Are Some Tweets More Interesting Than Others? #HardQuestion

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

ABSTRACT

Twitter has evolved into a significant communication nexus, coupling personal and highly contextual utterances with local news, memes, celebrity gossip, headlines, and other microblogging subgenres. If we take Twitter as a large and varied dynamic collection, how can we predict which tweets will be interesting to a broad audience in advance of lagging social indicators of interest such as retweets? The telegraphic form of tweets, coupled with the subjective notion of interestingness, makes it difficult for human judges to agree on which tweets are indeed interesting.

In this paper, we address two questions: Can we develop a reliable strategy that results in high-quality labels for a collection of tweets, and can we use this labeled collection to predict a tweet's interestingness? To answer the first question, we performed a series of studies using crowdsourcing to reach a diverse set of workers who served as a proxy for an audience with variable interests and perspectives. This method allowed us to explore different labeling strategies, including varying the judges, the labels they applied, the datasets, and other aspects of the task. To address the second question, we used crowdsourcing to assemble a set of tweets rated as interesting or not; we scored these tweets using textual and contextual features; and we used these scores as inputs to a binary classifier. We were able to achieve moderate agreement ($\kappa = 0.52$) between the best classifier and the human assessments, a figure which reflects the challenges of the judgment task.

INTRODUCTION

Over the last six years, Twitter has evolved into a significant source of news and entertainment; some have even referred to the microblogging service as a global network of human sensors. According to the official Twitter blog, as of March 2013, well over 200 million active users were producing over 400 million tweets per day. Naturally, the genre and reach of individual tweets varies dramatically, from inspirational sayings to celebrity gossip to quips to breaking news. Some tweets are original on-the-spot reports of events or situational humor; others are well-worn headlines or memes that have already received substantial Internet attention in other venues like newspapers or blogs. Still others are comparable to chat room fodder, conversations among circles of friends.

If we consider Twitter as a rich data source, as a large and varied collection of very short plain text documents, some of these documents will be more universally interesting than others; in other words, some tweets will interest more people than others do. Informed by a growing body of research on interestingness, we take interest to be a complex emotion, a confluence of plausibility, novelty, surprise, comprehensibility, and complexity[7]. In fact, in their efforts to identify interesting assertions on the web, Lin, Etzioni, and Fogarty have distilled interestingness into three qualities (specificity, distinguishability, and factual utility); their work demonstrates some of the challenges people have in reaching agreement about which assertions exhibit these qualities[19]. Silvia's research suggests that reverse measures may be easier for people to apply [28]. Thus, instead of breaking down interestingness into its constituent parts at the outset of our work, we are taking an exploratory approach: what is the best method for human judges to reach consensus on which tweets are interesting?

If human judges can label tweets' interestingness effectively, producing a training set that distinguishes between interesting and uninteresting tweets, it may then be possible to implement a classifier that uses predictive features to identify interesting tweets within a dynamic collection. We will thus have leading indicators of interest that will identify these tweets independently of social indicators of interest such as retweets, or personal indicators of interest such as favorites (both of which can be lagging indicators for some applications). These tweets can serve a variety of purposes: they can be used for content selection and indexing, so interesting tweets are stored longer than uninteresting ones; they can be used to produce a new user experience, so an applications can be built to surface interesting tweets or reconstruct a story from topically similar tweets; or they can be useful for serendipitous discovery, supporting information encountering [10] outside of the normal subscription structure. In short, they can provide new ways of managing tweets as a collection.

Naturally, judging whether a tweet is interesting or not is a subjective activity. Thus it is difficult to tell whether human assessors are working earnestly and effectively so that consensus among them produces a high-quality gold set of labeled tweets. There are multiple contingent factors at play: for example, the number of assessments per tweet; trade-offs between assessor diversity and judgment coherence; the nature of interestingness when it is applied to unfamiliar tweets (that is, tweets outside of a person's normal feed) and how to elicit it in the context of a crowdsourced task; how the test

dataset should be prepared; and what kind of labels to use. In short, the design of the judgment task crucially influences label quality. Like Aroyo and Welty, we expected some disagreement among assessors and hoped to find a way to use it productively [3] Our ultimate aim is to refine a method to create a high-quality training set for the classifier.

Thus, the paper is structured as follows: First we discuss related work, and how it serves to frame and inform our effort. Second, we report briefly on an open-ended exploration of what makes a tweet interesting to give us an informal sense of what the workers react to when they are labeling tweets. Next we describe our efforts to design the judgment task, as we vary its contingent aspects (the crowd, the labels, the dataset, and the work), focusing on the challenges we encountered and the lessons we learned. Finally we discuss a predictive model, and evaluate it as it compares to human performance. We conclude by summarizing our results and laying out research directions.

RELATED WORK

In this work, we are using Twitter as a data source, an evolving collection of very short documents; our aim is to identify leading indicators of interest by developing a dataset of tweets with high-quality labels. Three areas of related work have influenced our efforts to identify high-quality tweets: (1) investigation of Twitter as a social phenomenon to better understand the medium and its users; (2) techniques for filtering, analyzing, and presenting tweets based on topical relevance; and (3) methods for classifying tweets with no predefined information need.

User studies have grounded our expectations of how tweets are written and consumed. Zhao and Rosson [32] conducted one of the first studies on Twitter use; they found a great deal of content diversity. Other reports salient to our understanding of tweet features include coverage of conversational characteristics of tweets and retweets [5] and tweet credibility assessment [23]. A global perspective on detecting sentiment via language use is described in [25]; while language use and sentiment are outside of our immediate set of predictive features, there is potential to extend our approach to include characteristics human judges may be using. Furthermore, we acknowledge the importance of predefined categories in biasing interestingness judgment [4]; thus in this study we maintained a flexible notion of interestingness and explored its many interpretations.

Much prior work that uses Twitter as a data source is relevance-focused: there is a well-defined topic (e.g. a political situation or a natural disaster), and a *deus ex machina* approach to gathering the tweets related to the topic. This approach has been used to show topic evolution, including bursts of activity, and event detection. Notable examples of this relevance-focused work includes gathering and analyzing tweets relevant to the Iran elections [12] and Japan's earthquake reporting system [27]. By contrast, we consider the question of interestingness without pre-defining a topic or a genre, and we take a multiple perspective, human-centered approach to detecting these tweets.

