

On the Real-time Prediction Problems of Bursting Hashtags in Twitter

Shoubin Kong¹, Qiaozhu Mei², Ling Feng¹, Zhe Zhao², Fei Ye³

¹{kongsb09@mails., lingfeng@}tsinghua.edu.cn

²{qmei, zhezhao}@umich.edu

³feiye08@gmail.com

Abstract. Hundreds of thousands of hashtags are generated every day on Twitter. Only a few become bursting topics. Among the few, only some can be predicted in real-time. In this paper, we take the initiative to conduct a systematic study of a series of challenging real-time prediction problems of bursting hashtags. Which hashtags will become bursting? If they do, when will the burst happen? How long will they remain active? And how soon will they fade away? Based on empirical analysis of real data from Twitter, we provide insightful statistics to answer these questions, which span over the entire lifecycles of hashtags.

Keywords: hashtag, burstiness, real-time prediction

1 Introduction

As one of the leading platforms of social communications and information dissemination, Twitter has become a major source of information for common Web users. An overload of information is being diffused in real-time, which makes it easy for the users to obtain broad perspectives and quick updates about real world events, and in the meantime, makes it difficult for the users to filter useful and trending information from the noisy context.

Conversations on Twitter are featured with their “burstiness”, the phenomenon that a topic of discussion suddenly gains a considerable popularity, and then quickly fades away. Such bursting topics are usually triggered by breaking news, real world events, malicious rumors, or various types of behavior cascades such as campaigns of persuasion.

These bursting topics, usually referred to as trending topics, provide users with fresh discoveries and timely updates of events. Much study has also investigated the value of the bursting topics in a broader context. Bursts of topics, sentiments, and questions have been demonstrated to have a predictive power of product sales [5], stock market [2], search engine queries [26], outburst of diseases [17], elections [23], and even natural disasters [20]. Therefore, an earlier detection of such trending topics implies an increased revenue, a reduced damage, a timely treatment, and better decision-making in general. To help people discover the bursting topics in time, twitter deploys a list of trending topics as long as they are detected.

However, it may be already too late to react even if a burst can be *detected* in no time. On April 23rd, 2013, a false claim about explosions at the White House and the injury of the president sent by the hacked account of the Associated Press quickly became an explosive burst on Twitter ¹. Although the rumor was debunked and the hacked account was deleted as soon as the burst was detected, damage had been made - the bursting topic had shaken the stock market so badly that the Dow Jones Indices experienced a sudden drop of more than 100. If only we can *predict* the outbreak of a topic before it bursts! But can we?

Hashtags, user-specified strings starting with a # symbol, have been commonly used as identities of topics in Twitter. From a 10% random sample of the Tweet stream, we can identify about 400,000 new hashtags every day. However, only dozens of them become bursting. Among the dozens, there may be an even smaller proportion which one can predict in real-time. What are the proportions? How effective is the prediction, and what are the most important factors? How early can the prediction be done? In this study, we conduct the first *systematic* study of the real-time prediction of bursting hashtags. The key contributions of this paper include the following:

1. We take the initiative to provide formal definitions of a bursting hashtag as well as three key states in the lifecycle of a bursting hashtag. We define a series of real-time prediction problems that are concerned with these states of bursting hashtags.
2. We conduct a systematic study of these real-time prediction tasks, by exploring different solutions and different types of features, in particular novel time series features. We provide a comprehensive summary of the distribution of bursting hashtags and the effectiveness of real-time prediction.
3. Experiments are conducted on real datasets from Twitter to evaluate the performance of the proposed solutions. We also experimentally examine effectiveness of different features and analyze their contributions to the prediction performance.

2 Problem Setup

2.1 Definitions

The lifecycle of a hashtag can be formed as a time series $\langle c_1, c_2, \dots, c_t, \dots \rangle$. c_t denotes the count of tweets containing the hashtag at the t -th time interval. Considering the real-time characteristic of Twitter, the granularity of the time interval is set to 1 minute in this study. Definitions of bursting hashtags are as follows:

Definition 1. *Prediction-Trigger.* A clear majority of hashtags will never get burst, and a substantial number of bursting hashtags have a long dormant period before they burst. The average time before a hashtag gets burst is about **8.72**

¹ <http://www.foxnews.com/us/2013/04/23/hackers-break-into-associated-press-twitter-account/>

days since the hashtag first appears. Therefore we define a trigger to obtain a candidate set of hashtags to be predicted. For each hashtag, a five-minute sliding window is used to check the total count of tweets containing the hashtag within the consecutive five minutes, denoted as C_{slw} . If $C_{slw} > \delta$, the prediction is triggered.

Definition 2. Burst. We define the burst of a hashtag by referencing to the definition of spikes in [5]. Within 24 hours since the prediction was triggered, if c_t is greater than $\max(c_1 + \delta, 1.5c_1)$, t is defined as the onset of burst. δ , can be adjusted according to the statistics of real data. We have mentioned that in our dataset about 400,000 new hashtags can be identified every day. δ is set to 50 in this paper, which makes the ratio of bursting hashtags about 0.6%, i.e., about 25 bursting hashtags can be found each day. If a larger ratio is required, the value of δ should be set smaller, and vice versa.

