Crowdsourcing a Subjective Labeling Task: A Human-Centered Framework to Ensure Reliable Results

Omar Alonso
Microsoft Corp.
omalonso@microsoft.com

Catherine Marshall
Microsoft Corp.
cathymar@microsoft.com

Marc Najork*
Google
najork@acm.org

ABSTRACT

How can we best use crowdsourcing to perform a subjective labeling task with low inter-rater agreement? We have developed a framework for debugging this type of subjective judgment task, and for improving label quality before the crowdsourcing task is run at scale. Our framework alternately varies characteristics of the work, assesses the reliability of the workers, and strives to improve task design by disaggregating the labels into components that may be less subjective to the workers, thereby potentially improving inter-rater agreement. A second contribution of this work is the introduction of a technique, Human Intelligence Data-Driven Enquiries (HIDDEN), that uses Captcha-inspired subtasks to evaluate worker effectiveness and reliability while also producing useful results and enhancing task performance. HIDDEN subtasks pivot around the same data as the main task, but ask workers to perform less subjective judgment subtasks that result in higher inter-rater agreement. To illustrate our framework and techniques, we discuss our efforts to label high quality social media content, with the ultimate aim of identifying meaningful signal within complex results.

Categories and Subject Descriptors

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

General Terms

Experimentation, Human Factors

Keywords

Crowdsourcing, label quality, relevance assessment

1. INTRODUCTION

Relevance assessment and item classification form the cornerstone of most modern information retrieval approaches. Because information retrieval corpora are often very large, it is impossible to use humans to label the entire collection. Instead, representative parts of the collection are labeled by hand, and predictive features are identified based on this subset so that machine learning approaches can be used to process the rest. It is easy to see why it is essential to focus on eliciting reliable high quality labels early in the process: these labels, which are often difficult to assign and assess, will determine future performance and outcomes [1].

Frequently this core set of labels is obtained though human computation or crowdsourcing: workers label representative elements from the dataset using a fixed set of descriptors. Multiple judges may label the same item so that inter-rater agreement may be used to compensate for possible ambiguity or to resolve labeling discrepancies. But what happens if the workers fail to agree? Does this mean that the labeling activity is simply too subjective?

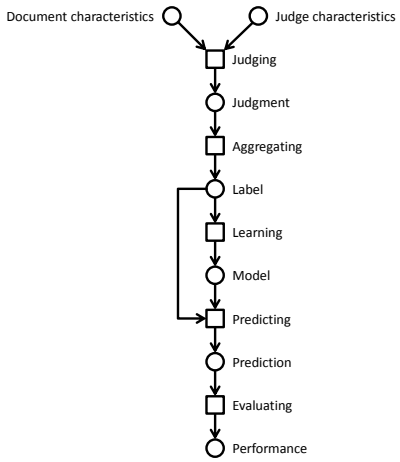

Figure 1: From document assessment to performance evaluation.

To formalize what we mean by a *subjective labeling task*, it is helpful to understand how labels are created and used in various information retrieval settings. Figure 1 illustrates such a process. Data are shown as circles, and functions are shown as boxes. Two factors contribute significantly to the labeling process: the characteristics of the documents that are to be labeled and ultimately evaluated by the retrieval system, and the characteristics of the judges that perform the labeling. Document characteristics may include their content, their provenance, creation date, consumption and citation statistics, and so on. Judge characteristics include

*Work was performed while author worked for Microsoft.

their demographic traits (e.g., their age and gender), knowledge (e.g., language skills) and predispositions (e.g., preferences and beliefs). Each judge is given a set of items to judge, for example to decide whether the item belongs to a certain class.

It is important to stress that each judgment depends on both the characteristics of the item *and* the characteristics of the judge. Once judgments have been obtained, the various judgments for each item are aggregated (for example averaged, or converted to a binary judgment by a majority-wins rule), producing a set of labeled items. These labeled items are being fed into a machine learning algorithm that attempts to infer the best-possible model explaining the labeling of the items based on their characteristics. The inferred model is used to predict the labels of a *holdout set* of items that have been labeled but not used during the model construction. The predicted and actual labels of these items form the input to an evaluation measure, which is used to quantify how well the inferred model captures the judgments of the human judges.

To understand how the judges' characteristics inform their judgments, and to convince ourselves that judgment-sets with close-to-zero inter-rater agreements can still produce models that have perfect prediction performance, consider the following example: A pool of 51 male and 49 female judges is given a set of 100 photos, 60 of them showing male and 40 showing female subjects' faces, and asked to decide whether the photo depicts a person of the opposite gender. The gender of each judge and the gender of each subject are characteristics that factor into the judgments. Assuming perfect judgments and aggregation by averaging, 60 photos will be labeled 0.49, and 40 will be labeled 0.51. By conventional measures, inter-rater agreement is close to 0, but most learning algorithms (say, a decision-tree based one) will determine that the subject's gender is the characteristic that is predictive of the label. The model will then perfectly predict the labels of the photos in the holdout set.

In the above example, the task given to the judges was not "subjective" in the colloquial sense – for each pairing of assessor and item, there was a "correct" answer. If we change the setting to a political poll, with 51 conservative and 49 liberal voters (judges) asked to approve 60 conservative and 40 liberal candidates (items), the task given to the judges sounds more subjective, but all of the observations we drew from the first example still hold: inter-rater agreement will be close to 0, and nonetheless we can derive a model that, for a candidate with a known political affiliation (characteristic) will predict the rate of approval (the label). So, for the purposes of this paper, we define a *subjective labeling task* to be one where the judgment process strongly depends on characteristics of the judge, whereas an *objective labeling task* is one where judgments solely depend on the document characteristics. Table 1 explains how our task, represented by the figure's bottom row, compares with other types of labeling efforts that have been described in the literature.

