Topical Engagement on Twitter: Using Consistency of Activity as a Means of User Segmentation

Aurora C. Schmidt, Clay Fink, and Nathan D. Bos
The Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Road, Laurel, MD 20723-6099
aurora.schmidt@jhuapl.edu, clayton.fink@jhuapl.edu, nathan.bos@jhuapl.edu

Abstract

In this paper we introduce a measure for the consistency of Twitter user behavior as a way of measuring engagement with a topic. Using this measure of topical engagement, we filter tweets collected from over 90,000 Nigerian users over a three-year period concerning that country's head of state, President Goodluck Jonathan. We show that this measure effectively identifies a small set of highly engaged users that produce a disproportionate amount of activity, both in tweet volume and mentions of other users. Additionally, we show preferential within-group activity, with these users disproportionately mentioning other highly engaged users. Lastly, we show the potential of using changes in the percentage of engaged users within the overall user set for gauging the significance of events that occurred in Nigeria over the 2011 to early 2013 period.

Introduction

Social media platforms such as Twitter give people the ability to register their opinions and evaluations of emerging and ongoing events. They also allow them to pass on information they consider to be of importance to their followers. Individual users may approach a particular topic in different ways. Some may be very engaged in the topic because of personal or professional interest and tweet about it consistently over a period of time. Others may only tweet episodically or when there is a major event associated with a topic.

Political scientists have long been interested in identifying the most politically engaged segments of a population. This segment may serve as agenda-setters or opinion leaders [Zaller, 1992]. While it is clear that a small proportion of social media users produce a disproportionate amount of content, the properties and functions of classes of users who are highly engaged politically are not well understood. We examine what a population of politically engaged Twitter users can tell us about a topic. In particular, can this population's activity be used to determine whether events associated with the topic are ephemeral or have true newsworthiness or significance?

So what is a good measure of topical engagement? One approach would be to look at the overall amount of content from users about a topic, and defining engaged users as those with the most tweets. Unfortunately, this would ignore users who tweet consistently about a topic, but not at a high volume, and would include users who tweeted at a high rate over a very short period of time - say, during a single day - but never return to the topic. Given such cases, a measure that strikes a balance between tweet volume and the consistency of tweeting for each user is preferred. We describe a measure of engagement based on the entropy of a user's daily tweet rate over a given interval of days that captures how a user's tweets are dispersed over time.

In the following sections we discuss related work; describe our Twitter data set; introduce our measure of engagement and define a technique for detecting anomalous events from tweet volume; investigate some properties of the users we classify as engaged, including how the proportion of engaged users can be used to flag significant events; and conclude with a summary of our results.

Related Work

How segments of social media users differentially engage with topics has not been well studied. There has been work on identifying nodes in a user network who are the most cost effective to monitor for information cascades corresponding to emerging news events [Leskovec et al., 2007][Zhao et al., 2013]. These approaches look at this as a sensor placement problem and apply social network analysis techniques to identify individual users that are most active propagating information, which in itself can be looked at as a measure of engagement. The approach we describe, in contrast, does not take into consideration network information, relying only on a user's daily tweeting frequency. To our knowledge, update frequency has not been investigated as a method for studying the topical engagement of social media users. However, update frequency has been used to identify automated accounts (i.e. "bots") on Twitter. [Chu et al., 2012] describe a entropy based measure for detecting the burstiness of user updates, assuming that users that are less bursty are likely to be automated accounts.

Data

The work described here focuses on Twitter data from the West African nation of Nigeria. Using the Twitter search

API, we collected tweets from nearby 45 Nigerian cites with populations of 100,000 or more. We then used the API to consistently poll the timelines of users returned from the geographic queries. Starting in April of 2010, we collected 219 million tweets from one million users through the beginning of June 2013. These counts reflect data from users that we verified as being from Nigeria by matching profile locations against a gazetteer of Nigerian place names or checking that the tweets returned with geotags were within the queried locations. Tweets are mostly in English and Nigerian Pidgin English.

We focus on tweets that referenced Nigerian President Goodluck Jonathan. Jonathan became president in May 2010 upon the death of President Umaru Yar'Adua, for whom he served as Vice President. He stood for election in April 2011 and won a four year term. Extracting tweets about Jonathan is problematic since his first name is often used in tweets as a salutation and his last name is a common first name. We resolve this issue by using an expansion query containing terms that co-occur significantly with Jonathan's full name and his initials (GEJ), scoring the results using a probabilistic information retrieval technique [Robertson and Jones, 1976]. Using a minimum query score of 12 (derived by observation of the results) we retrieved 596,750 tweets from 90,557 users. This represents the topical portion of tweets we analyze in this paper.