Definition 3. Off-Burst. Starting from $c_{t'}$, if all the values are smaller than $\max(c_1 + \delta, 1.5c_1)$ in the following 24 hours, t' is defined as the end of the burst. We can say the hashtag is off-burst since t' .

Definition 4. Death. The definition of “Off-Burst” corresponds to the definition of “Burst”. Analogously, the “Death” is defined corresponding to the definition of “Prediction-Trigger”. If a bursting hashtag could no longer satisfy the condition for triggering prediction in consecutive 24 hours, the hashtag is considered dead. In other words, a complete lifecycle of the bursting hashtag come to an end.

Fig. 1 shows several examples of bursting hashtags. It can be observed that they vary in when they burst and how long the bursting is sustained. Based on the definitions above, we propose a framework of the real-time prediction shown in Fig. 2, covering the entire lifecycles of hashtags. When a new hashtag

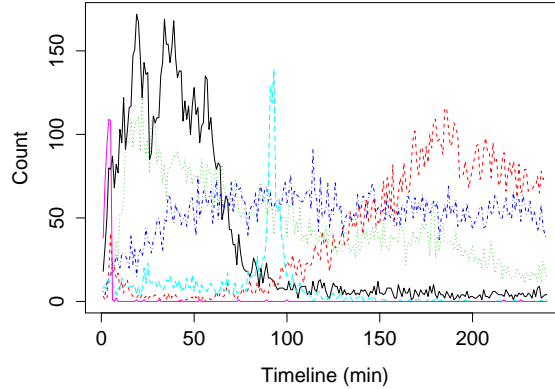


Fig. 1. Examples of bursting hashtags

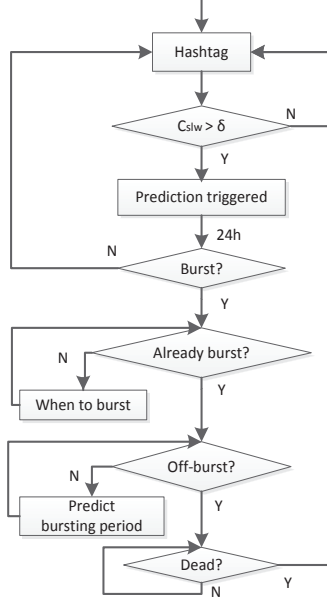


Fig. 2. The framework of real-time prediction

comes, a five-minute sliding window is used to constantly check whether it triggers prediction. When it satisfies the triggering condition, real-time prediction is triggered. The first prediction task is to predict whether it will be a bursting hashtag. If it will be, then we predict when it will burst, i.e., the **T**ime period **B**efore the onset of **B**urst (*TBB*). For a hashtag that has already burst, we skip the first two prediction tasks and directly predict when it will be off-burst, i.e., the **T**ime period that it can **R**emain **A**ctive (*TRA*). When this bursting hashtag is dead, it is taken as a new hashtag, entering the prediction process again. In other words, when a hashtag comes to the end of last lifecycle, it automatically starts next round of life.

2.2 Prediction Tasks and Solutions

Four prediction problems have been raised over the lifecycle of a hashtag. In this study, we focus on the first three problems closely related to “burst”.

Task 1. Will a hashtag be a bursting hashtag? This problem can be framed to a normal binary classification task.

Input: A set of candidate hashtags which triggered prediction $HT = \{ht_1, ht_2, \dots\}$

Output: A class label for each hashtag $L(h_i), h_i \in HT$, indicating whether it will be a bursting one.

Solution: We propose a weighted SVM-based method to solve this problem, whose dataset is unbalanced. An optimal weight for the positive class is needed to train the classification model. Algorithm 1 shows the process of optimizing

Algorithm 1 Training optimal classification model.**Input:** TNS : training set; TTS : training-test set; PNR : the ratio of negative samples to positive samples; $R_w = \{1, 2, \dots, 2PNR\}$: the range of the weight for positive class;**Output:** C_{opt} : the optimal classifier;

```

1:  $F_{max} \leftarrow 0$ ;
2: for all  $w \in R_w$  do
3:   Training a weighted SVM classifier  $C_w$  on  $TNS$ ;
4:   Compute  $F_1$ -score  $F_1^w$  by applying  $C_w$  to  $TTS$ ;
5:   if  $F_1^w > F_{max}$  then
6:      $C_{opt} \leftarrow C_w$ ;
7:      $F_{max} \leftarrow F_1^w$ ;
8:   end if
9: end for
10: return  $C_{opt}$ ;

```

the weight for the positive class. Since the dataset is unbalanced, F_1 -score is used as the criteria for training the optimal model. At the same time, we also tried several related methods to evaluate the performance. The evaluation results are demonstrated in Section 4.2.

Task 2. If a hashtag will be a bursting one, when will it get burst? This problem can be framed to a regression task.