In other words, our approach is not about information seeking, i.e. identifying those tweets that are relevant to a specified query or topic. Instead, identifying interesting tweets is a form of information encountering [10], which involves "accidental discovery of useful or interesting information." In other words, interestingness is serendipitous and relative to a person's implicit interests. Nonetheless, ranking algorithms for estimating a tweet's relevance to a given query employ many of the same signals as the ones we identified for predicting whether or not a tweet is interesting; see for examples the features identified by Metzler and Cai [21] for the TREC Microblogging task. This suggests that relevance prediction subsumes interestingness prediction: it identifies tweets that are responsive to a query as well as interesting to the information seeker. In other words, identifying interesting tweets is useful in many scenarios, including but not limited to search.

Very recently, automatic classification of tweets has received attention as a mechanism to consume content in better ways. [8] suggest using user categories; on the other hand, [33] concentrate on trending topics. As part of our tweet labeling and classification studies, we discuss and compare these proposals. [9] use a learning approach to rank; their approach combines content relevance, user authority and tweet-specific features to show the best and most relevant tweets. A personalized tweet ranking model that exploits retweets is presented in [30].

Several recent efforts are motivated by research questions that are well-aligned with ours. Hurlock and Wilson conducted a user study on different factors that make a tweet useful for a set of search tasks [15]; like us, they were looking for underlying rationales for tweet classification and for agreement among assessors. A crowdsourcing approach for detecting uninteresting content is presented in [1]; we have observed that uninteresting content is easier to identify than interesting content. A similar crowdsourcing scheme for assessing the perceived value of a tweet content is described in [2].

METHODS

Participants familiar with Twitter were recruited from two crowdsourcing platforms. First, workers we refer to with the prefix RJ were recruited from a crowdsourcing platform that specializes in relevance judgment; these workers were familiar with the labeling task. Second, workers we refer to with the prefix MT were from Amazon Mechanical Turk; these workers we saw as more of a proxy for Twitter users with diverse perspectives. Workers were paid at standard rates for the respective crowdsourcing platform. MT workers were paid 3 cents per HIT; RJ workers were paid about five times that amount as expert relevance judges.

As a prelude to the labeling exercises, we conducted an informal survey (discussed in the next section) using workers recruited from both platforms. After the survey, we conducted as series of labeling studies. Instructions defined the labels for the workers, then we asked the workers to apply the study's labeling scheme to each tweet. Tweets were presented to the workers as a profile name and photo adjacent to the text of the tweet, followed by a choice of labels. The label choices,

which will be described later, varied according to each study's goals.

Datasets for the studies were prepared by taking a random sample of tweets from the public English-language Twitter feed (aka the Twitter firehose) from a specific bounded time period. Each tweet in a sample dataset included the associated profile name, profile picture, number of followers, the tweet's publication time, and its unique ID. Any further preparation will be described in the individual sections.

WHAT MAKES A TWEET INTERESTING?

Before we used crowdsourcing methods to develop a labeled dataset of interesting tweets, we wanted to get an informal sense of what workers would be looking for, both to help us assess label quality and to help us anticipate judgment diversity. Although the survey is relatively small-scale (113 responses at varying levels of detail, from 3 words to 196 words, average length: 45 words), the responses can set the stage for subsequent labeling tasks.

Responses to the question "*In your opinion, what makes a tweet interesting? If you read a tweet, what specific characteristics make you think that the tweet is relevant?*" were open-coded [29] and divided into two categories: *Endogenous qualities* and *exogenous qualities*. Endogenous qualities refer to the tweet's content; exogenous qualities are associated with the tweet's context, such as its source or originality. Some of the characteristics (e.g. credibility, genre, and style) are similar to those attributed to longer documents; others are microblog-specific (e.g. the use of retweet conventions).

The tweet's source (in 55 responses) stands out as the most important exogenous quality. Many respondents say they evaluate tweets by who posted them, sometimes privileging tweets from celebrities (14 positive and 1 negative response), and other times seeking tweets from people they know (e.g. friends, family, and colleagues). Tweets from unknown sources are scrutinized more carefully; respondents claim they pay attention to profile photos and news agency logos. Of particular note, 4 respondents described situations in which they believed themselves to be privy to firsthand reports. For example, MT68 wrote: "*So a tweet is interesting if it ... updates the situation or experiences of a person I'm following (I have an aquaintance working for an NGO in Dafur who is really opening my eyes to what she is experiencing).*"

Of secondary importance are the tweet's originality (in 19 of the 113 responses) and timeliness (in 9 responses). For example, RJ 5 wrote, "*If the tweet has basic knowledge or something I already know then I will not find it as interesting as a tweet that tells me something new. This mostly applies to news and tech stories.*" Only one respondent mentioned paying attention to the number of retweets a tweet has garnered; three respondents said they noticed if a tweet was part of a longer story or conversation.

Endogenous qualities were sometimes general and hinged on a tweet's topical relevance. As we might expect from the interestingness literature, respondents reflected on the content's qualities — Is it useful information? Does it seem to be factual? — and genre: is the topic breaking news? Is it celebrity gossip? Some respondents mentioned specific topics that were trending around the time of the survey (for example, Occupy Wall Street); others listed areas of personal interest ("*entertainment, politics, clothing, and food*" [MT35]). Less often, a tweet had to contain certain elements (such as location) to guarantee its salience to the respondent.

Some endogenous qualities associated with genre, form, and style were controversial. Personal tweets are exceptionally divisive: 16 respondents said they were interesting; 10 said they were inherently trivial. For example, MT15 wrote, "*A tweet is interesting when it's relateable and actually comes from a real person talking to you or about their lives.*" By contrast, MT27 wrote, "*I am not interested in personal tweets. People can tweet their personal stuff on Facebook.*"

Embedded links are less controversial. 17 respondents said they contribute to the value of a tweet; two said links were reason to ignore the tweet. Brevity is sometimes cited as a positive attribute ("*I read most all the short tweets that I happen to come across in my feed*" [MT25]).

