Unfolding the Event Landscape on Twitter: Classification and Exploration of User Categories

Munmun De Choudhury*
Microsoft Research
munmund@microsoft.com

Nicholas Diakopoulos
Rutgers University
diakop@rutgers.edu

Mor Naaman
Rutgers University
mor@rutgers.edu

ABSTRACT
Social media platforms such as Twitter garner significant attention from very large audiences in response to real-world events. Automatically establishing *who* is participating in information production or conversation around events can improve event content consumption, help expose the stakeholders in the event and their varied interests, and even help steer subsequent coverage of an event by journalists. In this paper, we take initial steps towards building an automatic classifier for user types on Twitter, focusing on three core user categories that are reflective of the information production and consumption processes around events: organizations, journalists/media bloggers, and ordinary individuals. Exploration of the user categories on a range of events shows distinctive characteristics in terms of the proportion of each user type, as well as differences in the nature of content each shared around the events.

Author Keywords
Events, Social media, Twitter, User classification

ACM Classification Keywords
J.4 Computer Applications: Social And Behavioral Sciences

INTRODUCTION
As a computer mediated communication channel, Twitter has become a powerful destination for people posting content around events. Whether those events be the marriage of royalty, the destruction due to a tornado, or the launch of a product, there is valuable information about these events to be gleaned by examining the content shared by thousands of individuals. In particular, the composition of communities posting content for events is of interest. This paper takes initial steps towards building an automatic classifier for user types on Twitter, differentiating between organizations, journalists/bloggers, and ordinary individuals. We then use those classifications to characterize a series of diverse events.

Such study of intrinsic differences in events and the types of participants and reactions they induce is interesting in its own right, to develop a theory of event-related information production, and can enable the next generation of event content exploration tools. Such tools will improve the understanding of the stakeholders in an event and their actions and interests, and even help steer subsequent coverage of an event by the mass media. Understanding event participation in the light of user categories can also help estimate source credibility in reference to a particular topic of inquiry.

This work builds on that of Starbird et al. [7] and Vieweg et al. [8], that addressed the characterization of events by the types of Twitter users that posted messages about them. However, these studies rely on manual coding of users, and only examine events of one type (natural hazards). Other recent studies have considered characterizing users on Twitter outside the context of events. These efforts include using Twitter lists to identify elite users such as celebrities, bloggers, media, and organizations [9], classification of political affiliation [1], and classification of demographic attributes [3]. Other efforts have looked at characterizing events using various data such as content of tweets, the network structure of event participants, the temporal signature of event messages as well as user participation breakdown [6, 4].

In this work, we classify users according to three primary categories: *organizations*, *journalists/media bloggers*, and *ordinary individuals*, that may participate in events in ways that exhibit different interaction, content, and usage patterns. For instance, organizations may coordinate rescue efforts for emergency events, ordinary individuals may have "eyewitness" information [8], and journalists may engage in synthesis and amplification as second hand information providers [7]. In this light, the contributions of this work include: (1) developing a robust automatic classifier for user types on Twitter, thus overcoming the limitations of manual coding efforts; and (2) extending the understanding of participation in Twitter activity around events by using our classifier to examine participation in a diverse range of events.

CLASSIFYING USER CATEGORIES ON TWITTER
As noted above, we focus on three user archetypes: organizations, journalists/bloggers, and ordinary users.

Organizations are entities associated with some social, political, or business goal and often have a Twitter account for the sake of marketing, public relations, or customer service. Organizations can be commercial (e.g. a marketing firm), or non-profit (e.g. an NGO on breast cancer awareness) and could include a company, brand, product, or charity.

*This work was performed while the author was at Rutgers.

CSCW 2012, February 11–15, 2012, Seattle, Washington, USA.

Journalists/Media Bloggers include individuals who are associated with a mass media enterprise/news organization or who maintain a blog that reflects their professional interests on various local and global issues.

Ordinary Individuals are users likely to be on Twitter for a variety of reasons such as posting updates on their day-to-day life, expanding professional opportunities, maintaining contacts with their friends, or to discover relevant/useful content relating to their interests.

Other. This final category includes all users who do not satisfy the properties of any of the above three categories.

We note here that in certain cases, there could be Twitter users who belong to more than one category, or their category affiliation is ambiguous. For example, a user may be a low visibility journalist, not strictly associated with a media organization, making her Twitter behavior similar to both an ordinary individual and a journalist. For simplicity, in this paper we assume that each user can belong uniquely to *exactly one* category. Multi-category probabilistic assignments will be addressed in future research.

