Abstract. Topic popularity prediction in social networks has drawn much attention recently. Various elegant models have been proposed for this issue. However, different datasets and evaluation metrics they use lead to low comparability. So far there is no unified scheme to evaluate them, making it difficult to select and compare models. We conduct a comprehensible survey, propose an evaluation scheme and apply it to existing methods. Our scheme consists of four modules: classification; qualitative evaluation on several metrics; quantitative experiment on real world data; final ranking with risk matrix and MinDis to reflect performances under different scenarios. Furthermore, we analyze the efficiency and contribution of features used in feature oriented methods. The results show that feature oriented methods are more suitable for scenarios requiring high accuracy, while relation based methods have better consistency. Our work helps researchers compare and choose methods appropriately, and provides insights for further improvements.

Keywords: Topic Popularity Prediction; Social Network; Twitter; Features; Survey; Evaluation

1 Introduction

Social network has become an indispensable part of our life, where users have access to posting and reading messages to and from the public. Everyday, in social networks, there are millions of messages posted about a wide range of topics. However, after a period of information dissemination, only a few prevailing topics will burst. Prediction of popularity of topics has meaningful impacts on many services and applications, such as marketing, advertisement, search engine queries and recommendation system [1,2]. News media utilizes online social networks to maximize their news visibility [3]. Hence, predicting the popularity of topics in social network is significant for both academia and industry.

Formally, we define the topic’s popularity as the number of messages related to it [4] and the popularity fluctuates with time. The objective of topic popularity prediction problem is to detect topics and predict their popularities in the near future utilizing the obtainable information.

Various topic popularity prediction methods proposed by previous works have different properties such as accuracy and complexity, which leads to different
suitability and feasibility under diverse scenarios. They all claim that they have
great performance in the experiments. However, tested on different datasets and
evaluation metrics, those methods have little comparability between each other,
which makes it difficult to choose an appropriate one under specific scenarios.
The objective of this paper is to perform a fair and strict evaluation of each
topic popularity prediction method with a uniform evaluation scheme.

We have collected the literatures about topic popularity prediction in social
network published recently as completely as possible. After the survey, we, as
the third party, develop a four-module evaluation scheme in order to make a con-
victive comparison to the topic popularity prediction methods comprehensively.
We classify all the methods into two categories: feature oriented methods and
relation based methods in the first module. In the second module, we carry out a
qualitative evaluation representing the temporal, spatial, and platform universal-
ity. We also list some properties such as core algorithms and data preprocessing
procedures, which fix their position among various problem situations. We fur-
ther perform the scheme’s third module: a quantitative experiment utilizing a
dataset grabbed from Twitter. We test the methods based on the whole dataset
without any optimization and give the accuracy evaluation which is objective
and valuable to refer. In the fourth module, we rank the methods by MinDis
metric based on risk matrix [4]. With various demands under different scenarios,
we can use MinDis to evaluate models and choose an appropriate method.

Moreover, in feature oriented methods, features are the linchpin to ensure the
accuracy. However, extracting features costs a lot, which requires to use efficient
features to improve the accuracy and reduce the complexity at the same time.
Therefore, we also give an evaluation of features’ contribution.

The contributions of our work are listed as follows:

– We conduct a comprehensive survey on topic popularity prediction in social
network, which is shown in the qualitative evaluation section;
– We propose a four-module evaluation scheme to fairly compare and rank
existing or prospective methods under different scenarios;
– We give an evaluation of features’ contribution on this problem.

The paper is organized as follows. We first give some definitions in Section 2.
In Section 3, we introduce our evaluation scheme. In Section 4, we classify the
topic popularity prediction methods and give a qualitative evaluation of them.
Section 5 shows the quantitative comparison results of the methods. Section 6
talks about the final ranking module. Section 7 deals with the efficiency of the
features using in machine learning models. In Section 8 we give summaries and
conclusions.

2 Technical Preliminaries

Definitions: The social network is defined as a graph, \( SN = (U, E, M) \). \( U \)
is the set of the users (nodes) and \( E \) is the set of the edges in the graph de-
defined as \( (u_q, u_p) \), which is weighted by times that \( u_p \) interacts (mention or reply)
with \( u_q \). \( M \) is the set of messages posted by \( U \). A topic \( h \) can be described as \((U^h, E^h, M^h)\), where \( U^h \) have posted messages about \( h \), \( M^h \), and \( E^h \) are all the interactions at \( h \). We divide the time \( T \) into a series \( T_1, T_2, ..., T_t \). The time series of \( h \) is \( TS^h = TS^h_1, TS^h_2, ..., TS^h_t \). \( TS^h_t \) is \((U^h_t, E^h_t, M^h_t)\) which means a part of \((U^h, E^h, M^h)\) where \( M^h_t \) is the messages in \( M^h \) posted in \( T_t \) and \( U^h_t \) only contains the users who post messages in \( M^h_t \). Formally, the popularity \( P \) of \( h \) in time period \( T_t \), \( P^h_t \), is defined as \( |M^h_t| \).