Some common endogenous qualities — good writing, humor, or understandability — are subjective. Yet respondents often identified them as vital to interesting tweets. Over one-third of the respondents (38/113) specified humor as desirable (e.g. "*Some of the best tweets are just plain funny, or the kind of inside joke exchanges you might see between twitter users that only dedicated followers are aware of.*" [MT59]). Style and writing mattered to almost one-quarter of the respondents (e.g. "*The specific characteristic that makes a tweet relevant for me, more than anything else, is intelligence.*" [MT4]).

This informal survey reveals: (1) There is some consensus on the qualities that make a tweet interesting, but there is also diversity and disagreement, which suggests that label quality will be an area we need to address; (2) Some characteristics (e.g. humor) are more subjective than others; and (3) There will be shift from the survey's goal to identify characteristics of tweets that are interesting to the individual worker to tweets that are broadly interesting to more people. In other words, workers will need to accommodate to a task that is familiar (reading tweets), but which must be performed self-consciously (labeling tweets with this broader mandate).

JUDGING TWEETS

Many analyses of Twitter as a dataset rely on the extraction of a set of event-related tweets. Although these tweets may be gathered in real-time [20], usually they are collected after the event has occurred by using keywords and hashtags. For example, tweets have been collected about crises such as hurricanes Gustav and Ike [14], political events [14], the London riots [18], or the Arab Spring [6]), and sporting events such as soccer games [20]. The event may even reflect a Twitter-internal phenomenon, e.g. the rise and fall of a meme [31]. [24] discuss methods for subsetting tweets related to trending events.

Instead of focusing on specific events after they have occurred, we want to identify interesting tweets as they unfold; it is an approach more akin to information encountering than information seeking. Our initial survey revealed that

current events are only part of what's interesting in Twitter: people are also looking for tweets that are funny or opinionated; celebrity gossip; interesting facts; personal revelations – a diverse range that often falls outside of what we would normally think of as events or trending topics.

Given the wide range of perspectives on what kinds of tweets are interesting, human judgment seems like the most effective way to identify them. By asking different people which tweets are interesting, multiple perspectives can be represented. Furthermore, by assigning the same tweets to multiple judges, we can minimize the effects of judges with more obscure interests and identify the tweets that people generally find interesting. This approach, a variant on crowdsourced labeling methods that have been applied to relevance judgment, suggests that we explore three aspects of label assignment–the judges; the labeling scheme; and the labeling conditions–to arrive at a set of high-quality labels.

The studies we describe in this section will help us answer the following questions related to label quality: (1) Is it better to increase the number of judges, thus broadening the number of perspectives represented, or is it better to use a smaller number of expert judges, thus focusing on performance and reliability? (2) Is it better to have labels that categorize tweets, and use the categories as proxies for interest, or to ask about interestingness directly? and (3) Is it better to filter out tweets we strongly believe are uninteresting (for example, URL-only spam tweets), thus increasing the density of interesting tweets, or is it better to ask the judges to label all of the tweets, thus creating a richer model of uninteresting tweets?

We addressed these three questions through a series of crowdsourced labeling tasks. In addition to exploring the label quality issues outlined above, the most successful of these labeling tasks produced a set of labeled tweets to use in the development of a predictive model.

Study 1: Coded Classification

In our first labeling study, we began with a classification scheme from the literature; the classification scheme focused on tweet genres rather than on interestingness per se. This study was designed to help set expectations about worker consensus and identify potential problems with the labeling task.

We began with 9990 tweets, sampled at random from the public English-language Twitter firehose between August 7 and October 1, 2011. RJ workers classified sets of tweets (5 per task) using the 5 categories suggested by [33]: *news*, *current*, *meme*, *commemorative*, and *other*. We piloted the experiment with the workers who would be classifying the tweets; because they had difficulties labeling some of the tweets, we added two new labels that they suggested: *adult* and *thinking out loud* (saying things that might better remain as private thoughts).

Each tweet was judged by three expert workers, resulting in 29970 labels; individual workers judged between 1125 to 8444 tweets. Times to judge 5 tweets averaged between 37 seconds (about 7 seconds per tweet) and 2 minutes, 21 seconds (almost half a minute per tweet).

label	# of 3-way agreement	%	proportion of total
adult	18	1%	0.143
commemorative	1	0%	0.017
current	3	0%	0.014
meme	242	19%	0.095
news	7	1%	0.056
other	116	9%	0.046
tol	875	69%	0.208
total	1262	100%	0.126

Table 1. Instances of 3-way agreement per label (Study 1)

By using these classifications, we hoped to make the labeling task less subjective, thus promoting inter-assessor agreement. However, the judges did not reach consensus about how to label the majority of public tweets. Table 1 shows the instances of 3-way agreement on each label. The "proportion" column reflects how often workers agreed on a label. For example, workers used the label *adult* 377 times; 54 of them (18 tweets $\times$ 3 judgments) agreed. Consensus judgment was unusually high for the new labels *thinking out loud* (0.208) and *adult* (0.143), and an order of magnitude lower for *current* (0.014) and *commemorative* (0.017). Not only were these labels applied less frequently; they were also applied inconsistently.

The low level of inter-assessor agreement suggested that genre-based classification was not a successful strategy. Many tweets had three different labels (thus defeating the "majority wins" strategy). Sometimes labeling differences reflected cultural differences, varying sensibilities, or failed attempts at humor. For example, a satirical tweet such as, "Rive Gauche: The Bar bites back - UK Blawg roundup - Wife of MP convicted of nicking a kitten http://t.co/BusCJ3Fd", is fairly culturally-specific, and it is not surprising that the three judges assigned it three different labels (*current*, *news*, and *meme*). In other words, two workers did not recognize the tweet as satire, and could not distinguish between the two genre classifications, and the other worker decided *meme* was the best label to handle humor.

Study 2: Streamlined Labeling

As a result of Study 1, we decided to streamline judgments into binary decisions (interesting or not) and to increase the number of workers voting on each tweet. This change would enable us to pursue interestingness in a more direct way and would increase the breadth of judgment perspectives. We realized that if inter-assessor agreement continued to be low, the *true/false* labels would throw label quality into doubt, since it would be difficult to establish the accuracy of the labels.

