Hashtags on Twitter: Linguistic Aspects, Popularity Prediction and Information Diffusion

Pawan Goyal

CSE, IITKGP

July 24-28, 2014
What are #Hashtags?

Come under the general category of memes; a short unit of text that spreads from person to person within a culture.

**syntax**

Adding a hash symbol (#) before a string of

- letters
- numerical digits or
- underscore sign (_)
Why are Hashtags useful?

Hashtags are used to classify messages, propagate ideas and to promote specific topics and people. Allow users to create communities of people interested in the same topic, making it easier to find and share information related to it.
Hashtags are used to classify messages, propagate ideas and to promote specific topics and people.
Why are Hashtags useful?

- Hashtags are used to classify messages, propagate ideas and to promote specific topics and people.
- Allow users to create communities of people interested in the same topic, making it easier to find and share information related to it.
A new social event can lead to the simultaneous emergence of several different hashtags, each one generated by a different user. They can either be accepted by other members of the network or not. Some propagate and thrive, while others die eventually or immediately after birth, being restricted to a few messages.
A new social event can lead to the simultaneous emergence of several different hashtags, each one generated by a different user.
A new social event can lead to the simultaneous emergence of several different hashtags, each one generated by a different user. They can either be accepted by other members of the network or not.
A new social event can lead to the simultaneous emergence of several different hashtags, each one generated by a different user. They can either be accepted by other members of the network or not. Some propagate and thrive, while others die eventually or immediately after birth, being restricted to a few messages.
Novelty’s propagation process

Subgraphs from Twitter dataset showing two distinct moments in the process of spreading #musicmonday
Research Objectives

To understand the process of propagation of innovative hashtags in light of linguistic theories.

Interesting Questions

- Does the distribution of the hashtags in frequency rankings follow some pattern, as the words in the lexicon of a language?
- Is the length of a hashtag a factor that influences to its success or failure?
Dataset Used

- 55 million users
- 2 billion follow links
- 8% users ignored because the profile was private
55 million users
2 billion follow links
8% users ignored because the profile was private
More than 1.7 billion tweets between July 2006 and August 2009 were analyzed
What aspect of the tweets would be important?

- Must find interchangeable hashtags, i.e. different tags used for the same purpose, to characterize messages on the same topic.
What aspect of the tweets would be important?

- Must find interchangeable hashtags, i.e. different tags used for the same purpose, to characterize messages on the same topic.
- For example, #michaeljackson, #mj, #jackson refer to the same subject.
A minor base was built for each of the following topics:

- Michael Jackson (singer’s death widely reported during that period) → MJ
- Swine Flu (H1N1 epidemic as major issue of 2009) → SF
- Music Monday (a very successful campaign in favor of posting tweets related to music on Mondays) → MM
Forming the bases

Filtering tweets that included

▶ MJ: 'michael jackson'
▶ SF: 'swine flu' or '#swineflu'
▶ MM: '#musicmonday'

Hashtags on Twitter
July 24-28, 2014
Filtering tweets that included

- at least one hashtag

▶ MJ: 'michael jackson'
▶ SF: 'swine flu' or '#swineflu'
▶ MM: '#musicmonday'
Filtering tweets that included

- atleast one hashtag
- at least one of the following terms that was thought to be related to the topics:
  - MJ: ‘michael jackson’
  - SF: ‘swine flu’ or ‘#swineflu’
  - MM: ‘#musicmonday’
Summary information

- Number of tweets posted in that base
- Number of users who posted tweets
- Number of follow links among users of the base
- Number of different hashtags used in the tweets of the base
Summary information

- Number of tweets posted in that base
- Number of users who posted tweets
- Number of follow links among users of the base
- Number of different hashtags used in the tweets of the base

<table>
<thead>
<tr>
<th>Base</th>
<th>Tweets</th>
<th>Users</th>
<th>Follow links</th>
<th>Different hashtags</th>
</tr>
</thead>
<tbody>
<tr>
<td>MJ</td>
<td>221,128</td>
<td>91,176</td>
<td>3,171,118</td>
<td>19,679</td>
</tr>
<tr>
<td>SF</td>
<td>295,333</td>
<td>83,211</td>
<td>5,806,407</td>
<td>17,196</td>
</tr>
<tr>
<td>MM</td>
<td>835,883</td>
<td>196,411</td>
<td>7,136,213</td>
<td>16,005</td>
</tr>
</tbody>
</table>

Table 1. Summary information about the bases built.
Empirical Phenomenon

**Rich-get-richer phenomenon**

*Also known as ‘preferential attachment process’: In some systems, the popularity of the most common items tends to increase faster than the popularity of the less common ones.*

**Zipf’s Law**

Frequency of words in English or any other language follow a power law.

\[ f \propto \frac{1}{r} \]
Empirical Phenomenon

**Rich-get-richer phenomenon**

*Also known as ‘preferential attachment process’*: In some systems, the popularity of the most common items tends to increase faster than the popularity of the less common ones.