Input: A set of bursting hashtags which haven't burst $HB = \{hb_1, hb_2, \dots\}$.

Output: The time period (minutes) before the onset of each bursting hashtag $TBB(h_i), h_i \in HB$. Note that predicting the exact value of TBB is extremely difficult and is often not necessary. Therefore we relax the problem and predict the natural logarithm of TBB , $\log(TBB)$. In other words, we turn to predict the range of the time period.

Solution: We tried five different models to solve this problem, including Linear Regression(LR), Classification And Regression Tree(CART), Gaussian Process Regression(GPR), Support Vector Regression(SVR) and Neural Network(NN). The evaluation results can be found in Section 4.3.

Task 3. Once a hashtag get burst, how long will it remain active, i.e, when will it be off-burst? This problem can also be framed to a regression task.

Input: A set of bursting hashtags which have got burst $HB' = \{hb'_1, hb'_2, \dots\}$.

Output: The time period (minutes) that each bursting hashtag can remain active $TRA(h_i), h_i \in HB'$. Similar to task 2, we turn to predict the natural logarithm of TRA , $\log(TRA)$.

Solution: To solve this problem, we also tried the five different models used in Task 2. The evaluation results are shown in Section 4.4.

Table 1. Distribution of bursting hashtags over time

	5min	15min	30min	1h	3h	6h	24h	48h
RAB	30.32%	51.02%	68.03%	80.88%	92.68%	95.12%	100.00%	100.00%
ROB	4.35%	39.53%	47.29%	59.90%	81.15%	88.50%	96.11%	99.60%
RAD	0.43%	14.60%	19.38%	32.84%	61.95%	73.46%	88.60%	98.15%

Table 2. Proportion of bursting hashtags in the dataset

5min	15min	30min	1h	3h	6h
12.39%	9.24%	6.27%	3.58%	1.27%	0.80%

2.3 Statistics and Challenges

δ , the parameter in the definition of “burst”, can be adjusted according to the statistics of real data. We analyzed a two-month dataset from Nov 1, 2012, to Dec 31, 2012 and found that about 400,000 new hashtags are generated every day. Table. 1 shows the distribution of bursting hashtags over time. The three keys in the table, *RAB*, *ROB*, and *RAD*, are ratios defined as follows:

$$RAB = \frac{\#hashtags_already_burst}{\#bursting_hashtags}$$

$$ROB = \frac{\#hashtags_of_fburst}{\#bursting_hashtags}$$

$$RAD = \frac{\#hashtags_already_dead}{\#bursting_hashtags}$$

From Table. 1 we can obtain three observations. Since the time when the prediction was triggered, about 95% of bursting hashtags get burst within 6 hours; about 96% of bursting hashtags are off-burst within 24 hours; about 98% of bursting hashtags are dead within 48 hours.

The most challenging problem for bursting hashtag prediction comes from the unbalanced data. Table 2 shows the distribution of hashtags triggering prediction. It can be seen that, as time goes by the data becomes more and more skewed. The proportion of bursting hashtags in the dataset even goes down to **0.8%** at the 6th hour. It is quite challenging to precisely predict so few bursting hashtags from the data set.

3 Feature Space

In this section, we explore different types of features which may indicate the future trend of hashtags, including meme features, user features, content features, network features, hashtag features, time series features, and prototype features.

3.1 Meme Features

Tweet count. We use the number of tweets containing a hashtag to represent current popularity of the hashtag, instead of using the appearance count of the hashtag. This is because some tweets may use the same hashtag multiple times.

Author count. Besides tweet count for a hashtag, we also consider the unique number of authors who posted tweets containing the hashtag. This feature can be used to recognize those hashtags automatically posted by some fake accounts.

Retweet count. Retweeting is the typical way of information diffusion in Twitter. Interesting information can spread quickly and broadly through retweets. If a user retweeted a tweet, that means the content of the tweet successfully attracted the attention of this user and motivated him to share it. Besides indicating the interestingness of messages, the retweeting behavior of a user may also affect his followers.

Mention count. Mention is a directional sharing behavior in Twitter. Messages can be shared to a designated user using @ as the prefix of the user's name. If a user was mentioned in a tweet with a hashtag, he probably took part in the topic, especially when this mention came from his friends.

Url ratio. A url in Twitter can be a link of a picture, a song, a video, or a piece of news. High ratio of tweets with urls may indicate a topic about a good song, an interesting picture or video, or a piece of breaking news. For example, #GoodLife, a hashtag with a high url ratio, was about a great new song posted by the hippop musician Lyinheart on Memorial Day.

We also consider the ratio version of author count, retweet count and mention count.

3.2 User Features

Total follower count. In Twitter, if a user posts a tweet, this tweet will be shown on the personal pages of the user's followers. When any follower see this tweet (suppose it has a hashtag), he may retweet it or post a new tweet using this hashtag if he is interested in this topic. Therefore, the total count of followers seems to be the potential scale of future adoption of the hashtag.