For Study 2, we used three data sets (D1, D2, and D3) from different time periods during 2012. Table 2 describes these data sets, which were also used to identify the predictive features we describe in Section 6. To ameliorate some of the problems we observed in our earlier labeling experiment, RJ workers were asked to simply judge whether the tweets in the three datasets were interesting or not (i.e. they had to label them *true* or *false*). Each tweet received either three (D2) or five (D1, D3) judgments. We increased the number of workers for the final labeling exercise to check the effect of broadening worker perspectives.

tweet dataset description			crowdsourcing description		judgment summary				
dataset	tweet period	# tweets	# workers	avg. seconds/tweet	# judgments	*false*	*true*	*%false*	*%true*
D1	Jan 2-4	1870	9	7.74	9350	8241	1109	88%	12%
D2	Jan 13-14	9905	7	6.38	29715	24850	4865	84%	16%
D3	Apr 1-15	1995	23	4.56	9975	8584	1391	86%	14%

Table 2. Datasets used for Study 2. All data sets were drawn in 2012.

Dataset D1 consists of a uniform random sample of all public English language tweets from the January 2-4 time period; each tweet includes a profile name and image. To increase the likelihood that judges would encounter interesting tweets, we filtered dataset D2 so it only contained tweets over 70 characters that had at least one URL and at least one hashtag; we also eliminated tweets that mentioned another user. In the third judgment task (using D3, April's dataset), we again selected tweets that were likely to be substantive (those that were two or more words long and contained URLs) and removed tweets from profiles with fewer than 250 followers. We also removed tweets that *start* with @, i.e. are part of a conversation between two or more users, as opposed to those simply mentioning another user.

Table 3 summarizes the results by judge for the first of the three *true/false* labeling tasks. Each worker judged fewer tweets than in the prior study (6 of the 7 judges from the last study participated in this one; 3 new judges also joined the group). 88% of the tweets were judged to be uninteresting. Most of the workers who judged a large number of tweets adhered to approximately the same interesting/uninteresting ratio, although one worker (RJ 10) who participated in the last study, found the tweets to be almost universally uninteresting; another (RJ 14) went in the opposite direction and found 32% of the tweets to be interesting. The average time a worker spent judging 5 tweets ranged from about 12 seconds to 95 seconds, with an average of about a half minute per task, or about 6 seconds per judgment. We believe this variation to be normal; there was no immediate indication of worker fatigue, idiosyncratic patterns of responses, or low task completion time that would lead us to suspect the veracity of the workers' answers. Notice however that this judgment time is considerably lower than it was for the last study, when it ranged between 37 seconds and 2 minutes, 21 seconds to judge 5 tweets. We suspect this is due to the lowered overhead of having only two categories.

Table 4 shows the number of tweets with a given average rating, where 0 indicates that all judges rated the tweet as non-interesting, and 1 indicates that all found it interesting. Using a conservative "majority rules" strategy, only 113 tweets out of 1870, about 6%, might be regarded as interesting. Considering that most of the tweets were labeled *false*, we focused instead on the tweets that a majority of the judges labeled *true*, or interesting.

We examined the 6 tweets in D1 that all five judges deemed interesting as an informal check on label quality. As we would expect from the initial user study, these tweets are well formed (spelled correctly, not abbreviated, and grammatical). Furthermore, all 6 tweets contain links and are longer than average; none of the accounts seem to belong to overt spammers, although their follower numbers tend to be low. According to [31], low follower numbers do not necessarily signal that the tweets are spam. Furthermore, the link destinations seem to be legitimate articles or blog posts; for example, one article is about Chinese airlines' refusal to pay EU carbon tax; another is about rare dolphins being found in Bangladeshi waters. Although some workers may have followed the links, work times would indicate that workers usually did not. Thus, given the range of individual interests and the mostly-legitimate appearance of the tweets, we have no reason to directly impugn label quality.

Many of the other 107 tweets labeled as interesting by a majority of judges were similarly plausible examples of news or current events. 9 did not have links, indicating either primary news sources (e.g. a story about a student with a pellet gun being shot to death by police) or tweets the judges found inspirational (e.g. one was a religious aphorism) or humorous. These tweets again align with the criteria elicited by the initial user study. A significant number of these 107 tweets were items (e.g. laptop batteries and car parts) for sale under an Amazon affiliates program. A few links led to out-and-out phishing sites. Were we getting high-quality labels? Why were the judges choosing what they did? Would our judges have noticed Keith Urbahn's famous tweet breaking Osama bin Laden's death?

Our experiences with datasets D1 and D2 led us to conduct a third *true/false* labeling exercise using dataset D3. This time, we solicited additional workers familiar with Twitter to augment the efforts of the judges who had labeled D1 and D2. 23 judges participated in the third part of Study 3. The dataset itself was prepared differently too, as we described earlier, in an effort to include more interesting tweets. Would these changes matter?

worker	*false*	*true*	total judgments	*%false*	*%true*
RJ 4	527	18	545	97%	3%
RJ 9	1632	218	1850	88%	12%
RJ 10	1836	14	1850	99%	1%
RJ 14	306	144	450	68%	32%
RJ 17	60	10	70	86%	14%
RJ 21	1141	209	1350	85%	15%
RJ 22	1703	162	1865	91%	9%
RJ 32	1031	334	1365	76%	24%
RJ 33	5	0	5	100%	0%
total	8241	1109	9350	88%	12%

Table 3. Per-worker assessments of interestingness using dataset D1 (Study 2)

rating	# tweets
0.0	1178
0.2	441
0.4	138
0.6	66
0.8	41
1.0	6

Table 4. Number of tweets in D1 with a given average rating

use frequency	*General*		*Limited*		*NotImportant*		*ProbablySpam*		total
AllTheTime	71	9.45%	321	42.74%	188	25.03%	171	22.77%	751
Daily	49	12.25%	144	36.00%	165	41.25%	42	10.50%	400
OncePerWeek	10	16.95%	24	40.68%	15	25.42%	10	16.95%	59
Seldom	17	10.56%	58	36.02%	61	37.89%	25	15.53%	161
total	147	10.72%	547	39.90%	429	31.29%	248	18.09%	1371

Table 5. Use frequency v. label assignment (Study 3)

The third *true/false* labeling exercise produced results similar to the first and second. About 86% of the tweets were labeled *false* (from above, the first was 88%; the second was 84%). Workers judged anywhere from 10 tweets (two tasks) to 1765 tweets; the average number of tweets judged was 434, surely enough to get a sense of the relative merit of the tweets. Most workers spent several minutes judging the 5 tweets (the average was close to two minutes), although a few read and dispatched the tweets in an average of just over 5 seconds. There was nothing unusual about any of the workers' judgments. This time, however, no tweets had 5-way agreement that they were interesting. 10 tweets had 4 *true* judgments, and 57 tweets had 3-way agreement that they were interesting. Thus broadening the field of judges and narrowing the dataset indeed reduced the number of tweets judges agreed were interesting.