---

**Zipf’s Law**

Frequency of words in English or any other language follow a power law.

\[
f \propto \frac{1}{r}
\]

\[
\log(f) = \log(k) - \log(r)
\]
Distribution of Hashtags

- $i$-tweets hashtags: hashtags appearing in at most $i$ tweets
- $j$-tweet hashtags: hashtags that appear in at least $j$ tweets
Distribution of Hashtags

- $i$-tweets hashtags: hashtags appearing in at most $i$ tweets
- $j$-tweet hashtags: hashtags that appear in at least $j$ tweets

<table>
<thead>
<tr>
<th>Base</th>
<th>% of $i$-tweet hashtags inside the base $i=1$</th>
<th>$i=2$</th>
<th>$i=10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MJ</td>
<td>59%</td>
<td>72%</td>
<td>88%</td>
</tr>
<tr>
<td>SF</td>
<td>59%</td>
<td>73%</td>
<td>92%</td>
</tr>
<tr>
<td>MM</td>
<td>60%</td>
<td>74%</td>
<td>91%</td>
</tr>
</tbody>
</table>

Table 2. Distribution of less common hashtags of each base.

<table>
<thead>
<tr>
<th>Base</th>
<th>number of $j$-tweet hashtags inside the base $j=10,000$</th>
<th>$j=5,000$</th>
<th>$j=1,000$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MJ</td>
<td>3</td>
<td>6</td>
<td>28</td>
</tr>
<tr>
<td>SF</td>
<td>3</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>MM</td>
<td>2</td>
<td>3</td>
<td>28</td>
</tr>
</tbody>
</table>

Table 3. Distribution of most popular hashtags of each base.
Data from most used Hashtags

- Music Monday, slope = -1.163
- Michael Jackson, slope = -1.140
- Swine Flu, slope = -1.037
Volume of Tweets vs. its position in popularity ranking

<table>
<thead>
<tr>
<th>Base</th>
<th>Most used</th>
<th>2\textsuperscript{nd} most used</th>
<th>3\textsuperscript{rd} most used</th>
</tr>
</thead>
<tbody>
<tr>
<td>MJ</td>
<td>#michaeljackson</td>
<td>#michael</td>
<td>#mj</td>
</tr>
<tr>
<td></td>
<td>35,861</td>
<td>27,298</td>
<td>16,758</td>
</tr>
<tr>
<td></td>
<td>12.3%</td>
<td>9.3%</td>
<td>5.7%</td>
</tr>
<tr>
<td>SF</td>
<td>#swineflu</td>
<td>#h1n1</td>
<td>#swine</td>
</tr>
<tr>
<td></td>
<td>230,457</td>
<td>70,693</td>
<td>12,444</td>
</tr>
<tr>
<td></td>
<td>51.5%</td>
<td>15.8%</td>
<td>2.8%</td>
</tr>
<tr>
<td>MM</td>
<td>#musicmonday</td>
<td>#musicmondays</td>
<td>#music</td>
</tr>
<tr>
<td></td>
<td>824,778</td>
<td>11,770</td>
<td>5,106</td>
</tr>
<tr>
<td></td>
<td>79.7%</td>
<td>1.1%</td>
<td>0.5%</td>
</tr>
</tbody>
</table>
Zipf’s Other Laws: Word length and word frequency

Word frequency is inversely proportional to their length.

The length of the most popular hashtags was compared with the the less popular ones.

Main Findings

- Most popular ones are simple, direct and short
- Among those with little utilization, many are formed by long strings of characters
## Most Common Hashtags and most common 15-character hashtags

<table>
<thead>
<tr>
<th>Most common hashtags (number of tweets)</th>
<th>Most common hashtags with 15 or more characters (number of tweets)</th>
</tr>
</thead>
<tbody>
<tr>
<td>#michaeljackson (35,861)</td>
<td>#nothingpersonal (962)</td>
</tr>
<tr>
<td>#michael (27,298)</td>
<td>#iwillneverforget (912)</td>
</tr>
<tr>
<td>#mj (16,758)</td>
<td>#thankyoumichael (690)</td>
</tr>
<tr>
<td>#swineflu (230,457)</td>
<td>#swinefluhatesyou (1,056)</td>
</tr>
<tr>
<td>#h1n1 (70,693)</td>
<td>#crapnamesforpubs (145)</td>
</tr>
<tr>
<td>#swine (12,444)</td>
<td>#superhappyfunflu (124)</td>
</tr>
<tr>
<td>#musicmonday (824,778)</td>
<td>#musicmondayhttp (540)</td>
</tr>
<tr>
<td>#musicmondays (11,770)</td>
<td>#fatpeoplearesexier (471)</td>
</tr>
<tr>
<td>#music (5,106)</td>
<td>#crapurbanlegends (23)</td>
</tr>
</tbody>
</table>
## Average Length

<table>
<thead>
<tr>
<th>Topic</th>
<th>Average length of...</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>...the <em>k</em> most popular hashtags</td>
<td><em>k</em>=10</td>
<td><em>k</em>=20</td>
<td><em>k</em>=30</td>
<td><em>k</em>=40</td>
<td><em>k</em>=50</td>
</tr>
<tr>
<td>MJ</td>
<td>7.1</td>
<td>6.85</td>
<td>7.8</td>
<td>8.02</td>
<td>7.74</td>
<td>10.16</td>
</tr>
<tr>
<td>SF</td>
<td>5.3</td>
<td>7.35</td>
<td>7.17</td>
<td>7.2</td>
<td>7.04</td>
<td>10.3</td>
</tr>
<tr>
<td>MM</td>
<td>9.5</td>
<td>8.4</td>
<td>7.27</td>
<td>6.4</td>
<td>5.92</td>
<td>11.66</td>
</tr>
</tbody>
</table>

