Understanding Communities via Hashtag Engagement: A Clustering Based Approach

Orianna DeMasi University of California, Berkeley
Douglas Mason Pinterest
Jeff Ma Twitter, Inc.

ICWSM 2016

Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)

Understanding Communities via Hashtag Engagement: A Clustering Based Approach

Orianna DeMasi *
University of California, Berkeley
Berkeley, CA 94720
odemasi@eecs.berkeley.edu

Douglas Mason †
Pinterest
San Francisco, CA 94103
douglas@pinterest.com

Jeff Ma
Twitter, Inc.
San Francisco, CA 94103
jma@twitter.com

Abstract

We develop insight into community use of hashtags on social media and find that hashtags with behavior indicative of real world communities are more engaging. To do this, we study the relationship of hashtag usage with user engagement on Twitter. Hashtag engagement is useful as a surrogate measure of how active community members are. We develop a framework for describing hashtag temporal usage, show the existence of 4 broad classes of hashtags, and show that the engagement of a hashtag varies significantly between classes. Periodically used hashtags, such as for TV shows and weekly community chats, are the most engaging, while hashtags relating to events are the least engaging. Looking at how community dynamics vary within this framework reveals that a hashtag being used more frequently is not positively correlated with it being more engaging. We then explore the periodically used hashtags and find negative correlations with diversity of the user base, which implies concentrated communities are the most engaging. We conclude by studying a set of community conversation-oriented hashtags and find these hashtags to be more engaging than other hashtags, regardless of dynamic type. Our findings support the hypothesis that hashtags with stronger community behavior are more engaging.

Introduction

Hashtags have become a cornerstone of online social media. From their birth on the Twitter platform, hashtags have evolved from their basic form of a short string of text preceded by a pound symbol to be a tool for myriad purposes, e.g. ad campaigns and online chats (Yang et al. 2012). In addition to being versatile, they are exceedingly popular and have been adopted on nearly every social media platform.

While hashtags may have been introduced to catalog information, they now can enable users to rally social movements (e.g. #BlackLivesMatter a social movement for racial justice), disseminate public health campaigns (e.g. #ECigTruths from Chicago Department of Public Health), and connect to communities (e.g. #RunChat a community for runners). Finding and connecting users to relevant communities online is of paramount importance for improving user experience, and hashtags can potentially enable such connections. Even though researchers have acknowledged this possibility (Laniado and Mika 2010; Rasoo and Now 2010), little research thus far has targeted understanding the community use of hashtags.

In this study we focus on how engaging different types of hashtags are and consider relations with community metrics. We find that hashtags that have a stronger resemblance to real world communities are more engaging.

One of our contributions is a framework for understanding different dynamic types of hashtags. With this framework, we unify previous work on identifying periodic hashtags (Cook, Kenthapadi, and Mishra 2013) and event hashtags (Cha et al. 2010; Lehmann et al. 2012; Crane and Sornette 2008; Lin et al. 2013; Shamma, Kennedy, and Churchill 2011). This unification validates previous observations that there are coherent dynamic types of hashtags (Hsu, Chang, and Chen 2010; Romero, Meeder, and Kleinberg 2011).

Our second contribution is a set of analyses using this framework on a comprehensive cohort dataset. We include a comparison of engagement between hashtag types. Our analyses take steps in the direction of understanding engagement of hashtag types. This understanding is important, not just as a retrospective analysis, but as an actionable way for finding, connecting, and supporting communities.

One of our findings is that periodically recurring hashtags are the most engaging type of hashtag, on average. Previous work that has analyzed peaks in hashtag usage, i.e. events, is minimally actionable, as events are difficult to predict. In contrast, periodic events are predictable, so the ability to identify and understand periodically used hashtags has implications for how to design and implement new features for social media. Such new features could include weekly checkins on relevant content or community features built up around a periodic event. By showing that periodically used hashtags are the most engaging, our work implies that systems implemented around this periodic content would have the most impact. While our work focuses on hashtags, its broader implication is that systems designed to leverage periodic content to connect recurring communities will create a more engaging experience than connecting more ephemeral groups, e.g. groups that connect over events.

The structure of this paper is as follows. We begin by motivating the complexity of hashtags. We then propose a

*Work does as intern as Twitter, Inc.
†Work done as employee at Twitter, Inc.