We have been investigating a particular example of a subjective labeling exercise, identifying high-quality content in socially produced data. If we can create a classifier to identify high-quality content, it will enable us to create a range of practical applications. For example, the ability to identify interesting tweets will allow us to selectively index the feed, greatly compressing the size of the index and increasing the feed's utility to a wide range of users.

Nature of task	Aggregation Approach	Evaluation technique
Objective question has a *correct* answer (objective)	Reliable judge assigns appropriate label for an item	Evaluate workers by comparing individual results with gold set
Judgment question has a *best* answer (partially objective)	Inter-rater agreement determines label for an item	Evaluate workers by comparing individual results with consensus
Subjective question has *consistent* answer (subjective)	Repeatable polling determines probability of a label for an item	Evaluate workers by computing the consistency of results between groups

Table 1: A spectrum of labeling tasks, from objective to subjective.

Thus in a practical sense, we are still faced with the need to address the early phase of the end-to-end process shown in Figure 1, acquiring a high quality labeled dataset. We have identified three potential sources of quality shortfalls: (1) the work itself (e.g. the reduced dataset intended to represent the larger whole and the choice of labels the workers are asked to apply to it); (2) the workers doing the labeling (e.g. the crowd's reliability and collective expertise); and (3) the task design (the way the task is presented to workers). These three elements – work, workers, and task – are contingent on one another: adjusting one element may have a meaningful effect on the others. In our work, we explore the nature of the task, and ways to ensure that reliable labels have been assigned to a scalable subset of the larger dataset.

Through a series of labeling experiments aimed at assessing tweet interestingness, we set out to make the following contributions:

- Test a "work-workers-task design" framework to improve label quality when the task involves asking the crowd a subjective question to elicit the label. (i.e., a framework that addresses the third row of Table 1);
- Investigate a Captcha-inspired way to evaluate worker effectiveness and reliability that produces useful subresults while it enhances task performance;
- Develop a technique that allows us to assess content quality in such a way that we can identify meaningful signal to use in a machine learning setting.

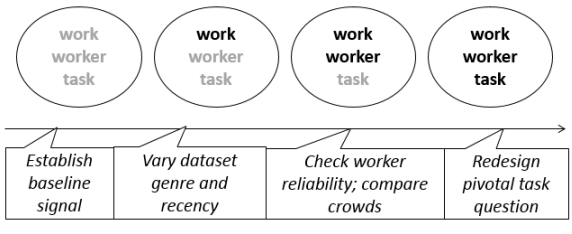

Figure 2: Framework and flow for investigating quality.

It is important to adjust and fully debug the labeling process before it is scaled up and put into continuous production; mistakes at scale are expensive and may be far-reaching. Figure 2 shows our general strategy for ensuring label quality and worker reliability. First we established a baseline by running an initial series of labeling tasks; this series showed us what to expect in the way of inter-rater agreement. Next we began to vary the dataset genre with the idea of improving the quality of the tweets we put in front of the judges. As the crowdsourcing literature suggests, we began to wonder if the workers were going too quickly, or if the crowd we were consulting was ill-suited to the task. In so doing, we discovered a new method for checking work reliability that would contribute to label quality and produce additional useful results. Finally, we adjusted the task design in an effort to reduce the subjectivity of the question, thereby potentially improving the inter-rater agreement. Figure 2 shows how we added in different elements of our framework into the test (they turn from gray to black). Ultimately, our aim is to produce a fully reliable outcome: the same process should yield the same results, reflecting an experimental maturity.

In this paper, we first review multiple areas of related work. Next we describe our general method for conducting these labeling experiments. We then discuss our results specific to each part of the framework (workers, task design, work). Finally we draw conclusions about this type of subjective crowdsourcing task and how to measure reliability.

2. RELATED WORK

We relied on four bodies of related work to inform our efforts. First, we consulted research that was close in intent to ours: using crowdsourcing to evaluate content quality, especially efforts to identify high-quality tweets or other very short documents, including closely related efforts within the TREC community. Second, we pursued efforts to develop statistical measures of agreement between multiple raters and to assess experimental reliability. Because we had framed our investigations of tweet quality in terms of interestingness, we turned to psychology literature on the nature of interestingness. Finally, we sought to improve our crowdsourcing process by borrowing techniques and findings from related work on crowdsourcing at scale, crowdsourcing in general, and validating worker reliability. We discuss each area in turn.

Evaluating content quality. The fundamental aim of this work is to identify high-quality content in very short documents, especially tweets and comments. André, Bernstein, and Luther [4] do this by relying on self-selected volunteers to accumulate ratings (worth reading, okay, and not worth reading) on tweets from accounts that they follow to characterize what makes a tweet worth reading. 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 similar crowdsourcing approach to labeling 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 we have based our research on. 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; we are building on these results to understand if we can improve on the automatic detection results.

Measuring agreement and experimental reliability. In general, much of the related work is contingent on identifying high quality content through worker agreement. What is an acceptable minimum signal upon which to base a binary classifier? Alonso, Marshall, and Najork were only able to achieve moderate agreement (kappa = 0.51) between crowd-labeled tweets and a classifier [3]. The work we describe in this paper has an emphasis on improving the signal by multiple means, while trying to establish a minimum acceptable signal. What happens if the workers disagree? Aroyo and Welty have investigated the idea of productively using disagreement among judges [5]; like them, we are exploring a crowdsourcing task that is sufficiently subjective that we do not expect the workers to reach consensus. In Section 3 of this paper, we discuss our current reliance on Krippendorff's alpha [10] and Fleiss's kappa [7] as baseline measures of agreement, with the idea that these measures must be supplemented with another metric.

