieve moderate agreement (Fleiss's $\kappa = 0.51$) between crowd-labeled tweets and a classifier [3]. Our work ultimately strives to increase classification accuracy by improving label quality; the first step in this process is to productively use low inter-rater agreement. Aroyo and Welty have investigated a technique for productively using disagreement among judges [5]; like them, we are exploring a crowdsourcing task in which we do not expect consensus labels. Others cope with low inter-rater agreement in the data cleaning process, for example, by removing items that cannot be labeled through consensus [16]; we are attempting to develop a process that will allow us to retain the entire signal.

3. APPROACH

Our overall approach involves rapid iteration of labeling tasks using small datasets, from 100 to 2,500 tweets, each judged by 5 workers; because we are focused on debugging the labeling process, we have decreased the dataset size to make it practical to test different aspects of the dataset, the workers, and the task itself. Most trends we are interested in are observable by comparing these results; otherwise it would be costly and slow to test combinations of multiple factors to pinpoint the source(s) of low inter-rater agreement (e.g. dataset recency or the question we

are asking the workers). Once this technique is refined, we can begin testing it at scale. Figure 2 shows our general approach to structuring the investigation. The circles in the top part of the diagram identify the portion of the work we are varying; the boxes along the bottom specify individual debugging exercises. Each debugging exercise may involve multiple datasets.

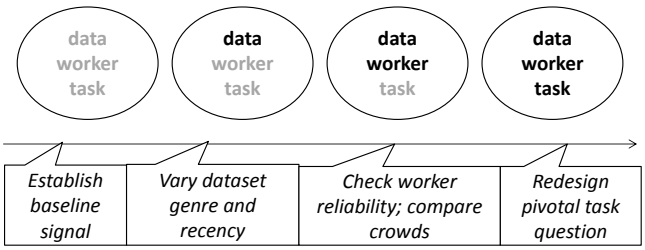

Figure 2. Structure, strategy, and flow for debugging a crowdsourced task with low inter-rater agreement.

For each investigation, we circulated and discussed a report summarizing the cumulative results, including the performance of individual judges, the outcome of subtasks, significant or surprising results, the level of inter-rater agreement using the measure described above, and other variations in the experiment. Examining the cumulative results each time allowed us to compare the factors we varied between individual runs.

We go on to describe specific aspects of our approach to debugging, including preparing the many datasets we used, discussing the difference in the crowdsourcing platforms, how we measured agreement, and how we established a baseline that represents our real task, judging the interestingness of tweets.

Preparing the datasets. Fourteen datasets were used to test each aspect of the flow described in Figure 2, and two additional datasets were used to set baseline worker agreement values and interest level. All datasets were sampled from the Twitter firehose, and filtered according to the needs of the investigation. For example, news tweets were filtered from random tweets according to the account issuing the tweet and the tweet's date.

Necessarily this process suggests we use a large number of small datasets to compare the effects of different factors on inter-rater agreement, and to identify the sources that might be influencing the outcome. As we progressed through the process described in Figure 2, we sometimes increased the number of factors that we compared to resolve their individual effects. After we became confident we understood a particular effect, we could once again eliminate a particular factor from the process and return to our original goal of identifying interesting tweets from a dataset of random tweets. Table 1 describes the sixteen datasets used in the debugging process. We will remind the reader of what the datasets contain when we present individual results; Table 1 just shows the lay of the land. Datasets B1-B2 were drawn to set a baseline; G1-G3 were drawn to test changes in the dataset contents; W1-W7 were drawn to test worker reliability and expertise; and T1-T4 were drawn to test the task design.

dataset IDs	tweets	genre	date (relative to task)
B1	2,000	random	Two months prior to task
B2	2,000	random	Recent
G1	2,500	news	Two years prior to task
G2	2,500	news	One year prior to task
G3	2,500	news	Recent
W1, W2, W5, T1	100	news	Recent
W3, W6, T2	100	news	One year prior to task
W4, W7,T3	100	news	Two years prior to task
T4	100	random	Two years prior to task

Table 1. Datasets used in the debugging process. In all cases, each tweet was judged by 5 workers, so 100 tweets yields 500 labels. The dataset IDs correspond to each debugging phase.

Datasets B1 and B2, each containing 2,000 random tweets, were used to establish a 10,000 judgment baseline. Datasets G1, G2, and G3, each containing 2,500 tweets, were used to test effects of both genre (specifically news) and recency. After we noticed recency effects from the baseline set, we drew subsequent datasets that were either contemporaneous with the task, or one or two years old. Datasets W1 through W7, each containing 100 tweets, were used to test the effects of worker reliability and expertise. Finally, datasets T1 through T4, each containing 100 tweets, were used to test the effects of a new task design.