Table 6. Average length of the most and the less popular hashtags. The samples with the less popular hashtags were formed by 50 randomly selected hashtags among those which appeared only in one tweet of each base.
Average Number of Characters

- Michael Jackson
- Music Monday
- Swine Flu

Number of characters (mean ± sd)

- K = 10
- K = 20
- K = 30
- K = 40
- K = 50
- Random

Pawan Goyal (IIT Kharagpur)

Hashtags on Twitter

July 24-28, 2014
## Underscores in Hashtags

A table showing the number of underscored hashtags and the percentage of these hashtags among the first $i$ tweets for different bases.

<table>
<thead>
<tr>
<th>Base</th>
<th>Number of underscored hashtags</th>
<th>% of underscored hashtags among $i$-tweet hashtags</th>
</tr>
</thead>
<tbody>
<tr>
<td>MJ</td>
<td>251 (1.2%)</td>
<td>$i=2$ 89%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$i=10$ 97%</td>
</tr>
<tr>
<td>SF</td>
<td>155 (0.9%)</td>
<td>$i=2$ 87%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$i=10$ 97%</td>
</tr>
<tr>
<td>MM</td>
<td>143 (0.9%)</td>
<td>$i=2$ 89%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>$i=10$ 98%</td>
</tr>
</tbody>
</table>
Underscores in Hashtags

Distribution of hashtags containing the sign “_”

- 97% of _-hashtags are used in 10 or less tweets
- #michael_jackson: position 248, only 128 tweets
- #swine_flu: position 67, only 246 tweets
- #music_monday: wasn’t even used

User behavior seems to indicate rejection of this sign
Objective

Given an idea/meme $m$, and a time frame $t$, can we predict the acceptance of $m$ in the community (social network)?
Content-based Popularity Prediction

Objective

Given an idea/meme $m$, and a time frame $t$, can we predict the acceptance of $m$ in the community (social network)?

Interesting Questions

- Can we accurately predict the acceptance of a meme based solely on the meme’s content?
- Does the meme’s context improve the prediction?
- Relation between graph topology and the content and how do they integrate for efficient propagation?
Corpus Used

400 million tweets, tweeted between June-December, 2009.
Filtering and Normalization

Filtering

- Filtered tweets containing non-Latin characters, to maintain a corpus of English tweets only
- Hashtags that appear over 100 times

Normalization

The same hashtag could have got different counts because of being introduced in a different week.

\[ N(ht) = \sum_{j \in \text{weeks}} \text{count}(ht_{ij}) \cdot w_j \]
### Filetering and Normalization

#### Filtering
- Filtered tweets containing non-Latin characters, to maintain a corpus of English tweets only
- Hashtags that appear over 100 times

#### Normalization
The same hashtag could have got different counts because of being introduced in a different week.

\[
N(ht^i) = \sum_{j \in \text{weeks}} \frac{\text{count}(ht^i_j) w_1}{w_j}
\]
Fresh Hashtags

Hashtags that did not get popular before the corpus was collected.
Fresh Hashtags

Hashtags that did not get popular before the corpus was collected.

**Definition**

A hashtag is defined as *fresh* if it did not appear in the first week or if its normalized count in the first week is less than 10% of its normalized count in its peak.
Figure 2: Four typical temporal trends (unnormalized counts).
Identifying the distinct words in a hashtag

#thankyousachin
Identifying the distinct words in a hashtag

#thankyousachin

Issues with hashtags: #savethenhs, #weluvjb

- Matching hashtags against a lexicon of English words
- Exploiting redundancy of hashtags that differ only orthographically
Preprocessing

Identifying the distinct words in a hashtag

#thankyousachin

Issues with hashtags: #savethenhs, #weluvjb

- Matching hashtags against a lexicon of English words
- Exploiting redundancy of hashtags that differ only orthographically

matches tuples like #freeiran, #FreeIran; performs segmentation as per the capital letters
The Target Function

\[ f(ht) = n \]

*ht* is a vector space representation of a given hashtag

*n* is the normalized count of its occurrences in a time frame.
The Target Function

\[ f(ht) = n \]

- \( ht \) is a vector space representation of a given hashtag
- \( n \) is the normalized count of its occurrences in a time frame.
- Transformed target function: \( f'(ht) = \log(n) \).
**Prediction Model**

### The Target Function

\[ f(ht) = n \]

- \( ht \) is a vector space representation of a given hashtag.
- \( n \) is the normalized count of its occurrences in a time frame.

Transformed target function: \( f'(ht) = \log(n) \).

### Regression Model

Training set: \((X, Y) = \{x_i, y_i\}\), where for a hashtag \( h_{ti} \):

- \( x_i \): feature vector representation of \( h_{ti} \)
- \( y_i = \log(n_i) \), where \( n_i \) is the normalized count of occurrences of \( h_{ti} \)
Prediction Model

The Target Function

\[ f(ht) = n \]

\(ht\) is a vector space representation of a given hashtag
\(n\) is the normalized count of its occurrences in a time frame.
Transformed target function: \(f'(ht) = \log(n)\).