A psychological notion of interestingness. Although interestingness is a notion that is used intuitively by much of the research we have cited above (for example, Lin, Etzioni, and Fogarty refer to it as a social construct, best identified symptomatically, e.g. by retweets [11]), the psychology literature takes a more nuanced look at interestingness as a complex human emotion [6, 16]. Colton and Bundy tie interestingness to plausibility, novelty, surprisingness, comprehensibility, and complexity [6]; Silvia adds curiosity-provoking to that list, and also suggests that reverse measures of these properties are useful for triangulation [16]. In Section 4.3, we describe how we apply these insights to our task design.

Crowdsourcing techniques. Crowdsourcing at scale has been the subject 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 [1]. The reliability of (and indeed humanness) of workers has generally been a focus of von Ahn, Blum, and Langford's Captcha research [17]; this work has also branched off in a direction in which the Captchas produce useful work as re-Captchas [18]. Captcha-like techniques were first introduced to crowdsourced user studies by Kittur, Chi, and Suh [9]. Because user studies tend to involve more effort per HIT than labeling tasks, we were informed by this prior work, but we needed to take a slightly different approach than all of these predecessors to ensure worker reliability and to enhance our labeling efforts; this approach is described in Section 4.2.

3. METHOD

In this section, we discuss the method we have used for eliciting labels from workers, and how it has evolved through the course of the project. In earlier research, we assumed that inter-rater agreement would yield reasonable labels, given a representative dataset of tweets. Hence we used datasets that contained between 2,000 and 10,000 tweets drawn at random from recent samples of the Twitter firehose. Each tweet was labeled by 5 judges to arrive at an appropriate label by consensus. We also explored a variety of label sets, taken from the literature (and modified if

necessary) in an effort to design labels that were both more expressive and a better match for the data; our thought was that labels that better matched the data would make them easier for the workers to assign. We quickly realized there was a discrepancy between how the judges reacted to the data (in human terms), and how we saw the data (in analytic or computational terms). These labeling exercises neither improved inter-rater agreement, nor provided us with the high-quality labeled data we would need if we were going to take a machine learning approach. However, they highlighted the importance of debugging the different aspects of the task to enable us to extract a meaningful signal.

Thus we set out on this phase of our research with more focused contributions in mind (beyond simply identifying high quality content). To do this, we scaled back the dataset size, and ran more iterations of the individual experiments, gradually adjusting elements from each of the three contingent parts of the framework.

The first explorations we performed were directed at the datasets we were asking the workers to label. Perhaps we were giving workers tweets that were too wide-ranging (from conversations and bon mots to breaking news and product endorsements), and of too low overall quality: reading tweets in those early datasets was discouraging, and we sensed judging so many obviously low quality tweets might be frustrating for workers as well. Possibly the low number of positives would either lower the workers' standards, or cause them to miss high quality content when they saw it. We decided to shift our attention to tweets that were, by virtue of the accounts that tweeted them, in the news genre. Because we thought workers might have various biases toward (or against) particular news agencies (e.g. CNN or the Wall Street Journal), we "de-branded" them by removing the account name before we asked workers to judge them. Figure 3 shows the baseline task design as we shifted data genres. We also simplified the labels to a binary labeling scheme.

Please read the tweets below and mark the ones that you think are interesting. You can mark more than one. Also, if you think that none of them are interesting, please don't make any selections. Tip: Some tweet will be hard to label. Please try to be consistent.
Task: Please mark the tweets that you think are interesting. Feel free to give us more feedback in the box near the respective tweet.

[] *EU to consider listing Hezbollah as terrorist group http://t.co/6WTKAPOz* [input feedback]
[] *Sheepdogs Sophie and Sarah Become Viral Stars http://t.co/QPBxNpZg* [input feedback]
[] *Corruption case threatens Spain's ruling party - and its economy http://t.co/aAc5ZjYe* [input feedback]
[] *All beef products must be tested for #horsemeat by next Friday after reports of #Findus contamination* [input feedback]
[] *The great economic experiment of 2013: Ben Bernanke vs. austerity* [input feedback]

Figure 3: Task design for judging news tweets.

Subsequently, we realized that our datasets were still unnecessarily large to resolve the research questions we were pursuing; the datasets could be scaled once we'd fully debugged each contingent element of our framework. In other words, we could perform a greater number of smaller experiments, and still find sufficient statistical significance to debug them. We reduced the dataset size for each investigation to 100 tweets (500 judgments) and changed the statistical measures we were using as we better understood the judgment's subjectivity. Table 2 shows an overview of the investigations described in the remainder of this paper, highlighting which aspects of the framework we varied in accordance with the flow described in Figure 2. Specific outcomes and details of the changed framework elements are discussed in the appropriate sections.

When we turned our attention to the workers, we used two different crowds: workers recruited from an internal crowdsourcing platform which specializes in relevance judgments and workers recruited from Amazon Mechanical Turk (AMT). There were distinct trade-offs between the two platforms in addition to worker expertise. Workers from the internal crowdsourcing platform were paid more than AMT workers (according to the market rate on each), and workers on AMT were more numerous, so we could get faster turn-around on our debugging investigations. In all cases, we recruited judges who were interested in doing labeling tasks and who were familiar with Twitter data. But were we really taking advantage of worker expertise the higher cost platform offered us?