Classification Methodology. We use a standard machine learning framework to classify Twitter users into our categories. For the classification features, each user is represented as a vector of the following features, selected to capture the differences between categories:

- *Network/structural features* consist of the indegree (#followers) and outdegree (#friends) of a user.
- *Activity features* consist of the number of posts by a user until the time of the crawl and the number of posts they marked as favorites.
- *Interaction features* include the patterns of how a user engages with their Twitter "audience". The specific features are: the fraction of re-tweets among all the posts from a user, the fraction of replies directed to other users, the fraction of @-mentions of other users in a user's posts and the fraction of posts with URLs in them.
- *Named entities* capture the presence or absence of named entities in the content of the posts from a user (like a place name, or a company). The named entities are derived using the natural language toolkit OpenCalais.
- *Topic distribution* is the topical association of the history of a user, to a set of 18 broad themes from the IPTC Media Topic News Codes, again derived using OpenCalais.

Next we compared eight different classifiers to empirically determine the best suitable classification technique. We also compared to a baseline method that used the profile description of a user's Twitter account to infer categories. The baseline checked for the existence of terms specific to organizations (e.g. company, brand, charity) or journalists/media bloggers in the profile text. The best performing classifier was found to be k Nearest Neighbors with $k = 10$ (that also far outperformed the baseline by $\sim$30%). Due to limited space we report only on this classifier here.

TRAINING AND VALIDATION

We begin by describing our training model which we developed by gathering and labeling users from the Twitter public timeline, as well as collecting users from pre-existing labeled directories on Twellow and Muckrack.

We first gathered a random sample of 1,850 users from the Twitter public timeline during April 2011(http://twitter.com/public_timeline). The users from this source were categorized using *Amazon Mechanical Turk* (http://www.mturk.com). We set up the tasks so that they could only be performed by Turkers located in the U.S. with a more than 95% approval rating. Each Turker was presented with a Twitter user's screen name, linked to the user's Twitter profile page. We had each user's profile labeled by two Turkers (Scott's PI was moderate at 0.55) and then excluded exemplars where there was no agreement (16.4% of sample) to train our model on less ambiguous data.

On top of the timeline we also used the *Twellow* directory (http://www.twellow.com/) to collect a list of 1,532 organizations and the *Muckrack* directory (http://muckrack.com/) to collect a list of 1,490 self-identified journalists and media bloggers on Twitter.

Combining the three different sources above, we obtained a training dataset comprising 4,932 Twitter users, who are labeled into one of the four categories. For each labeled user in our training dataset, we used the Twitter API to collect the data required to compute the features described above. We used each user's 200 most recent Twitter status updates (through May 2011) to compute the named entities and topic distribution features.

Actual \ Predicted	Organizations	J/MB	OI	Other
Organizations	**1341**	240	24	6
J/MB	42	**1957**	19	5
OI	9	27	**1208**	11
Other	2	7	22	**12**

Table 1. Confusion matrix combined over all the five folds of cross validation on the training dataset.

We validated our training model using five-fold cross validation. The confusion matrix of classification is shown in Table 1. We note that our training model performs extremely well for "Journalists/Media bloggers" (J/MB) and "Ordinary Individuals" (OI). Table 2 further shows the balanced accuracy[1], precision, recall and F1 measure for the label predictions. We observe high values of the four metrics across the categories, which suggests that our training model is suitable to be used for categorizing users on Twitter. The poor performance for the 'Other' category can be explained by its lower number of users, as well as the fact that the category is loosely defined—comprising users who do not fit into the other categories.

TESTING ON EVENTS

Since our goal is to understand the user categories spanning different events on Twitter, we now describe the various

[1]Balanced accuracy curbs skewness in category sizes, and is given as the mean of sensitivity (true positive rate) and specificity (true negative rate).

Category	#users	Bal. acc.	Precision	Recall	F1
Org	1611	94.21%	0.96	0.82	0.89
J/MB	2023	92.41%	0.86	0.97	0.91
OI	1255	96.45%	0.94	0.96	0.95
Other	43	71.01%	0.42	0.27	0.33

Table 2. Balanced accuracy, precision, recall and F1 measure of the results of predicting user category labels using our classifier. Results are averaged across the five cross validation folds.

events used in this work, and the categorization of users in them using our classifier. We first collected data on a set of events from Twitter. For each event we manually identified keywords, and then used the Twitter streaming API to download a large sample of English-language tweets corresponding to the keywords that are (presumably) representative of the event's discussion on Twitter. Table 3 provides information about the number of posts and number of users in each event sample.

EVENT NAME	#POSTS	#USERS
Bonnaroo 2011	22,159	13,685
Search keywords: #bonnaroo		
Earth day 2011	587,750	359,131
Search keywords: earth day, #earthday		
2011 Egyptian Revolution	4,798,828	530,629
Search keywords: #egypt, egypt riots, egypt revolution		
Release of iPad 2	151,356	78,768
Search keywords: #ipad2		
Osama Bin Laden's death	3,258,319	1,656,967
Search keywords: #OBL, osama, osama bin laden		
Academy Awards 2011	854,154	293,601
Search keywords: #oscars		
Super Bowl XLV 2011	831,172	399,209
Search keywords: #superbowl2011, superbowl		
Wikileaks	72,448	26,143
Search keywords: #wikileaks, #cablegate, #assange		

Table 3. Description of the eight events used in this paper.