Regression Model

Training set: \((X, Y) = \{x_i, y_i\}\), where for a hashtag \(ht_i\):
\(x_i\): feature vector representation of \(ht_i\)
\(y_i = \log(n_i)\), where \(n_i\) is the normalized count of occurrences of \(ht_i\)

\[ Y = b + w^T X \]
L_1 Regularization with Stochastic Gradient Descent

\[ L_r(b, w) = \frac{1}{2} \sum_i \left( y_i - (b + \sum_j w_j x_i^j) \right)^2 + \frac{1}{2} \lambda \|w\| \]
L1 Regularization with Stochastic Gradient Descent

Parameter update for Stochastic Gradient Descent (SGD)

\[
L_r(b, w) = \frac{1}{2} \sum_i \left( y_i - (b + \sum_j w_j^T x_i^j) \right)^2 + \frac{1}{2} \lambda \|w\| 
\]

\[
\Delta b = \eta_t (y_i - (b + w^T x_i)) \\
\Delta w_i = \eta_t (y_i - (b + w^T x_i)x_i - \lambda w_i)
\]
**Model Features**

- **Hashtag content**: features that can be extracted from the hashtag itself.
- **Global tweet features**: features related to the content of the tweets containing the hashtag.
- **Graph topology features**: features related to graph topology and retweet statistics.
- **Global temporal features**: features related to temporal pattern of the use of the hashtag.
Character Length

7 bins were used: 2, 3, 4, 5, 6-9, 10-14, >14 characters
## Character Length

7 bins were used: 2, 3, 4, 5, 6-9, 10-14, >14 characters

## Number of words

55% of the hashtags were compounds of more than one word, e.g. #freelran, #GoogleGoesGaga.

**Four bins:** 1 word, 2-3 words, 4 words, >4 words
Orthography

Hashtags can be written in capital letters, contain some capital and/or digits, e.g. #myheart4JB. ‘right’ writing style may make it readable: savethenhs vs. saveTheNHS

Attributes: no caps, some caps, all caps, contain digits
Lexical Items

Hashtag/its words are matched against five predefined lists:

- A general lexicon containing all words from a large portion (612MB) of English Wikipedia
- A list of proper names taken from the name list compiled at the US census of 1995
- A list of celebrity names compiled from Forbes' 'The Celebrity 100' lists of 2008-2010.
- A short list of holidays and days of the week
- A list of all the world's countries
**Lexical Items**

Hashtag/its words are matched against five predefined lists:

- a general lexicon containing all words from a large portion (612MB) of English Wikipedia
Lexical Items

Hashtag/its words are matched against five predefined lists:

- a general lexicon containing all words from a large portion (612MB) of English Wikipedia
- a list of proper names taken from the name list compiled at the US census of 1995
Lexical Items

Hashtag/its words are matched against five predefined lists:

- a general lexicon containing all words from a large portion (612MB) of English Wikipedia
- a list of proper names taken from the name list compiled at the US census of 1995
- a list of celebrity names compiled from Forbes’ ‘The Celebrity 100’ lists of 2008-2010.
Lexical Items

Hashtag/its words are matched against five predefined lists:

- a general lexicon containing all words from a large portion (612MB) of English Wikipedia
- a list of proper names taken from the name list compiled at the US census of 1995
- a list of celebrity names compiled from Forbes’ ‘The Celebrity 100’ lists of 2008-2010.
- a short list of holidays and days of the week
**Lexical Items**

Hashtag/its words are matched against five predefined lists:

- a general lexicon containing all words from a large portion (612MB) of English Wikipedia
- a list of proper names taken from the name list compiled at the US census of 1995
- a list of celebrity names compiled from Forbes’ ‘The Celebrity 100’ lists of 2008-2010.
- a short list of holidays and days of the week
- a list of all the world’s countries

Each of the five attributes is an attribute in the vector.
**Location**

Location of a hashtag can give an indication of the way it is used:
For instance, if located in the middle of the tweet, hashtag also serves as part of the sentence and not only as a meta tag.
**Location**

Location of a hashtag can give an indication of the way it is used:
For instance, if located in the middle of the tweet, hashtag also serves as part of the sentence and not only as a meta tag.
Three locations: *prefix, infix, suffix*

Ex: “AP: Report: #Iran’s paramilitary launches cyber attack http://is.gd/HiCYJU #iranelections #freeiran”
Hashtag Content Features

**Location**

Location of a hashtag can give an indication of the way it is used: For instance, if located in the middle of the tweet, hashtag also serves as part of the sentence and not only as a meta tag.

Three locations: *prefix*, *infix*, *suffix*

Ex: “AP: Report: #Iran’s paramilitary launches cyber attack http://is.gd/HiCYJU #iranelections #freeiran”

Last two hashtags are considered suffixes
**Hashtag Content Features**

**Location**

Location of a hashtag can give an indication of the way it is used:
For instance, if located in the middle of the tweet, hashtag also serves as part of the sentence and not only as a meta tag.