102

Hashtag Engagement

- **Contributions**
  - Temporal usage of hashtags
  - C**ategorizing** hashtag types and analyzing the implications for engagement and communities
  - A framework for clustering hashtags based on temporal usage
  - Proposing a metric of engagement
    - compare this metric of engagement between the hashtag types

Categorizing Hashtag Usage: Features

- **h**: hashtag
- **T(h)** : set of Tweets contain **h**
- Vol, popularity: **V(h)** = log(|T(h)|)
- **f(h,t)** : % of tweets that were made with ***h*** during a given time, ***t***

Features:

- Max(f(h,t)) in an hour
- t= 24h centered around peak
- #Tweets in 4h around peak / #Tweets in 24h centered around peak
- Max(f(h, t)) , t= each day of the week
- An indicator of whether every hour of the study period had a low percentage of the volume
- An indicator of whether every day of the week had a low percentage of the total volume

#Clusters: using ***silhouette*** metric

Community Metrics

- Two measures:
  - **Engagement E(h)**
    - To quantify how engaging is a hashtag
    - h has “received an engagement” if it has been either **Retweeted** or **Favoured**
    - E(h) is the proportion of Tweets with a hashtag that have received an engagement
    - it is robust to the phenomena of a hyper popular Tweet receiving thousands or millions of engagements.
  - **Diversity D(h)**

$$E(h) = \frac{\sum_{\tau \in T(h)} \mathbb{I}_{[\tau \text{ has Retweet or Favouring}]}}{|T(h)|}$$

Community Metrics II

- Two measures:
  - **Diversity D(h)**
    - To quantify how broadly a hashtag is adopted

$$D(h) = \frac{|U(h)|}{|T(h)|}.$$

    - is the reciprocal of the average number of times a user Tweets with the hashtag
    - Abnormally low diversity is indicative of a spammer or bot driving the hashtag usage

Dataset

- All Tweets that:
  - have **English language**
  - are from users in the **United States**
  - Using a **hashtag at least once** during the 30 day study

- Study period starting

  January 15, 2015

Users	19,197,367
Tweets	2,529,886,239
Tweets with #	437,167,710
Hashtag occurrences	801,850,909
Unique hashtags	18,149,314
Popular hashtags	34,500

Dataset : Removing spammers and bots

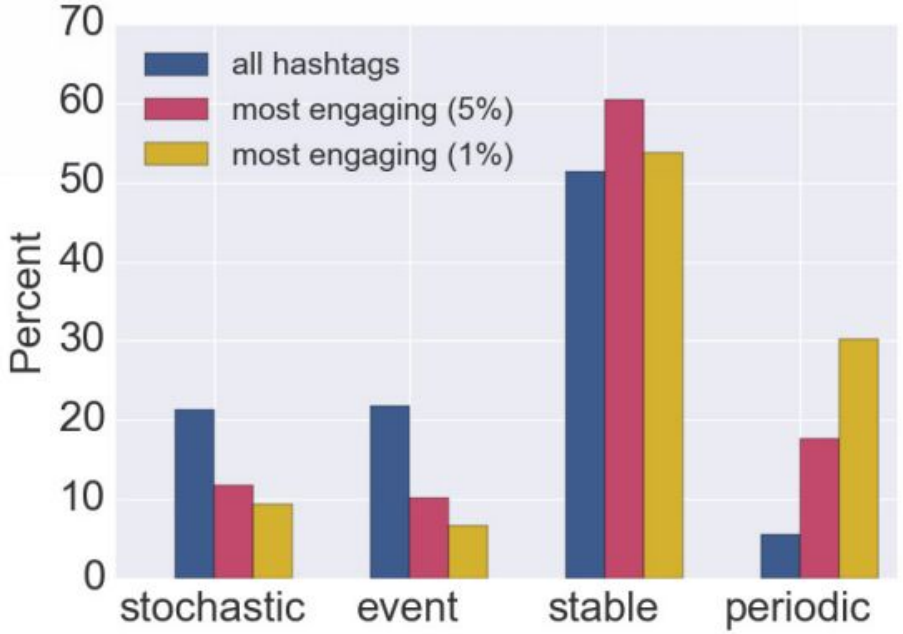
- Minimal adoption
  - h with extremely low diversity ($D(h) \leq .02$)
  - 1,581 hashtags
  - mostly represented advertisers of pornography