Maximum follower count. Within the users who adopted a hashtag, if there is one whose follower count is much larger than others, e.g. a celebrity, the followers of this user may play a leading role in the potential adoption of this hashtag. Besides, for two hashtags which have the same total follower count, the maximum follower count may break the tie.

Passivity. Active users often post or retweet tweets following some hashtags. On the contrary, passive users rarely do so unless the topics are attractive enough. The passivity of a user is defined as the reciprocal of average number of tweets posted by this user per day, which is formed as:

$$Psv(u_i) = \frac{N_d(u_i)}{1.0 + N_t(u_i)}$$

where $N_d(u_i)$ denotes the number of days since the user account was created, and $N_t(u_i)$ denotes the total number of tweets posted by this user.

3.3 Content Features

Special signals. Casual language is commonly used in Twitter. Users often use repeated letters in words to strengthen the mood, for example, goooooood, ple-asssssse. Besides, repeated punctuation marks can also indicate users' emotion strength. Repeated exclamation mark (!!!) and question mark (???) are also considered as special signals in this study. The tweets with special signals are counted and used as a feature.

Word-level sentiment strength. For different kinds of hashtags, users tends to use words of varying sentiment strength in tweets. Whether for positive or negative sentiment, those words of strong sentiment usually mean bursting hashtags. For example, in those tweets containing the bursting hashtag #songsiwillalwayslove, some words of strong positive sentiment were used, such as beautiful, awesome, amazing, favorite, etc. While in tweets containing the bursting hashtag #LondonRiots, some words of strong negative sentiment were used, such as terrible, sad, sorry, etc. Using SentiWordNet [1], a lexical resource for supporting sentiment classification and opinion mining applications, average positive score of words can be computed for each hashtag, as well as average negative score. These two scores are used as features indicating word-level sentiment strength.

Emoticon count. An emoticon is a metacommunicative pictorial representation of a facial expression, usually constructed by punctuation marks or traditional alphabetic. Emoticons are commonly used to express a person's feelings in social media, which fall into two typical types. One type is happy/winking emoticons, the other type is sad/disappointed emoticons. We adopted the collections of emoticons in [15], which focus on Twitter part-of-speech tagging. Here are some examples of emoticons.

- Happy emoticon: :) :-) :') :] =]
- Sad emoticon: :(:-(:'(:[=[

We count the tweets with happy emoticons and the tweets with sad emoticons separately, and use them as two independent features.

3.4 Network Features

As mentioned above, retweets and mentions can accelerate the diffusion of a hashtag. A user network for a hashtag can be constructed by use of retweets and mentions, which is a directed graph $G = (V, E)$. Then we can extract several features from this retweet-mention network.

The order of the graph. The order of the graph is $|V|$, i.e. the number of vertices. Each vertex represents a user involved in the retweet-mention network.

Density. Density of the graph is defined as the total number of edges divided by the total number of possible edges, which is used to describe the general level of connectedness in the graph:

$$Density = \frac{|E|}{|V| \times (|V| - 1)}$$

Average degree. Average degree of the graph is another feature measuring the connectedness in the graph, which is formed as:

$$AveDegree = \frac{2|E|}{|V|}$$

Entropy of degree distribution. Entropy of degree distribution can be used to measure the heterogeneity of the network, which is defined as:

$$Entropy = - \sum_{k=1}^{|V|} p(k) \log(p(k))$$

where $p(k)$ is computed using frequencies of node degree.

3.5 Hashtag Features

Length of the hashtag. The length of a hashtag is defined as the number of characters in the hashtag. For example, the length of #3peopleulove is 12. A hashtag which is too short or too long tends not to become a bursting hashtag.

Case-sensitive hashtag count. The case of a hashtag may be changed during diffusion. For example, #3peopleulove, #3PeopleuLove and #3PeopleULove are the same hashtag. Popular hashtags tend to have more versions of different cases.

Co-occurrence times with other hashtags. Sometimes, some hashtags are not used individually, but are used together with other hashtags, e.g. #boston#explosion. Here the co-occurrence times is calculated by the number of tweets with two or more hashtags, one of which is the hashtag to be predicted.

3.6 Time Series Features

Dormant period. Dormant period refers to the time period before the prediction was triggered. For different hashtags, the dormant periods vary from several seconds to several weeks. When predicting how long a bursting hashtag will remain active, a similar feature of how long this hashtag has been bursting is also considered.