We made an effort to exclude automated accounts, or bots, since this content is not from actual individuals. We filtered out bots based on the assumption that bots will show a low level of interaction with other users via mentions and retweets. Based on the complete volume of tweets from all users we identified as topical, we calculated the percentages of their tweets that were retweets or contained mentions. We label as bots those users for which we had 100 or more tweets and had both retweet and mention rates below ten percent. This identified 1,326 accounts. Manual inspection of a sample of these users confirmed that this was a reasonable method for identifying bots; many of the accounts identified were for major Nigerian media sources or were obviously bots.

Methods

Recent work on the usage behavior of online resources, such as Hulu video streaming, show users tend to behave in a bursty or clumpy manner, [Zhang, Bradlow, and Small, 2013]. We also observed frequent clumpiness in a single user's Twitter activity, especially as pertains to our topic of interest. By clumpy, we mean the user's activity may rapidly fluctuate from frequent tweeting to much lower or even nonexistent tweeting.

For the purpose of characterizing a user's engagement with a topic, we use a measure of consistency of behavior over a window of 60 days. A reliable metric for consistency is the entropy of the fraction of tweets that arrived each day during the fixed time period; *tweet fraction entropy*. For an N-day window, the user's daily tweet fractions, denoted as $\mathbf{f} = [f_1, f_2, \ldots, f_N]$, may be written as,

$$\begin{bmatrix} \frac{c_1}{C_T} & \frac{c_2}{C_T} & \cdots & \frac{c_N}{C_T} \end{bmatrix} \tag{1}$$

where c_i is the number of tweets made by the user on day i, for $i \in [1, N]$, and C_T is the total number of tweets over all N days; i.e., $C_T = \sum_{i=1}^{N} c_i$. The tweet fraction entropy of user k for the window ending at time t is then defined by

$$e_k(t) \;=\; H(\mathbf{f}_t) \;=\; -\sum_{i=1}^{N} f_i \log_2(f_i) \tag{2}$$

where $\mathbf{f}_t$ is the daily fractions of tweets for the N-day interval ending at time t. The entropy in (2) may be recognized as the entropy of a discrete variable with N possible values, [Cover and Thomas, 2006]. The tweet fraction entropy can be interpreted as the entropy for the distribution of a single tweet having arrived on each day of the period. We will refer to engagement and tweet fraction entropy interchangeably.

We use a sliding 60-day window over the entire three year period of topically relevant tweets to measure each user's time-varying engagement with the topic. This measure of consistency ranks a user's engagement as high if they tweet similar numbers of topical tweets each day, rather than looking at their total, possibly bursty, activity. We note that tweet fraction entropy reaches a maximum when the user tweets exactly the same, nonzero, number of topical tweets on every day of the period, resulting in an entropy of $\log_2(N)$.

Tweet fraction entropy is a way of decoupling consistency from the raw volume generated by a user. However, to achieve a particular entropy, for example three bits, a user must have tweeted a minimum of eight times in the period. In the next section, Experiments, we compare the metrics of average rate of topical tweets to the engagement metric. This shows the lower bound on activity needed to achieve a certain level of engagement, but also shows that variation among users makes the two metrics different.

User Segmentation

One use for our measure of users' topical engagement is to segment users into groups according to their overall level of engagement. This potentially allows us to identify the groups of highly politically engaged users that we speculate about in the introduction. Our approach to user segmentation is to compute the average of the user's engagement metric over all 60-day sliding windows ending in the time period of interest. Since our dataset was collected over three years, and during a time in which Twitter was rapidly growing in number of users and popularity, we needed to account for users that joined or left Twitter during the time period. We use the time of each user's first and last tweet (not necessarily topical) as an estimate of their join and leave dates, and only compute their expected engagement metric over windows overlapping with their time of 'existence'.

After ranking all users by their average engagement, we then split users with average engagement above zero into four quartiles. A fifth group contains all users with expected engagement exactly equal to zero. This last group has many more users than the first four and is due to the power-law activity of users in a social network; i.e., there are many users with very low activity and a small set of users with high activity. The set of bots we identified –as described in the data section above –are treated as the sixth segment. The result of this segmentation is summarized in Table 1.

Segment	Avg. Engagement Range	Number of Users
1	[0.37, 4.15] bits	9061
2	[0.18, 0.37] bits	9061
3	[0.09, 0.18] bits	9061
4	[0.00, 0.09] bits	9060
5	[0, 0] bits	52644
Bots	[0, 4.75] bits	1326

Table 1: Segments of users based on first four quartiles of average engagement plus the segment of users with exactly zero average engagement and automated users.

Volume-Based Event Detection

A major motivation for tracking user activity in Twitter is to detect events that affect a population. We here describe a simple method for measuring the anomalousness of aggregated user activity based on volume statistics, construing such differences in activity as being markers for an event.