Clearly, 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 two standard measures of inter-rater agreement: Krippendorff's alpha and Fleiss' kappa. Both produce values between 1 and -1. A value of 1 indicates perfect agreement among workers, a value of 0 indicates that workers are assigning labels randomly; and a negative value indicates that disagreements are systematic. Krippendorff's alpha has the advantage to handle data sets where the number of raters per item varies, which is the case for some of crowdsourcing experiments as we will see later. As explained in Section 1, both measures are meaningful only for objective labeling tasks, where characteristics of the workers do not factor into the labeling tasks.

4. FRAMEWORK

In this section, we will discuss the results of each set of explorations, with the ultimate aim of validating the framework that is an important part of our research outcome. As we discussed in Section 3, our first variations were aimed at discovering whether we could elicit higher agreement by narrowing the data genre to news. We then turned our attention to the workers: could we develop a method to evaluate the quality of their work and to improve the efficacy of their results? Finally, we scrutinized the task design: would a user-centered process of label assignment, one that considered the emotional components of interest, reduce the task's cognitive load as well as improve inter-rater agreement?

4.1 Work: narrowing the data genre

Our first area for investigation was the dataset genre: if we started with a dataset containing only very recent news tweets, would the limited genre be more likely to result in agreement? Although people have differing levels of interest in some types of news stories (e.g. in our initial

Flow stage	Additional framework elements affected	HIT IDs
Establish baseline signal	None (dataset recency)	B1-B2
Vary dataset genre and recency	Work (dataset genre)	G1-G3
Check worker reliability	Work, worker (platform)	W1-W7
Redesign pivotal task question	Work, worker, task (task design)	T1-T4

Table 2: Associating HITs with framework and flow.

probe, some workers said they were attracted by celebrity stories; others said that celebrity stories were beneath contempt), 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 news sources: *@latimes*, *@reuters*, *@nytimes*, *@WSJ*, *@USATODAY*, *@washingtonpost*, *@csmonitor*, *@ABC*, *@BloombergNews*, and *@BBCNews*. An example of a typical tweet from one of these sources (in this case Reuters) is: *EU to consider listing Hezbollah as terrorist group http://t.co/6WTKAPOz*. Because we had previously observed that recency may influence the workers' assessment of the tweets, we drew 2,500 random news tweets for dates in February 2011, 2012, and 2013. Naturally, some were more time-sensitive than others; likewise, the news subgenres varied from human interest tweets: *Sheepdogs Sophie and Sarah Become Viral Stars http://t.co/QPBxNpZg* (from *@ABC*) to sports *Each team to have warm-up period once power restored: NFL officials* (from *@BloombergNews*), business *Zip codes don't make entrepreneurs. Why startups can thrive anywhere: http://t.co/rkJjfqjc* (from *@WSJ*), and hard news *Corruption case threatens Spain's ruling party - and its economy http://t.co/aAc5ZjYe* (from *@csmonitor*). These results we could now compare with our baseline labels from previous judgment tasks. Table 3 shows these results.

Confining the tweets to news, especially recent news, improved the quality of the dataset as far as the judges were concerned (over 29% of the recent news tweets were assessed as interesting, dropping to 21% when the news was several years old, as compared to under 17% for fresh random tweets, and 14% for older random tweets). Inter-rater agreement also improved from random tweets to news (Krippendorff's α went up), and from older tweets to more recent (this effect seemed even more pronounced in the dataset of random tweets). However, even with this genre shift in the dataset, the inter-rater agreement was still fairly low.

Were the workers reading the tweets carefully? We had begun this work with a fairly optimistic perspective on whether the judges would agree on the interestingness of most tweets, thus allowing us to not only label the tweets, but also to evaluate the workers. Workers who consistently disagreed with the norm, we felt, could be eliminated as working too carelessly or reading too quickly; most people would be able to identify interesting tweets just as surely as they could detect relevance. By this time, we were questioning this assumption – apparently, even reliable workers were tending to disagree – and we were ready to take a closer look at the workers, especially with an eye toward vetting their work.

4.2 Workers: quality and platform

The crowdsourcing literature urges us to focus attention on the workers and the quality of their output. Unreliable performance, either as a result of fatigue, frustration, carelessness, or out-and-out fraud needed to be ruled out. 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?

To reliably assess workers' diligence, we drew on two existing forms of worker checks. First we considered the attention or comprehension checks designed for crowdsourced user studies [9]: these studies recruit participants from the crowd, and must therefore ensure that workers are completing surveys or questionnaires in good faith. To do this, they sometimes interrupt the worker with unrelated questions that simply make sure the worker is still paying attention; the workers themselves have dubbed these *attention checks*. A clever attention check not only catches workers who have fallen asleep on the job; if they are designed correctly, they also serve to engender good will between worker and requestor [12]. Sometimes, if a HIT involves substantial instructions or task-related reading, it will include a *memory check* (again, a worker designation). These are questions directly stemming from the reading; they ensure that the worker is reading careful and may act as a speed bump to keep workers from working more quickly than is prudent for comprehension. Neither of these checks seemed appropriate for our work as-is, since our work involves many separate judgment tasks rather than one sizable HIT (e.g., a survey). Furthermore, since the content of the HIT changes each time, we can't design a comprehension-based memory check, and any attention check would be counterproductive, because it would add an annoying burden to each task.

To address these shortfalls, we considered the notion of Captchas, in which spam detection relies on the results of a microtask, such as OCR correction [17]. Successful completion of a Captcha makes it likely that a real human is at the keyboard rather than a bot. Thus spam detection using Captchas has the beneficial side effect of completing useful work.

