

Who Gives A Tweet? Evaluating Microblog Content Value

Paul André^{1,2}, Michael S. Bernstein³, Kurt Luther⁴

¹Carnegie Mellon University
Pittsburgh, PA
paul.andre@cmu.edu

²Electronics &
Computer Science
Uni. Southampton, UK

³MIT CSAIL
Cambridge, MA
msbernst@mit.edu

⁴Georgia Institute
of Technology
luther@cc.gatech.edu

ABSTRACT

While microblog readers have a wide variety of reactions to the content they see, studies have tended to focus on extremes such as retweeting and unfollowing. To understand the broad continuum of reactions in-between, which are typically not shared publicly, we designed a website that collected the first large corpus of follower ratings on Twitter updates. Using our dataset of over 43,000 voluntary ratings, we find that nearly 36% of the rated tweets are worth reading, 25% are not, and 39% are middling. These results suggest that users tolerate a large amount of less-desired content in their feeds. We find that users value information sharing and random thoughts above me-oriented or presence updates. We also offer insight into evolving social norms, such as lack of context and misuse of @mentions and hashtags. We discuss implications for emerging practice and tool design.

ACM Classification Keywords

H5.m. Information interfaces and presentation (e.g., HCI).

General Terms

Design; Human Factors; Measurement

INTRODUCTION

Microblogging has been found to have broad value as a news and communication medium [3,6], but little is known about fine-grained content value. Existing studies focus on signals of positive or negative reactions like retweets [10] and unfollowing [5], but these signals capture only extreme reactions. Users' reactions to their feeds are often varied: items can bore, can spur interest, can be funny. However, there are no existing public signals for investigating users' more nuanced reactions at a large scale. If we could better understand what users do and do not value, and why, we could: 1) derive *design implications* for better tools or automatic filters, and 2) develop insight into *emerging norms and practice* to help users create and consume valued content.

This work contributes an analysis of microblog content from the reader's point of view, powered by a novel design for collecting large numbers of voluntary ratings. We developed a website that encourages Twitter users to give

anonymous feedback to accounts they follow in exchange for feedback from their own followers and other users. Using our corpus of approximately 43,000 ratings, we ask: 1) *What* content do Twitter users value? For example, do users value personal updates while disliking opinions? We then ask: 2) *Why* are some tweets valued more than others?

Conventional wisdom exists around these questions, but to our knowledge this is the first work to rigorously examine whether the commonly held truths are accurate. Further, by collecting many ratings, we are able to quantify effect sizes. A better understanding of content value will allow us to improve the overall experience of microblogging.

BACKGROUND

Twitter content analysis has identified a number of different categories of user and message [9], which we use in our analysis to assess perceived value by category. Previous investigations of specific scenarios have identified positive evaluations related to retweeting [10] and search evaluation [4], especially for facts or useful links. Negative outcomes have been explored in the context of unfollowing, and have been linked to network structure and 'bursts' of (uninteresting) tweets [5]. Readers may also value content for mere social communication and awareness (the equivalent of saying 'hello' in the hallway [8]) rather than for any substantive content.

An analysis of tweets by visual attention [2] suggests that interfaces can direct users to high-value content, e.g., by highlighting infrequent authors. This prior work used interest judgments to conclude that there are no externally-visible markers like replies or retweets for many of the tweets that users found interesting. This paper extends previous work by gathering a large set of judgments that contain not just value but also *content* and *reason*.

DESIGN

To understand the perceived value of Twitter content, we needed a corpus of tweet ratings. We capitalized on Twitter users' curiosity of how others view them to build this corpus. We designed a web site, called *Who Gives a Tweet* (WGAT), that delivers anonymous feedback from followers and strangers in exchange for rating tweets.

