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?

Why are Hashtags useful?

- 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.

Social Phenomena

Social Phenomena

- A new social event can lead to the simultaneous emergence of several different hashtags, each one generated by a different user.

Social Phenomena

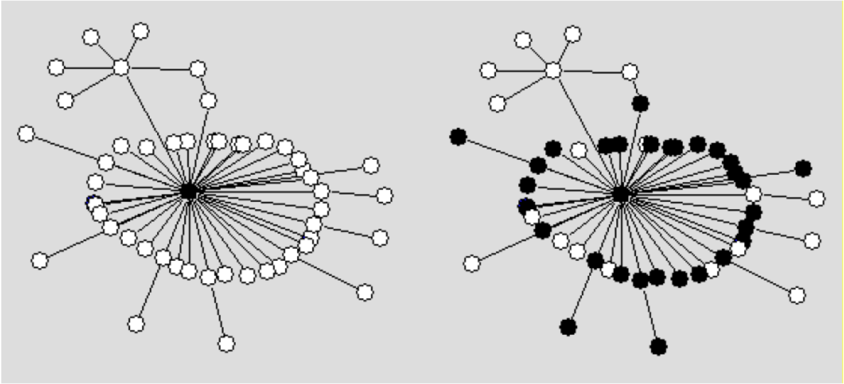
- A new social event can lead to the simultaneous emergence of several different hashtags, each one generated by a different user.
- They can either be accepted by other members of the network or not.

Social Phenomena

- 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 propogation process

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



First Reference

Evandro Cunha, Gabriel Magno, Giovanni Comarela, Virgilio Almeida, Marcos André Goncalves, and Fabrício Benevenuto. 2011. *Analyzing the dynamic evolution of hashtags on Twitter: a language-based approach*. In Proceedings of the Workshop on Languages in Social Media (LSM '11). Association for Computational Linguistics, Stroudsburg, PA, USA, 58 - 65.

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

Dataset Used

- 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.

Topics Selected

A minor base was built for each of the following topics:

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

Forming the bases

Filtering tweets that included

Forming the bases

Filtering tweets that included

- atleast one hashtag

Forming the bases

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

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

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

Distribution of Hashtags

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

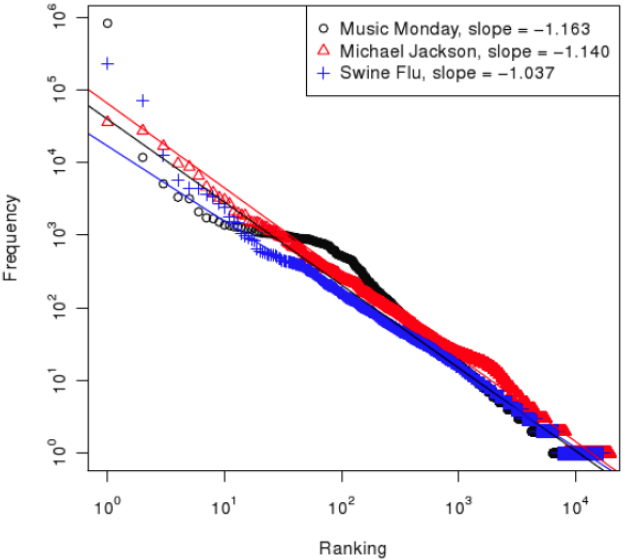
Base	% of *i*-tweet hashtags inside the base		
	i=1	*i*=2	*i*=10
MJ	59%	72%	88%
SF	59%	73%	92%
MM	60%	74%	91%

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

Base	number of *j*-tweet hashtags inside the base		
	j=10,000	*j*=5,000	*j*=1,000
MJ	3	6	28
SF	3	4	14
MM	2	3	28

Table 3. Distribution of most popular hashtags of each base.

