IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 2, 2020, accepted May 3, 2020, date of publication May 18, 2020, date of current version June 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2995411

Forwarding Behavior Prediction Based on Microblog User Features

CHUNLONG FU, YAJUN DU, BINYAN LYU, QIAOYU ZHOU, RUILIN HU, PENG JIA, AND YUJIAN ZHOU

School of Computer and Soft Engineering, Xihua University, Chengdu 610039, China

Corresponding author: Yajun Du (duyajun@mail.xhu.edu.cm)

This work was supported in part by the National Nature Science Foundation under Grant 61872298, Grant 61532009, Grant 61802316, and Grant 61902324, and in part by the Sichuan Science and Technology Service Industry Demonstration Project under Grant 2019GFW115.

ABSTRACT In microblog networks, when a user posts a microblog, other users may forward the post, and then the forwarding process will bring about the rapid dissemination and diffusion of information. In this paper, we propose a comprehensive and novel approach to predict user forwarding behavior. Firstly, we build the feature sets that affect the microblog forwarding, such as interest topic,geographic location, user aggregation coefficient,neighborhood overlap and so on. These features are classified into four categories: user characteristics, microblog features, network structure features, and interactive behavior characteristics. Secondly, we establish a feature selection model based on Filtering and Wrapping for predicting the forwarding behavior of users. The model includes three aspects: (1)ANOVA(Analysis of variance): The value of each feature is analyzed by variance analysis. If the feature variance is small, the feature provides less information. (2)χ^2 test and point-two-column correlation analysis: They filter discrete and continuous features, respectively. (3)Wrapper analysis: In order to solve the strong correlations between the features, we use LVW(Las vagas wrapper) algorithm to analyze the above feature sets, and then obtain the optimal feature combination. Finally, we propose the forwarding prediction model based on AdaBoost(Adaptive boosting) algorithm. Experimental results demonstrate that the model has the highest precision and F1 score than Naive Bayes, Logistic Regression, Random Forest and SVM(Support vector machine), and the F1 score reached 0.885. Among different topics, our proposed AdaBoost prediction model has good recall and F1 scores for different topics. In addition, by using different feature sets for comparison experiments, it is found that the optimal features selected in this paper are very effective.

INDEX TERMS Ensemble learning, feature selection, interest drift, microblog forwarding, topic model.

I. INTRODUCTION

Due to the popularity of social networks, the communications become more and more convenient. In the social networks, online users can get the information what they want. Meanwhile, they can publish their own views and interact with other online users. So without going outdoors, they know all the world's affairs. Meanwhile, a large number of social network platforms,such as Sina microblog, Twitter, Facebook, etc. have a strong influence on the social networks, generating a large number of topics and events, which show the most important or hot news in time. Therefore, we study that microblog users' forwarding prediction can grasp the forwarding rules accurately and effectively, hold the information dissemination and control public opinion analysis. Meanwhile, online public opinion products are in great demand in the market. So, Effective public opinion analysis products are of great value to the development of various fields. The following part includes the research status at home and abroad and the contributions of this paper.

The associate editor coordinating the review of this manuscript and approving it for publication was Sabah Mohammed.

A. FACTORS AFFECTING FORWARDING BEHAVIOR

Fan *et al.* [1] find that Sina microblog is very different from Twitter, it has its own characteristics between the network structures and the user behavior. Ma *et al.* [2] extract seven information in a microblog features from hashtag strings and tweets sets containing hashtag, and extract 11 history microblog features from social graphs formed by users