Crowdsourcing platforms. To test the influence of platform during our debugging process, we recruited workers from two different platforms: a Microsoft-internal crowdsourcing platform which specializes in relevance judgments (UHRS) and Amazon Mechanical Turk (AMT). Workers were paid according to the market rate on each; workers are paid relatively more on UHRS than they are on AMT because of the presumed expertise of UHRS workers. On AMT, we used recognized best practices, and required that workers had 97% past success and had completed at least 50 past tasks; this tends to eliminate most spammers. UHRS has no equivalent ability to set worker qualifications; thus we vetted the workers by blocking spammers as they appeared. Tasks were monitored when they were underway to ensure that workers understood the instructions and were able to use the templates to label the collections. Unless we otherwise specify, the tasks were done by UHRS workers.

Measuring agreement. Obtaining trusted data is vital to our approach. Because we expect to run crowdsourcing jobs continuously, it is important to show that the data produced by each step is reliable. We rely on a standard measure of inter-rater agreement, Krippendorff's α, which produces values between 1 and -1. A value of 1 indicates perfect agreement among workers, a value of 0 suggests that workers may be assigning labels randomly. A negative value further indicates that disagreements are systematic. The literature on inter-rater reliability provides recommendations on cutoffs for data to be considered reliable. Values above 0.8 suggest that the judges are in almost perfect agreement whereas values between 0.61 and 0.8 suggest substantial agreement. Lower values indicate progressively less agreement and hence lower reliability. In Section 4, we report values for the disaggregation of interest at the ends of the spectrum ("worthless" and "important"; see Table 7, T1) that fall under the rubric of fair agreement. Many of the values for inter-rater agreement in this paper suggest only slight agreement. It is

this problem—how to pinpoint the source of low agreement values—that we address in this paper. Krippendorff's α is capable of handling datasets where the number of raters per item varies, which is the case for some of the label sets we produced. Generally, inter-rater agreement will be higher when there are fewer categories to choose from; in our investigations, we are asking workers to make binary judgments.

Establishing a baseline. Before we started running debugging probes, we needed to set a baseline for worker performance. This baseline used datasets B1 and B2, each containing 2,000 random tweets (drawn from the Twitter firehose, as we explain above). Each dataset was used to establish a 10,000 judgment baseline. Recall that B2's tweets were recent and B1's were older. As we discovered earlier [3], random tweets vary in quality and topic. However, a tweet that looks obviously worthless to us might match an individual judge's interests in a completely subjective way. For example, in earlier investigations, when asked about brief cryptic tweet, a judge who labeled it as *Interesting* explained that it was recent gossip about Kim Kardashian (who was referred to in the tweet by a nickname none of us were aware of), and that he or she loved Kim Kardashian. It is easy to see why inter-rater agreement on random tweets might be low.

Figure 3 shows an annotated version of our baseline task for judging tweets. The judges were given brief instructions about the judgment, but no further definition of interesting. They were asked if the tweet—as it was shown—was interesting to a broad audience.

Figure 3. Sample baseline task for judging tweets. A tweet is displayed and the main judgment question is posed.

Table 2 shows the results from our two baseline tasks. The percentages have been averaged over all 10,000 judgments. Krippendorff's α expresses inter-rater agreement. We compute % interesting by using majority vote.

	B1 (older, random)	B2 (recent, random)
% interesting	16.7%	14.3%
Krippendorff's α	0.013	0.052

Table 2. Baseline values for 5 workers labeling random tweets.

In our baseline case, the age of the tweets appears to have little influence on the overall level of interest that the judges express.

4. RESULTS

In this section, we will discuss the results of each set of explorations, with the ultimate aim of both debugging this particular tweet labeling task before we scale it, as well as developing a reliable debugging process in general. The general debugging process is an important research outcome for us.

As Figure 2 suggests, our first variations were aimed at discovering whether we could elicit higher agreement by narrowing the data genre to news. In other words, was it the nature of the task that elicited such low agreement, or was it the type of data we were using?

We then turned our attention to the workers: because the judgment task is subjective, a gold set or agreement can't be used to validate the workers' labels. So we needed to develop a different type of method to evaluate the quality of their work. Because regular quality checks produce so much drag (they are creating more work for the workers), we were also interested in using this method to not just evaluate their work, but also to improve the quality of their work.

Finally, we scrutinized the task design: would a user-centered process of label assignment, one that considered different emotional components of what made something interesting, reduce the task's cognitive load as well as improve inter-rater agreement?

4.1 Effects of changing the dataset

Our first area to debug was the effects of dataset genre: would limiting the dataset to recent news tweets improve inter-rater agreement? Knowing this effect helps set our expectations for the results of our interestingness question. In other words, is our low inter-rater agreement the result of the subjectivity of the question, or does it stem from the quality of the data?

Although people have differing levels of interest in some types of news stories, we thought it likely that the workers would agree that some stories were of more universal importance. Thus for our first investigation, we extracted tweets from ten recognized top US newspapers, news services, and broadcast news: the *Los Angeles Times* (*@latimes*), Reuters newswire (*@reuters*), the *New York Times* (*@nytimes*), the *Wall Street Journal* (*@WSJ*), *USA Today* (*@USATODAY*), the *Washington Post* (*@washingtonpost*), the *Christian Science Monitor* (*@csmonitor*), ABC News (*@ABC*), Bloomberg News (*@BloombergNews*), and BBC News (*@BBCNews*). Because our baseline measurements suggested that the age of the tweets may influence workers' assessment, we drew 2,500 random news tweets for a date contemporaneous with the task, and dates one and two years older.