Study 3: Rationale and Expertise

Our third round of labeling studies was motivated by a desire to pinpoint the causes of low levels of label agreement: did workers simply disagree on which tweets were broadly interesting, or were we experiencing label quality problems? This time, we focused on the workers: could they articulate why they were assigning the labels they did? Did the workers' familiarity with Twitter influence the labels they used? To conduct this set of studies, we developed new labels so judges could: (1) distinguish between tweets of general interest (*General*), tweets interesting to a narrower audience (*Limited*), and tweets that aren't interesting (*NotImportant*); and (2) reflect on the possible presence of spam (*ProbablySpam*).

The studies used two datasets of random tweets from the public English language Twitter feed, harvested about a month apart. One was from the first two weeks of March; the other is from the first two weeks of April (the labeling took place in late April). The tweets were filtered in an effort to increase the density of interesting tweets. Each contained a link and at least one word, and originated from a profile with at least 250 followers. Responses beginning with an @ were omitted to eliminate conversations. The datasets included tweets from a dozen profiles with more than half a million followers, including OfficialAdele (4.5M followers), BreakingNews (3.8M followers), and NASA (2M followers). About a quarter of the tweets were between 130 and 140 characters long.

To complete this task, workers first labeled a tweet according to our new scheme (*General*, *Limited*, *NotImportant*, *ProbablySpam*). They then recorded the rationale for assigning this label, either by responding to an open-ended question, or by choosing among label-specific reasons for the label (for example, to justify a label of *Limited*, we asked whether the tweet was interesting only to a specific audience, interesting only in a specific geographic region, or interesting only to the account's followers). Finally they were asked to characterize their own Twitter use: were they constantly reading their Twitter feed (*AllTheTime*); reading it daily (*Daily*); reading it at least once a week (*OncePerWeek*); or referring to it infrequently (*Seldom*). We asked workers to record their Twitter use frequency each time they judged a tweet (rather than only once); this turned out to be a useful way of catching spammers in Study 3, since spammers tended to choose a random response each time they answered this question.

To this end, we discarded judgment data if (a) workers reported their own Twitter use inconsistently; (b) workers left the label or experience question unanswered; or (c) workers spent too little time on the task. After we performed this quality assessment, we were left with 1371 labeled tweets that were coupled with self-reported use frequency from 85 workers. Table 5 shows this breakdown. Judges who were the most frequent and least frequent Twitter users were apt to find fewer tweets interesting. Unsurprisingly spam perception seemed to be more acute among the most frequent users, probably due to prior exposure.

In spite of the fact that most workers could quickly assess whether a tweet was interesting to a broad audience, they seemed to find it difficult to describe why (before they were offered a set of specific rationale to choose from). In the open-ended version of the rationale question, to explain why a tweet about an athletic team's (the Buffalo Bills) press conference was interesting, RJ 42 offered the (possibly intentional) non-sequitur, "*A historical figure is always interesting.*" In fact, *General*, our label for tweets of broad interest, was the label most apt to be assigned without a rationale, even if the judges completed the rest of the task.

On the other hand, workers had less difficulty explaining why a tweet is only interesting to a limited audience. For example, RJ 43 labeled a tweet offering a coupon for a local restaurant (Don't miss today's Groupon - Up to 51% Off Mediterranean Fare at Falafel King: http://t.co/ivincKm3) as *Limited* with the sensible rationale, "*only interesting if you live in minneapolis area.*" Similarly, in the open-ended case, judges who dismissed tweets as *NotImportant* and *ProbablySpam* were seldom at a loss for a rationale, although knowledge of Twitter conventions and practices played a role in determining whether some of this reasoning made sense. For example, RJ 44, who reported less familiarity with Twitter, cited a short-

label	frequency	mismatch	blank
General	84	6	1
Limited	302	21	0
NotImportant	295	10	41
ProbablySpam	144	63	5
Total	825	42	105

Table 6. Labels assignments and rationale mismatches (Study 3)

ened link: *Don't trust short links* as a reason for labeling a tweet as *ProbablySpam*.

There were 825 instances in which judges labeled tweets and rationalized the labels with the canned responses. Table 6 shows the frequency of each label's use, along with the number of rationales left blank and the number in which the rationale and label were a mismatch. As we might expect, the labels were assigned in roughly the same proportions as they were when judges were asked to come up with their own rationale. As before, the most frequent label, *Limited*, was never left blank (although there were 21 mismatches between label and rationale). Tweets labeled *NotImportant* and *ProbablySpam* have an appreciably smaller proportion of inappropriate rationales, but both have a large number (41 and 63 respectively) that were left blank. Thus we might think of these labels as falling under the rubric of "I'll know it if I see it."

This rationale and experience-focused labeling exercise showed us that complicating the labeling was apt to frustrate the judges; this was the only study we did that resulted in an appreciable number of unfinished tasks and cases of worker spam. It also demonstrated that it is probably easier to identify and reason about uninteresting and limited tweets than it is to find and rationalize interesting ones, supporting Silvia's theory that reverse measures may be easier for people to apply [28]. The need to supply rationale revealed more pronounced effects of worker expertise: not surprisingly, much is hidden behind the labels.