To construct our exploration dataset on each of the eight events, we randomly sampled 5,000 users from the set of posts for each event. For each such user, we collected the data required to construct the features used in the classifier. The previously discussed nearest neighbor classifier was then utilized to label each user (of the 5,000 users in each event) to a particular category. For validation, we generate ground truth labels for 100 random users per event using Mechanical Turk, as before. The results of comparing the user labels indicate that our classification technique yields high accuracy over all the events (88.73%, averaged over the three core categories); demonstrating that our training model is efficient in categorizing Twitter users over a diverse set of events.

EXPLORING USER CATEGORIES

Participatory Behavior of User Types

We show here how users of different categories participated in each of the eight events. Participation is evaluated using the relative representation of each category in the events, in terms of users (the proportion of the number of users in the category to the number of users captured in our test set), and in terms of content (the proportion of content posted by users of each particular category for the event). The results are shown in Figure 1, and discussed below. We wish to emphasize here that verification across a much larger set of events may be necessary for the purposes of generalization.

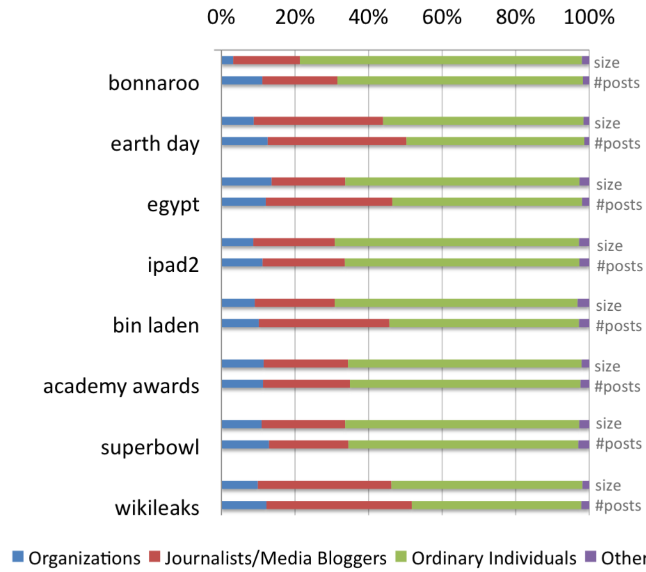

Figure 1. Distribution of sizes and posts from different user categories.

Figure 1 shows, perhaps unsurprisingly, that localized events such as Bonnaroo Music Festival are characterized by a large number of ordinary individuals that post a large proportion of the content. On the other hand, national events, such as Earth Day and the Wikileaks news event seem to involve a large proportion of organizations and journalists/media bloggers. Earth Day, for example, is more likely to involve organizations and journalists/bloggers spreading environmental awareness. Similarly, Wikileaks was an event of great interest to news organizations and political journalists/ media bloggers: consequently these categories seem to have posted a larger proportion of content for Wikileaks than their numbers predict. Events manifesting as 'breaking news', such as the Egyptian revolution and the death of Bin Laden appear to gather even more participation and attention, in terms of category size and disproportionate fraction of posts, from organizations and journalists. Planned events scheduled to happen at a fixed date and time or at a known location/set of locations (e.g. release of iPad 2, the Academy Awards and Superbowl XLV) appear to involve large participation from ordinary individuals, perhaps due to the fact that some of these events are nationally televised [6].

Content Characteristics of User Categories

The classification also allows us to examine the differences in terms of characteristics of content posted by different categories of users. Due to lack of space, we demonstrate this potential for only two of our events, a local event (Bonnaroo), and a mass-media news event (Wikileaks). For each event and user category, we report numbers that reflect the proportion of posts on the event having URLs, that are @-replies to other users, is a re-tweet, reflect an exclamatory sentiment (presence of the exclamation sign) and are questions. These five aspects of the posts are indicative of various types of content typically shared on Twitter.