Genre. Table 3 compares the results of judging a recent news dataset with the agreement and interestingness of the recent baseline dataset of random tweets. Again, each tweet was judged by 5 workers. Although there are more than twice as many tweets judged to be interesting in the recent tweets, the inter-rater agreement shows relatively little improvement. Genre thus may have less effect than we had thought it might.

	B2 (recent, random)	G3 (recent news)
% interesting	14.3%	29.3%
Krippendorff's α	0.052	0.068

Table 3. Comparison of inter-rater agreement for baseline (recent random tweets) and news tweets.

Recency. We might think recency would have a more profound effect on news tweets than it did on our baseline random tweets, since the interestingness of news stories decays quickly. Table 4 compares three datasets of news tweets, G1, G2, and G3. G1 is the oldest of the news datasets (the tweets are two years old relative to when the task was put out for judgment). G2 contains one year old tweets and G3 contains tweets contemporaneous with the task. B1 is the non-recent baseline random tweets. We can see that the old news tweets become less interesting over time, but that there's still slightly better inter-rater agreement on which ones are interesting. There is also a decrease in inter-rater agreement, but it is not as profound as we expected it to be.

The genre constraint did not improve inter-rater agreement as much as we had expected. Thus we have to believe the fault lies with either the workers or the question itself.

	B1	G1 (oldest)	G2	G3 (recent)
% interesting	16.7%	21.3%	27.8%	29.3%
Krippendorff's α	0.013	0.037	0.074	0.068

Table 4. Effect of dataset age on interestingness and on inter-rater agreement.

4.2 Checking worker reliability and expertise

Our debugging efforts next shifted to the workers: were they continuing to pay attention as they judged tweet after tweet? Would we see better performance on one crowdsourcing platform than another? The crowdsourcing literature urges us to focus on the quality of worker output as a likely culprit when the results don't meet expectations. Unreliable performance, either as a result of fatigue, frustration, carelessness, lack of requisite expertise, or out-and-out fraud needs to be ruled out as part of our debugging process. But how could we eliminate poor quality work without a gold set to spot-check workers' performance or high inter-rater agreement to identify normative answers?

Worker reliability. To assess workers' reliability, we drew on two existing forms of worker checks, attention checks and memory checks. Attention checks ensure crowdsourced user studies are completed in good faith [9]. Workers are periodically asked unrelated questions to test whether they are still paying attention. Clever attention checks may also engender good will between worker and requestor [12]. Unlike attention checks, which are not task-related, memory checks evaluate reading comprehension. Although both checks address worker reliability, adaptation was necessary to make them suitable to our needs, since our HIT content varied, and our microtasks were much smaller than a conventional user study that employs memory checks.

Thus we supplemented the checks with the notion of reCaptchas [18], in which spam detection relies on the results of a useful microtask, such as OCR correction [17]. Successful completion of a reCaptcha makes it likely that a real human is at the keyboard and has the beneficial side effect of completing useful work for some ancillary task. We also added a third goal, improving data quality, since we were adding drag to the process by significantly increasing the size of the task. Thus, instead of being unrelated to the HIT content, the additional microtasks were designed to focus the workers' attention on the tweets they were about to judge. We call these specialized within-task reCaptchas Human Intelligence Data-Driven Enquiries (HIDDENs). HIDDENs allowed us to assess task performance in the absence of a gold set and achieved two quality-related goals:

1. To complete the judgment task, workers had to read the tweet three times, attending to different aspects of the short post for each microtask (first, its topic; second, whether it was about a specific person; and finally the original judgment task). The design goal was to avoid distracting the worker from the primary task; instead the each part built up to assigning the label.
2. Each of the embedded questions was designed to represent increasing levels of subjectivity. The first question we asked was objective and computable; the second was less objective, relying on worker agreement to determine an acceptable answer; and the third was our original more subjective judgment task.

Specifically in our initial implementation of HIDDENs, the first embedded microtask asked workers to count the tweet's hashtags. Although this is computable, it required workers to read the tweet a first time (albeit superficially). The second embedded microtask, assessing whether a tweet was about a specific person (i.e. contained a proper noun that was a person's name), required additional thought. Instructions specified that the name could neither be an account name (@name), nor a hashtag (#name). Figure 4 shows the resulting task template.

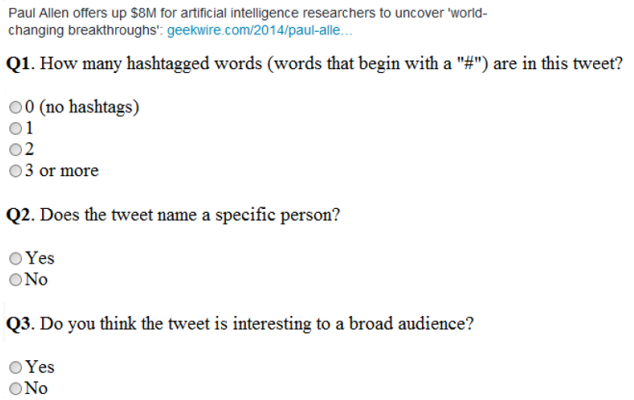

Figure 4. Task template with HIDDENs (Q1 and Q2). Q3 is the original judgment question shown in Figure 3.