Three locations: *prefix, infix, suffix*

Ex: “AP: Report: #Iran’s paramilitary launches cyber attack http://is.gd/HiCYJU #iranelections #freeiran”

Last two hashtags are considered suffixes
#Iran is considered Infix
Hashtag Content Features

Location

Location of a hashtag can give an indication of the way it is used:
For instance, if located in the middle of the tweet, hashtag also serves as part
of the sentence and not only as a meta tag.
Three locations: prefix, infix, suffix
Ex: “AP: Report: #Iran’s paramilitary launches cyber attack http://is.gd/HiCYJU
#iranelections #freeiran”
Last two hashtags are considered suffixes
#Iran is considered Infix

Collocation

Whether it collocates with other hashtags?
Value 1 if more than 40% of the occurrences are collocated with other
hashtags.
Cognitive Dimension

Some words trigger specific emotions and encourage specific behavior and this psychological dimension can influence its spread.
Cognitive Dimension

Some words trigger specific emotions and encourage specific behavior and this psychological dimension can influence its spread. LIWC project assigns words to a number of emotional and cognitive dimensions.
Cognitive Dimension

Some words trigger specific emotions and encourage specific behavior and this psychological dimension can influence its spread. LIWC project assigns words to a number of emotional and cognitive dimensions.

Ex: positive sentiment, negative sentiment, optimistic, self, anger etc.
Global Tweet Features

- 1000 most frequent words the hashtag co-occurred with were extracted.
- This list is mapped to the 69 LIWC categories
Graph Topology Features

**Average Number of followers**

Average number of followers of users, who used the hashtag, is divided to 19 bins on a sub logarithmic scale.
Graph Topology Features

**Average Number of followers**

Average number of followers of users, who used the hashtag, is divided to 19 bins on a sub logarithmic scale.

**Max Number of followers**

Max number of followers of users, who used the hashtag, is divided to 19 bins on a logarithmic scale.
Graph Topology Features

**Average Number of followers**

Average number of followers of users, who used the hashtag, is divided to 19 bins on a sub logarithmic scale.

**Max Number of followers**

Max number of followers of users, who used the hashtag, is divided to 19 bins on a logarithmic scale.

**Retweets ratio**

Tendency of a hashtag to appear in retweeted messages.
The normalized weekly count of each hashtag was sampled in four time stamps: $w_i, i \in \{t, t + 1, t + 2, t + 6\}$
The normalized weekly count of each hashtag was sampled in four time stamps: $w_i, i \in \{t, t+1, t+2, t+6\}$

$t$ is the first week of occurrence and $t+j$ is the $j$-th week after the first occurrence.
The normalized weekly count of each hashtag was sampled in four time stamps: \( w_i, i \in \{ t, t + 1, t + 2, t + 6 \} \)

\( t \) is the first week of occurrence and \( t + j \) is the \( j \)-th week after the first occurrence.

Three lag values are obtained \( d_{k \in 1, 2, 3} \), where \( d_k \) is the ratio of change from the previous time stamp. (stickiness and persistence)
The normalized weekly count of each hashtag was sampled in four time stamps: \( w_i, i \in \{t, t + 1, t + 2, t + 6\} \)

- \( t \) is the first week of occurrence and \( t + j \) is the \( j \)-th week after the first occurrence.
- Three lag values are obtained \( d_{k \in 1, 2, 3} \), where \( d_k \) is the ratio of change from the previous time stamp. (stickiness and persistence)
- 17 bins on a logarithmic scale (-200% to 200% change)
Experimental Setup

**Learning three aspects in the prediction**

- what is the attribute combination that yields the best prediction?
- what are the strongest attributes and how do they complement each other?
- how does the prediction accuracy change given different time frames?
Experimental Setup

Learning three aspects in the prediction

- what is the attribute combination that yields the best prediction?
- what are the strongest attributes and how do they complement each other?
- how does the prediction accuracy change given different time frames?

Performance measurement
**Experimental Setup**

**Learning three aspects in the prediction**

- what is the attribute combination that yields the best prediction?
- what are the strongest attributes and how do they complement each other?
- how does the prediction accuracy change given different time frames?

**Performance measurement**

- Experiments were executed in a 10-fold cross validation manner.
Experimental Setup

**Learning three aspects in the prediction**
- what is the attribute combination that yields the best prediction?
- what are the strongest attributes and how do they complement each other?
- how does the prediction accuracy change given different time frames?

**Performance measurement**
- Experiments were executed in a 10-fold cross validation manner.
- Performance is measured by the mean square error (MSE).
Table 1: MSE of basic models and the hybrid model in horizons. $MSE_n$ indicates results for acceptance prediction in an $n$ weeks time frame.