**Problem Statement:** Given a social network \( SN \), the problem of topics popularity prediction is to predict \( P \) of the topics in \( T_{t+1} \) using the information, or features in \((U, E, M)\) before \( T_t \). In our evaluation, we specify the purpose of the prediction to be predicting whether a topic will be popular.

## 3 The Evaluation Scheme

In this section, we elaborate our Topic Popularity Prediction Evaluation Scheme, which can be divided into 4 modules: Classification, Qualitative Evaluation, Quantitative Experiment, and Final Ranking. The scheme is shown in Figure 1.

### 3.1 Classification and Qualitative Evaluation Overview

According to our investigation, we classify the methods briefly into two categories: **Feature oriented methods** and **Relation based methods**. **Feature oriented methods** use raw features and universal classification or regression models to predict. **Relation based methods** aim at building a specific model by finding the mathematics relationship between the features and the popularity.

Qualitative evaluation is performed to learn about the characteristics of each method. Three main metrics: accuracy\(^1\)(\(a\)), complexity (\(c\)), and universality (\(u\)), as well as other properties such as denoising processes and algorithm categories, are evaluated.

We divide the three metrics into 3 levels and we provide a criteria table (Table 1) for them. These three metrics describe the synthetic performance of a method and we can use this criteria for all methods.

---

\(^1\) Accuracy is measured by F1 score.
Table 1. Criteria on Qualitative Evaluation

| Level  | Accuracy | Complexity² | Universality |
|--------|----------|-------------|--------------|
| HIGH   | (60%, 1) | If the complexity of preprocessing and extracting more than 20% of the features is \(O(n^2)\) or the main model’s complexity is over \(O(n^2)\). | Easy to transplant to other platforms (social network) and can deal with huge data stream in any time period. |
| MEDIUM | (30%, 60%) | Few parts of the methods have a complexity over \(O(n^2)\) but they are not in an absolute dominant role. | Have some limits to transplant to other platforms or can deal with common cases of data. |
| LOW    | (0, 30%) | If most preprocessing procedures, feature extracting, and the main model can be done within \(O(n)\). | Only be used in some specific situations or constricted to data scale. |

3.2 Quantitative Experiment Overview

Qualitative evaluation gives out accuracy level of the methods. In this module we evaluate precision more elaborately, excavate deep level information and achieve more precise and fairer evaluation.

Based on the unified big dataset from Twitter, we perform the experiments on topic popularity prediction methods, which makes the result comparable and gets rid of possible “cherry-picking” problems.

3.3 Final Ranking Overview

With the results of previous steps, we gain insights from risk management and employ risk matrix [4] to generate the weights due to the nature that a model’s final score should be considered by the category of probability against the category of importance. Then we rank models by presenting a novel metric, dubbed as \(\text{MinDis}\).

We use \(w_{a_1}, w_{a_2}, w_{rm}, w_c, \text{ and } w_u\) to represent the importance weights of \(\text{Macro-F1}, \text{Micro-F1}, \text{RMSE}, \text{complexity}, \text{and universality}.

\(\text{MinDis}\) is a metric proposed to describe how far a method is away from the ideal method according to the Euclidean Distance for a specific application scenario. It is formulated as:

\[
\text{MinDis} = \sqrt{\text{Dis}_{F1_{\text{macro}}} + \text{Dis}_{F1_{\text{mic}}} + \text{Dis}_{\text{rm}} + \text{Dis}_c + \text{Dis}_u} \quad (1)
\]

In the equation \(\text{Dis}_{\text{metrics}}\) represents \(w_{\text{metrics}} (E(\text{metrics}) - \text{metrics})^2\). This shows the Euclidean Distance of a metrics between the evaluated method and the perfect one in theory. \(\text{MinDis}\) can be used to rank all the methods under a specific scenario to select the most appropriate method.

4 Classification and Qualitative Evaluation

We first classify the methods and give a qualitative evaluation of them, which is aimed at comprehensively analyzing the method characteristics and giving an introduction of their merits and drawbacks.