Before comparing volumes of activity over a long period, we must first renormalize the activity by the estimated growth in Twitter popularity and usage. To estimate the growth in Twitter we fit a linear trend to the total topical volume over the full 3 years. The cost function used for this fit, rather than least squares, was the ℓ_1-distance, in order to reduce sensitivity to the frequent spikes in activity and better fit the overall trend. These growth parameters were solved using the convex optimization software package, cvx, [Grant and Boyd, 2013], to compute the scalar parameters a and b given by

$$\text{minimize} \quad \|\mathbf{v} - a\mathbf{t} - b\|_1 \quad \text{subject to} \quad b \geq 0 \tag{3}$$

where $\mathbf{v}$ is the vector of daily topical volume and $\mathbf{t}$ represents the corresponding day numbers from day 1 to day 1148. Based on this, we estimated a topical growth rate of $a = 0.59$ additional tweets per day.

After dividing the aggregate daily volume by the growth rate of Twitter, we next learn a baseline variance of activity for daily tweets. Using a sliding 30-day window we compute a variance of the activity over time, $\sigma^2(t)$. To find the baseline variance x, since the activity is bursty, we again use ℓ_1-norm as cost to reduce sensitivity to outliers.

$$\text{minimize} \quad \|\sigma^2(t) - x\|_1 \quad \forall\ t \tag{4}$$

Using the baseline variance for activity, we then compare increases above a weighted 7-day average to the baseline variance to characterize how anomalous that activity was. Specifically, the rescaled volume is first filtered by a 7-day averaging window with exponential weights. The filter used has the impulse response, $h(t) = [0.553, 0.248, 0.112, 0.050, 0.023, 0.010, 0.005]$, for $t = 1, 2, \ldots, 7$. We then compare each day's rescaled volume to the filtered 7-day averages to compute deviations from the average. The anomalousness of each positive deviation is measured by the negative log of the chi-squared probability of each deviation.[1] As a simple threshold we use a anomaly score of 10 or above to detect potential events.

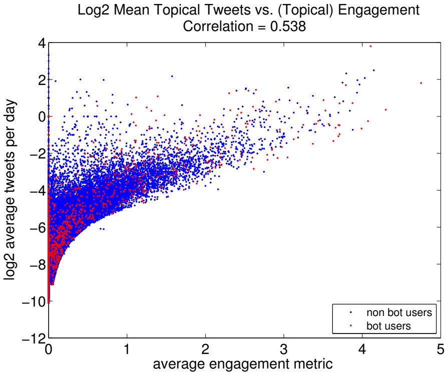

Figure 1: Comparison of user's average engagement (in bits) versus the log of their average topical tweet rate.

Experiments and Results

Engagement vs. Tweet-rate

We first compared the chosen engagement metric, 60-day tweet-fraction-entropy, to each user's mean tweet rate, also over the same 60-day window. Figure 1 shows the comparison of average engagement with the mean tweet rate in log space, converted to match the units of the engagement metric, which is in bits. We can clearly see the lower bound on tweets required to achieve each level of engagement. However, at each engagement level, there remains variation in the population of how many tweets they make. This is due to variations in the consistency of each user's behavior, and is reflected in the correlation of these two statistics, $\rho = 0.538$.

Volume of Activity of Each Segment

We found that the level of engagement of users was predictive of the amount of topical volume the users generated. Examining the segment of the most highly topically engaged users, we found they made up a disproportionate amount of the total volume of tweets mentioning President Jonathan.

The most engaged segment of users made up 52.5% of the topical volume despite being only 10% of the total population. However, these highly engaged users only make up approximately 0.38% of all tweets, topical and otherwise, showing that engagement with the topic is not explained simply by engagement with Twitter. On average topical tweets made up 9.6% of our entire database of tweets.

Inter-Segment Mentions

We examined the network behavior of the six groups of users during the period of the 2011 presidential campaign from January to April. Table 2 shows the group-to-group mentions. Users in the most engaged category were responsible for over 23% of all mentions, despite being only 10% of the population. In addition, they were most likely to mention members of their own group, with over 26.7% of all mentions being of other users in group one. The next highest proportion of mentions originating from group one was of group five, at 23.5%, due to the dominating size of that group (i.e. users with a tweet-fraction-entropy of zero). The

[1] In Matlab, this probability is computed using 1 - the chi2cdf of the normalized deviation.

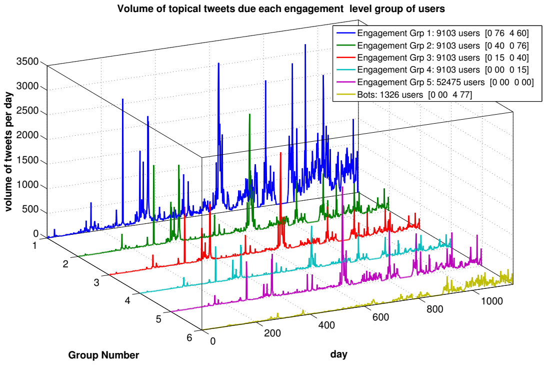

Figure 2: Topical tweet volume produced by each engagement segment.

disproportionate level of intragroup mentions for highly engaged users suggests that this group may have interesting network-level properties worth further investigation.