Q1 relies only on characteristics of the tweet; Q2 relies on the worker's knowledge (e.g., being aware that Mubarak is a person, not a place); and Q3 is our original judgment question. We anticipated high agreement on the first subtask (otherwise the worker was suspect, either because he or she wasn't paying close attention to the task or because he or she wasn't consulting the tweet). We expected good agreement on the second, depending on the breadth of a worker's awareness. The HIDDENs allowed us to evaluate worker reliability in the absence of a gold set and gave a way to assess the individual worker's attentiveness and skill. Repeated anomalies in Q1's answer across workers may reveal problems in the dataset (e.g. a hashtag that's part of an emoticon) or can inform dataset analytics. Because Q2 relies on some expertise, its results can be useful to a colleague, or may be used to perform a related task (this way work can be interlocked, so spam detection on one task can provide useful results for another). By design, Q1, Q2, and Q3 are tied together by the single piece of content being judged.

Table 5 shows the results of the first investigations that used the HIDDENs. Both W1 and W2 contained 100 current news tweets. The rows labeled ***all*** contain the results of all the judgments; the rows labeled ***cleaned*** use the performance on Q1 to remove workers with accuracy of under 0.9. Agreement on Q1 (counting hashtags) was high across both labeling investigations. Q2 (finding names) elicited expectedly less agreement, but higher than the main task (judging tweet interestingness). Most importantly, removing the work of the judges whose accuracy on Q1 was below 0.9 did not improve agreement on Q2 and Q3. In other words, performance on the HIDDENs revealed that the workers were doing the task in good faith.

	W1 (news, recent)	W2 (news, recent)
Q1 α (all)	0.779	0.775
Q1 α (cleaned)	0.824	0.888
Q2 α (all)	0.722	0.734
Q2 α (cleaned)	0.731	0.708
Q3 α (all)	0.050	0.157
Q3 α (cleaned)	0.045	0.160

Table 5. Check of worker reliability using HIDDENs (Q1 and Q2). The rows labeled *cleaned* use the performance on Q1 to remove workers with accuracy of under 0.9.

There are several other phenomena of interest shown in Table 5. Removing the work of the judges who perform poorly on Q1 does not necessarily result in comparable performance improvements on Q2 (e.g. compare the ***all*** and ***cleaned*** columns for W2). Furthermore, the inter-rater agreement on Q1 and Q2 was roughly comparable (and high) for both datasets, but the performance on Q3 continued to show low inter-rater agreement and variability.

Comparing platforms to check expertise. Performance on Q2 suggests that it requires some expertise to do it correctly (in addition to attentiveness). In diagnosing the worker portion of our task, it seems wise to check the performance across platforms too. What would happen if we tried our task template plus HIDDENs (i.e. the template shown in Figure 4) on the AMT crowdsourcing platform? We would expect UHRS workers to perform better on Q2 (their specialty in relevance judgment would suggest they have the broad knowledge Q2 requires). Using different crowdsourcing platforms also allowed us to further explore the efficacy of the HIDDEN technique.

Table 6 documents this cross-platform comparison. All datasets consist of news tweets, drawn from different periods. As in the other tasks, each tweet has been judged by 5 workers. The table does not exclude any workers based on their performance on the HIDDEN subtasks, so we can compare worst-case workers. The data from W1 and W2 in Table 5 has been carried forward to Table 6 for the sake of easy comparison.

The performance on the first two questions (Q1 and Q2) should tell us something about worker reliability (Q1) and expertise (Q2). Table 6 shows that inter-rater agreement is similar for Q1 on both platforms, ranging from 0.800 to 0.876 on AMT and from 0.775 to 0.882 on UHRS. The difference on Q1 is negligible; workers on both platforms must care about appearing reliable. Because Q2 requires more specialized knowledge—workers must not only follow instructions, they also had to be familiar with a range of world leaders, celebrities, and other newsmakers—we expected UHRS workers to perform better on Q2 (especially since they are routinely exposed to similar tasks). Table 6 shows this was by and large not the case. AMT workers mostly did as well or better.