² \(n\) is the number of tweets or users, and we only consider the number of traversal times in the complexity.
4.1 Feature Oriented Methods

In general, the topic popularity prediction problem for these methods can be divided into several parts: data preprocessing, denoising, feature extracting, and machine learning. For features are the key factors, in order to find efficient features, we briefly divide all the features further into six categories [6] (shown in Table 2).

In Table 2 we list all the metrics used to investigate the feature oriented methods in our evaluation scheme. Some methods treat hashtags as topics and utilize directly obtained features to train a classifier [5,6], while other methods pre-process the features to detect the topics and then train the machine learning algorithms [20]. The metrics in the evaluation scheme are described below:

Accuracy, Complexity, and Universality: Refer to Table [1].

F. Categories: Features impact the accuracy and efficiency directly. In Table 2, we list the categories of features a method adopts and the size of the feature space (Number of Features). Due to the limited space, we use 1 - 6 to denote six kinds of features: Content, User, Network, Meme, Hashtag, and TimeSeries respectively. We further evaluate feature effectiveness in Section 7.

Denoise: Denoising is an influential procedure to promote the performance of prediction. Some methods generate an equal internal time series (TS) to obtain information in early time intervals. Some use critical drop point to extract potential burst keywords. Others detect emerging topics (DET), and predict whether they will burst.

| Method | Paper Publish | Accuracy | Complexity | Universality | Algorithm | Denoise |
|--------|---------------|----------|------------|--------------|-----------|---------|
| F-I    | AIS&T         | Medium   | Medium     | High         | * * * * *  | 18      |
| F-II   | SIGIR        | Low      | High       | High         | * * * * *  | 26      |
| F-III  | ITAIC        | High     | High       | Medium       | * *        | 6       |
| F-IV   | ARCS         | -        | High       | Medium       | * *        | 3       |
| F-V    | SIGIR        | High     | High       | Medium       | * *        | 6       |
| F-VI   | ICWSM        | High     | High       | Medium       | * * * * *  | Many    |
| F-VII  | IJCAI        | -        | High       | Low          | * *        | -       |
| F-VIII | IC-BSNMT     | High     | Low        | High         | * * * *    | 8       |

4.2 Relation Based Methods

Different from feature oriented methods, relation based methods regard the social network as heterogeneous distributions.

Geographical records of social networks work as sensors to find regions’ interests. Thus, [15] proposed a geographical method, Latent Geographical Topic
Analysis (LGTA), which combines contents and locations. A much more popular type of methods is hashtag-based\cite{7}. Performances of these methods are more stable than others due to the empirical studies of burst hashtags patterns. Many other models are proposed based on Natural Language Processing (NLP), probability theory, graph theory, and so on. Clustering is another interesting approach.\cite{27} clustered users into groups and predicted the group-level popularity which improves the accuracy.

The evaluation results are listed in Table 3. The chosen metrics in the evaluation platform can be described as follows:

**Accuracy, Complexity, and Universality** are the same as above.

**Recognition Time**: This metric depicts how long a method can predict, such as R-IV can predict a topic two days prior to its burst on the civil unrest datasets.

**Data Preprocessing**: Some methods preprocess data before prediction, thus have high dependences on how well data are preprocessed.

**Field**: We summarize the field of knowledge the methods base on. Geographical methods are accessible in most platforms and languages, with a relative low accuracy. Methods based on graph theory and NLP can detect hot topics much earlier than others. However, they have a heavy dependence on preprocessing procedure.

| Method | Published in | Model | Accuracy | Complexity | Recognition Time | Universality | Data Preprocessing | Field |
|--------|--------------|-------|----------|------------|------------------|--------------|-------------------|-------|
| R-I \cite{7} | CIKM | IMA | Medium | Low | - | High | Probability (hashtag) |
| R-II \cite{9} | AAAI | Pre-Whether | High | Low | Medium | Medium | Probability |
| R-III \cite{8} | Sci. W. J. | Pre-recognition Model | Medium | Medium | High | High | * |
| R-IV \cite{13} | SIGKDD | NPHGS | High | High | High | Medium | * Graph Theory |
| R-V \cite{15} | IW3C2 | LGTA | Low | Low | Low | High | Geographical |
| R-VI \cite{16} | WWW | Temporal and Spatial Model | Medium | Medium | Medium | Medium | Geographical |
| R-VII \cite{17} | CIKM | Hybrid Model | High | Medium | Low | High | * Probability |
| R-VIII \cite{18} | Neurocomputing | Diffusion Model | Medium | Medium | Medium | High | Graph Theory |
| R-IX \cite{19} | NIPS | Latent Source Model | High | Low | Low | High | Probability |
| R-X \cite{20} | TWEB | SPIKEM | High | Medium | - | High | Probabilistic |
| R-XI \cite{21} | WWW | GPOP | High | Medium | High | High | * Tensor decomposition |
| R-XII \cite{22} | JIIS | EdgeMRF | - | Medium | - | High | MRF |

### 5 Quantitative Experiment

We perform a quantitative experiment based on a real world dataset to guarantee the fairness. We choose five methods to evaluate: F-I, F-II, R-I, R-II and R-III,
for they are all proposed recently, contain relatively complete feature spaces, and have great universality.