The results are shown in Table 4. We observe that in these events, organizations have a relatively higher number of posts with URLs. For Bonnaroo, the organizations tend to be

Category	Frequent Keywords	Category-specific Keywords
	bonnaroo	
Organizations	festival, live, video, headline, news, tickets	#madness, #newfavoritesong, rock, videos
Journalists/Bloggers	you, festival, live, see, artist, listen	#partyplane, #coachella, fans, stars, volunteer
Ordinary Indv.	my, festival, going, live, we, excited, awesome	seeing, performing, bad, interested, haha, epic
	wikileaks	
Organizations	#cablegate, they, government, censorship, news	documents, statement, threat, report, classified
Journalists/Bloggers	why, #cablegate, who, secret, media, blog	#obama, secrecy, rumors, diplomats, president
Ordinary Indv.	assange, us, truth, my, secret, government	interpol, law, #iamwikileaks, information

Table 4. High frequency keywords (stopword eliminated, based on tf–idf computation) associated with different user categories and category-specific keywords corresponding to two events.

more 'interactive', as shown by the higher proportion of @-replies. Organizations and journalists appear to re-tweet more extensively in the Wikileaks dataset, perhaps propagating released information and news updates by re-tweeting. Ordinary individuals are characterized in these events by a low proportion of URLs in their posts, and a higher proportion of interactive posts that are @-replies to others.

In both events, organizations appear to pose very few questions ("?" column), whereas journalists/media bloggers and ordinary individuals exhibit the opposite behavior, perhaps indicating interactive behavior seeking feedback, and thoughts of others on the events. The expression of exclamations are also more frequent in the case of all categories for Bonnaroo, compared to Wikileaks, which provides evidence that there are inherent differences among events, and that user types respond differently in the context of different events.

Category	URL	@	RT	!	?
		bonnaroo			
Organizations	0.77	0.58	0.07	0.33	0.01
Journalists/Bloggers	0.52	0.48	0.16	0.28	0.13
Ordinary Individuals	0.37	0.56	0.13	0.19	0.11
		wikileaks			
Organizations	0.71	0.31	0.41	0.05	0.06
Journalists/Bloggers	0.63	0.39	0.37	0.05	0.09
Ordinary Individuals	0.63	0.52	0.29	0.08	0.09

Table 5. Different content types shared by user categories corresponding to two events. The columns are not mutually exclusive, i.e., a post with a '?' which is also an 'RT' is considered for both the columns.

In the final segment of our exploratory analysis on content characteristics of user categories, Table 5 shows high frequency keywords and keywords appearing more frequently in a particular category (left) compared to the rest (right). We chose to compare the same two events, Bonnaroo and Wikileaks. Again, there are differences in the content shared across the categories on the two events. Organizations tend to use words such as headline and news frequently, reflecting that there could be several news organizations reporting on the events. Journalists/bloggers tend to engage in more interactive behavior by using terms such as 'you' and by reflecting their interests with words like media and blog. Finally, for ordinary individuals we observe the presence of sentiment words (e.g., excited, awesome, bad) and heavy reference to themselves in the first person (e.g., my, us).

CONCLUSIONS AND FUTURE WORK

In this paper, we took initial steps towards categorizing the stakeholders of events on Twitter. We developed a classifier that automatically identified users whose behavior corresponded to three core categories: organizations, journalists/media bloggers and ordinary individuals. Exploration of events from Twitter provided several interesting insights. First, we showed that different events gather different degrees of participation (number of users from each category) and attention (proportion of posts from each category). Second, we demonstrated differences in the nature of the content that were shared by the categories on different events. In our examples, while organizations tend to frequently point to external information sources via URLs on their posts, ordinary individuals appeared to be more reflective of their personal experiences and sentiments on the events, and were observed to engage in greater interaction with others.

ACKNOWLEDGMENTS

This work was supported by a Google Research Award, and by the National Science Foundation under award IIS-1017845.

REFERENCES

1. J. An, M. Cha, K. Gummadi, and J. Crowcroft. Media landscape in twitter: A world of new conventions and political diversity. In *Proc. ICWSM '11*.
2. M. De Choudhury, S. Counts, and M. Czerwinski. Identifying relevant social media content: Leveraging information diversity and user cognition. In *HT '11*.
3. D. Gayo-Avello. All liasons are dangerous when all your friends are known to us. In *HT '11*.
4. M. Naaman, H. Becker, and L. Gravano. Hip and trendy: Characterizing emerging trends on twitter. *JASIST*, 62(2):902–918, 2011.
5. M. Naaman, J. Boase, and C.-H. Lai. Is it really about me?: message content in social awareness streams. In *Proc. CSCW '10*.
6. D. A. Shamma, L. Kennedy, and E. F. Churchill. Conversational shadows: Describing live media events using short messages. In *Proc. ICWSM*, 2010.
7. K. Starbird, L. Palen, A. Hughes, S. Vieweg. Chatter on the red: What hazards threat reveals about the social life of microblogged information. In *CSCW '10*.
8. S. Vieweg, A. L. Hughes, K. Starbird, and L. Palen. Microblogging during two natural hazards events: What twitter may contribute to situational awareness. In *Proc. CHI*, 2010.
9. S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on twitter. In *Proc. WWW '11*.