The results on Q3 tell us more about worker diversity than about worker reliability; if the platform attracts more diverse workers, we might expect lower inter-rater agreement. Our expectations are borne out by the Krippendorff's α scores for Q3: AMT workers are likely to be more diverse (inter-rater agreement is lower for comparable datasets). Thus, for our purposes, we might evaluate specific trade-offs between the platforms depending on a project's goals; worker reliability on both platforms appears good.

	platform	Q1 α	Q2 α	% int	Q3 α
W1 (news, recent)	UHRS	0.779	0.772	43.8%	0.050
W2 (news, recent)	UHRS	0.775	0.734	57.0%	0.157
W3 (news, older)	UHRS	0.882	0.752	48.8%	0.157
W4 (news, oldest)	UHRS	0.819	0.774	53.4%	0.190
W5 (news, recent)	AMT	0.850	0.843	55.0%	0.105
W6 (news, older)	AMT	0.800	0.840	51.0%	0.030
W7 (news, oldest)	AMT	0.876	0.734	40.2%	0.085

Table 6. Comparison of workers from UHRS and AMT platforms. Krippendorff's α is used to assess inter-rater agreement. Percent judged interesting is also shown for Q3.

4.3 Redesigning the task

Now that we have established that inter-rater agreement is not significantly improved by changes in the dataset or by worker reliability or expertise, we turn our attention to the work itself: Is there a way of formulating the task and redesigning the template such that we can improve inter-rater agreement?

In our earliest efforts, we assumed that workers would reach consensus that a small subset of tweets were inherently interesting because they referred to global events, major celebrities, or culturally pervasive memes. But it proved to be difficult to judge tweets' interestingness out of context.

Could we render the task more accessible by breaking down this disembodied notion of interestingness? The psychology literature considers interestingness to be a complex human emotion [6]. Colton and Bundy tie interestingness to plausibility, novelty, surprisingness, comprehensibility, and complexity [6]; Silvia adds curiosity-provoking to that list, and suggests that reverse measures of these properties are useful for triangulation [15]. Although it was impractical to use this literature to fully disaggregate interestingness into its component parts, specific characteristics could be used to evaluate the efficacy of this approach in redesigning the task.

The trade-off was thus to ask six simpler questions in the place of one complex judgment (interesting or not?) with the hope that greater specificity would make the work easier to do and would improve inter-rater agreement. Workers were asked to make the six decisions independently (although in practice several were mutually exclusive and could be used for triangulation).

Using our interpretation of the interestingness literature, we gave workers the ability to specify whether each tweet is some combination of: (a) worthless, (b) trivial, (c) funny, (d) piquing my curiosity, (e) useful information; or (f) important news. Although these characteristics are not comprehensive, they may be ordered along a spectrum from negative (worthless) to positive (important news). They were incorporated in our existing template with the HIDDEN questions (Q1 and Q2) to vet the workers. Figure 5 shows the redesigned template. The new template design was tested against workers from AMT, since they performed well in our previous tests of worker reliability and expertise.

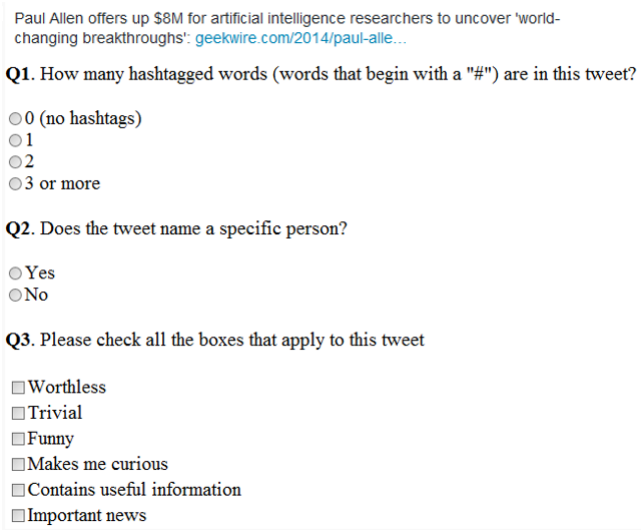

Figure 5. Redesigned task template showing HIDDENs (Q1 and Q2) and disaggregated interest characteristics (Q3).

Table 7 summarizes the results of testing the interestingness characteristics against four different datasets. T1 was the pilot for four investigations using the new task design; so we could compare it with other results, we began with a dataset of recent news tweets. T2 and T3 showed the workers older news tweets (comparable to W3 and W4 in Table 6); T4 returned to our original goal of having the workers judge random tweets; these again were old tweets. Aggregate values are computed by assigning each of the six Q3 characteristics a positive or negative value (worthless and trivial are -1; funny, curious, useful, and important are 1) and adding them together. If the value is greater than 0, the tweet is considered to be interesting. If the value is less than or equal to 0, the tweet is considered uninteresting. Krippendorff's α is calculated as it was in previous tables.

Question	T1 (news, recent)	T2 (news, older)	T3 (news, oldest)	T4 (random, oldest)
Q1	0.910	0.907	0.954	0.974
Q2	0.758	0.728	0.618	0.843
Q3 (aggregate)	0.137	0.063	0.014	0.088
Q3, worthless	0.384	0.033	0.045	-0.023
Q3, trivial	0.097	0.043	-0.061	0.025
Q3, funny	0.134	-0.016	0.169	0.049
Q3, curious	0.056	0.026	0.130	0.061
Q3, useful	0.079	0.048	0.014	0.160
Q3, important	0.314	0.207	0.000	0.170

Table 7. Krippendorff's α for four datasets after a task redesign based on a decomposition of interestingness (using concepts from [6] and [15]).