Polynomial coefficients. Time series can reflect the revolution of a hashtag. Since the granularity of the time series is 1 minute, sometimes the length of the complete time series is too long, which is unreasonable and unnecessary to be directly used as a feature. We adopted two methods to represent the shape of time series. One is polynomial curve fitting, whose function is formed as:

$$f(x, W) = \sum_{k=0}^{\beta} w_k x^k,$$

subject to

$$\beta = \begin{cases} t_p - 1 & \text{if } t_p \leq 6 \\ 6 & \text{others.} \end{cases}$$

Table 3. Derivative features from time series

Feature Name	Mathematical Presentation	Description
mean_value	$E(c) = 1/(t_p - s + 1) \times \sum_{j=s}^{t_p} c_j$	mean value of the time series
std_value	$\sqrt{1/(t_p - s + 1) \times \sum_{j=s}^{t_p} (c_j - E(c))^2}$	standard deviation of the time series
d_last_first	$c_{t_p} - c_s$	d-value between the last point and the first one
d_last_max	$c_{t_p} - \max(c_j), j \in \{s, s+1, \dots, t_p\}$	d-value between the last point and the maximum one
d_last_min	$c_{t_p} - \min(c_j), j \in \{s, s+1, \dots, t_p\}$	d-value between the last point and the minimum one
idx_max	$m c_m = \max(c_j), j \in \{s, s+1, \dots, t_p\}$	index of the maximum point
mean_fod	$E(fod) = 1/(t_p - s) \times \sum_{j=s}^{t_p-1} c_{j+1} - c_j $	mean value of the absolute first-order derivative
std_fod	$\sqrt{1/(t_p - s) \times \sum_{j=s}^{t_p-1} (c_{j+1} - c_j - E(fod))^2}$	standard deviation of the absolute first-order derivative
last_fod	$c_{t_p} - c_{t_p-1}$	last value of the first-order derivative
max_fod	$\max(c_{j+1} - c_j), j \in \{s, s+1, \dots, t_p-1\}$	maximum value of the first-order derivative
d_pfod_nfod	$pfod - nfod;$ if $c_{j+1} \geq c_j$, $pfod++$; else $nfod++$	d-value between positive and negative first-order derivative

where β denotes the order of the polynomial function, and t_p denotes the time (by minute) since the prediction was triggered. The upper bound of β is set to 6, because larger values lead to over-fitting. W , coefficients of polynomial curve fitting is used as the features representing the shape of time series.

Symbolic sequences. The other method to recognize the shape of time series is symbolic representation by use of SAX [10], a state-of-the-art method in time series analysis. SAX is leveraged to reduce the complete time series to a symbolic sequence, e.g., ACBF. Then 3-grams can be generated from the sequence. Note that 3-grams here are different from traditional ones at two aspects. First, the last item must be included, as it represents the latest status at that moment. Second, 3-grams here needn't be continuous. Therefore the 3-grams for the symbolic sequence ACBF are {ACF, ABF, CBF}. We can get a ranking list of 3-grams by processing the time series from bursting hashtags in the training set. The top-5 3-grams are used as features.

In addition, some derivative features are defined to describe the characteristics of the time series. Assume the time series obtained is formed as $\langle c_s, c_{s+1}, c_{s+2}, \dots, c_{t_p} \rangle$, where s denotes the starting minute when the prediction was triggered, and t_p denotes that minute when making a real-time prediction. Then details of the derivative features from time series are given in Table 3.

3.7 Prototype Features

Prototypes here refer to the similar historic hashtags to the new one to be predicted. When searching prototypes from historic hashtags, all the features but symbolic sequences are used as the representation of hashtags. Polynomial coefficients and symbolic sequences are both features representing the shape of time series. It would be redundant to use both of them. Symbolic sequences are excluded because the experimental results show that it's more effective to use polynomial coefficients. Based on Euclidian Distance, the similarity between hashtags is defined as follows:

$$Sim(h_i, h_j) = \frac{1}{1 + \sqrt{\sum_{n=1}^{\alpha} (F_n^i - F_n^j)^2}}, \quad (1)$$

where $h_i = \langle F_1^i, F_2^i, \dots, F_{\alpha}^i \rangle$, $h_j = \langle F_1^j, F_2^j, \dots, F_{\alpha}^j \rangle$.

Prototype features are different for the three prediction tasks defined in section 2. For prediction task 1, top- k prototypes for the hashtag to be predicted can be found from the historic dataset. In these top- k prototypes, the number of bursting hashtags is used as a feature. While for prediction task 2 and 3, top- k historic **bursting** prototypes should be found, whose weighted average TBB or TRA is used as a feature. Taking prediction task 2 as an example, the prototype feature (PF) is formed as:

$$PF(h_p) = \frac{\sum_{i=1}^k Sim(h_i, h_p) \times TBB(h_i)}{\sum_{i=1}^k Sim(h_i, h_p)}. \quad (2)$$

Considering k from 1 to 10, we can get 10 prototype features.

4 Evaluation

4.1 Experimental Settings

Dataset. Our datasets are collected through the Twitter stream API with Gardenhose access, containing roughly 10% of all public statuses on Twitter. We

Table 4. Datasets

Dataset	Time Period	#Positive Samples	#Negative Samples
Historic Set	2012.9 - 2012.10	1544	6681
Training Set	2012.11 - 2013.1	2382	11494
Test Set	2013.3	672	3153

have collected three datasets according to the requirements of the prediction problems. Table 4 shows the statistics of these datasets. Historic set is used to find the prototypes of hashtags in training and test sets. Training set is used to training prediction models. Once a prediction model is built, test set is used to evaluate the performance of the model.