<table>
<thead>
<tr>
<th>Model</th>
<th>$MSE_{10}$</th>
<th>$MSE_{15}$</th>
<th>$MSE_{20}$</th>
<th>$MSE_{25}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>4.988</td>
<td>3.796</td>
<td>3.125</td>
<td>2.698</td>
</tr>
<tr>
<td>HT_{all}</td>
<td>4.380</td>
<td>3.410</td>
<td>2.902</td>
<td>2.565</td>
</tr>
<tr>
<td>TW_{content}</td>
<td>4.776</td>
<td>3.509</td>
<td>2.743</td>
<td>2.221</td>
</tr>
<tr>
<td>Graph</td>
<td>4.295</td>
<td>3.144</td>
<td>2.404</td>
<td>1.923</td>
</tr>
<tr>
<td>Temporal</td>
<td>3.294</td>
<td>2.893</td>
<td>2.507</td>
<td>2.112</td>
</tr>
<tr>
<td>Hybrid_{all}</td>
<td>2.584</td>
<td>2.098</td>
<td>1.685</td>
<td>1.315</td>
</tr>
</tbody>
</table>
### Table 3: MSE and correlation coefficient for various combinations of feature types for a 15 weeks time frame.

<table>
<thead>
<tr>
<th>Model</th>
<th>MSE</th>
<th>Corr–coeff</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>3.796</td>
<td>-0.021</td>
</tr>
<tr>
<td>$H_{t\text{all}}$</td>
<td>3.410</td>
<td>0.319</td>
</tr>
<tr>
<td>$TW_{\text{cont}}$</td>
<td>3.509</td>
<td>0.275</td>
</tr>
<tr>
<td>Graph</td>
<td>3.144</td>
<td>0.414</td>
</tr>
<tr>
<td>Temporal</td>
<td>2.893</td>
<td>0.487</td>
</tr>
<tr>
<td>$HT_{\text{cont}} + TW_{\text{cont}}$</td>
<td>2.967</td>
<td>0.467</td>
</tr>
<tr>
<td>$HT_{\text{cont}} + TW_{\text{cont}} + \text{Graph}$</td>
<td>2.546</td>
<td>0.573</td>
</tr>
<tr>
<td>$HT_{\text{cont}} + TW_{\text{cont}} + \text{Temp}$</td>
<td>2.321</td>
<td>0.6234</td>
</tr>
<tr>
<td>$\text{Graph} + \text{Temporal}$</td>
<td>2.450</td>
<td>0.594</td>
</tr>
<tr>
<td>Hybrid$_{\text{all}}$</td>
<td>2.098</td>
<td>0.669</td>
</tr>
</tbody>
</table>
Results

<table>
<thead>
<tr>
<th>Model</th>
<th>MSE</th>
<th>Corr–coeff</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d_1$</td>
<td>3.236</td>
<td>0.383</td>
</tr>
<tr>
<td>$d_2$</td>
<td>3.39</td>
<td>0.326</td>
</tr>
<tr>
<td>$d_3$</td>
<td>3.44</td>
<td>0.303</td>
</tr>
<tr>
<td>$d_1 + d_2$</td>
<td>3.088</td>
<td>0.431</td>
</tr>
<tr>
<td>$d_1 + d_3$</td>
<td>2.97</td>
<td>0.464</td>
</tr>
<tr>
<td>$d_2 + d_3$</td>
<td>3.19</td>
<td>0.398</td>
</tr>
<tr>
<td>$d_1 + d_2 + d_3$</td>
<td>2.893</td>
<td>0.487</td>
</tr>
</tbody>
</table>

Table 5: MSE and correlation coefficient for different number of lags and different distances between sampling points in 15 weeks horizon. $d_i$ indicates the the $i$-th lag described in Section 4.3.4.
What is Information Diffusion?

Online Information Diffusion
Understanding the tendency for people to engage in activities such as forwarding messages, linking to articles, joining groups, purchasing products, or becoming fans of pages after some number of their friends have.
What is Information Diffusion?

Online Information Diffusion

Understanding the tendency for people to engage in activities such as forwarding messages, linking to articles, joining groups, purchasing products, or becoming fans of pages after some number of their friends have.

Objectives of this research

- Widespread belief that different kinds of information spread differently online.
What is Information Diffusion?

Online Information Diffusion
Understanding the tendency for people to engage in activities such as forwarding messages, linking to articles, joining groups, purchasing products, or becoming fans of pages after some number of their friends have.

Objectives of this research

- Widespread belief that different kinds of information spread differently online.
- To study this issue on Twitter, analyzing the ways in which Hashtags spread on a network defined by interactions among Twitter users.
Twitter data crawled from August 2009 until January 2010.
Twitter data crawled from August 2009 until January 2010.

Collected over 3 billion messages from more than 60 million users.
Twitter data crawled from August 2009 until January 2010.

- Collected over 3 billion messages from more than 60 million users.
- Graph construction via @-messages: $X \rightarrow Y$ if $X$ directed at least 3 @-messages to $Y$. 

Graph size: 8.5 million non-isolated nodes, 50 million links

Studies 500 most used hashtags
Twitter data crawled from August 2009 until January 2010.
Collected over 3 billion messages from more than 60 million users.
Graph construction via @-messages: $X \rightarrow Y$ if $X$ directed at least 3 @-messages to $Y$.
Graph size: 8.5 million non-isolated nodes, 50 million links.
Twitter data crawled from August 2009 until January 2010.

Collected over 3 billion messages from more than 60 million users.

Graph construction via @-messages: $X \rightarrow Y$ if $X$ directed at least 3 @-messages to $Y$.

