Understanding Communities via Hashtag Engagement: A Clustering Based Approach

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

ICWSM 2016
Hashtag Engagement

● Contributions
  ○ Temporal usage of hashtags
  ○ Categorizing 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)**

\[
E(h) = \frac{\sum \mathbb{1}_{[\tau \text{ has Retweet or Favouring}]} }{|T(h)|}
\]
Community Metrics II

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

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

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
<td>19,197,367</td>
</tr>
<tr>
<td>Tweets</td>
<td>2,529,886,239</td>
</tr>
<tr>
<td>Tweets with #</td>
<td>437,167,710</td>
</tr>
<tr>
<td>Hashtag occurrences</td>
<td>801,850,909</td>
</tr>
<tr>
<td>Unique hashtags</td>
<td>18,149,314</td>
</tr>
<tr>
<td>Popular hashtags</td>
<td>34,500</td>
</tr>
</tbody>
</table>
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

\[
\begin{array}{c|c}
\text{Cluster: Event} & \text{coeff.} \\
\hline
V(h) & -0.0113 \\
D(h) & -0.0541 \\
\% \text{ links} & -0.0704 \\
\% \text{ mentions} & -0.0408 \\
\# \text{ hashtags} & -0.0138 \\
\end{array}
\]
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

<table>
<thead>
<tr>
<th>Cluster</th>
<th>chat $\mu$</th>
<th>non-chat $\mu$</th>
</tr>
</thead>
<tbody>
<tr>
<td>event</td>
<td>0.213</td>
<td>0.095</td>
</tr>
<tr>
<td>stochastic</td>
<td>0.240</td>
<td>0.134</td>
</tr>
<tr>
<td>stable</td>
<td>0.223</td>
<td>0.185</td>
</tr>
<tr>
<td>periodic</td>
<td>0.414</td>
<td>0.179</td>
</tr>
</tbody>
</table>
The End