Metrics. Since the dataset for prediction task 1 is largely unbalanced, accuracy is not a reasonable metric to evaluate the performance. The classifier can easily get high accuracy by predicting all samples into the class which plays the dominant role in the dataset. Precision, Recall and F-score are reasonable metrics for this prediction task.

$$Precision = \frac{true_positive}{true_positive + false_positive}$$

$$Recall = \frac{true_positive}{true_positive + false_negative}$$

$$F_\beta - score = (1 + \beta^2) \times \frac{Precision \times Recall}{\beta^2 Precision + Recall}$$

Since prediction task 2 and 3 are both regression tasks, the common metric Mean Squared Error(*RMSE*) is used to evaluate the prediction performance. The smaller *RMSE* is, the better the performance is.

4.2 Performance of Prediction Task 1

Performance Comparison Table 5 shows the performance comparison of prediction task 1, bursting hashtag prediction. Predictions were made at six representative time, which can be divided into three stages, early stage (5min, 15min), middle stage (30min, 1h) and late stage (3h, 6h). Predictions after 6 hours are not considered because about 95% of bursting hashtags get burst within 6 hours since the prediction was triggered. Conclusions can be drawn as follows:

Table 5. Performance comparison

Method	F_1 -score					
	5min	15min	30min	1h	3h	6h
NaiveBayes	0.299	0.342	0.312	0.197	0.132	0.081
C4.5	0.211	0.238	0.235	0.118	0.043	0
LR	0.068	0.244	0.209	0.190	0.044	0.077
NN	0.170	0.202	0.210	0.250	0.185	0.194
SVM	0.207	0.308	0.329	0.277	0.187	0.125
LSM [14,4]	0.184	0.145	0.161	0.097	0.014	0.022
Our Method	0.326	0.372	0.368	0.360	0.290	0.250

LR: Logistic Regression NN: Neural Network

- Our method significantly outperforms the other related methods in terms of F_1 -score at any prediction stages. Compared to the best performance of other related methods, the improvement by our method is more significant at later prediction stages.
- From 5-minute prediction to 6-hour prediction, the F_1 -score obtained by any method but LSM increases first and decreases afterwards. It's known to all that the later the prediction is made, the more information can be used for prediction. But why does the performance drop at later prediction stages? The reason is that as time goes by, the dataset become more and more skewed. As shown in Table 2, before 15 minutes, the proportion of positive samples in the dataset is not very low. But after 15 minutes, this proportion becomes lower and lower so that the increase of information used for prediction cannot compensate the proportion decrease of bursting hashtags in the dataset. Therefore the F_1 -score increases first and decreases afterwards.

In practical applications, there may be different requirements on precision and recall. For a bursting topic recommendation application, high precision seems to be preferred to ensure the quality of recommendation; while for a public opinion monitoring application, high recall tends to be required because it hopes to miss as few bursting topics or events as possible. These requirements can be satisfied by training optimal prediction models using different types of F -scores. For example, $F_{0.5}$ -score can be used when high precision is preferred, while F_2 -score can be used when high recall is preferred. Fig. 3 shows the performance comparison for optimizing prediction models using different F -scores. It can be seen that for all the cases, the precision is higher when optimizing $F_{0.5}$ -score, while the recall is higher when optimizing F_2 -score.

Analysis of Feature Contribution We defined and used 7 types of features for predicting bursting hashtags, meme features, user features, content features,

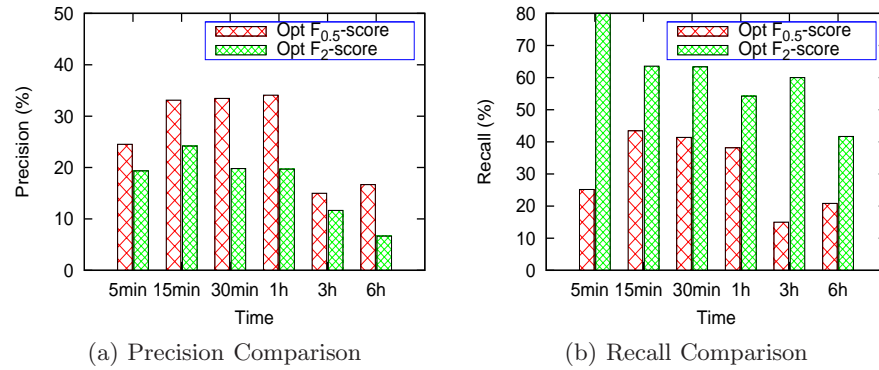


Fig. 3. Performance comparison for optimizing prediction models using different F -scores.