5.1 Introduction of the Experiments

Our dataset is grabbed from Twitter containing about 2 million tweets from August 1, 2015 to January 31, 2016. After filtering, all the data we used are tweets post in English because different languages may need different tokenization methods. Table 4 shows the detailed statistics about the dataset. In our evaluation, the time period is set to be 1 day.

Table 4. Detailed Information about the Dataset

| Statistic                  | Value | Statistic                        | Value |
|----------------------------|-------|----------------------------------|-------|
| Number of Tweets           | 2.5 M | Number of Followers              | 4.5 M |
| Number of Users            | 1.5 M | Number of Tweets with Hashtag    | 35K   |
| Number of Retweet          | 0.7 M | Fraction of Tweets with Hashtag  | 1.4%  |

We conduct 10-fold cross validation on all the data to train the methods using machine learning models. All the features are extracted without any unfair optimization. We choose 5 metrics: Precision, Recall, Macro-F1, Micro-F1, and RMSE to measure their performance. F1 is an metric that synthesizes precision and recall. Macro-F1 is impacted mainly by the prediction accuracy of rare classes (popular topics). Micro-F1 is similar to Macro-F1, but it is impacted mainly by the prediction accuracy of common classes (non-popular topics). We employ RMSE to uncover integral constancy of the methods.

5.2 Results of the Quantitative Experiment

We choose the data in 7 days to fairly compare both two kinds of methods. Feature oriented methods are further experimented with the half-year data, dubbed as (Origin). In Table 5 we list the accuracy of each method and we compare them in Figure 2.

Comparison between Feature Oriented Methods and Relation Based Methods: From Figure 2 we can find that feature oriented methods perform better than relation based methods in precision, recall, and Macro-F1. This meets expectations, for feature oriented methods take more factors into account and the process of training with big data can give a more exact relationship between the factors and popularity. While RMSE and Micro-F1 scores shows that feature oriented methods have a lower performance in prediction of non-popular topics, which may be because the model is over-fitting.

Comparison within Feature Oriented Methods: From Figure 2 we can see that when predicting for only 7 days, F-I with a smaller feature space has a better performance. It gives an expression to the importance of feature choice and the process to get rid of noise. For denoising, in F-I, only information in
time interval $t$ is used to denoise. While in F-II, time interval $t-4$, $t-3$, ..., $t-1$ are also used. It is possible that F-I does a better job in denoising in this short period prediction because the information used for denoising in F-I is closer to the predicted time period and can reflect the features more accurately. F-I (7 Day) has a higher accuracy than F-I (Origin). However, F-II (7 Day) is worse than F-II (Origin). This shows F-I is more suitable in short period prediction than F-II, while F-II is more suitable in predicting rare class in both situations for its Macro-F1 value is higher. In long period prediction, both of them have great accuracies.

Comparison within Relation Based Methods: From Figure 2 we find that R-II consistently achieves better performance than the other two relation based methods. The three latent features: Sum, Average Rate of Change, and Standard Deviation help the model to detect hot topic patterns in a universal way. However, R-I uses three features detected from experiential studies. Setting a class of miscellaneous does not always perform well because their patterns are not constant and will have different performances in different datasets. In our experiments, R-I is too sensitive to find enough hot topics. More effective features should be implemented to better classify patterns of events.

R-III is a language-related solution. Its performance displayed in Figure 2 is quite different from the other two, especially on Micro-F1 and RMSE. Micro-F1 proves that NLP methods are apt to predict with a large scale of data, or "rich" data. For its experiments are based on microblogging data, we find it defective to process Twitter data and this method highly depends on pretreatment of posts, and users should manually reprocess the clustering results. A high RMSE reveals its precariousness.

6 Final Ranking

With the evaluation results above, we rank the methods by MinDis metric to fit different demands. We implement the ranking steps under 4 classical scenarios respectively.
I: This scenario is under a balance requirement, where the importances of every metric are the same.

II: This scenario is under a complexity oriented requirement which is widely used in real time prediction. Low complexity and high universality are required to deal with large flow data stream from different platform.