Graph size: 8.5 million non-isolated nodes, 50 million links

Studies 500 most used hashtags
Manually identified 8 broad categories with at least 20 HTs in each.

Authors and 3 volunteers independently annotated each hashtag.

Levels of agreement was high.

<table>
<thead>
<tr>
<th>Category</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Celebrity</td>
<td>mj, brazilwantsjb, regis, iwantpeterfacinelli</td>
</tr>
<tr>
<td>Music</td>
<td>thisiswar, mj, musicmonday, pandora</td>
</tr>
<tr>
<td>Games</td>
<td>mafiawars, spymaster, mw2, zyngapirates</td>
</tr>
<tr>
<td>Political</td>
<td>tcot, glennbeck, obama, hcr</td>
</tr>
<tr>
<td>Idiom</td>
<td>cantlivewithout, dontyouhate, musicmonday</td>
</tr>
<tr>
<td>Sports</td>
<td>golf, yankees, nhl, cricket</td>
</tr>
<tr>
<td>Movies/TV</td>
<td>lost, glennbeck, bones, newmoon</td>
</tr>
<tr>
<td>Technology</td>
<td>digg, iphone, jquery, photoshop</td>
</tr>
</tbody>
</table>
Exposure Curve: Defining $p(k)$

**Neighbor Set of $X$**

For a given user $X$, the set of other users to whom $X$ has an edge.
Exposure Curve: Defining $p(k)$

**Neighbor Set of X**

For a given user $X$, the set of other users to whom $X$ has an edge.

**When does $X$ start mentioning a hashtag $H$?**

How do successive exposures to $H$ affect the probability that $X$ will begin mentioning it?
Exposure Curve: Defining $p(k)$

**Neighbor Set of $X$**

For a given user $X$, the set of other users to whom $X$ has an edge.

**When does $X$ start mentioning a hashtag $H$?**

How do successive exposures to $H$ affect the probability that $X$ will begin mentioning it?

- Look at all users $X$ who have not mentioned $H$, but for whom $k$ neighbors have
- $p(k)$: fraction of users who adopt the hashtag *direct* after their $k^{th}$ exposure, given that they hadn’t yet adopted it.
Average Exposure Curve for 500 most-mentioned hashtags

A ramp-up to the peak value, reached relatively early ($k = 2, 3, 4$),

Decline for larger values of $k$.
Average Exposure Curve for 500 most-mentioned hashtags

- A ramp-up to the peak value, reached relatively early ($k = 2, 3, 4$)
- Decline for larger values of $k$
Persistence and Stickiness

**Stickiness**

The maximum value of $p(k)$ (probability of usage at the most effective exposure)

### Diagram

- **Persistence**
  - A measure of the decay of exposure curves.
  - The ratio of the area under the curve $P$ and the area of the rectangle of length $\text{max}(P)$ and width $\text{max}(D(P))$.

- **Stickiness**
  - The maximum value of $p(k)$.
Persistence and Stickiness

**Stickiness**
The maximum value of \( p(k) \) (probability of usage at the most effective exposure)

**Persistence**
A measure of the decay of exposure curves.
**Persistence and Stickiness**

**Stickiness**
The maximum value of $p(k)$ (probability of usage at the most effective exposure)

**Persistence**
A measure of the decay of exposure curves.
The ratio of the area under the curve $P$ and the area of the rectangle of length $\max(P)$ and width $\max(D(P))$. 
Approximating Exposure Curves via Stickiness and Persistence

- Are Persistence and Stickiness the adequate pair of parameters for discussing the curves’ overall approximate shapes? Yes.
- Given the stickiness $M(P)$ and the persistence $F(P)$ of exposure curve $P$, we find an approximation $\tilde{P}$ to $P$ in the following way:

1. Let $\tilde{P}(0) = 0$
2. Let $\tilde{P}(2) = M(P)$
3. Now we will let $\tilde{P}(K)$ be such that $F(\tilde{P}) = F(P)$. This value turns out to be $\tilde{P}(K) = \frac{M(P) \times K \times (2 \times F(P) - 1)}{K - 2}$
4. Make $\tilde{P}$ piecewise linear with one line connecting the points $(0, 0)$ and $(2, M(P))$, and another line connecting the points $(2, M(P))$ and $(K, \tilde{P}(K))$. 

![Graph showing actual and approximation curves](image-url)
Idioms and Music have lower persistence than a random subset of hashtags of the same size.

Politics and Sports have higher persistence than a random subset.
Comparison of Hashtags based on Stickiness

Technology and Movies have lower stickiness than a random subset.

Music has higher stickiness than a random subset.