We built on these ideas (vetting workers through meaningful microtasks embedded in the work), but we also added a third goal, improving the quality of the work. Instead of being orthogonal to the work we were requesting, the additional microtasks were designed to focus the workers' attention on the tweets they were about to judge. We call these specialized within-task Captchas Human Intelligence Data-Driven Enquiries (HIDDEN), because they achieved the following two quality-related goals, while allowing the task to be performed reliably in the absence of a gold set.

1. To complete the HIDDEN microtasks, workers were asked to read the tweet three different times, attending to different aspects of the short post. In other words, the worker had to reflect on the tweet in multiple ways. What did the tweet's author want to emphasize? Was the tweet about a specific person? The subtasks did not distract the worker from the primary task; instead they built up to it.

	News			Random	
	2011 G1	2012 G2	2013 G3	6/18/2012 B1	8/20/2012 B2
# workers	46	47	49	19	30
# tweets	2500	2500	2500	2000	2000
%interesting	21.3%	27.8%	29.3%	16.7%	14.3%
Krippendorff's α	0.037	0.074	0.068	0.013	0.052

Table 3: Comparison of inter-rater agreement for news tweets and random tweets; each tweet has been judged by 5 workers.

2. Each of the embedded microtasks should represent a different row from Table 1. The first subtask should be completely objective (as well as computable); the second should be partially objective (so worker agreement was sufficient to determine a consensus answer); and the third, the problematic original task, could be much more subjective.

The first of the two embedded microtasks elicited results that could be reliably computed: the worker was asked to count the number of hashtags in the tweet. The results of this microtask brought some data anomalies to our attention, and required the worker to read through the tweet. The second embedded microtask required additional thought and judgment; workers were to assess whether a tweet was about a specific person, signaled by the presence of a proper noun in the tweet. We further stipulated that the name could neither be an account name (@name) nor a hashtag (#name), possibly forcing the worker to look more carefully at the instructions.

Results of HIDDEN work. Q1 relies only on characteristics of the tweet (i.e., it is an objective question), Q2 relies on the worker's knowledge (i.e., being aware that Mubarak is a person, not a place), and Q3 probes their preferences. We anticipated very high agreement on the first subtask (repeated disagreement with the norm meant the worker was suspect) and good agreement on the second, depending on the breadth of a worker's awareness. These HIDDEN subtasks allowed the task to be performed reliably in the absence of a gold set and gave a way to assess the individual worker's skill. Figure 4 summarizes the new task structure. Note that repeated anomalies in the answer to Q1 may reveal problems in the dataset or can inform data science or analytics questions. Q2 may be a partially objective question whose results are useful to a colleague, or to a related task (this way work can be interlocked, so spam detection on one task can be useful results for another). Q3 is the original subjective question. By design, Q1, Q2, and Q3 should be tied together by a single piece of content being judged.

Did the new task structure identify workers who weren't performing up to par? Table 4 shows the first investigations that used it. The first important result the HIDDEN questions revealed was that the workers were not the problem. Table 4 shows that the agreement on the first subtask (counting hashtags) was indeed high across both labeling experiments using similar datasets on the same platform, and the second was expectedly less, but nowhere near as low as the main task, judging the interestingness of the tweet. More importantly, removing the work of the judges who failed Q1 (accuracy < 0.9) did not improve agreement on Q2 and Q3.

In the past, we have relied on an internal crowdsourcing platform, UHRS, which recruits workers with specific relevance judgment expertise; in fact, we have come to rely on a small subset of UHRS workers who have done such tasks for us in the past. How much does this expertise influence the labeling results? These judges are generally paid more than AMT workers who do similar tasks, but who may have fewer specific qualifications. Hence as part of our worker quality investigations, we compared the inter-rater agreement and other aspects of task performance between the two platforms. Table 5 documents this comparison. The datasets all consist of News tweets, drawn from different periods. Labels are true (interesting) and false (not interesting). Each tweet has been judged by 5 workers. Note that the table does not exclude any workers based on performance on the HIDDEN subtasks; this allows us to compare between worst-case workers.

The performance on the first two questions (Q1 and Q2) should tell us something about worker reliability (Q1) and expertise (Q2). From Table 5, we can see that inter-rater agreement is similar for the first question, ranging from 0.799 to 0.875 on AMT and from 0.775 to 0.881 on UHRS, indicating most of the workers got it "right" on both platforms. This is unsurprising; it is a question requiring little expertise. Q2 requires more specialized knowledge: workers must not only follow instructions (they were told not to count names that appeared as hashtags or as account names, preceded by an "@"); they also had to be familiar with a range of world leaders, celebrities, and other newsmakers. This is a question that we would expect UHRS workers to perform better on since they are routinely exposed to similar tasks. This was by and large not the case. AMT workers did as well or better in most cases.

The subjectivity of the third question (Q3) tells us more about worker diversity than about worker reliability; if the platform attracts more diverse workers, we might expect less inter-rater agreement. Indeed our expectations are borne

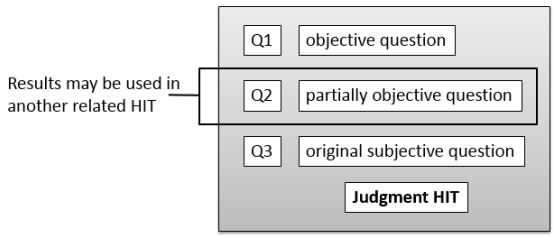

Figure 4: Basic structure of HIDDEN work. Q1 is a structural task with a computable answer that may inform data science or analytics questions; Q2 may be designed to be useful for another application; and Q3 is the original subjective judgment.

