A Human-Centered Framework for Ensuring Reliability on Crowdsourced Labeling Tasks

Omar Alonso, Catherine C. Marshall, Marc Najork
Microsoft
{omalonso,cathymar,najork}@microsoft.com

Abstract

This paper describes an approach to improving the reliability of a crowdsourced labeling task for which there is no objective right answer. Our approach focuses on three contingent elements of the labeling task: data quality, worker reliability, and task design. We describe how we developed and applied this framework to the task of labeling tweets according to their interestingness. We use in-task CAPTCHAs to identify unreliable workers, and measure inter-rater agreement to decide whether subtasks have objective or merely subjective answers.

Traditional labeling tasks such as relevance assessment or classification have a mature process for developing training data to use in machine learning applications. Labels are elicited from human judges, often as a crowdsourced task; a small number of these labels are compared with a gold set to ensure that workers are completing the labeling satisfactorily. If the judgment is difficult, worker consensus is used to determine the most appropriate label.

But what happens if there is no single best answer to a difficult judgment task? How can we arrive at a reliable, high-quality labeled training set? We have investigated a crowdsourcing task that we believe will become increasingly common as data sources such as Twitter evolve: a labeling task in which there is no gold set for assessing reliability and judges may not arrive at a single reproducible right answer.

Crowdsourcing this type of labeling task raises three potential sources of quality problems: the workers (the crowd's reliability and collective expertise), the task design (the way the task is presented to workers), and the work (the dataset and the labels to be applied to it). These three elements are contingent on one another: adjusting one may have a meaningful effect on the others.

To perform feature-based prediction of which Twitter posts (aka tweets) are consistently interesting to a broad audience, we set about acquiring labels for a training set of random tweets. Using familiar techniques influenced by the information retrieval (IR) literature, we created a structure for presenting the task, drew sample datasets of random tweets from the Twitter firehose, and tested several different label sets. We expected to end up with a set of labeled tweets, some interesting, others not, and to be able to develop a classifier with reasonable predictive performance, bounded by the degree to which the human judges agreed.

From an initial study, we knew that interestingness was a complex and somewhat idiosyncratic concept. Yet even given our tempered expectations, the labeling results were disappointing. Over the course of various adjustments to the label set and the data, we acquired 250,000 labels for over 50,000 tweets. But regardless of our adjustments, not only was the rate of inter-assessor agreement low (at best, Krippendorff's alpha hovered around 0.1), but also the tweets that were labeled as interesting seemed inexplicably random, even given the potential diversity of the workers' interests. How could a classifier be expected to outperform these unsatisfactory results? Indeed, given our best efforts at assembling a set of labeled tweets, and using majority vote as inputs to a binary classifier, we were only able to achieve moderate agreement (Fleiss' kappa = 0.52) between input and prediction (Alonso, Marshall, and Najork 2013).

We had begun this effort with fairly large labeling tasks (the early datasets contained between 2,000 and 10,000 tweets, each judged by 5 judges), but scaled back for debugging purposes. Subsequent experiments investigated the effects of changes to representative aspects of the framework.

Our first area for investigation was 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 degrees of interest in some types of news stories, we thought it likely that the judges would agree that some stories were of more universal importance. Indeed, judges found a larger fraction of tweets to be interesting (29.3% vs. 7.9%), but inter-rater agreement did not improve significantly (Krippendorff's alpha = 0.068 vs. 0.045).

The second area was the judges' expertise. We wondered if there was any difference in the performance of the relevance judgment experts (UHRS) and the more diverse, less expert, and lower-paid judges we recruited from Amazon Mechanical Turk (MTurk). The pool of judges had no noticeable effect on either the fraction of news tweets labeled as interesting (50.8% on UHRS vs. 48.7% on MTurk) nor on inter-rater agreement (0.139 on UHRS vs. 0.073 on MTurk).

In accordance with the crowdsourcing literature, we next focused our attention on the workers and the quality of their

Crowd	judges	interesting	Q1 α	Q2 α	Q3 α
UHRS	14	43.8%	0.779	0.722	0.050
	11	40.6%	0.917	0.735	0.049
UHRS	20	57.0%	0.775	0.734	0.157
	14	59.1%	0.946	0.692	0.145
UHRS	20	48.8%	0.882	0.752	0.157
	19	47.4%	0.883	0.742	0.125
UHRS	12	53.4%	0.819	0.774	0.190
	9	50.6%	0.930	0.822	0.130
MTurk	11	40.2%	0.876	0.734	0.085
	8	34.8%	0.931	0.708	0.049
MTurk	10	55.0%	0.850	0.843	0.105
	7	56.8%	0.886	0.864	0.065
MTurk	9	51.0%	0.800	0.840	0.030
	8	47.7%	0.942	0.869	-0.040

Table 1: Experiments investigating worker quality.

output. Unreliable performance, either as a result of fatigue, frustration, or carelessness, needed to be ruled out. But how would we eliminate poor quality work without a gold set?