Table 6. Evaluation of Feature Contribution

Feature Removed	$\Delta F_1\text{-score}/F_1\text{-score}$					
	5min	15min	30min	1h	3h	6h
Meme Features	-2.19%	1.95%	-2.68%	-4.35%	-18.08%	-8.57%
User Features	1.79%	2.76%	-4.35%	-6.08%	-41.14%	-9.86%
Content Features	1.74%	-1.07%	-4.24%	-3.18%	-21.34%	-12.09%
Network Features	-2.65%	2.55%	-1.88%	-1.33%	-38.12%	-17.76%
Hashtag Features	-1.15%	2.79%	-3.79%	-13.19%	-6.37%	-21.31%
Prototype Features	-3.89%	0.71%	-0.99%	0.00%	-0.72%	-5.88%
Time Series Features	-5.64%	-10.30%	-14.95%	-32.66%	-49.40%	-72.41%

network features, hashtag features, time series features and prototype features. In this section, we conduct experiments to examine the contributions of these features to prediction performance. During these experiments, we apply our method a number of times, each time removing only one type of features and recording the changes of prediction performance. The experimental results are demonstrated in Table 6. From Table 6, it can be observed that:

- Time series features are the most universal and effective features. When they are removed, the drop of F_1 -score is the most significant for all the cases. Prototype features seems to be the least effective features. For most cases, the drop of F_1 -score when they are removed is smaller than that when other features are removed.
- As time goes by, the contribution of time series features increases progressively. When they are removed, F_1 -score decreases by 5.64% for 5-minute prediction; while for 6-hour prediction, the drop of F_1 -score reaches 72.41%. The reason is that as time goes by, more and more negative samples show downtrend on time series. When removing any other features, the largest drop of F_1 -score appears at late prediction stage. This is because the dataset is so skewed at late prediction stage that different types of features are more necessary to be used together to train a synthetic prediction model.

4.3 Performance of Prediction Task 2

For this regression task, we evaluated five different models, Linear Regression, Classification And Regression Tree, Gaussian Process Regression, Support Vector Regression and Neural Network. Table 7 shows the results of performance comparison, measured by the typical metric $RMSE$. It can be seen that SVR gets the best performance for most cases. Only for the 15-minute and 1-hour prediction, is SVR slightly worse than GPR. From the 5-minute prediction to the 6-hour prediction, the performance by SVR drops first and rises afterwards. Obviously, the best performance is achieved at late stage, and the performance

Table 7. Performance for prediction task 2

Method	<i>RMSE</i>					
	5min	15min	30min	1h	3h	6h
LR	1.397	1.390	1.664	1.593	1.504	3.238
CART	1.413	1.413	1.675	1.761	1.618	1.295
GPR	1.379	1.335	1.570	1.540	0.890	1.205
SVR	1.375	1.345	1.566	1.586	0.806	1.177
NN	1.568	1.754	2.020	2.023	1.485	1.228

of early stage is better than middle stage. This implies that the uncertainty of middle-stage predictions is larger for predicting when a hashtag will burst.

4.4 Performance of Prediction Task 3

For this prediction task, we also evaluated the five different models tried in the prediction task 2. Table 8 shows the results of performance comparison. It can be observed that the performance of GPR and SVR is better than other methods. GPR achieves the best performance for 5-minute, 30-minute and 1-hour prediction; while SVR achieves the best performance for the other three cases.

5 Related Work

5.1 Analysis of Hashtag Adoption and Diffusion

Lin *et al.* [11] studied the dynamics of hashtag adoption by analyzing their growth and persistence. Based on empirical analysis of how the dual role of a hashtag affects hashtag adoption, Yang *et al.* [24] predicted whether a user would adopt a hashtag that he never used before. Romero *et al.* [18] studied the

Table 8. Performance for prediction task 3

Method	<i>RMSE</i>					
	5min	15min	30min	1h	3h	6h
LR	1.752	1.791	1.970	1.941	1.583	1.751
CART	2.073	1.893	1.870	1.880	1.661	1.767
GPR	1.728	1.692	1.711	1.777	1.376	1.691
SVR	1.791	1.663	1.721	1.789	1.315	1.672
NN	2.175	2.290	2.099	2.342	1.898	1.946

diffusion mechanics of hashtags of different types and topics. Chang [3] applied diffusion of innovation theory to explain the spread of hashtags on Twitter, i.e., hashtag usage and adoption. Romero [19] studied the interplay between social relationships and hashtag adoption and found that the social relationships between the initial hashtag adopters can predict future adoption of the hashtag.

5.2 Popularity Prediction in Microblogging Platforms

Item-level prediction. In microblogging services, interesting tweets are often retweeted by many users. There have been a number of works on popularity prediction of tweets. Suh *et al.* [21] performed analytics on factors which may impact retweeting in Twitter. Petrovic *et al.* [16] tried to predict whether a new tweet will be retweeted in the future through binary classification. Yang *et al.* [25] proposed a factor graph model to predict users' retweeting behaviors. Hong *et al.* applied multi-class classification to predict the popularity of tweets [7], and used Co-Factorization Machines to address the problem of predicting users' retweeting decisions [8]. Kong *et al.* [9] predicted lifespans of popular tweets based on their static characteristics and dynamic retweeting patterns.