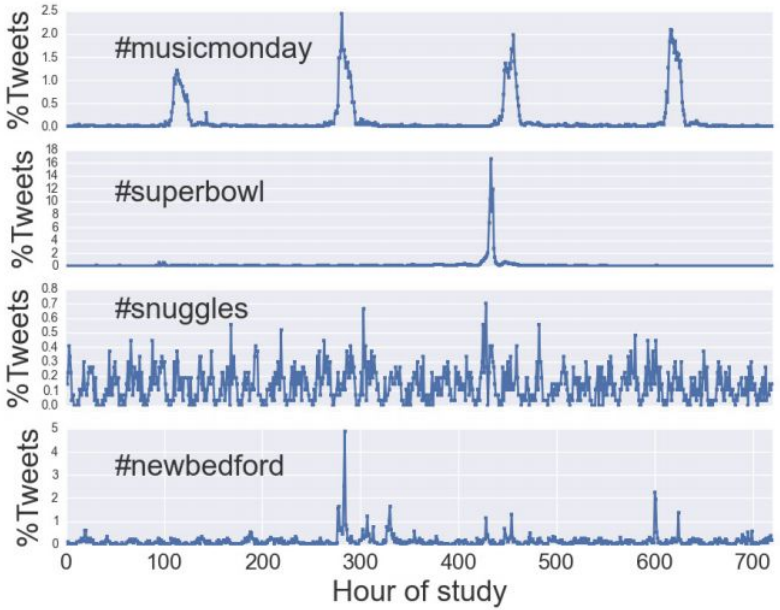
- Zero engagement
  - $E(h)=0$
  - 1,745 hashtags
  - mostly represented by Islamic propaganda

Dynamic Types

Using K-means clustering

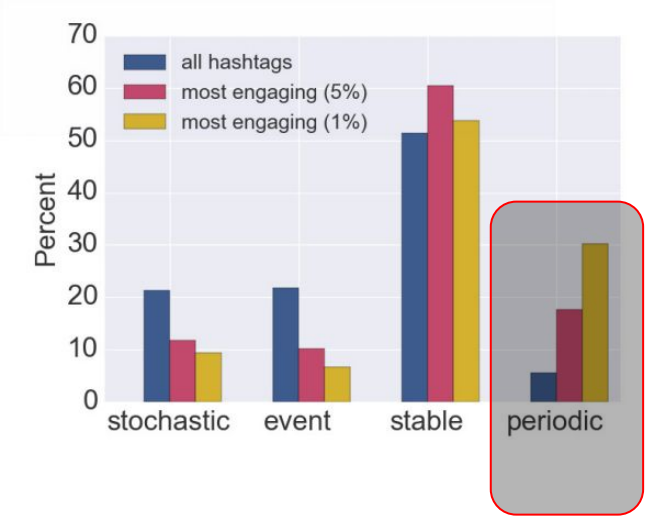
The clusters were validated by extensive manual inspection of a randomly selected subset of hashtags.





Refined periodically recurring subtypes

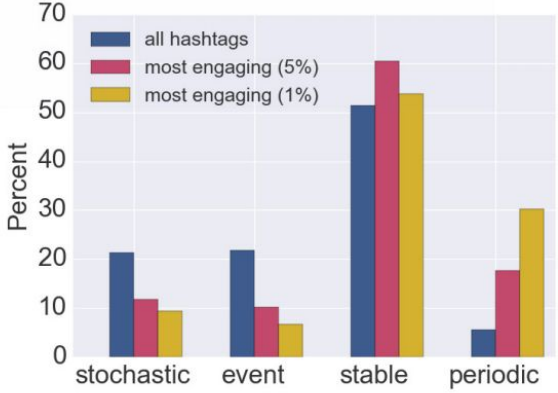
The clusters were validated by extensive manual inspection of a randomly selected subset of hashtags.