Workers performed well on the first HIDDEN subtask (counting hashtags) and acceptably on the second (identifying whether a person appeared in the tweet by name); inter-rater agreement for both HIDDEN subtasks was high. As we found in the last series of tests (see Table 5), eliminating the judgments of workers who performed poorly on the HIDDEN subtasks had minimal effect on inter-rater agreement, although this could mean that the HIDDENs had the desired effect of slowing the workers down and causing them to reflect on the tweets. We still believe this technique to be a useful addition to our arsenal of crowdsourcing tactics, especially since the HIDDEN subtasks are designed to produce useful secondary results.

Inter-rater agreement for the new decomposition of interestingness varied, depending on the age and genre of the tweets. For example, for T1 (fresh news tweets), both ends of the spectrum (worthless and important) elicited relatively high agreement. In between, agreement was lower. As the tweets aged, consistent with our other results, agreement also dropped. Interestingly, for very old news tweets (dataset T3), agreement was higher for those tweets that were *evergreen* (a journalistic term for stories that can be published at any time): funny tweets and those that provoked reader curiosity. On the other hand, the weakest signal appeared at the positive end of the spectrum, which tweets were useful or important. For random tweets (T4) inter-rater agreement was highest when workers were picking out important or useful tweets.

In other words, workers seemed to agree on the tweet characteristics that were in high contrast to the rest of the dataset. For current news, these would be the tweets that were most important and most worthless. For old news, this would be the tweets that were evergreen. Finally, the positive characteristics of old random tweets (tweets that were useful or important) stuck out against a sea of undistinguished content.

4.4 Effects of changing the task

To debug the crowdsourced labeling task, we made significant changes to three elements: the data, the workers, and the task design (see Figure 2). These adjustments all had effects on the task template, the work we asked the crowd to do. Our initial task template was simple; it displayed a tweet and posed the central question, "Do you think the tweet is interesting to a broad audience?" (see Figure 3). The next set of changes involved only the type of tweet the workers had to judge, in this case, tweets pulled from news sources. A subsequent change to the task template, the addition of HIDDENs, had a more profound effect on the work (see Figure 4). Finally, the disaggregation of interestingness characteristics into the task template shown in Figure 5 changed the work once again: for better or for worse, one judgment turned into six semi-independent judgments. (A tweet could be both trivial and worthless, or both important news and curiosity-provoking, for example).

We might expect this series of changes to the task template and the data characteristics to influence how long workers spent on an individual HIT. The aim of switching from random tweets to news was to improve the quality of the material the judges were working with (with the ultimate goal of improving inter-rater agreement). Thus we might expect the judges to spend more time pondering the relative merit of an individual tweet.

The HIDDENs implemented deliberate speed bumps. They were designed to slow the workers down and to cause them to read each tweet three times, each time with a different perspective. Although the HIDDENs were not difficult, we might expect them to add time to a worker's task performance.

The final change, the disaggregation of "interesting" into six characteristics might add a bit of drag to the work. But how much? The worker has more decisions to make (six times as many), but they should be easier decisions. We did not expect a significant change in the time spent on the task.

Table 8 compares the workers' times to perform the judgement task using the three task templates (the basic interestingness question, the template with the added HIDDENs, and the template with the disaggregated central judgment). We present a sensible progression of judgment tasks to illustrate the effects of the changes. Because the two crowdsourcing platforms may have minor differences in basic overhead (e.g. time to fetch and submit the HIT), we pivot on this change too. We use both average and median times (in seconds) to characterize task performance, specifically to compensate for some known variations in working style. Some workers are working as fast as they can; others are working against a backdrop of deliberate distractions like TV. We have chosen to use only the datasets of recent tweets in each case, just to eliminate one more factor that might influence judgment time. We refer to the judgment-only template shown in Figure 3 as "judgment only"; the template with the HIDDENs shown in Figure 4 is called "judgment+HIDDENs"; and the complete template shown in Figure 5 with disaggregated judgment and HIDDENs is called "HIDDENS+disagg". Although a few workers judging tweets in datasets B2 and G3 allowed the task to time out, there were sufficiently few who employed this strategy that it had minimal effect on the average or median times. Interestingly, once the judgments involved HIDDENS, no workers allowed the task to time out.

dataset	description: genre, platform, template	Avg. time/ tweet (sec.)	median time per tweet (sec.)
B2	random; UHRS; judgment only	2	1
G3	news; UHRS; judgment only	3	2
W1	news; UHRS; judgment+HIDDENS	14	11
W5	news, AMT, judgment+HIDDENS	18	13
T2	news, AMT, HIDDENS+disagg	26	18

Table 8. Comparison of time (in seconds) to perform judgment tasks using the three templates.

The values in Table 8 reveal several important differences. First, as we expected, it takes longer to judge higher quality content. A comparison of the values in rows B2 and G3 illustrates this difference: it might take almost twice as long to work with higher quality content. Once we shifted the focus to news tweets, it was no longer easy for workers to quickly discard tweets as uninteresting.