with hashtag. Suh *et al.* [3] find that URLs and hashtags have a strong relationship with forwarding ability in content characteristics. Li *et al.* [4] propose five features of interest similarity, user activity, content importance, user influence and user intimacy. SVM algorithm is used to predict the size of microblog forwarding. The prediction accuracy of the experiment reaches 86.63%. Zhang and Cai [5] find that there are different types of links in the Microblog network, and put forward the characteristics of homogeneity, micronetwork structure, geographical distance and gender. Chen *et al.* [6] focus on user attributes, message attributes and microblog user attributes, and establish a forwarding prediction model for hot topics by quantifying characteristics of forwarding activity, forwarding interest. Cao *et al.* [7] and Wang *et al.* [8] put forward three kinds of characteristics based on microblog content, user attributes and social relations and study the topological structure of the relationship of microblog users' interest network and propose a probabilistic cascade model. Tang *et al.* [9] use the idea of ''microeconomics'' to study the redistribution behavior of individual users and the relay relationship between users through the similarity between users, and finally transform the prediction problems into multi-task learning problems. Can *et al.* [10] analyze the image features besides the content and structure of tweets, and predict the number of forwarding tweets. Liu *et al.* [11] propose a method based on user activity and time window forwarding behavior, unreceived behavior and neglected behavior, and they propose a user forwarding rate and interaction frequency. Boyd *et al.* [12] analyze the psychological motivation of microblog users when forwarding posts. Lee and Sundar [13] study the credibility of information dissemination in Twitter. Zhong *et al.* [14] propose a method to discover user interest based on the characteristics of Microblog network. Xiao *et al.* [15]analyzed the factors affecting user behavior and found that implicit links play a very important role in user behavior. Michelson *et al.* [16] use knowledge base to eliminate the ambiguity of entities in tweet and classify them. Guo *et al.* [17] analyze Sina Microblog data and find that the factors affecting microblog forwarding are divided into three categories: microblog author, microblog heat and microblog interest.

B. FORWARDING BEHAVIOR OF SOCIAL NETWORKS

Zhang *et al.* [18] study how the friends affect a microblog user's forwarding behavior in a self-centered network. Galuba *et al.* [19] track and analyze the spread of URLs in Twitter social networks, and propose a propagation model that predicts which URLs a users may refer to. Tian *et al.* [20] analyze the factors affecting the dissemination of micro-blog information in regular network, random network and micro-blog information dissemination network. Bagdouri and Oard [21] propose a discriminant model to predict the possibility of users replying or retweeting on Twitter networks. Fan *et al.* [22], Yan *et al.* [23] quantify and analyze the information diffusion and network structure in Microblog. Tang *et al.* [24] combine personal and global characteristics, and establish the microblog forwarding model IRBLRUS based on user similarity. Yao *et al.* [25] make the statistics in the large-scale network structure of Sina Microblog from a macro perspective. They find that forwarding makes the network highly linked.

C. INFORMATION DISSEMINATION

It is important to study the dissemination of information in social networks for microblog forwarding. Pastor-Satorras and Vespignani [26] propose a dynamic model for spreading infection on a scale-free network, and find the average lifetime and persistence of the virus on the Internet. Xiao *et al.* [27]analyzed the factors that affect information dissemination through a hot spread model based on user's multidimensional attributes and evolutionary games. Liu *et al.* [28] propose an improved rumor propagation model SEIR(susceptible-exposed-infected-recovered), and use clustering algorithm to study the impact of user communities on the spread of microblog rumors. Kanavos *et al.* [29] propose a prediction model for Tweet retweeting depth and width using data mining technology. Ma *et al.* [30] use Sina Microblog data to study the microblog popularity. Tsur and Rappoport [31] propose a hybrid method based on linear regression to predict the propagation of an idea in a given time frame. The combination of Twitter content and topological structure features with time series can minimize the prediction error. Jenders [32] discuss important issues related to Twitter information dissemination, and analyze the impact of tweet posts and user characteristics on the dissemination. Gao *et al.* [33] propose an extended enhanced Poisson process model with time mapping process, and predict its future trends. Kupavskii *et al.* [34] study the number of forwarding times for the posts in Twitter within a fixed time period T. Bild *et al.* [35] study the factors affecting information dissemination on Facebook platform. Li *et al.* [36] summarize the information dissemination of online social networks.

D. CONTRIBUTION

In this paper, we mainly analyse the factors that affect the forwarding of microblogs, and formalize these factors to obtain the complete feature sets. Then these features are filtered to get the optimal features. Finally, the algorithm based on the integrated learning framework is used to predict whether the microblog is forwarded or not. Our main contributions are listed as follows:

- Representation of features. Firstly, by analyzing the factors affecting microblog forwarding, we first characterize the features and classify them into four categories according to their attributes: user characteristics, microblog characteristics, network structure characteristics and user interaction characteristics. Secondly, these features are formally expressed and quantified to obtain numerical features. Then, in order to solve the differences between the types and sizes of the eigenvalues, the maximum and minimum normalization method is

used to map the values of all the features to the [0,1] interval, and the complete set of features is obtained.