Data from most used Hashtags



Volume of Tweets vs. its position in popularity ranking

Base	Most used	2nd most used	3rd most used
MJ	#michaeljackson 35,861 12.3%	#michael 27,298 9.3%	#mj 16,758 5.7%
SF	#swineflu 230,457 51.5%	#h1n1 70,693 15.8%	#swine 12,444 2.8%
MM	#musicmonday 824,778 79.7%	#musicmondays 11,770 1.1%	#music 5,106 0.5%

Hashtag Length and Frequency

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

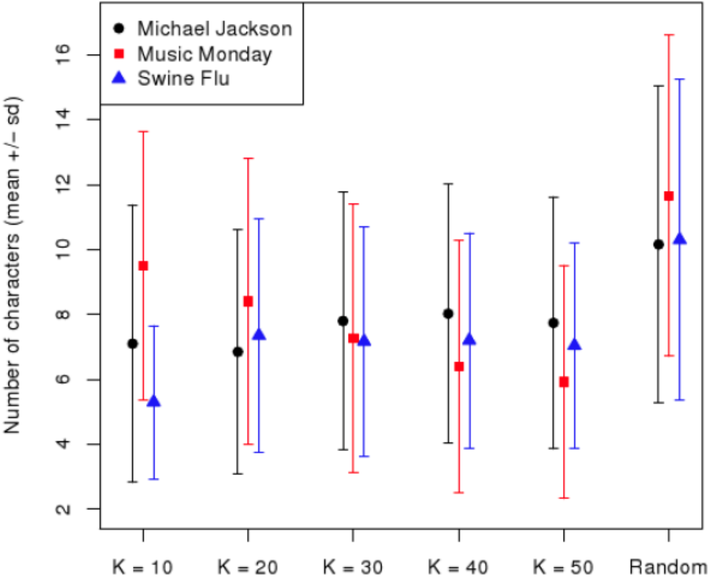
Most common hashtags (number of tweets)	Most common hashtags with 15 or more characters (number of tweets)
#michaeljackson (35,861)	#nothingpersonal (962)
#michael (27,298)	#iwillneverforget (912)
#mj (16,758)	#thankyoumichael (690)
#swineflu (230,457)	#swinefluhatesyou (1,056)
#h1n1 (70,693)	#crapnamesforpubs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

Average Length

Topic	Average length of...					
	...the *k* most popular hashtags					...the less popular hashtags
	k=10	*k*=20	*k*=30	*k*=40	*k*=50	
MJ	7.1	6.85	7.8	8.02	7.74	10.16
SF	5.3	7.35	7.17	7.2	7.04	10.3
MM	9.5	8.4	7.27	6.4	5.92	11.66

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.

Avergae Number of Characters



Underscores in Hashtags

Base	Number of _-hashtags	% of _-hashtags among *i*-tweet hashtags	
		i=2	*i*=10
MJ	251 (1.2%)	89%	97%
SF	155 (0.9%)	87%	97%
MM	143 (0.9%)	89%	98%

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

Second Reference

Oren Tsur and Ari Rappoport. 2012. *What's in a hashtag?: content based prediction of the spread of ideas in microblogging communities*. In Proceedings of the fifth ACM international conference on Web search and data mining (WSDM '12). ACM, New York, NY, USA, 643-652.

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)?

Content-based Popularity Prediction

Objective

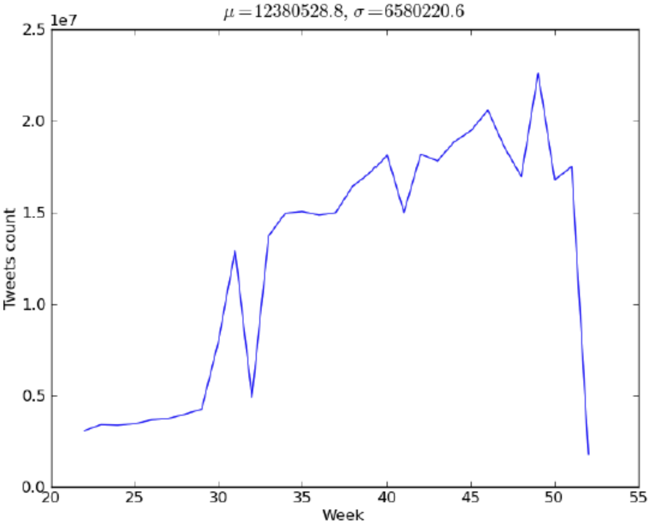
Given an idea/meme m, and a time frame t, can we predict the acceptance of m in the community (social network)?

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.