A comparison of the values in rows G3 and W1 gives us a sense of the magnitude of the speed bump we have introduced through the HIDDENs. Answering the two extra questions they pose more than triples the average time a worker spends on each HIT. Rather than seeming inefficient, this difference demonstrates that workers must read each tweet a little more carefully to complete the work. If the HIDDENs are well-designed—that is, they contribute to the overall data-related processing we want to do—they are worth the extra time they take. Workers should simply be paid somewhat more to undertake the redesigned task with the expectation that the quality of the work is improving.

The values in rows W1 and W5 represents the same work completed using different crowdsourcing platforms. Evidently it takes slightly (but not significantly) longer to complete the HIT using Amazon Mechanical Turk (AMT) than it does on Microsoft's internal crowdsourcing platform (UHRS). This difference could reflect simple bookkeeping differences (e.g. When does timing start? When does it end?) or network speed. In any event, the values look comparable using the two platforms.

How much does the expanded template—the version of the task that mandates six simple judgments in place of one complex subjective judgment—increase the time to perform the task? The values in row W5 and T2 help us sort out the answer to this question. To our surprise, while the final version of the template adds a bit of drag to the task, it does not increase it to an unworkably large extent. Again, considerations like this should be factored in to calculating a fair payment for the work.

Overall, the largest and most important difference is added by the introduction of the HIDDENs, which we hope will result in useful secondary data curation and produce higher-quality labels.

5. DISCUSSION

We set out to identify high quality tweets as an instance of an important problem: how to use the tremendous volume of socially produced content, much of it in the form of non-traditional documents. Tweets are very short and often cryptic; sometimes they make sense only to their immediate audience (the author's followers), and sometimes they can be presented to a much broader audience because they are perceived as interesting and important. Although there are social mechanisms for surfacing content in individual services, we are investigating a more general crowdsourced judgment process that is both timely and thorough. Initially we hoped to use a standard relevance judgment crowdsourcing process (which identifies the desired content through inter-rater agreement, and vets the workers by the degree to which their judgment aligns with their peers), but our experience demonstrated the need for a new way of handling this type of problem. Subjective assessment is key to working with socially produced data.

We have investigated a "data-workers-task design" process by varying each of these three crucial elements in turn to evaluate their influence. Our goal was to develop an effective way of debugging this type of crowdsourced labeling task, one with low inter-rater agreement. We also sought to develop a cost-effective and respectful way of vetting workers, since agreement with a norm was not going to be an effective way of assessing worker reliability and expertise.

5.1 A process for debugging

The key to our diagnostic process is to adjust each element in turn using small datasets. This allows us to try multiple combinations without worrying about whether to keep or discard the data; production runs are costly, and this technique enables us to debug the process beforehand.

Our first step was to pick detectable genres that are apt to yield a greater density of interesting content. If the interesting content is sparse and its overall utility is ambiguous, the task may begin to seem meaningless to the workers [3]. Although we ultimately intended to return to random tweets (since that is what the real data source consists of), switching to news tweets revealed that the low inter-rater agreement could not be attributed to the dataset.

Next we combined the news genre with worker-oriented changes: Were the workers working in good faith? We were aware that we were asking a subjective question, but it was the type of question

that we needed to ask. Many questions about social media have to do with value, and labels associated with content value or relevance can be subjective. Furthermore, we were interested in finding out if it was necessary to use the expertise-rich UHRS crowdsourcing platform, or determining whether AMT's worker diversity was advantageous for our task. Answering this question about platforms enabled us to experiment without further burdening the more costly (and slower turn-around) resource.

Finally, we discovered that it was useful to redesign the task itself. In our example task, teasing apart what we meant by "interesting" can reduce the task's complexity (workers were asked to judge qualities that are psychological components of the higher-level emotional concept). It may be easier for a worker to say whether a tweet is funny or potentially useful, than it is to assess the content's overall interestingness. The interestingness literature suggests two other variations we have yet to try. The first is to ask the questions as a negative rather than a positive. Research in this area tells us it may be more straightforward to detect the absence of a characteristic rather than its presence [15]. In other words, it may be easier for workers to tell us that content is useless, than to say that it is useful, or that it is not funny, rather than funny. The characteristics worthless and trivial are examples of a negative approach, although aggregating these two qualities may prove to be more effective. Economics literature also suggests that we ask the question in a way to distance it from the individual's own judgment. In other words, it might be better to ask, "Will others find this content useful?" rather than asking, "Do you find this content useful?"

Although we focused on tweets as an example of a socially-produced data source and on interestingness as the judgment characteristic, we believe that other complex judgement concepts (e.g. value) can be similarly disaggregated.

5.2 HIDDENs

Our second goal was to develop a method for assessing worker reliability that:

(1) incorporates worker-independent judgments to achieve inter-rater agreement on a per-microtask basis;

(2) contributes to the main question we were asking (in other words, doing this extra work can improve label quality and will provide a single point of focus, the content under judgment);

(3) does not seem like a meaningless attention check to the workers (although the workers expect checks like this, they appreciate work that is more meaningful or is designed to recognize their humanness [12]); and

(4) produces useful results, possibly for a colleague, for follow-up research.