III: Accuracy is heavily weighted and Macro-F1 has the biggest weight, which shows that the precise prediction of rare classes is important under this scenario. This situation conforms to recommendation systems' requirement.

IV: We highly consider consistency with the goal of choosing a method with consistent performance.

Table 6. Risk Matrix of Scenario III

| Certain | Likely | Possible | Unlikely |
|---------|--------|----------|----------|
| Negligible | Marginal | Critical | Catastrophic |

Table 7. Weights under Each Scenario

| Scenarios | \(w_c\) | \(w_u\) | \(w_a\) | \(w_r\) |
|-----------|--------|--------|--------|--------|
| I         | 0.200  | 0.200  | 0.200  | 0.200  |
| II        | 0.286  | 0.285  | 0.143  | 0.143  |
| III       | 0.182  | 0.091  | 0.363  | 0.182  |
| IV        | 0.125  | 0.125  | 0.125  | 0.5    |

We adopt risk matrix to determine the weight of metrics under different scenarios. The columns show different levels of severity if the metrics are too bad while the rows show how possible and frequent each good metric can be achieved in all methods. Each level has an interval of 20%. According to our survey, 50% methods have low accuracy, thus, the likelihood of Micro-F1 and Macro-F1 is Possible. Similarly, the likelihood of high complexity, low universality and high RMSE are likely, possible and possible respectively. Table 6 shows the risk matrix of scenario III. The weights of metrics under it can thus be specified as in Table 7. We do the same for scenario II, III and IV. Users can make up their own scenarios.

We use the weights and previous evaluation results to calculate MinDis. The values of low, medium, and high level are 0.4, 0.5 and 0.6 because they can keep the relatively balance relationship among the metrics and guarantee that there is no one metric overwhelming in the final results. The results are listed in Table 8. F-I (7 Day) outperforms others under three scenarios while R-II (7 Day) stands out under scenario IV showing a better consistency.

Table 8. Final Ranking of All the Methods in Each Scenario

| Scenarios | F-I (Origin) | F-II (Origin) | F-I (7 Day) | F-II (7 Day) | R-I (7 Day) | R-II (7 Day) | R-III (7 Day) |
|-----------|-------------|--------------|-------------|--------------|-------------|--------------|--------------|
| I         | 0.3190      | 0.2991       | 0.1848      | 0.5018       | 0.2471      | 0.2170       | 0.4730       |
| II        | 0.2697      | 0.2528       | 0.1638      | 0.4241       | 0.2221      | 0.2019       | 0.4916       |
| III       | 0.3603      | 0.3282       | 0.1979      | 0.4849       | 0.3194      | 0.2799       | 0.5112       |
| IV        | 0.3943      | 0.3656       | 0.2466      | 0.6521       | 0.2031      | 0.5786       |              |

7 Evaluation of Effectiveness of the Features

In this section, we analyze the features widely used in this problem, and evaluate their efficiency based on numerical experiments in order to choose appropriate
features. LibSVM is used as the machine learning model. We train the model using the whole dataset, and then 7-day data. 10-fold cross validation is used to separate training and testing data.

Altogether we take 34 features into account. The features are classified into six categories [6]. We list all the features in Table 9 with some qualitative analyses: Complexity which means the complexity to extract the feature, Temporal representing whether this feature varies with time, and Graphic telling us whether this feature describes graph \( (V_i^h, E_i^h) \).

The features are described as follows:

| No. | Name             | Description                                                                 | Complexity | Temporal | Graphic |
|-----|------------------|-----------------------------------------------------------------------------|------------|----------|---------|
| F_{c1} | NumEmo          | The total number of emoticons included in \( T_i^h \)                       | Medium     | *        |         |
| F_{c2} | NumSpeSig       | The total number of special signals like “!!!” or “Gooooood!” included in \( T_i^h \) | Medium     | *        |         |
| F_{c3} | SentiOHash      | The value of positive and negative points calculated by SentiStrength model  | Medium     |          |         |
| F_{c4} | Topics          | 20-Dimension topic distribution vector derived from \( T_i^h \) using LDA model | High       |          |         |
| F_{c5} | ActOUser        | Average activities of \( U_i^h \) gained by PageRank [22]                  | Medium     | *        | *       |
| F_{c6} | MaxOF           | The max number of followers of users in \( U_i^h \)                         | Medium     | *        | *       |
| F_{c7} | AvOF            | The average number of followers of users in \( U_i^h \)                     | Medium     | *        | *       |
| F_{c8} | Length          | The length of the hashtag                                                  | Low        |          |         |
| F_{c9} | multiFreq       | Fraction of tweets having more than one hashtags                           | Medium     |          | *       |
| F_{c10} | Clarity         | Kullback Leibler divergence of word distribution between \( T_i^h \) and tweets collection \( T \) | High       |          |         |
| F_{c11} | ExtClarity      | The extension of \( F_{c2} \)                                              | Quite High |          |         |
| F_{c12} | NumInHash       | Whether a number is contained in the hashtag                               | Low        |          |         |
| F_{c13} | NumOWord        | The number of words in a hashtag                                            | Low        |          |         |
| F_{c14} | Degree          | The average degree of nodes in \( G_i^h \)                                 | High       | *        | *       |
| F_{c15} | Density         | The density of \( G_i^h \)                                                 | High       | *        | *       |
| F_{c16} | Order           | The order of \( G_i^h \)                                                    | High       | *        | *       |
| F_{c17} | EntroODD        | The entropy of degree distribution of \( G_i^h \)                          | High       | *        | *       |
| F_{c18} | NumOBUser       | The number of border users of \( G_i^h \)                                 | Quite High |          |         |
| F_{c19} | ExpVec          | Exposure vector of border users                                             | Quite High |          | *       |
| F_{c20} | CompFrac        | Ratio between number of connected components and number of nodes in \( G_i^h \) | High       | *        | *       |
| F_{c21} | Weight          | The average weight of edges in \( G_i^h \)                                | High       | *        | *       |
| F_{c22} | TriFrac         | Fraction of users forming triangles in \( G_i^h \)                         | High       | *        | *       |
| F_{c23} | NumUser         | The number of users in the \( U_i^h \)                                   | Medium     | *        |         |
| F_{c24} | PracUser        | The fraction of users who have posted tweets with hashtag \( h \) in the \( G_i \) | High       |          |         |
| F_{c25} | NumOT           | The number of tweet in the \( T_i^h \)                                     | Medium     | *        |         |
| F_{c26} | FracOT          | The fraction of tweet having URL in it in the \( T_i^h \)                  | Medium     | *        |         |
| F_{c27} | Mn              | The mean value of the time series                                           | High       |          |         |
| F_{c28} | MnD             | The standard deviation value of the time series                            | High       |          |         |
| F_{c29} | Sd              | The mean value of the absolute first-order derivative of time series       | High       |          |         |
| F_{c30} | SdD             | The standard deviation value of the absolute first-order derivative of time series | High       |          |         |

Content Features \( F_c \): They reflect the meaning and sentiment of posts. Previous studies show that negative topics are more likely to be popular [11]. Therefore, the features describing sentiment (\( F_{c1}, F_{c2}, F_{c3} \)) are taken into account.
User Features $F_u$: Because a tweet posted by a more active user is more likely to be popular, $F_{u1}$ is listed in. $F_{u2}$ and $F_{u3}$ are extracted to reflect the number of followers who affect the dissemination of tweets.

Hashtag Features $F_h$: Hashtag Features show the number of words ($F_{h6}$), clarity ($F_{h3}$), and some other information. For instance, “livemusic” and “Epl” are two hashtags. “livemusic” has two words and a clearer meaning.

Network Features $F_n$: The graph $(U^h, E^h)$ is described by network features. Border users represent those who are not in $U^h$ but follow the users in $U^h$. $F_{n6}$ is a 15 dimension vector and dimension $D_i$ denotes that the number of border users following $i$ users in $U^h$ is $D_i$. A three nodes cycle in the graph is a triangle and let the set of triangles be $\Delta^h$. $F_{n2}$ is $|\Delta^h| / (||U^h||(|U^h| - 1)(|U^h| - 2) / 3)$. Let $C^h_i$ stands for the set of disconnected parts of the graph. $F_{n7}$ is defined as $|C^h_i| / |U^h|^i$.

Meme Features $F_m$: Meme Features are mainly about “numbers” and “fractions” in tweets such as the number and fraction of tweets with “@”. Time Series Features $F_t$: Time series can be described by the polynomial fitting curve and the absolute first-order derivative of the fitting curve. Their mean value and standard deviation value are chosen as features.

Some of these methods are multidimensional such as Topics ($F_{c4}$) whose dimension is 20, so the whole feature space has 68 dimensions. We remove each feature from the feature space use the new feature space to train the classifier respectively, and rank all the removed features by its relative contribution (RC), which is:

$$RC = -1000(A_i - A_s)$$

$A_i$ is the average value of Macro-F1 and Micro-F1 when $F_i$ is removed from the feature space and $A_s$ is the standard average value with the whole feature space.