- Optimizing the features. In order to avoid the influence of excessive or invalid features on the performance of this prediction model, we deletes the indicators that can not improve the performance of microblog forwarding prediction model. The first step is the single feature analysis. First, we use variance analysis to screen out features with small range of variation of eigenvalue. Then we use χ^2 test and point-two-column correlation analysis to screen out the features that have little correlation with the microblog forwarding, and get the primary feature set. The second step is the multi-feature analysis: combine the useful features obtained in the first step and delete the redundant features by using LVW feature combination algorithm, so that each feature is relatively independent. Finally, the optimal feature set is obtained.
- Predicting of microblog forwarding. We put forward the forwarding prediction model based on AdaBoost algorithm.In the training process, the AdaBoost algorithm changes the weights of sample data and the weak classifiers continuously, and finally combines all weak classifiers in some way to get the desired strong classifiers. In order to prove the superiority of our proposed feature-based microblog forwarding prediction algorithms, we first preprocess the microblog data sets, and then compare the proposed algorithm with the classical algorithm in the same data set to find the best prediction model. At the same time, we also study the microblog forwarding under different users and different topics.

II. RELATED WORKS

A. FEATURE SCREENING

At present, the existing feature selection algorithms can be divided into three categories: Filter mode, Wrapper mode and Embedded mode [37].

The feature filtering model only scores each dimension features according to divergence or correlation, and each score represents the importance degree of the features. Then, the model sets a threshold to select features by the number of features or by score ranking. Representative methods include analysis of variance, information gain, χ^2 test, and correlation coefficient method.

The feature wrapper model scores the feature set by using selected algorithms, and obtains the corresponding importance degree according to the scores of these feature combinations, and then selects the optimal feature set. Such subset selection can be regarded as an optimization process, that is, through the continuous heuristic method to search the feature subset, and the feature subset of each search is put into the learning model for training, comparing the features according to the evaluation indicators obtained by the model. The time complexity of the wrapper feature selection algorithm is closely related to the size of the feature set. The more the number of features in the feature set is, the higher the complexity of the algorithm becomes. This is because the larger the feature set is, the more the combinations of the corresponding feature subsets are. The more feature subsets are, the more the number of original set features are. The number of feature subsets increase exponentially.

The feature embedded model is a method of using a model combined with machine learning, in which the process of feature selection and the learning process of the model are completed simultaneously. For example, in the process of classification using logistic regression algorithm, the algorithm itself combines a regular term to constrain the number of features. When there are more features, the larger the regular term is, the larger the loss function is. So it is mutual restraint. The embedded model method relies on machine learning algorithms and is not widely used.

B. ENSEMBLE LEARNING

Ensemble Learning [38] is a machine learning method that uses multiple (usually homogeneous) learners to construct a strong classifier to solve the same problems. It can be divided into four categories: Bagging, Boosting and Stacking.

- **Bagging**: Bagging is a technique that repeats and returns samples from a data set based on an uniform distribution theory. The sub-trainers required for each base model are composed of the returned samples, each of which is as large as the original training set. There is no strong dependency between individual learners, that is, the classification results of each weak classifier are combined in some way to obtain the final output result. The most representative method is Random Forest.
- **Boosting**:The sample will continuously adjust the sample weight during the training process. The training set of the base model will be converted every time according to a certain strategy. Each time, the weight of the data set that was previously misclassified is increased, and the prediction is made. Classifiers with high accuracy increase weight. There is a strong dependency between individual learners, that is, the next classifier performs training or testing on instances where the previous classifier prediction is not accurate enough. Representative methods are AdaBoost (Adaptive Boosting), Gradient Boosting Machines (GBM), and Gradient Boosted Regression Trees (GBRT).
- **Stacking**: The stacking method is to train one model to combine other models. First, we train several different models, and then train a model with the output of each model as input to get a final output.