Filetering and Normalization

Filtering

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

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.

$$N(ht^i) = \sum_{j \in weeks} count({ht^i}_j)\frac{w_1}{w_j}$$

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.

Fresh hashtags

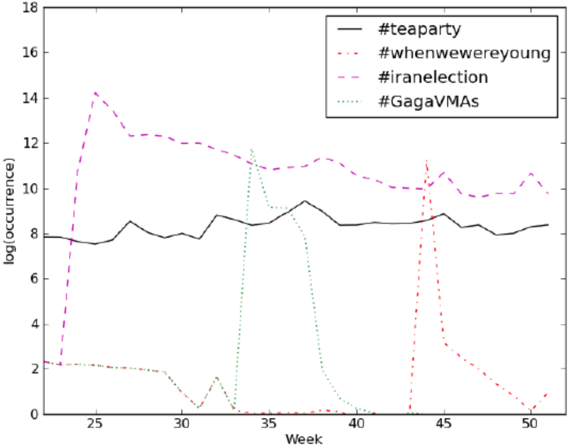


Figure 2: Four typical temporal trends (unnormalized counts).

Preprocessing

Identifying the distinct words in a hashtag

#thankyousachin

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

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

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.

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: $f'(ht) = log(n)$.

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

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

$$Y = b + w^T X$$

L1 Regularization with Stochastic Gradient Descent

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

L1 Regularization with Stochastic Gradient Descent

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

Parameter update for Stochastic Gradient Descent (SGD)

$$\begin{aligned}
\Delta b &= \eta_t\big(y_i - (b + w^T x_i)\big) \\
\Delta w_i &= \eta_t\big(y_i - (b + w^T x_i)x_i - \lambda w_i\big)
\end{aligned}$$

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.

Hashtag Content Features

Character Length

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

Hashtag Content Features

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. #freeIran, #GoogleGoesGaga.

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

Hashtag Content Features

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

Hashtag Content Features

Lexical Items

Hashtag/its words are matched against five predefined lists:

Hashtag Content Features

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

Hashtag Content Features

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

Hashtag Content Features

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.

Hashtag Content Features

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

Hashtag Content Features

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

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.

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”

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.

Hashtag Content Features

Cognitive Dimension

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

Hashtag Content Features

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.

Hashtag Content Features

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.

Graph Temporal Features

- The normalized weekly count of each hashtag was sampled in four time stamps: $w_i, i \in \{t, t+1, t+2, t+6\}$

Graph Temporal Features

- 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.

Graph Temporal Features

- 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 otained $d_{k\in 1,2,3}$, where d_k is the ratio of change from the previous time stamp. (stickiness and persistence)

Graph Temporal Features

- 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 otained $d_{k \in 1,2,3}$, 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).

Results

Model	MSE_{10}	MSE_{15}	MSE_{20}	MSE_{25}
baseline	4.988	3.796	3.125	2.698
HT_{all}	4.380	3.410	2.902	2.565
$TW_{content}$	4.776	3.509	2.743	2.221
Graph	4.295	3.144	2.404	1.923
Temporal	3.294	2.893	2.507	2.112
$Hybrid_{all}$	2.584	2.098	1.685	1.315

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.

Results

Model	MSE	Corr–coeff
baseline	3.796	-0.021
$\mathbf{Ht_{all}}$	3.410	0.319
$\mathbf{TW_{cont}}$	3.509	0.275
Graph	3.144	0.414
Temporal	2.893	0.487
$\mathbf{HT_{cont} + TW_{cont}}$	2.967	0.467
$\mathbf{HT_{cont} + TW_{cont} + Graph}$	2.546	0.573
$\mathbf{HT_{cont} + TW_{cont} + Temp}$	2.321	0.6234
Graph+Temporal	2.450	0.594
$\mathbf{Hybrid_{all}}$	2.098	0.669

Table 3: MSE and correlation coefficient for various combinations of feature types for a 15 weeks time frame.

Results

Model	MSE	Corr–coeff
$\mathbf{d_1}$	3.236	0.383
$\mathbf{d_2}$	3.39	0.326
$\mathbf{d_3}$	3.44	0.303
$\mathbf{d_1 + d_2}$	3.088	0.431
$\mathbf{d_1 + d_3}$	2.97	0.464
$\mathbf{d_2 + d_3}$	3.19	0.398
$\mathbf{d_1 + d_2 + d_3}$	2.893	0.487

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.