HIT ID	Platform	Datasets	Q1 α (all)	Q1 α (bad removed)	Q2 α (all)	Q2 α (bad removed)	Q3 α (all)	Q3 α (bad removed)
W1	UHRS	News-2013	0.779	0.824	0.722	0.731	0.050	0.045
W2	UHRS	News-2013	0.775	0.888	0.734	0.708	0.157	0.160

Table 4: Initial check of worker performance using the HIDDEN work (Q1 and Q2).

HIT ID	Platform	Datasets	Q1 κ	Q1 α	Q2 κ	Q2 α	% true	Q3 κ	Q3 α
W1	UHRS	News-2013	0.778	0.771	0.772	0.772	43.8%	0.048	0.050
W2	UHRS	News-2013	0.775	0.775	0.734	0.734	57.0%	0.155	0.157
W3	UHRS	News-2012	0.881	0.882	0.752	0.752	48.8%	0.156	0.157
W4	UHRS	News-2011	0.819	0.819	0.774	0.774	53.4%	0.188	0.190
W5	AMT	News-2011	0.875	0.876	0.734	0.734	40.2%	0.083	0.085
W6	AMT	News-2013	0.850	0.850	0.843	0.843	55.0%	0.103	0.105
W7	AMT	News-2012	0.799	0.800	0.840	0.840	51.0%	0.028	0.030

Table 5: A comparison of performance between workers from UHRS and AMT platforms, labeling news datasets from varying years. Fleiss's κ and Krippendorff's α are used to assess inter-rater agreement on each of the three questions (the two HIDDEN questions, Q1 and Q2, and the primary interestingness question, Q3).

out by the kappa and alpha scores for Q3: workers are likely to be more diverse on AMT (inter-rater agreement is lower across the board for comparable datasets). Thus, for our purposes, we might evaluate specific trade-offs between the platforms depending on the goals of a particular project; worker reliability on both platforms appears good.

4.3 Task design: reducing cognitive load

By the time we had reached this point in our experimentation, we were aware that we were asking a difficult and subjective question. We'd thought early on that a small subset of tweets would pop out of the dataset as inherently interesting because they referred to big events, major celebrities, or culturally pervasive memes, and that the crowdworkers would reach some core consensus about what was most universally interesting. After all, regular Twitter users do just that when they retweet or favorite a tweet in their own feed. They are able to distinguish what is interesting in an impressionistic "I'll know it when I see it" way, although research has shown that they are taking into account the perceived interests and tastes of their followers [3]

So again we reflected on what we were asking workers to do. Why was it so difficult for them to agree? Was the task simply too dependent on an individual's perspective? We began to look into the nature of interestingness: what makes something interesting? How can an understanding of the psychological concept of interestingness be reflected in our task design? Interestingness, according to the psychology literature, is a complex emotion [6, 16]. Furthermore, others have used similar tactics to establish whether an item is interesting or not [11].

We already knew that workers looked for (or avoided) a variety of characteristics in their own Twitter feeds (e.g. personal tweets are variously sought and shunned by readers [3]), but these characteristics were too numerous and detailed (and sometimes controversial) to ask workers about one-by-one, and did not reflect the psychological interpretation of interestingness as an emotion. Perhaps some specific characteristics of tweets could be used in conjunction with a more generic sense of what makes something interesting to design a better, more usable template for workers to assess the tweets by incorporating questions that were either easier to answer (closer to the worker's initial response) or less subjective. Although we could not cover the complete ground of what makes a tweet interesting, we could extract a certain number of distinctive properties and ask about them individually to see if we could reach better inter-rater agreement on them.

The trade-off was thus to ask six judgment questions where we had begun with one (interesting or not?) with the hope that greater specificity and a closer match to the components of interestingness would ease the task's cognitive load for the workers. Workers were asked to make the six decisions independently, with the idea that one characteristic would not preclude another (although several were, in fact, mutually exclusive). From our interpretation of the interestingness literature, we gave workers the ability to specify whether each tweet: (a) is worthless, (b) is trivial, (c) is funny, (d) piques my curiosity, (e) is useful information; or (f) is important news. Although these characteristics are by no means comprehensive, they formed a rough ordering from very negative (worthless) to very positive (important), and gave us a basis for trying out the new approach in conjunction with our results to date (in other words, we included the HIDDEN questions to vet the workers and varied the tweets' genre). Figure 5 shows the evolved task template, as contrasted with the initial template shown in Figure 3.

Indeed we found improved inter-rater agreement, especially on a few of the characteristics: there is signal hidden in the noisy evaluation of interestingness. Table 6 summarizes the results of this set of experiments. T1 was the pilot for the four investigations of the new lower overhead task design, and used current news stories. T1, T2, and T4 all showed the workers news tweets exclusively; T3 presented random tweets. T3 and T4 were both older tweets – from 2011 and earlier. The T2 dataset consisted of news tweets from 2012. The answers to the first two questions on each task – our HIDDEN subtasks – revealed that most workers took the time to come up with the right answer for the first subtask (which involved counting hashtags) and a reasonable