- All day events (#monday)
- Weekly events (TV shows)
- Periodic events with strong imbalance between events or less than weekly
- Events more frequent than weekly, or significant support on some days of a week (daily chats)

Engagement varies between dynamic types

- Comparing distributions of hashtag engagement
  - The dynamics of how a hashtag is being used is related to how engaging the hashtag is
  - Periodically recurring hashtags cluster is the **most engaging**
  - Cluster of **event** hashtags is **least engaging**
  - periodic content could be leveraged to connect users with more engaging content



Volume and diversity

- There is a lack of positive correlation of **engagement** with popularity for all dynamic hashtag types
  - Volume does not increase engagement
  - Lower diversity can be more engaging

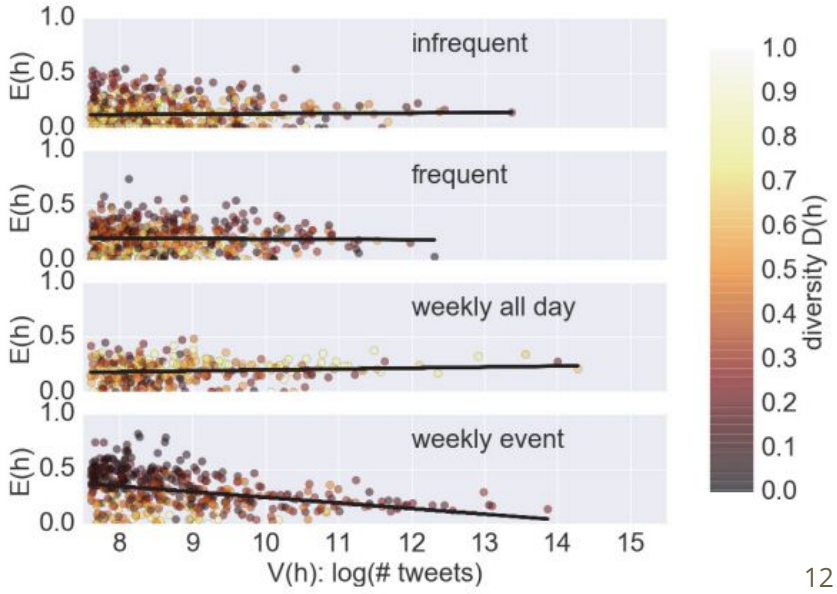
Measures:

- % tweets with Link and Mention
- Average number of hashtags in tweets

	coeff.
Cluster: Event	
$V(h)$	-0.0113
$D(h)$	-0.0541
% links	-0.0704
% mentions	-0.0408
# hashtags	-0.0138

Subclusters of periodically recurring

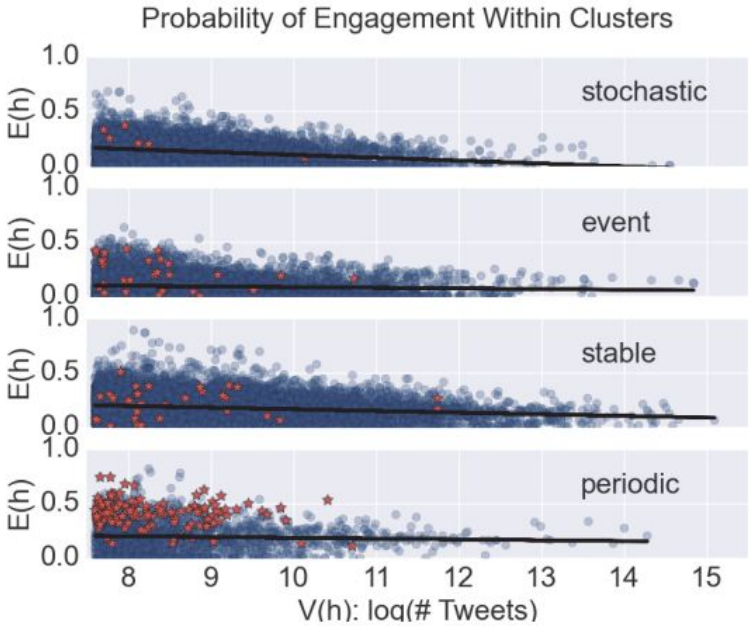
- Low diversity implies a concentrated user group.
- Hashtags for weekly events with low diversity are the most engaging
  - Lower diversity with higher engagement hints at **focused community structure**.
  - This result shows that to find engaging weekly event hashtags, looking at the size of the user base or # Tweets is insufficient.



Community-oriented “chats” are more engaging

- Chat hashtags: hashtags contain "chat", e.g., #dadChat , #phdChat
  - Engagement is higher for chats even though they do not have a relatively large volume
  - These observations support the broader observation that community-oriented hashtags are more engaging.
  - They also indicate that different types of periodically occurring hashtags exist

Cluster	chat μ	non-chat μ
event	0.213	0.095
stochastic	0.240	0.134
stable	0.223	0.185
periodic	0.414	0.179



The End