After signing in to the website, users see a list of ten tweets that they must rate before receiving feedback on their own. This mechanism capitalizes on anticipated reciprocity: users are performing an action (that provides information to the site) in the hopes of receiving something they value (ratings). WGAT finds tweets to rate from Twitter accounts

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CSCW'12, February 11–15, 2012, Seattle, Washington.
Copyright 2012 ACM 978-1-4503-1086-4/12/02...\$10.00.

that the user follows, preferring accounts that have signed up for WGAT, and filtering out @replies. The user then rates each tweet as *Worth Reading*, *OK*, or *Not Worth Reading*. Users could also skip rating any tweet. Tweets are displayed with the corresponding author name and avatar to simulate the real-world experience of reading (and judging) a Twitter feed. Users may optionally explain why they chose that rating. We collected these details via checkboxes and a freetext response. The checkbox options were: *funny*, *exciting*, *useful*, *informative*, or, *arrogant*, *boring*, *depressing*, *mean*. These adjectives were adapted from previous work [1] and iterated with pilot studies.

METHOD

After Who Gives A Tweet launched, popular news sites like Mashable, TechCrunch, OneForty and CNN wrote about the site, and the link went viral. The subsequent spike in traffic provided us with a significant number of users and ratings from many different parts of the Twitter network. We base our analysis on this data from the period of 30 Dec 2010 to 17 Jan 2011. The dataset includes 43,738 tweet ratings from 1,443 users. These users rated the accounts they follow, an even broader population of 21,014 Twitter users. All analysis is drawn only from follower ratings.

Category Labeling

We gathered a sample of 4,220 ratings from users who rated at least ten tweets. For each tweet, we determined a content category using an adapted version of Naaman's [9] tweet categorization scheme, which includes categories like Me Now (current mood or activity), Presence Maintenance (e.g., "*Hullo twitter!*"), Self Promotion (e.g., sharing a blog post the author just published), and Information Sharing. We edited the typology by removing anecdote categories as they were very rare in both datasets. We also added a Conversation category to capture many discussion-oriented tweets that did not fit cleanly into the typology.

To apply content labels, we used the paid crowdsourcing service Crowdfunder. Crowdfunder provides a high-quality result by using questions with known answers (ground truth data) to discard the submissions of workers who do not substantially agree with the ground truth. The paper authors built a ground truth dataset by following Naaman's manual tagging scheme and gave the ground truth labels to Crowdfunder along with the full set of tweets to label. Cohen's kappa between Crowdfunder's labels and a held-out set of tweets labeled by the paper authors was 0.62: moderate to strong agreement. Experimentation indicated agreement would be as high as 0.81 if ground truth included multiple categories per tweet as Naaman did.

Sample Bias and Limitations

Given the naturalistic growth of the site and the technology-centric media attention, we began by investigating potential biases in our dataset. We compared our distribution of tweet categories to Naaman et al.'s random sample [9] and found two main differences: our dataset contained more

Information Sharing tweets (49% vs 22%), and fewer Me Now tweets (10% vs 40%). This difference is likely attributable to the TechCrunch and Mashable demographic, and the inclusion of marketers and organizations in our dataset (unlike Naaman et al., who removed them).

Thus, our analysis should generalize to a population of information-sharing Twitter users, but may not apply evenly to all subpopulations on Twitter. For example, information sharing might be regarded more highly than in other samples. However, our sample is similar to previous analyses (e.g., [2, 4]), or broader than them due to viral spread, and it is orders of magnitude larger. We believe this sample is broad enough to draw valuable conclusions.

Regression Analysis

To examine the effect of tweet category on rating, we used an ordered logistic regression. Standard linear regression assumes that the outcome is ratio or interval. However, we did not believe that users' psychological distance between Not Worth Reading and Neutral was necessarily the same as the distance between Neutral and Worth Reading. For example, it is possible that the bar for tweets Not Worth Reading is lower than the bar for tweets Worth Reading. An ordered logistic regression is non-parametric and does not make this assumption, so we use it instead. We used content category as the predictor, holding out the Presence Maintenance category (the most disliked) as a baseline for the categorical dummy variables, and controlled for rater.