We first evaluate with the whole half-year training data. The fifteen most efficient features (Rank1-Rank15) and the fifteen least efficient features (Rank54-Rank68) are listed in Table 10. We observe that, first, meme features are more efficient, $F_{m1}$ and $F_{m7}$ can reflect the present incidence of the topics. A tweet with URL contains more abundant contents that people are more interested in, so $F_{m8}$ is also an efficient feature. Second, border users can decide how widely the topic will be disseminated, so $F_{n5}$ is the most efficient feature. Third, there are 5 dimensions of Topic Vector among the Top 15, which reflects an actual phenomenon that some specific topics have a relatively high possibility to be popular. Fourth, there is no Time Series Feature in Top 15. Fifth, $F_{n6}$ does not perform as efficient as expected. It is likely that the low dimensions of the vector are too undistinguished to help classify the popularity. While the high dimensions of the vector need a relatively complete dataset to work efficiently. Sixth, except for the $F_{c4}$, most of the efficient features are temporal and graphic. However, the complexities of them are commonly high. In the evaluation with only 7-day training data, the results are similar. We also find that time series features are more suitable in short period training to describe the latest trends.
### Table 10. The 15 Most Efficient Features and the 15 Least Ones

| Rank | Feature | Rank | Feature |
|------|---------|------|---------|
| 1    | \(F_{m8}\): Number of Border User | 54   | \(F_{n3}\): Co-occurrence Frequency |
| 2    | \(F_{m1}\): Number of User | 55   | \(F_{n6}\): Fraction of Retweet |
| 3    | \(F_{m8}\): Fraction of URL | 56   | \(F_{n6}\): D-3 of Exposure vector |
| 4    | \(F_{m7}\): Number of Tweet | 57   | \(F_{t2}\): D-4 of Topic Vector |
| 5    | \(F_{n3}\): Clarity | 58   | \(F_{n4}\): D-4 of Topic Vector |
| 6    | \(F_{m2}\): Fraction of User | 59   | \(F_{n6}\): D-6 of Topic Vector |
| 7    | \(F_{n4}\): D-15 of Topic Vector | 60   | \(F_{n1}\): Count of Emoticons |
| 8    | \(F_{m4}\): D-17 of Topic Vector | 61   | \(F_{n4}\): Extension Clarity |
| 9    | \(F_{m2}\): Fraction of User | 62   | \(F_{n6}\): D-4 of Exposure vector |
| 10   | \(F_{m1}\): Activity of User | 63   | \(F_{n2}\): Graph Density |
| 11   | \(F_{n7}\): Component Fraction | 64   | \(F_{n6}\): D-10 of Exposure vector |
| 12   | \(F_{n4}\): D-5 of Topic Vector | 65   | \(F_{n4}\): D-17 of Topic Vector |
| 13   | \(F_{n4}\): D-12 of Topic Vector | 66   | \(F_{n3}\): Average of Followers |
| 14   | \(F_{n4}\): D-16 of Topic Vector | 67   | \(F_{n8}\): Edge Weight |
| 15   | \(F_{n4}\): Max of Followers | 68   | \(F_{n3}\): Sentiment of Hashtag |

---

### 8 Conclusion

This work proposes a rigorous evaluation scheme for forecasting the rise of certain topics in social media and surveys the existing methods based on it. Prior methods’ experiments are based on different datasets, which results in low comparability among them and there is no prior research performing the evaluation. In our scheme, we first classify the methods into two categories: \textit{Feature Oriented Methods} and \textit{Relation Based Methods}. We then perform qualitative evaluation and quantitative experiment based on a real-world dataset. We give out the final ranking values to greatly reflect their comprehensive performance under given scenarios. The evaluation results show that \textit{Feature Oriented Methods} are more suitable for accuracy dominant scenarios and \textit{Relation Based Methods} have better consistency. It also provides several detailed findings. First, denoising essentially suits models to different scenarios. Second, the number of features and data samples have no special relation to the prediction performance. Therefore, choosing efficient features is important. Third, \textit{Feature Oriented Methods} have a higher accuracy but they suffer from heavy overhead such as feature extracting. In addition, we evaluate the efficiency of 34 features used in \textit{Feature Oriented Methods}. The results show that features about graph and temporal information are more efficient. In all, our work reveals different characteristics of the methods, provide guidelines on selecting appropriate methods under different scenarios, and help select efficient features for applications.
References