Third Reference

Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. 2011. *Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter.* In Proceedings of the 20th international conference on World wide web (WWW '11). ACM, New York, NY, USA, 695-704.

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 and Graph Construction

- Twitter data crawled from August 2009 until January 2010.

Twitter Data and Graph Construction

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

Twitter Data and Graph Construction

- 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.

Twitter Data and Graph Construction

- 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 and Graph Construction

- Twitter data crawled from August 2009 until January 2010.
- Collected over 3 billion messages from more than 60 million users.
- Graph construction via @-messages: $X \to Y$ 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

Hashtag Categories

- Manually identified 8 broad categories with atleast 20 HTs in each
- Authors and 3 volunteers independently annotated each hashtag.
- Levels of agreement was high

Category	Examples
Celebrity	mj, brazilwantsjb, regis, iwantpeterfacinelli
Music	thisiswar, mj, musicmonday, pandora
Games	mafiawars, spymaster, mw2, zyngapirates
Political	tcot, glennbeck, obama, hcr
Idiom	cantlivewithout, dontyouhate, musicmonday
Sports	golf, yankees, nhl, cricket
Movies/TV	lost, glennbeck, bones, newmoon
Technology	digg, iphone, jquery, photoshop

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

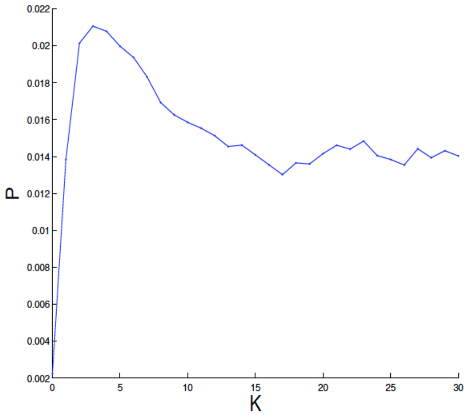
For a given user X, the set of other users to whom X has an edge.

When does X start mentioning a hashtag H?

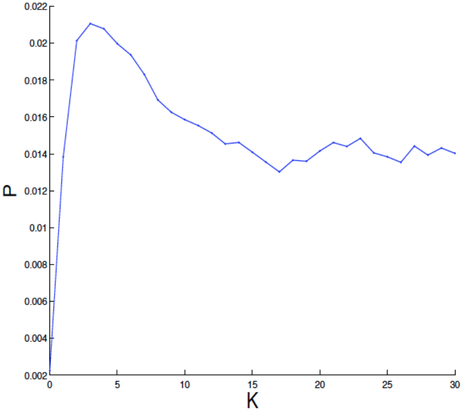
How do successive exposures to H affect the probability that X will begin mentioning it?

- 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

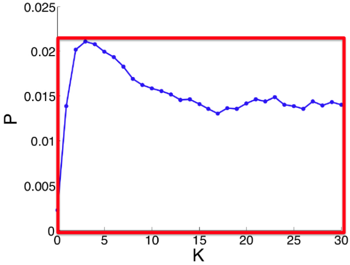


Average Exposure Curve for 500 most-mentioned hashtags



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

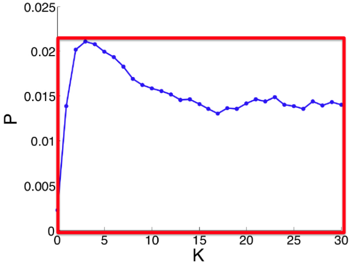
Persistence and Stickiness



Stickiness

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

Persistence and Stickiness



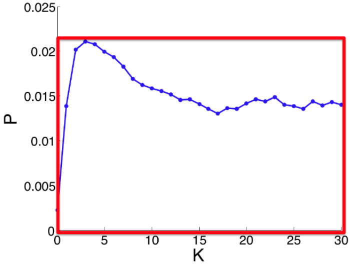
Stickiness

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

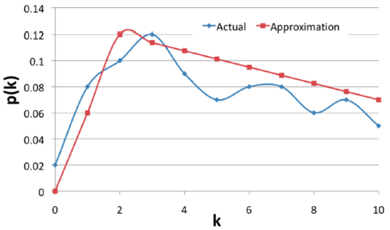
The maximum value of $p(k)$ (probability of usage at the most effective exposure)