Grp.	1	2	3	4	5	Bots
1	151969	106476	97210	76806	133807	2370
2	95850	89552	85326	65757	128493	1271
3	85385	84451	84075	69661	134527	1048
4	58049	59166	62351	55887	110914	681
5	79764	91758	103269	93814	203502	918
Bots	383	221	225	164	260	78

Table 2: Num. mentions by segment (row) of users in (col).

Gauging Event Response

In Methods, we described an anomaly-based detector for identifying events. The lower panel of figure 3 shows detected events, circled, based on the total volume of topical tweets. Looking at the peak volumes alone, we might think an event that occurred on March 17, 2011 was most significant. This was the reaction to an interview of Jonathan by the Nigerian pop star, D'banj. Despite generating a huge volume of traffic on that day, the event is ephemeral in terms of significance. The top panel of fig. 3 shows the fraction of users whose engagement was above 2 bits as a function of date. Judging from the level of user engagement, the two events with the most significance occur in mid-April of 2011 and in January of 2012. The first period corresponds to the presidential election, which was followed by riots in the country's northern regions. The second occurs after the removal of a fuel subsidy that led to mass protests and a major downturn in Jonathan's popularity. We may view the fraction of engaged users as a population-level statistic, whose computation scales with the size of the population monitored as well as time-period, but shows the potential for tracking engagement to improve the measure of the significance of events to a population over simply the volume of activity.

Conclusions and Future Work

We describe a measure that is useful in finding users that are highly engaged with a topic. These users are responsible for a disproportionate share of the content and tend to preferentially mention each other. Also, we find that the proportion of highly engaged users is potentially useful for flagging significant vs. ephemeral events as compared to tweet volume alone. Follow-on work will include investigating the relationship between user engagement and the level of positive or negative sentiment toward a particular entity.

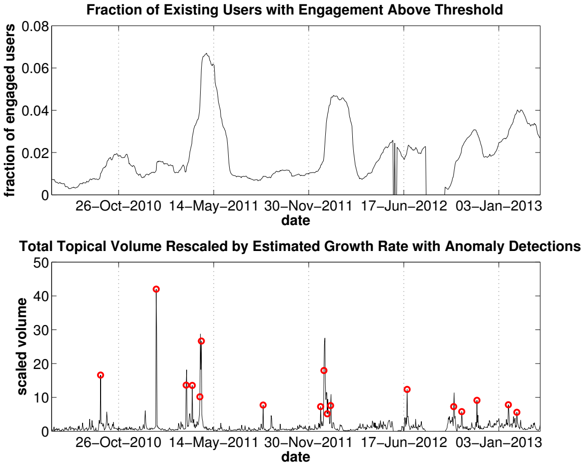

Figure 3: Comparison of the fraction of users with engagement metric above 2 bits versus total topical volume-based anomaly event detections.

Acknowledgments

This work was funded under Office of Naval Research (ONR) grant N00014-10-1-0523.

References

Chu, Z.; Gianvecchio, S.; Wang, H.; and Jajodia, S. 2012. Detecting automation of Twitter accounts: Are you a human, bot, or cyborg? *IEEE Trans. on Dependable and Secure Comp.* 9(6):811–824.

Cover, T. M., and Thomas, J. A. 2006. *Elements of Information Theory*. John Wiley and Sons, Inc., second edition.

Grant, M., and Boyd, S. 2013. CVX: Matlab software for disciplined convex programming, version 2.0 beta. http://cvxr.com/cvx.

Leskovec, J.; Krause, A.; Guestrin, C.; Faloutsos, C.; VanBriesen, J.; and Glance, N. 2007. Cost-effective outbreak detection in networks. In *Proceedings of the 13th ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining*, 420–429. ACM.

Robertson, S. E., and Jones, K. S. 1976. Relevance weighting of search terms. *Journal of the American Society for Information science* 27(3):129–146.

Zaller, J. R. 1992. *The Nature and Origins of Mass Opinion*. Cambridge University Press.

Zhang, Y.; Bradlow, E. T.; and Small, D. S. 2013. New measures of clumpiness for incidence data. *Journal of Applied Statistics* 40(11):2533–2548.

Zhao, J.; Lui, J.; Towsley, D.; Guan, X.; and Wang, P. 2013. Social sensor placement in large scale networks: A graph sampling perspective. *arXiv preprint arXiv:1305.6489*.