To reliably assess workers' diligence, we created in-task CAPTCHAs, a technique similar to the objective, verifiable questions used in crowdsourced user studies (Kittur, Chi, and Suh 2008), as well as in other spam detection settings (v. Ahn, Blum, and Langford 2004). The in-task CAPTCHAs added two ancillary judgment subtasks to our primary task (Q3), labeling a tweet as interesting or not. The first (Q1) is a simple task that can be verified computationally: counting the number of hashtags in the tweet. The second ancillary subtask (Q2) had a single correct answer but required more thought and judgment: assessing whether a tweet contained a person's name that was neither an account (@name) or a hashtag (#name). We anticipated very high agreement on the first subtask (repeated disagreement with the truth meant the worker was suspect) and good agreement on the second, depending on the breadth of a worker's awareness (e.g. Mubarak is person, not a place). Q1 and Q2 not only provided a window onto the worker's reliability, they also meant that the worker had to read (or at least scan) the tweet twice before performing the main judgment task, thus potentially improving the quality of the results.

Table 1 shows the results of a representative sample of these experiments. Each experiment used 100 tweets belonging to the "news" genre; each tweet was presented to 5 of the judges; and each task consisted of the three questions described above. The first row of each group shows statistics based on all judges; the second is based on only those judges who answered at least 95% of the CAPTCHAs correctly. Eliminating judges who underperformed on Q1 does not increase inter-rater agreement on Q2 or Q3 (measured by Krippendorff's alpha), nor does it increase the percentage of "interesting" labels. Thus, we can be confident that the fault did not lie with the workers.

Finally, we set out to diagnose the pivotal element of the task, Q3. We knew we were asking a subjective question, but our initial belief was that the judges would reach some core consensus about which tweets were universally interesting. After all, Twitter users do just that when they retweet or favorite other peoples' tweets. They are able to distinguish what is interesting in an *I'll know it when I see it* way.

	News		All	
	labels	α	labels	α
worthless	33	0.384	241	0.014
trivial	154	0.097	193	-0.061
funny	10	0.134	28	0.169
curiosity-provoking	121	0.056	16	0.130
useful information	111	0.079	29	0.014
important news	120	0.314	1	0.000

Table 2: Experiments investigating task quality.

Others have used similar tactics to establish whether an item is interesting or not (Lin, Etzioni, and Fogarty 2009).

So we began to look into the nature of interestingness: what makes something interesting, and how can that understanding be reflected in our task design? Interestingness, according to the psychology literature, is a complex emotion (Silvia 2005). 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.

Thus we disaggregated interestingness into 6 preliminary characteristics, some of which could be true in conjunction with one another, and others that were mutually exclusive; workers could specify whether a tweet is (a) worthless, (b) trivial, (c) funny, (d) curiosity-provoking, (e) useful information, or (f) important news. Although these characteristics are by no means comprehensive, they formed a rough ordering from negative (worthless) to positive (important), and gave us a basis for trying the new approach. As shown in Table 2, inter-assessor agreement increased for the extrema ("worthless" and "important news") of news tweets, but not for random tweets.

In the end, what impressed us the most was the importance of debugging and fine-tuning every aspect of the framework, the workers, the work, and the task template. It is easy to blame the workers for one's woes — or even to side with them and say "if only we paid them more, they would surely read closely enough to reach agreement" — or to dismiss the task concept as too ambiguous (or ambitious) to tackle this way. But of course, this is the very reason we might bring humans into the judgment process.

References

Alonso, O.; Marshall, C. C.; and Najork, M. 2013. Are some tweets more interestingthan others? #hardquestion. *Microsoft Research Technical Report MSR-TR-2013-83*.

Kittur, A.; Chi, E. H.; and Suh, B. 2008. Crowdsourcing user studies with mechanical turk. In *Proc. of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, 453–456. New York, NY, USA: ACM.

Lin, T.; Etzioni, O.; and Fogarty, J. 2009. Identifying interesting assertions from the web. In *CIKM*, 1787–1790.

Silvia, P. J. 2005. What is interesting? exploring the appraisal structure of interest. *Emotion* 5:89–102.

v. Ahn, L.; Blum, M.; and Langford, J. 2004. Telling humans and computers apart automatically. *CACM* 47(2):57–60.