RESULTS

How Much of the Twitter Feed is Valued?

Followers described 36% of the rated tweets as Worth Reading (WR), thought that 25% were Not Worth Reading (NotWR), and remained neutral about the other 39%. Given that users *actively choose* to follow these accounts, it is striking that so few of the tweets are actively liked. On a per-user basis, we find that the average user finds 41% (sd=20%) of their rated tweets Worth Reading. This wide variation in quality suggests that the analyses to follow can have a large impact on the Twitter experience.

What Categories of Tweets are Valued?

It might be reasonable to assume that information sharing tweets are particularly valued, given Twitter's emphasis on real-time news. Or, it might be feasible that followers enjoy personal status updates, since they separate Twitter from other information sources like RSS feeds.

We investigated the impact of tweet category on rating using the tweets categorized by Crowdfunder. Figure 1 summarizes category ratings. The results of our ordered logistic regression analysis are in Table 1. The odds ratios in the table can be interpreted as: *Question to Followers* had 2.83 times the odds as being rated Worth Reading instead of Neutral, in comparison to a *Presence Maintenance* tweet. Figure 1 illustrates these differences graphically. For example, Presence Maintenance tweets had a 45%

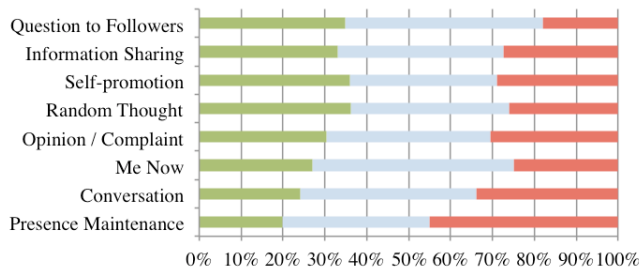


Figure 1. Ratings of a 4,220-tweet subset of our corpus. From left to right, colors indicate percentages of Worth Reading, Neutral, and Not Worth Reading. (Ordered by Odds Ratio in Table 1.)

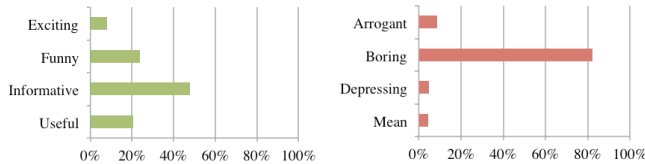


Figure 2. Based on 17,557 ratings from checkboxes, informative leads the reasons for liking a tweet, while boring dominates the reasons for disliking.

probability of being Not Worth Reading, compared to just 18% of Question to Followers tweets.

The three most strongly disliked categories were Presence Maintenance, Conversation, and Me Now (the tweeter’s current status). For example, a Me Now tweet had just 25% chance of being Worth Reading. Odds of being Worth Reading (vs. Neutral) were just 1.89 times that of Presence Maintenance; $z=1.94$, $p\approx.05$. One might reasonably expect followers to be interested in personal details. However, this does not seem to be the case. Analyzing the freetext responses to understand the reasons, we found many cases in which the follower was not interested by the tweeter’s life details, e.g., “sorry, but I don’t care what people are eating”, “too much personal info”, “He moans about this ALL THE TIME. Seriously.” There is a special hatred reserved for Foursquare location check-ins: “foursquare updates don’t need to be shared on Twitter unless there’s a relevant update to be made”, or, more simply: “4sq. ffs...”

Presence Maintenance tweets (e.g., “Hullo twitter!”) were the most strongly disliked. These tweets had variously 1.5-2.5 times worse odds than any other category. Freeform text indicated that these pieces of phatic communication were generally considered contentless: “I have one word for one word tweets: BORING”, or “useless.”