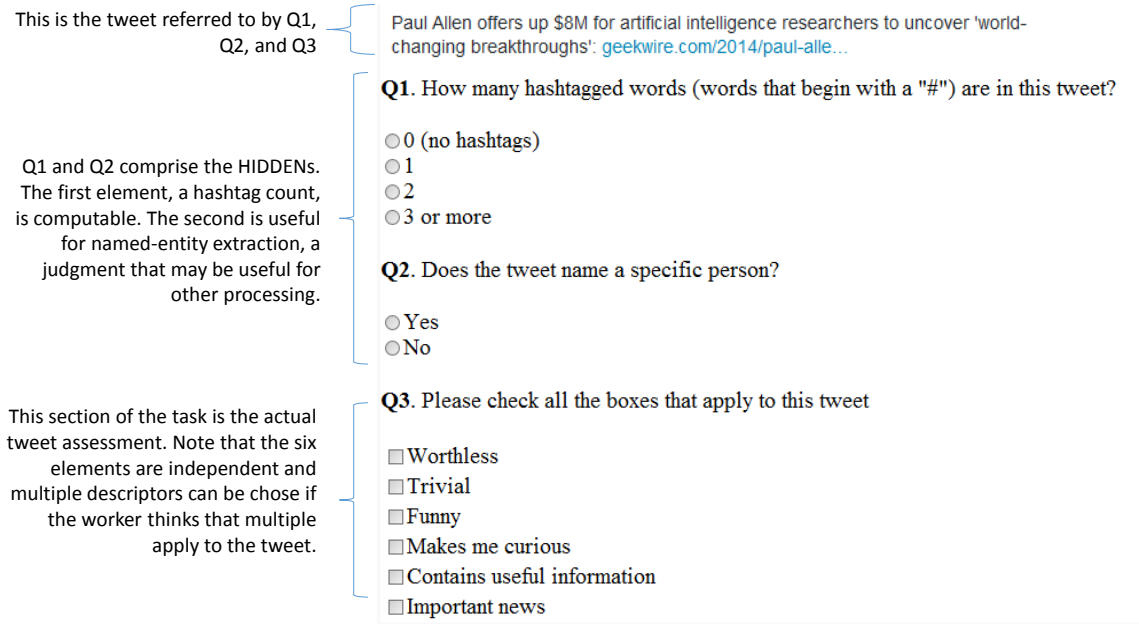

Figure 5: Task template including interest characteristics and HIDDEN questions.

answer for the second subtask (which involved identifying whether a person appeared in the tweet by name); inter-rater agreement for both HIDDEN subtasks was high. If we look at the inter-rater agreement for the new decomposition of interestingness, we can see that the components seemed to involve different degrees of subjectivity, and were easier or more difficult to assign depending on both the age and genre of the tweets. For example, in the case of T1 (fresh news tweets), there was relatively high agreement at both ends of the spectrum (worthless news and important news), with less agreement in between, suggesting greater subjectivity. As the tweets got older, agreement dropped. For random older tweets (T3), subjectivity was generally higher. The only weak signal appeared at the positive end of the spectrum, which tweets were useful or important. Note that inter-rater agreement for T4, the very old news tweets, was generally low, and only high for those tweets that were apparently "evergreen" (a journalistic term for stores that can be published at any time): funny stories and those stories that provoke reader curiosity.

Notice that eliminating the judgments of the workers who performed poorly on the HIDDEN subtasks – the original intent of the addition – had less effect than we would have hoped, and in some cases did not improve task performance. We believe this technique to still be a useful addition to our arsenal of crowdsourcing tactics, especially since the HIDDEN subtasks are designed to produce useful, albeit less subjective results.

5. DISCUSSION

We set out to identify high quality tweets as an exemplar of an important problem: how to use the tremendous volume of socially produced content, much of it in the form of non-traditional documents. In the case of tweets, the documents are very short and often cryptic; sometimes they make sense only to their immediate audience (the author's followers), and sometimes they are able to be scaled up to a much broader audience because they are perceived as interesting and important. Certainly there are social mechanisms for identifying these short pieces of content in individual services, but we are investigating a more general crowdsourced judgment process that can be both timely and thorough. Initially we'd 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 as time went on, we realized just how subjective the question was that we were asking the workers, and we began to see the need for a new way of thinking about this sort of problem; we feel that subjective assessment is going to be key for working with socially produced data

Thus in this paper, our goal was to reflect on the "work-workers-task design" framework, looking more carefully at how we were varying these three important elements, and how we could tell we were on the right track. We were also out to develop a cost-effective and respectful way of vetting the workers, now that we knew that sheer consensus with a norm was not going to be an effective way of assessing their reliability. Finally, we knew all too well that our results were not going to achieve consensus and that we were going to need to develop a different aggregation technique, one that either reflects the characteristics of the judges or takes their variance into account.

Framework. As far as our framework goes, it seems like the key is to adjust each element in turn with much smaller datasets. This allows us to try various combinations with-