Pawan Goyal (IIT Kharagpur)

Hashtags on Twitter

July 24-28, 2014
Comparison of Hashtags based on Stickiness

- Technology and Movies have lower stickiness than a random subset
- Music has higher stickiness than a random subset
Persistence vs. Stickiness

Idioms and Politics: Same stickiness but opposite persistence

Music has high stickiness but low persistence

Stickiness does not explain the diffusion well by itself
Idioms and Politics: Same stickiness but opposite persistence
Persistence vs. Stickiness

- Idioms and Politics: Same stickiness but opposite persistence
- Music has high stickiness but low persistence
- Stickiness does not explain the diffusion well by itself
Sample curves for #cantlivewithout (blue) and #hcr (red)
## Comparison of Hashtag by Mention and User Counts

<table>
<thead>
<tr>
<th>Type</th>
<th>Mentions</th>
<th>Users</th>
<th>Mentions/User</th>
</tr>
</thead>
<tbody>
<tr>
<td>All HTS</td>
<td>93,056</td>
<td>15,418</td>
<td>6.59</td>
</tr>
<tr>
<td>Political</td>
<td>132,180</td>
<td>13,739</td>
<td>10.17</td>
</tr>
<tr>
<td>Sports</td>
<td>98,234</td>
<td>11,329</td>
<td>9.97</td>
</tr>
<tr>
<td>Idioms</td>
<td>99,317</td>
<td>26,319</td>
<td>3.54</td>
</tr>
<tr>
<td>Movies</td>
<td>90,425</td>
<td>15,957</td>
<td>6.57</td>
</tr>
<tr>
<td>Celebrity</td>
<td>87,653</td>
<td>5,351</td>
<td>17.68</td>
</tr>
<tr>
<td>Technology</td>
<td>90,462</td>
<td>24,648</td>
<td>5.08</td>
</tr>
<tr>
<td>Games</td>
<td>123,508</td>
<td>15,325</td>
<td>6.61</td>
</tr>
<tr>
<td>Music</td>
<td>87,985</td>
<td>7,976</td>
<td>10.39</td>
</tr>
</tbody>
</table>

Table: Median Values
## Comparison of Hashtag by Mention and User Counts

<table>
<thead>
<tr>
<th>Type</th>
<th>Mentions</th>
<th>Users</th>
<th>Mentions/User</th>
</tr>
</thead>
<tbody>
<tr>
<td>All HTS</td>
<td>93,056</td>
<td>15,418</td>
<td>6.59</td>
</tr>
<tr>
<td>Political</td>
<td>132,180</td>
<td>13,739</td>
<td>10.17</td>
</tr>
<tr>
<td>Sports</td>
<td>98,234</td>
<td>11,329</td>
<td>9.97</td>
</tr>
<tr>
<td>Idioms</td>
<td>99,317</td>
<td>26,319</td>
<td>3.54</td>
</tr>
<tr>
<td>Movies</td>
<td>90,425</td>
<td>15,957</td>
<td>6.57</td>
</tr>
<tr>
<td>Celebrity</td>
<td>87,653</td>
<td>5,351</td>
<td>17.68</td>
</tr>
<tr>
<td>Technology</td>
<td>90,462</td>
<td>24,648</td>
<td>5.08</td>
</tr>
<tr>
<td>Games</td>
<td>123,508</td>
<td>15,325</td>
<td>6.61</td>
</tr>
<tr>
<td>Music</td>
<td>87,985</td>
<td>7,976</td>
<td>10.39</td>
</tr>
</tbody>
</table>

### Table: Median Values

Political and Idioms are among the most mentioned, but Idioms are used by twice the number of people that use Politics
Let $G_m$ be the subgraph induced by the first $m$ users of a given hashtag.

Let the \textit{border} of $G_m$ be the set of nodes not in $G_m$ with at least one edge to a node in $G_m$.

Let the \textit{internal degree} of a node in $G_m$ be the number of neighbors it has in $G_m$.

Let the \textit{entering degree} of a node in the border of $G_m$ be the number of neighbors it has in $G_m$. 
Structure Comparison for Political Hashtags ($G_{500}$)

<table>
<thead>
<tr>
<th>Type</th>
<th>Internal Degree</th>
<th>Triangle Num</th>
<th>Entering Deg.</th>
<th>Border Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>All HTS</td>
<td>1.41</td>
<td>384</td>
<td>1.24</td>
<td>13425</td>
</tr>
<tr>
<td>Political</td>
<td>2.55</td>
<td>935</td>
<td>1.41</td>
<td>12879</td>
</tr>
<tr>
<td>Upper Error Bar</td>
<td>1.82</td>
<td>653</td>
<td>1.32</td>
<td>15838</td>
</tr>
<tr>
<td>Lower Error Bar</td>
<td>1.00</td>
<td>112</td>
<td>1.16</td>
<td>11016</td>
</tr>
</tbody>
</table>
Structure Comparison for Political Hashtags ($G_{500}$)

<table>
<thead>
<tr>
<th>Type</th>
<th>Internal Degree</th>
<th>Triangle Num</th>
<th>Entering Deg.</th>
<th>Border Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>All HTS</td>
<td>1.41</td>
<td>384</td>
<td>1.24</td>
<td>13425</td>
</tr>
<tr>
<td>Political</td>
<td>2.55</td>
<td>935</td>
<td>1.41</td>
<td>12879</td>
</tr>
<tr>
<td>Upper Error Bar</td>
<td>1.82</td>
<td>653</td>
<td>1.32</td>
<td>15838</td>
</tr>
<tr>
<td>Lower Error Bar</td>
<td>1.00</td>
<td>112</td>
<td>1.16</td>
<td>11016</td>
</tr>
</tbody>
</table>

The early adopters of a political hashtag message more with each other, create more triangles, and have a border of people with more links into the early adopter set.