Persistence

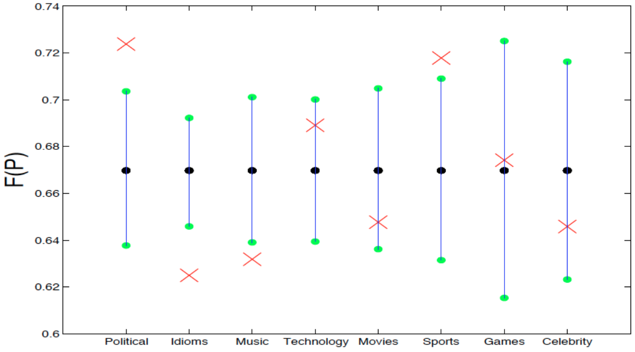
A measure of the decay of exposure curves.
The ratio of the area under the curve P and the area of the rectangle of length $max(P)$ and width $max(D(P))$.

- 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 $\widetilde{P}$ to P in the following way:

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

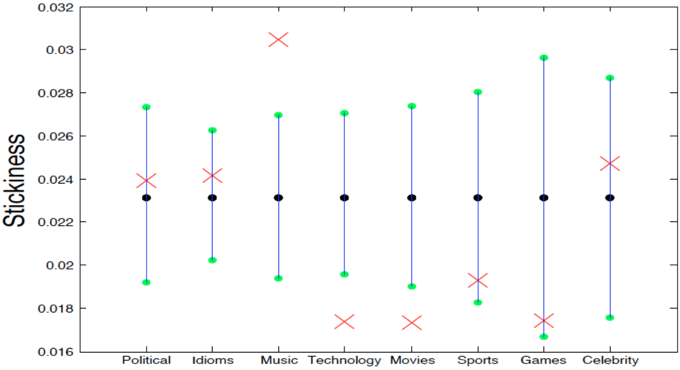


Comparison of Hashtags based on Persistence

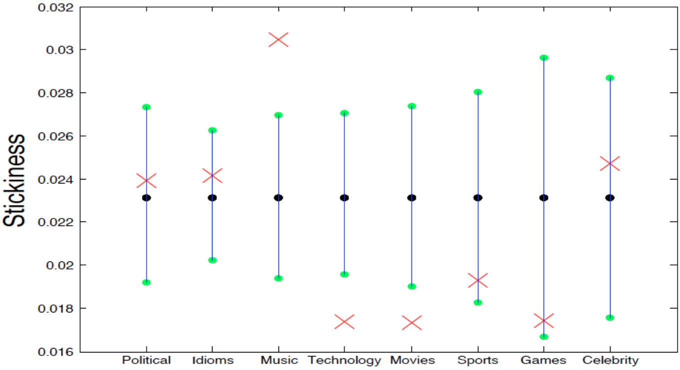


- Idioms and Music have lower persistence than a random subset of hashtags of the same size
- Politics and Sports have higher persistence than a random subset

Comparison of Hashtags based on Stickiness

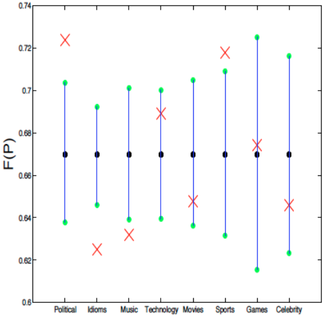


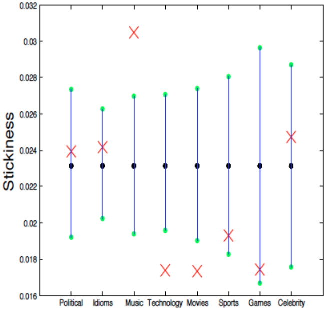
Comparison of Hashtags based on Stickiness