Lessons from the Three Studies

We knew from the outset that identifying interesting tweets would be a difficult enterprise, one that necessarily varies with the judges' own interests and proclivities. To investigate ways of improving label quality, we varied the number of judges, the selection of labels, and the judgment conditions. In so doing, we attempted to identify the tensions inherent to the task, and to resolve these tensions in a way that would result in the most reliable labeled dataset. By creating the labeled set of tweets, the workers were setting the bar for how well the predictive features could work; after all, the performance of the predictive features will never be able to surpass human judgment.

In the end, we discovered that the simplest labeling scheme (*true/ false*) was the most tractable for the workers; they were able to work quickly and intuitively, and to handle larger datasets. Adding judges improved coverage – diverse perspectives on interestingness were represented – but decreased consistency. In evaluating the trade-offs, we discovered that the best performance came from a small number of very experienced judges rather than a large number of diverse judges. Finally, we established the folly of filtering the tweets. Although initially we felt that a smaller dataset with more valuable tweets would produce more positive labels, and more agreement on those positive labels, over time we came to understand that without sufficient exposure to patterns in the data (especially spam tweets), workers are unable to identify either interesting tweets and spam correctly. In other words, the effects of filtering may amplify poor performance, and confound any subsequent efforts to create a predictive classifier.

Interestingness is a Subjective Notion

Our initial (and perhaps naive) belief was that some tweets would stand out as more universally interesting and judges would be able to agree on their interestingness. To test this hypothesis, we computed Krippendorff's alpha [16], a statistical measure of inter-rater agreement.[1] An alpha value of 1 indicates perfect agreement among judges; a value of 0 indicates that judges are assigning labels randomly; and a negative value indicates that disagreements are systematic, for instance because some judges hold different opinions than others. The Krippendorff alpha value for dataset D1 is 0.037; in other words, inter-rater agreement was no higher than one would attribute to chance.

We attributed this low level of inter-rater agreement to the diversity of tweets, coupled with the varying interests of the judges. Judges were unable to agree whether a random tweet was interesting, but maybe they would be able to agree if we restricted the genre to a specific domain of broad interest, such as news. To test this hypothesis, we assembled three datasets of tweets authored by ten well-known news organizations (ABC, BBC, Bloomberg, Christian Science Monitor, Los Angeles Times, New York Times, Reuters, USA Today, Washington Post, and the Wall Street Journal), and randomly sampled 2500 tweets per dataset from the tweets issued by these organizations in February 2011, 2012 and 2013, respectively. We had workers on the RJ crowdsourcing platform label these tweets, showing the workers only the tweets but no profile name or image (so the credibility of a recognizable source would not further confound the judgment). We discovered that compared to D1, judges found a larger percentage of the tweets interesting, and that this percentage increased with the recency of the tweets: the percentage of interesting tweets was 11.9% for D1, 21.3% for 2011 news, 27.8% for 2012 news, and 29.3% for 2013 news. However, this higher level of interest did not result in higher inter-rater agreement: the Krippendorff alpha values were 0.037 for 2011 news, 0.074 for 2012 news, and 0.068 for 2013 news. In other words, even though judges found news tweets to be more interesting on average than random tweets, agreement on what is interesting did not improve.

To test whether such low inter-rater agreement could be attributed to judgment spam, we ran a series of seven additional streamlined labeling experiments, four on the RT platform and three on the MT platform. Each experiment used 100 randomly sampled tweets from the news genre. This time, we presented one tweet at a time to each judge, and we asked three questions: (1) how many hashtags does this tweet contain; (2) does the tweet contain a person's name; and (3) is this tweet interesting? Because question (1) can be answered algorithmically, we can use it to determine if any of the workers are spammers, selecting random answers even for clear

[1]We chose Krippendorff's alpha over other inter-rater agreement measures (such as Fleiss' kappa [11]) because it is able to handle data sets where the number of raters per item varies and some of our crowdsourced experiments contain a small number of holes.

and easy questions; this is the quality control method used by other types of crowdsourcing tasks. For example, Momeni et al. [22] employed objective questions to ensure label quality while crowdsourcing the task of labeling user comments on museum assets as useful or not.

Unfortunately, what we were experiencing was not a quality control issue. Not only did we find that there were a few spammers among the workforce, but also that they typically stopped after having completed fewer than a dozen tasks. More importantly, we found that eliminating the work of judges who did poorly on question (1) did not affect inter-rater agreement on questions (2) and (3) in any statistically significant way.

The low inter-rater agreement on whether a tweet is interesting or not – even for genres that should be broadly interesting, and even after controlling for poorly performing workers – left us convinced that interestingness is indeed a fully subjective notion; it is difficult to identify any tweets that evoke universal interest. In other words, there is little hope in constructing a classifier that identifies such tweets. Even so, it may still be possible to identify tweets that are interesting to a majority of users. An analogy would be electoral polls: while it is impossible to predict the preference of any individual voter (absent information about that individual), pollsters are doing quite well in predicting the aggregate preference of the electorate. The next section explores the possibility of such a classifier.

IDENTIFYING PREDICTIVE FEATURES

Next, we explored how well various computable signals correlate with the streamlined labels assigned by our workers. We studied 13 features and their effectiveness as signals of tweet interestingness. Six of these features are variations of the features used by Duan et al. [9] for ranking tweets as to their relevance with respect to a given query. The features are:

1. Tweet length in characters, similar to the "tweet length in words" in [9].
2. Tweet length in characters after removing "@user" mentions.
3. Follower count of tweet's author, identical to the "follower score" feature in [9].
4. Number of mentions of a tweets' author, how many times the author of a given tweet has been mentioned in other tweets via the "@user" convention.
5. Presence of mention in tweet, identical to the "Reply" feature in [9].
6. Presence of URL in tweet, identical to the "URL" feature in [9].
7. Presence of hashtag in tweet, similar to the "hash tag score" feature in [9].
8. Tweet starts with "RT", indicating a retweet.
9. Ratio of letter and digits to characters.
10. Fraction of words starting with capital letters.
11. Fraction of misspelled words.
12. Hashtag-SALSA score.
13. BM25 score of tweet averaged over all queries in a query set, similar to the BM25 feature in [9].