Task identifier/platform	T1-AMT	T2-AMT	T3-AMT	T4-AMT
Dataset	News-2013	News-2012	Random-2011	News-2011
HITs with 0 choices	1	0	1	0
HITs with 1 choice	444	439	434	492
HITs with 2 choices	45	57	61	8
HITs with 3 choices	5	4	4	0
Frequency of worthless	33	10	23	241
Frequency of trivial	154	184	116	193
Frequency of funny	10	9	8	28
Frequency of makes me curious	121	127	159	16
Frequency of contains useful information	111	115	158	29
Frequency of important news	120	120	104	1
Fleiss's κ for Q1, all judges	0.909	0.907	0.974	0.954
Fleiss's κ for Q2, all judges	0.758	0.728	0.843	0.618
Fleiss's κ for Q3, worthless (all judges)	0.383	0.031	-0.025	0.043
Fleiss's κ for Q3, trivial (all judges)	0.095	0.041	0.023	-0.063
Fleiss's κ for Q3, funny (all judges)	0.132	-0.018	0.047	0.168
Fleiss's κ for Q3, curious (all judges)	0.054	0.024	0.059	0.128
Fleiss's κ for Q3, useful (all judges)	0.077	0.046	0.158	0.012
Fleiss's κ for Q3, important (all judges)	0.313	0.205	0.168	-0.002
Krippendorff's α, Q1*	0.910/0.931	0.907/0.907	0.974/0.974	0.954/0.954
Krippendorff's α, Q2*	0.758/0.765	0.728/0.728	0.843/0.843	0.618/0.618
Krippendorff's α, Q3 aggregate*	0.137/0.128	0.063/0.063	0.088/0.088	0.014/0.014
Krippendorff's α, Q3, worthless*	0.384/0.383	0.033/0.033	-0.023/-0.023	0.045/0.045
Krippendorff's α, Q3, trivial*	0.097/0.088	0.043/0.043	0.025/0.025	-0.061/-0.061
Krippendorff's α, Q3, funny*	0.134/0.134	-0.016/-0.016	0.049/0.049	0.169/0.169
Krippendorff's α, Q3, curious*	0.056/0.057	0.026/0.026	0.061/0.061	0.130/0.130
Krippendorff's α, Q3, useful*	0.079/0.066	0.048/0.048	0.160/0.160	0.014/0.014
Krippendorff's α, Q3, important*	0.314/0.303	0.207/0.207	0.170/0.170	0.000/0.000

Table 6: A task redesign based on a decomposition of interestingness. Starred elements include values with all judges, and with unreliable judges ($\alpha < .9$ for Q1) omitted.

out 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 we sensed are apt to yield a greater density of interesting content. If the interesting content is sparse and its overall importance is ambiguous, the task may begin to seem meaningless.

Next we combined the new genre with worker-oriented changes: was it necessary to use the more expensive crowdsourcing platform, or were results equivalent when we used AMT? Answering this question about platforms enabled us to experiment without further burdening the more costly platform.

Finally, it was useful to redesign the task itself: teasing apart what we mean by "interesting" can reduce subjectivity and can make the task less cognitively burdensome. It may be easier to say whether a tweet is funny or potentially useful, than it is to assess the vaguer quality of 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 [16]. 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. In some sense, we tried that with the characteristics worthless and trivial. It may be that these were too subtle a distinction; lumped together they may have been 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?"

HIDDENs. Our second goal was to develop a method for assessing worker reliability that (1) incorporates less subjective judgments; (2) contributes to the main question we were asking (in other words, doing this extra work improves label quality and provides a single point of focus, the data 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 orthogonal to our original purpose, for follow-up research.

Recent crowdsourcing work has found that workers like being given a HIT that's essentially a break in the action, a task that's just fun (read this cartoon), which gives them a chance to catch their breath and go back to the central work refreshed [15]. We also need to establish whether the order that we ask the HIDDENs has any unanticipated effects: Would we get the same results if we asked the named entity detection task first? Is it annoying to do the first task (counting the hashtags), a task that is clearly computable

to most sophisticated crowdworkers. We can design other sorts of subtasks that are on a spectrum of subjectivity.

HIDDENs subtasks are most useful if they can be disaggregated in a way that is recomposable. In other words, we want the results of the HIDDENs to be useful to others on our team who are working with the same data set. How can we create a library of these questions so that they not only vet the workers, but also create useful results for others using the dataset.

Subjectivity. Our third goal at its essence is an invitation to take a closer look at how to handle subjective assessments and how to use the varied signal they produce. Rather than simply saying, "this is a subjective question and is thus inappropriate for a crowdsourced approach", we would rather develop alternative ways of looking at the results and measuring their reliability. Past efforts have raised the possibility of thinking of the questions like polls [3]: If 7 out of 10 judges think this is an interesting tweet, can we replicate the proportion of individual judgments and use polling measures to assess reliability? What is the best way to collect and use this type of more nuanced data?

6. CONCLUSIONS AND FUTURE WORK

Although we have not solved the problem of identifying the high quality content that is hidden in the large volume of user-created social media, we feel that we have made substantial headway. The work-workers-task design framework we have investigated has shown itself to be effective in rapid-turnaround debugging of a difficult labeling task; the HIDDENs, in-task Captchas, show promise of being an avenue for both measuring worker reliability and potentially improving label quality; and finally, we are beginning to understand how to work within the confines of what is essentially a subjective question.

In addition to the research we identified in Section 5, we are also 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 subjective data (labels) we are gathering. Taken together, advances in all of these areas will be an important step in improving access to social media data, and in using subjective judgments in a variety of data-driven applications.

7. REFERENCES

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

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

[3] O. Alonso, C. Marshall, and M. Najork. Are Some Tweets More Interesting than Others?#HardQuestion. In *Proc. of HCIR* 2013.

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

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

[6] S. Colton, A. Bundy, and T. Walsh. On the Notion of Interestingness in Automated Mathematical Discovery. *Int. J. of Hum.-Comput. Stud.*, 53(3), 2000, 351–375.

[7] J. L. Fleiss. Measuring Nominal Scale Agreement Among Many Raters. *Psychol. Bull.* 76(5), 1971, 378–382.

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

[9] A. Kittur, E. Chi, and B. Suh. Crowdsourcing User Studies with Mechanical Turk. In *Proc. of CHI* 2008, 453–456.

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

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

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

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

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

[15] J. Rzeszotarski, E. Chi, P. Paritosh, and P. Dai. Inserting Micro-Breaks into Crowdsourcing Workflows. In *Proc. of HCOMP* 2013.

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

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

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