The most-liked categories were Questions to Followers, Information Sharing, and Self-Promotion (often sharing links that you created). To some extent, these results may reflect our sample bias of Twitterers. However, they also suggest that the Twitter ecosystem values learning about new content. For example, “The headline arouses my curiosity” or “Wow. Didn’t know that was happening. Thanks for informing me.” Questions to Followers were

Predictor	Odds Ratio	z value
Question to Followers	2.83	2.94*
Information Sharing	2.69	3.05*
Self-Promotion	2.69	2.61*
Random Thought	2.47	2.89*
Opinion / Complaint	2.05	1.93~
Me Now	1.89	1.94~
Conversation	1.57	1.26
Presence Maintenance	N/A	N/A

Table 1. Odds ratios of the ordered logistic regression on rating. Presence Maintenance is the baseline condition. (e.g., Question to Followers had 2.83 times the odds as being rated Worth Reading instead of Neutral, in comparison to a Presence Maintenance tweet.) $N=4220$. * $p<.01$, ~trend $p\approx.05$

often liked either because the follower thought “this is a good use of Twitter” or because of an interest in the topic itself “gives one pause to think about the question posted.”

Why are Tweets (Not) Valued by Followers?

The previous section covered *what* was valued about tweets; this section elaborates on *why*. This analysis uses our entire 43,738 tweet dataset as rated by followers. When WGAT users rated a tweet as worth reading (WR) or not (NotWR), they could also select reasons, and enter free text. Of tweets rated WR, 67% were tagged with at least one reason; 38% of those NotWR had a reason.

Not Worth Reading: Being boring, repeating old news, cryptic, or using too many # and @ signs

Being boring is far more prevalent a problem than expected. It was the standout reason for rating NotWR, accounting for 82% of all explanations (Figure 2). Because Twitter emphasizes real-time information, tweeting old information led to Boring responses like “Yes, I saw that first thing this morning” or “I’ve read this same tweet so many times.” Some users offered suggestions: “since your followers read the [New York Times] too, reposting NYT URLs is tricky unless you add something.” Boredom is also associated with banal or prosaic tweets, leading to responses like “and so what?” or “it’s fine, but a bit obvious.”

Users often complained when the tweet did not share enough context to be understandable or worthwhile. Many updates linked to a photo or blog without any other explanation: “just links are the worst thing in the world.” Local updates were a point of contention: “don’t live there, don’t care.” Negative sentiments or complaints were not worth reading: “Kinda negative :-((”, “whining.”

Twitter-specific syntax was a common source of complaint, particularly the overuse of hashtags and @mentions: “Too many tags – can hardly find the real content.” Users also disliked tweets mentioning someone rather than just @replying or Direct Messaging them: “dm thanks for rts is better”, “Twitter’s fault; feels like listening in on a private conversation.” Sometimes the extra syntax was appreciated: “If you dropped in a hashtag, I could save the search and find out the answer later.”

Our users also rated tweets by celebrities or organizations they followed. There was tension between expecting a professional insight, and getting personal ones: “*I unfollowed you for this tweet. I don’t know you; I followed you b/c of your job.*” News organizations should consider the tension between giving all the information in a tweet, and piquing a user’s curiosity: “*Newsy, and all the news I want is here. Not much of a tease.*”

Worth Reading: Information, humor, conciseness

Ratings revealed that our users primarily valued Twitter as an information medium. Tweets worth reading were often informative (48%) or funny (24%), as seen in Figure 2. These tags had very little overlap: a tweet was often one or the other, but not both. Information links were valued for novelty or an appealing description: “*interesting perspective on something I know nothing about*”, “*makes you want to know more.*” Humor was a successful way to share random thoughts or opinions especially: “*it’s witty and snarky. worth the read.*” In keeping with Twitter’s focus on short messages, followers appreciated conciseness: “*few words to say much, very clear.*” A human aspect was also appreciated: “*personal, honest and transparent.*”

LIMITATIONS & FUTURE WORK

Our volunteer population was skewed towards technologists or “informers” [9]. Though our results provide insight into this user base, it will be important in future work to address all types of users and understand which results generalize, particularly whether there are different communities in Twitter with different value judgments.