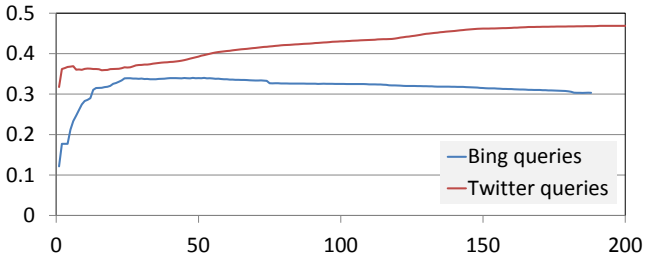

Figure 1. Correlation between average ratings and average BM25 for D1, depending on the temporal extent (in days) of the query set, for queries issued to Bing Social and to Twitter

The first four features are integer-valued (the number of characters in a tweet and the number of followers of an author). The next four features are valued either 1 or 0 (depending on the presence or absence of an "RT", a URL, a hashtag or a mention of another user). The last five features are real-valued. Feature 12 is computed by constructing a bipartite graph between tweets containing hashtags and hashtags contained in tweets, and computing SALSA hub and authority scores [17] for the tweets and hashtags, respectively. Feature 13 is based on BM25 [26], a scoring function for estimating the relevance of a document with respect to a query. We conjecture that a tweet is generally interesting if it is relevant to the information needs of a significant fraction of the audience, and we view query logs as a manifestation of that information need. We compute the BM25 relevance score of a tweet for each query in a given query log, and average these scores over the entire query log. In this study, we used two sources of query logs: Bing's Twitter search service, and Twitter's own search service, observed by mining web browsing logs.

We computed these features on data set D1, described earlier in Table 2. For the BM25 feature, we separately considered queries issued to Bing's Twitter search as well as Twitter's own search, and we averaged BM25 scores of a uniform random sample of such queries in a time window leading up to the tweet, eliminating duplicate queries. As evidenced by Figure 1, the duration of the time window affects the quality of the signal. In the following, we set the query time window to the optimum duration.

As a very basic test of each of the above thirteen features, we computed the Pearson product-moment correlation coefficient over all tweets between a given feature and the average rating of the tweet. The results are summarized in Table 7. We observe that the "presence of URL", "tweet length", and "average BM25" features are fairly strongly correlated, while the "presence of mention" feature has a moderate negative correlation (i.e. assessors find tweets that are a reply to another person less interesting – it is not exciting to read a single exchange of an ongoing conversation). Counter to common intuition, there is no correlation between the average rating of a tweet and the number of users following its author. This observation is consistent with a survey by Gayo-Avello [13] that found that high follower count is often indicative of low-quality "spam" accounts.

Categorizing tweets as interesting or not is a classic classification problem. Thus, we next explored whether these features

feature	Pearson's ρ
Tweet length	0.424
Tweet length without mentions	0.494
Follower count	0.045
Author's mentions	-0.103
Presence of mention	-0.241
Presence of URL	0.557
Presence of hashtag	0.074
Tweet starts with "RT"	0.004
AlphaNum/character ratio	0.045
Fraction of capitalized words	0.152
Fraction of misspelled words	-0.127
Hashtag-SALSA	0.065
Avg. BM25 on Bing queries	0.337
Avg. BM25 on Twitter queries	0.423

Table 7. Correlation coefficient between average ratings and individual features for data set D1

can provide useful signals to classifiers. We considered each tweet in D1 that was rated as interesting by a majority of assessors to be truly interesting, and we trained and tested the sixty-odd binary classifiers contained in the Weka machine learning toolkit on all features, using standard ten-fold cross validation. Classifier performance is quantified using Fleiss' kappa [11], a measure of agreement between classifier and assessor, where a value of 1 indicates perfect agreement and a value of 0 indicates a level of agreement that can be attributed to chance. Among the seventy-one binary classifiers included in Weka 3.6, multinomial logistic regression performed best, with Fleiss' kappa of 0.52, indicating "moderate agreement". 1654 of the 1757 uninteresting and 84 of the 113 interesting tweets were correctly classified.

CONCLUSION

Taken as a whole, Twitter is significant as a robust human sensor network. Yet the universality and accessible meaning of individual tweets varies. Naturally, some tweets will be more interesting and important to a broader audience. Likewise, others will be in coded language or directed at a very limited or local audience. We began this project with the idea that in spite of the inherent subjectivity of the endeavor, people will be able to discover, label, and agree on a set of interesting tweets.

After eliciting some characteristics that workers associate with tweets that are interesting to them as individuals, we began investigating reliable methods of getting high quality labeled data, with the aim of identifying predictive features that may be used to classify tweets. Although human curation eventually catches some of these tweets, we believe it is important to go beyond the social mechanisms that are already in place (such as retweeting) to identify interesting content; unlike retweets, this judgment task is intended to locate leading indicators, and it eliminates some types of audience bias, since judgment takes place outside of the context of the feeds a reader subscribes to.

But to classify tweets automatically, we first need to augment a training set with high-quality labels. This is by no means straightforward, since interests vary, interpretations vary, and the labeling task itself is at the same time both tedious and taxing. Through multiple labeling exercises, we investigated three contingent elements of crowdsourced tweet labeling: the workers (Can we assess the reliability of a judge's performance? Is it better to have more judges with a greater variety of interests or fewer judges with known expertise and interests?), the work (Will less subjective labels be easier to apply or will an intuitive "I'll know it when I see it" approach work? How do the recency and genre of the dataset affect the results?), and the task design.

Although at first blush, labeling tweets looks like a standard relevance judgment task, albeit with very short documents. Indeed, we began with a task design that had been successful for other labeling applications, and with a less subjective label set that others had used successfully to describe types of tweets. But we discovered that it is only through iterative tuning of these contingent elements that we can arrive at reliable labels. Along the way, we encountered various bumps in the road: there is a relationship between the overall number of judges and the achievable level of agreement; filtering works at cross-purposes to any future efforts to create a predictive classifier; and simple, intuitive labeling schemes (e.g. *true* and *false*) accelerate acquisition, but at a cost of hiding rationale and making label quality more difficult to assess.