Recent crowdsourcing work has found that workers also like being given a HIT that's essentially a break in the action, a task that's just fun (e.g. read this cartoon), which gives them a chance to catch their breath and go back to the central work refreshed [7]. Although our current task design is not particularly entertaining, because workers do a lot of these judgment tasks, working quickly, this second type of break may prove to be useful.

To fully evaluate the HIDDENs, we need to establish whether the order that we ask the questions has any unanticipated effects: Would we get the same results if we asked the more difficult named entity detection task first? Are workers put off by the computability of the first task (counting a tweet's hashtags)? It would also be useful to develop other types of subtasks along a spectrum of subjective judgment and future utility.

6. CONCLUSION AND FUTURE WORK

Our data-workers-task design process has proven effective in rapid-turnaround debugging of a difficult labeling task. The HIDDEN technique also shows promise as an avenue for measuring worker reliability and potentially improving label quality. Finally, we are beginning to understand how to work within the confines of a more subjective question that depends on the judges' preferences and backgrounds.

In addition to the research we identified in the Discussion section, we are pursuing a more sophisticated incentive structure that will allow us to reward workers who are adept at predicting how their peers will label data. We are also investigating how the HIDDENs can draw reliability questions from a library of pre-categorized subtasks (that is, the subtasks must address the same data sources, and must either be somewhat objective, or subjective only to the extent that the answer can be readily determined by inter-rater agreement). Finally, we are drawing on multi-disciplinary literature to learn how to test the reliability of the data (labels) we are gathering. Taken together, advances in all of these areas will be an important step in improving access to social data, and in using subjective judgments in a variety of data-driven applications.

7. REFERENCES

[1] Alonso, O. Implementing Crowdsourcing-based Relevance Experimentation: An Industrial Perspective. *Information Retrieval*, 16(2), 2013, 101-120.

[2] Alonso, O., Carson, C., Gerster, D., Ji, X., and Nabar, S. Detecting Uninteresting Content in Text Streams. *Proceedings of CSE 2010*, 39-42.

[3] Alonso, O., Marshall, C. and Najork, M. Are Some Tweets More Interesting Than Others? #HardQuestion. *Proceedings of HCIR 2013*, pp. 2:1--2:10.

[4] André, P., Bernstein, M., and Luther, K. Who Gives a Tweet?: Evaluating Microblog Content Value. *Proceedings of CSCW 2012*, 471-474.

[5] Aroyo, L. and Welty, C. Harnessing Disagreement in Crowdsourcing a Relation Extraction Gold Standard. Tech. Rep. RC25371, IBM Research, 2013.

[6] Colton, S., Bundy, A., and Walsh, T. On the Notion of Interestingness in Automated Mathematical Discovery. *International Journal of Human-Computer Studies*, 53(3), 2000, 351-375.

[7] Dai, P., Rzeszotarski, J., Paritosh, P., and Chi, E. And Now for Something Completely Different: Improving Crowdsourcing Workflows with Micro-Diversions. *Proceedings of CSCW 2015*, 628-638.

[8] Josephy, T., Lease, M., and Paritosh, P. (Eds.). Crowdsourcing at Scale Workshop. *Proceedings of HCOMP 2013*.

[9] Kittur, A., Chi, E., Suh, B. Crowdsourcing user studies with Mechanical Turk. *Proceedings of CHI 2008*, 453-456.

[10] Krippendorff, K. *Content Analysis*. Sage, 2004.

[11] Lin, T., Etzioni, O., and Fogarty, J. Identifying Interesting Assertions from the Web. *Proceedings of CIKM 2008*, 1787-1790.

[12] Marshall, C.C. and Shipman, F.M. Experiences Surveying the Crowd: Reflections on Methods, Participation, and Reliability, *Proceedings of WebSci 2013*, 234-243.

[13] Metzler D. and Cai, C. USC/ISI at TREC 2011: Microblog track. *Proceedings of TREC 2011*.

[14] Momeni, E., Tao, K., Haslhofer, B., and Houben, G. Identification of Useful User Comments in Social Media: A Case Study on Flickr Commons. *Proceedings of JCDL 2013*, 1-10.

[15] Silvia, P. What is Interesting? Exploring the Appraisal Structure of Interest. *Emotion*, 5, 2005, 89-102.

[16] Sultan, M.A., Bethard, S., Sumner, T. Towards automatic identification of core concepts in educational resources. *Proceedings of JCDL 2014*, ACM, 379-388.

[17] von Ahn, L., Blum, M., and Langford, J. Telling Humans and Computers Apart Automatically. *CACM*, 47(2), 2004, 57-60.

[18] von Ahn, L., Maurer, B., McMillen, C., Abraham, D., and Blum, M. reCaptcha: Human-Based Character Recognition via Web Security Measures. *Science*, 321(5895), 2008, 1465-1468.