We asked users to rate tweets, but not rate the person who tweeted. There may be a social obligation to follow people whose tweets are perhaps not personally valued. Long-term ratings of one’s feed would enable detailed analysis on a per-user basis. An analysis of users *no longer* followed would provide another perspective on the value discussion, as would self-ratings on users’ own tweets. We would also like to consider the effect of potentially unvalued tweets actually having a meta-level value in maintaining awareness and relationships.

DISCUSSION & CONCLUSION

Social media technologies present both new opportunities for connection, as well as new tensions and conflicts [7]. As a first step at answering questions of microblog content value, we designed a website to collect the first large corpus of follower ratings on Twitter updates. Using 43,000 volunteer ratings on tweets, we asked what is (or is not) valued, and why.

Distribution. Our sample of Twitter users rated 36% of tweets as worth reading, 25% as not, and 39% as middling. The average user rated 41% (sd=20%) of tweets as worth reading. In a personally curated stream, it may be surprising that so few rated tweets were considered worth reading.

Content. Information sharing, self-promotion (links to personally created content) and questions to followers were valued highly, while presence maintenance, conversational and ‘me now’ statuses were less valued.

Emerging Practices. Our analysis suggests: embed more context in tweets (and be less cryptic); add personal commentary, especially if retweeting a common news source; don’t overuse hashtags and use direct messages (DMs) rather than @mentions if more appropriate; happy sentiments are valued and “whining” is disliked, and questions should use a unique hashtag so followers can keep track of the conversation.

We see two directions for utilizing these results, and a comparison to other sites with social media updates. Facebook, for example, has invested significant time and experimentation to determine who and what to show in one’s newsfeed. Twitter, on the other hand, has been successful despite, or because of, very simple presentation (essentially viewing all updates). Thus, the first direction is *technological* intervention: design implications to make the most of what is valued, or reduce or repurpose what is not. The second focuses more on Twitter’s simplistic view at the moment and taking a *social* intervention approach: helping to inform users about perceived value, audience reaction and emerging norms, but ultimately leaving users in control of what they share and what is seen. Both approaches have the potential to address issues of value and audience reaction, improving the experience of microblogging for all.

ACKNOWLEDGMENTS

We thank Robert Kraut, m.c. schraefel, Jaime Teevan, Ryen White, Sarita Yardi, and anonymous reviewers for helpful feedback on this and earlier versions of this paper.

REFERENCES

1. Barash, V., Duchenaus, N., Isaacs, E. & Bellotti, V. Faceplant: Impression (Mis)management in Facebook Status Updates. *Proc. ICWSM 2010*.
2. Counts, S., & Fisher, K. Taking It All In? Visual Attention in Microblog Consumption. *Proc. ICWSM 2011*.
3. Ehrlich, K. & Shami, S. Microblogging Inside and Outside the Workplace. *Proc. ICWSM 2010*.
4. Hurlock, J., & Wilson, M.L. Searching Twitter: Separating the Tweet from the Chaff. *Proc. ICWSM 2011*.
5. Kwak, H., Chun, H. & Moon, S. Fragile online relationship: a first look at unfollow dynamics in twitter. *Proc. CHI 2011*.
6. Kwak, H., Lee, C., Park, H., & Moon, S. What is Twitter, a Social Network or a News Media? *Proc. WWW 2010*.
7. Marwick, A. E. & boyd, d. Twitter Users, context collapse, and the imagined audience. *New Media & Society* 13(1), 2011.
8. Miller, V. New Media, Networking and Phatic Culture. *Convergence* 2008, 14:387.
9. Naaman, M., Boase, J. & Lai, C. Is it really about me? Message Content in Social Awareness Streams. *CSCW 2010*.
10. Suh, B., Hong, L., Pirolli, P. & Chi, E. H. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network. *Proc. SocialCom 2010*.