As a result of our investigations, we discovered that the limitations of judge performance sets a ceiling on how well the classifiers can perform. Traditional IR relevance assessment approaches, while seductive, are unlikely to be appropriate for an information encountering task. Additional research will establish the optimal balance between fatigue and familiarity, and inter-assessor consistency and variability of the crowd's perspectives. Furthermore, since interestingness is complex construct, it would be valuable to weave some of its more important constituents into the judgment task.

As a final step in our investigation, we cast the detection of interesting tweets as a classification problem and examined the correlation between 13 predictive features and a tweet dataset, labeled by the crowd workers using the least problematic set of labels, *true* and *false*. In this exercise, we replicate previous findings that a link's presence is a strong signal of interestingness. We also showed that features such as tweet length (without @ mentions) and average BM25 on Twitter queries are also important indicators of quality.

One promising avenue of future work is to exploit certain query characteristics such as high temporal locality (e.g. news-driven queries) and low temporal locality (e.g. topics of long-standing interest) for fine-grained correlation of time window and signal quality. Once we are able to reliably identify interesting tweets, we also plan to investigate how they may be aggregated and used for other purposes, such as reconstructing complex events that unfold over time. By not prejudging the tweets using filters at the outset–thus embedding an irrecoverable interpretive spin–we hope to arrive at robust, comprehensive sets of interesting tweets.

ACKNOWLEDGMENTS

We would like to thank Miles Efron for his helpful guidance, as well as the anonymous HCIR reviewers for their constructive comments.

REFERENCES

1. Alonso, O., Carson, C., Gerster, D., Ji, X., and Nabar, S. U. Detecting Uninteresting Content in Text Streams. In *CSE* (July 2010), 39–42.
2. André, P., Bernstein, M. S., and Luther, K. Who gives a tweet?: evaluating microblog content value. In *CSCW* (2012), 471–474.
3. Aroyo, L., and Welty, C. Harnessing disagreement in crowdsourcing a relation extraction gold standard. Tech. Rep. RC25371 (WAT1304-058), IBM Research, 2013.
4. Bowker, G. C., and Star, S. L. *Sorting Things Out: Classification and Its Consequences*. MIT Press, 1999.
5. Boyd, D., Golder, S., and Lotan, G. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *HICSS* (2010), 1–10.
6. Choudhary, A. N., Hendrix, W., Lee, K., Palsetia, D., and Liao, W.-K. Social media evolution of the Egyptian revolution. *CACM 55*, 5 (2012), 74–80.
7. Colton, S., and Bundy, A. On the notion of interestingness in automated mathematical creativity. In *AISB* (1999), 82–91.
8. De Choudhury, M., Diakopoulos, N., and Naaman, M. Unfolding the event landscape on twitter: classification and exploration of user categories. In *CSCW* (2012), 241–244.
9. Duan, Y., Jiang, L., Qin, T., Zhou, M., and Shum, H.-Y. An empirical study on learning to rank of tweets. In *COLING* (2010), 295–303.
10. Erdelez, S. Information encountering: a conceptual framework for accidental information discovery. In *ISIC* (1997), 412–421.
11. Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychol.l Bull. 76*, 5 (1971), 378–382.
12. Gaffney, D. #iranelection: quantifying online activism. In *WebSci10* (2010).
13. Gayo-Avello, D. Nepotistic relationships in twitter and their impact on rank prestige algorithms (2013). 1250–1280.
14. Hughes, A. L., and Palen, L. Twitter adoption and use in mass convergence and emergency events. *IJEM 6*, 3/4 (2009), 248–260.
15. Hurlock, J., and Wilson, M. L. Searching twitter: Separating the tweet from the chaff. In *ICWSM* (2011).
16. Krippendorff, K. *Content Analysis*, 2nd ed. Sage Publications, 2003.
17. Lempel, R., and Moran, S. Salsa: the stochastic approach for link-structure analysis. *TOIS 19*, 2 (2001), 131–160.
18. Lewis, P. Reading the riots: Investigating England's summer of disorder. *The Guardian*, September 5, 2011.
19. Lin, T., Oren, Etzioni, and Fogarty, J. Identifying interesting assertions from the web. In *CIKM* (2009), 1787–1790.
20. Marcus, A., Bernstein, M., Badar, O., Karger, D., Madden, S., and Miller, R. Twitinfo: aggregating and visualizing microblogs for event exploration. In *CHI* (2011), 227–236.
21. Metzler, D., and Cai, C. USC/ISI at TREC 2011: Microblog track. In *TREC* (2011).
22. Momeni, E., Tao, K., Haslhofer, B., and Houben, G.-J. Identification of useful user comments in social media: A case study on Flickr commons. In *JCDL* (2013), 1–10.
23. Morris, M. R., Counts, S., Roseway, A., Hoff, A., and Schwarz, J. Tweeting is believing?: understanding microblog credibility perceptions. In *CSCW* (2012), 441–450.
24. Naaman, M., Becker, H., and Gravano, L. Hip and trendy: characterizing emerging trends on twitter. *JASIST 62*, 5 (2011), 902–918.
25. Poblete, B., Garcia, R., Mendoza, M., and Jaimes, A. Do all birds tweet the same? characterizing twitter around the world. In *CIKM* (2011), 1025–1030.
26. Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M., and Gatford, M. Okapi at TREC-3. In *TREC* (1994).
27. Sakaki, T., Okazaki, M., and Matsuo, Y. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW* (2010), 851–860.
28. Silvia, P. J. What is interesting? exploring the appraisal structure of interest. *Emotion 5*, 89–102.
29. Strauss, A., and Corbin, J. *Basics of Qualitative Research*. Sage Publications, 1998.
30. Uysal, I., and Croft, W. B. User oriented tweet ranking: a filtering approach to microblogs. In *CIKM* (2011), 2261–2264.
31. Yardi, S., Romero, D., Schoenebeck, G., and Boyd, D. Detecting spam in a twitter network. *First Monday 15*, 1 (2010).
32. Zhao, D., and Rosson, M. B. How and why people twitter: the role that micro-blogging plays in informal communication at work. In *GROUP* (2009), 243–252.
33. Zubiaga, A., Spina, D., Fresno, V., and Martínez, R. Classifying trending topics: a typology of conversation triggers on twitter. In *CIKM* (2011), 2461–2464.