1. Wu, B. and et al.: Sequential Prediction of Social Media Popularity with Deep Temporal Context Networks. In: IJCAI, pp. 3062–3068 (2017)
2. Kong, S. and et al.: On the Real-time Prediction Problems of Bursting Hashtags in Twitter. Comput. Sci. 109(3), 246–248 (2014)
3. Villi, M. and et al.: Recommend, Tweet, Share: User-Distributed Content (UDC) and the Convergence of News Media and Social Networks. Springer Berlin Heidelberg (2016)
4. Talbot, J.: What’s right with risk matrices?. Jakeman Business Solutions
5. Ma, Z. and Sun, A. and Cong, G.: On predicting the popularity of newly emerging hashtags in twitter. J. Am. Soc. Inf. Sci. Tec. 64(7), 1399–1410 (2013)
6. Kong, S. and Mei, Q. and Feng, L. and Ye, F. and Zhao, Z.: Predicting bursts and popularity of hashtags in real-time. In: SIGIR, pp. 927–930 (2014)
7. Cui, A. and Zhang, M. and Liu, Y. and Ma, S. and Zhang, K.: Discover breaking events with popular hashtags in twitter. In: CIKM, pp. 1794–1798 (2012)
8. Zhu, T. and Yu, J.: A Prerecognition Model for Hot Topic Discovery Based on Microblogging Data. Sci. W. J. 2014, 360934 (2014)
9. Deng, Z. and et al.: Effectively predicting whether and when a topic will become prevalent in a social network. In: AAAI, pp. 210–216 (2015)
10. Zhang, X. and Li, Z. and Chao, W. and Xia, J.: Popularity Prediction of Burst Event in Microblogging. In: WAIM, pp. 484–487 (2014)
11. Thelwall, M. and Buckley, K. and Paltoglou, G.: Sentiment in Twitter events. J. Am. Soc. Inf. Sci. Tec. 62(2), 406–418 (2011)
12. Li, C. and Sun, A. and Datta, A.: Twevent: segment-based event detection from tweets. In: CIKM, pp. 155–164 (2012)
13. Chen, F. and et al.: Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In: SIGKDD, pp. 1166–1175 (2014)
14. Gao, S. and Ma, J. and Chen, Z.: Effective and effortless features for popularity prediction in microblogging network. In: WWW, pp. 269–270 (2014)
15. Yin, Z. and Cao, L. and Han, J. and Zhai, C. and Huang, T.: Geographical topic discovery and comparison. In: WWW, pp. 247–256 (2011)
16. Sakaki, T. and Okazaki, M. and Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors. In: WWW, pp. 851–860 (2010)
17. Cataldi, Amodeo, G. and Blanco, R. and Brefeld, U.: Hybrid models for future event prediction. In: CIKM, pp. 1981–1984 (2011)
18. Du, Y. and He, Y. and Tian, Y. and Chen, Q. and Lin, L.: Microblog bursty topic detection based on user relationship. In: ITAIC, pp. 260–263 (2011)
19. Ishikawa, S. and Arakawa, Y. and Tagashira, S. and Fukuda, A.: Hot topic detection in local areas using Twitter and Wikipedia. In: ARCS, pp. 1–5 (2012)
20. Chen, Y. and Amiri, H. and Li, Z. and Chua, T.: Emerging topic detection for organizations from microblogs. In: SIGIR, pp. 43–52 (2013)
21. Becker, H. and Naaman, M. and Gravano, L.: Beyond Trending Topics: Real-World Event Identification on Twitter. In: ICWSM, pp. 11: 438–11: 441 (2011)
22. Zhang, X. and et al.: Event detection and popularity prediction in microblogging. Neurocomputing 129, 1469–1480 (2015)
23. Wang, X. and Wang, C. and Ding, Z. and Zhu, M. and Huang, J.: Beyond Trending Topics: Real-World Event Identification on Twitter. J. Intell. Inf. Syst. (2017)
24. Page, L. and Brin, S. and Motwani, R. and Winograd, T.: The PageRank citation ranking: bringing order to the web. Stanford InfoLab (1999)
25. Chen, G. H. and Nikolov, S. and Shah, D.: A Latent Source Model for Nonparametric Time Series Classification. In: NIPS - Volume 1, pp. 1088–1096 (2013)
26. Kong, S. and Mei, Q. and Feng, L. and Zhao, Z. and Ye, F.: Nonlinear Dynamics of Information Diffusion in Social Networks. ACM Trans. Web 11(2), 1–40 (2017)
27. Hoang, M. X and et al.: GPOP: Scalable Group-level Popularity Prediction for Online Content in Social Networks. In: WWW, pp. 725–733 (2017)
28. Liu, Y. and Han, W. and Tian, Y. and Que, X. and Wang, W.: Trending topic prediction on social network. In: IC-BNMT, pp. 149–154 (2013)