Topic-level prediction. In [6], Gupta *et al.* used regression, classification and hybrid approaches to predict future popularity of current popular events. Given a popular event at the time interval t , they predicted the status of this popular event at $t + 1$. Analogously, Ma *et al.* [12,13] predicted popularity of hashtags in daily granularity. Concretely, they predicted the range of popularity using classification methods. Tsur *et al.* [22] studied the effect of content on the spread of hashtags in weekly granularity. Contrary to these works, this study performs real-time (minute-level) prediction for hashtags. Most germane to this work are two studies from the same group [14,4], which focused on predicting trending topics by time series classification. The authors collected some historic trending topics, and selected the slice of time series centered at the trend onset of these topics as reference signals. Then new hashtags were classified by comparing observed time series to these signals. Different from all of these works, our work provides the first definition of bursting hashtags and studies comprehensive real-time prediction problems covering their entire lifecycles.

6 Conclusion

In this paper, we take the initiative to propose a real-time prediction framework of bursting hashtags. We define a series of interesting but challenging prediction tasks covering the entire lifecycles of bursting hashtags. Features of different types, as well as solutions are proposed to solve the prediction tasks. Evaluation experiments are conducted on real datasets from Twitter, and the results show that the proposed features and solutions are effective to the prediction tasks.

References

1. S. Baccianella, A. Esuli, and F. Sebastiani. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 2200–2204, 2010.
2. J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.
3. H. C. Chang. A new perspective on twitter hashtag use: diffusion of innovation theory. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–4, 2010.
4. G. H. Chen, S. Nikolov, and D. Shah. A latent source model for nonparametric time series classification. In *Advances in Neural Information Processing Systems*, pages 1088–1096, 2013.
5. D. Gruhl, R. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 78–87. ACM, 2005.
6. M. Gupta, J. Gao, C. Zhai, and J. Han. Predicting future popularity trend of events in microblogging platforms. *Proceedings of the American Society for Information Science and Technology*, 49(1):1–10, 2012.
7. L. Hong, O. Dan, and B.D. Davison. Predicting popular messages in twitter. In *Proceedings of the 20th international conference companion on World Wide Web*, pages 57–58. ACM, 2011.
8. L. Hong, A. S. Doumith, and B. D. Davison. Co-factorization machines: modeling user interests and predicting individual decisions in twitter. In *Proceedings of the sixth ACM international conference on Web Search and Data Mining*, pages 557–566. ACM, 2013.
9. S. Kong, L. Feng, G. Sun, and K. Luo. Predicting lifespans of popular tweets in microblog. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1129–1130. ACM, 2012.
10. J. Lin, E. Keogh, S. Lonardi, and B. Chiu. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11. ACM, 2003.
11. Y. R. Lin, D. Margolin, B. Keegan, A. Baronchelli, and D. Lazer. # bigbirds never die: Understanding social dynamics of emergent hashtag. *arXiv preprint arXiv:1303.7144*, 2013.
12. Z. Ma, A. Sun, and G. Cong. Will this# hashtag be popular tomorrow? In *Proceedings of the 35th ACM SIGIR international conference on Research and Development in Information Retrieval*, pages 1173–1174. ACM, 2012.
13. Z. Ma, A. Sun, and G. Cong. On predicting the popularity of newly emerging hashtags in twitter. *Journal of the American Society for Information Science and Technology*, 2013.
14. S. Nikolov. *Trend or No Trend: A Novel Nonparametric Method for Classifying Time Series*. PhD thesis, Massachusetts Institute of Technology, 2012.
15. O. Owoputi, B. OConnor, C. Dyer, K. Gimpel, N. Schneider, and N. A. Smith. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL-HLT*, pages 380–390, 2013.
16. S. Petrovic, M. Osborne, and V. Lavrenko. Rt to win! predicting message propagation in twitter. *Prof. of AAAI on Weblogs and Social Media*, 2011.

17. J. Ritterman, M. Osborne, and E. Klein. Using prediction markets and twitter to predict a swine flu pandemic. In *1st international workshop on mining social media*, 2009.
18. D. M. Romero, B. Meeder, and J. Kleinberg. 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*, pages 695–704. ACM, 2011.
19. D. M. Romero, C. Tan, and J. Kleinberg. On the interplay between social and topical structure. In *Proc. 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2013.
20. T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.
21. B. Suh, L. Hong, P. Pirolli, and E.H. Chi. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Second IEEE international conference on Social Computing*, pages 177–184. IEEE, 2010.
22. O. Tsur and A. Rappoport. 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*, pages 643–652. ACM, 2012.
23. A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welp. Election forecasts with twitter how 140 characters reflect the political landscape. *Social Science Computer Review*, 29(4):402–418, 2011.
24. L. Yang, T. Sun, M. Zhang, and Q. Mei. We know what@ you# tag: does the dual role affect hashtag adoption? In *Proceedings of the 21st international conference on World Wide Web*, pages 261–270. ACM, 2012.
25. Z. Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su. Understanding retweeting behaviors in social networks. In *Proceedings of the 19th ACM international conference on Information and Knowledge Management*, pages 1633–1636. ACM, 2010.
26. Z. Zhao and Q. Mei. Questions about questions: an empirical analysis of information needs on twitter. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1545–1556. International World Wide Web Conferences Steering Committee, 2013.