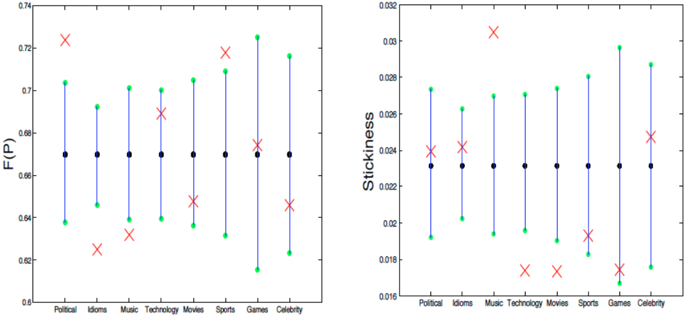
- Technology and Movies have lower stickiness than a random subset
- Music has higher stickiness than a random subset

Persistence vs. Stickiness



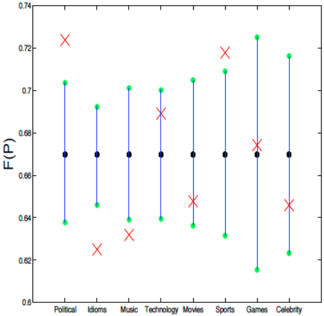


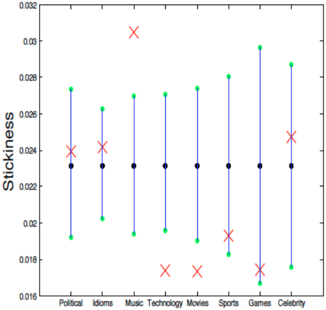
Persistence vs. Stickiness



- Idioms and Politics: Same stickiness but opposite persistence

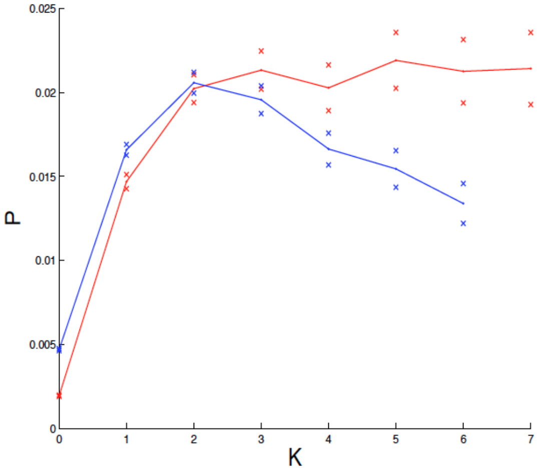
Persistence vs. Stickiness





- 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

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

Table: Median Values

Comparison of Hashtag by Mention and User Counts

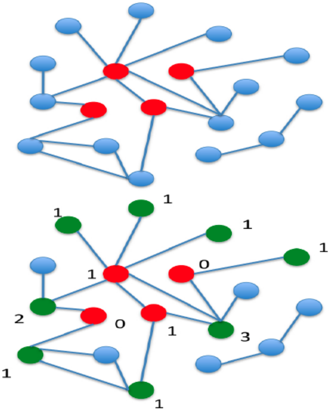
Type	Mentions	Users	Mentions/User
All HTS	93,056	15,418	6.59
Political	**132,180**	**13,739**	10.17
Sports	98,234	11,329	9.97
Idioms	**99,317**	**26,319**	3.54
Movies	90,425	15,957	6.57
Celebrity	87,653	5,351	17.68
Technology	90,462	24,648	5.08
Games	123,508	15,325	6.61
Music	87,985	7,976	10.39

Table: Median Values

Political and Idioms are among the most mentioned, but Idioms are used by twice the number of people that use Politics

The Structure of Initial Sets

- Let G_m be the subgraph induced by the first m users of a given hashtag.
- Let the *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 *internal degree* of a node in G_m be the number of neighbors it has in G_m.
- Let the *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})

Type	Internal Degree	Triangle Num	Entering Deg.	Border Nodes
All HTS	1.41	384	1.24	13425
Political	2.55	935	1.41	12879
Upper Error Bar	1.82	653	1.32	15838
Lower Error Bar	1.00	112	1.16	11016

Structure Comparison for Political Hashtags (G_{500})

Type	Internal Degree	Triangle Num	Entering Deg.	Border Nodes
All HTS	1.41	384	1.24	13425
Political	2.55	935	1.41	12879
Upper Error Bar	1.82	653	1.32	15838
Lower Error Bar	1.00	112	1.16	11016

- 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.