For the training of several base classifiers, it is necessary to make a comprehensive judgment of these output results, and there are some rules for classification problems of numerical classes. Multiple basic classifiers perform classification prediction, and then vote according to the classification result, and different voting methods are used according to different voting rules. Common methods include One ticket veto, By all votes, Less obey most, and threshold voting. The One ticket veto means that as long as there is a veto in all

classifier prediction results, the final classification result is veto; the By all votes indicates that all the classifier prediction results are positive votes, then the final classification result is positive;the Less obey most indicates that in all the classifier prediction results, the affirmative and negative votes have the highest number of votes, then the final classification result is divided into which category, usually the number of base classifiers is odd; the threshold voting indicates The ratio of positive votes to negative votes in all classifier prediction results. If the ratio reaches a certain threshold, it is divided into corresponding categories, otherwise it is divided into another category.

III. FEATURES SETS AFFECTING MICROBLOG FORWARDING

In Microblog social network, users can post microblog posts and browse other people's posts. When users see a post, they can like, comment and forward. By focusing on other users, each microblog posts published by other users can be seen by the user, and the user may forward the post. The forwarding behavior between the users constitutes a microblog user forwarding network. In the microblog user forwarding network, if the user i pays attention to the user j. The user i is referred to as a downstream user (*fan*), and the user j is referred to as an upstream user (*Idol*). Effectively formalizing the definition of this phenomenon will help to study the forwarding behavior between users. Whether a microblog user forwards a particular microblog post can be formally defined as $F = f(u_i, w_{jk})$, $F \in \{-1, 1\}$. If $F = 1$, then the downstream user i forwards the upstream user j. If $F = -1$, then it is not forwarded. The definition of the downstream user and the upstream user is determined according to the concern cases. For a certain user, in general, he will pay attention to other users, other users will pay attention to him. So in real life, the vast majority of Miceoblog users are both downstream users and upstream users. Factors affecting whether Microblog posts are forwarded are closely related to Microblog users, Microblog content.

In Fig. 1, whether the downstream user forwards the upstream user's microblog is related not only to the upstream user but also to the microblog posts posted by the upstream user. These correlations can be roughly described by similarity. The relationship between the downstream user and the upstream user includes the similarity between the basic attributes of the users, such as whether the Microblog users are authenticated, whether the sexualities are the same, the similarities between the named entities, neighborhoods of the geographical locations of the users, similarities of the user and the user interests. The interesting similarity between users is constructed by the LDA topic model based on interest drift to construct relative entropy, and then converted into similarity. The similarity includes the forwarding strength of downstream users to upstream users. The construction of these features will be transformed into similarities between feature vectors. The downstream users are also closely related to the microblog posts posted by the upstream users. First, whether the downstream users forward the microblog posts is related to the importance and novelty of the microblog itself; and is also closely related to the Microblog topics. If it is the same as the interests of downstream users, then downstream users will tend to forward the Microblog posts of upstream users.

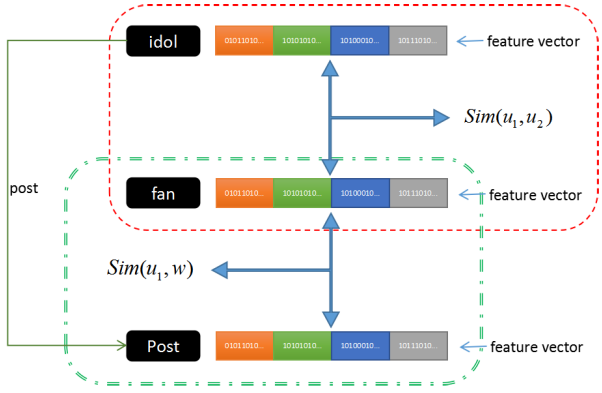

FIGURE 1. Similarity between downstream users, upstream users and Microblog posts.

A. USER FEATURES

For a specific microblog post published by a user, many factors affect the user forwarding the microblog post. Whether microblog post can be forwarded or not is closely related to the user's characteristics. We quantify the characteristics of microblog users that affect the forwarding of microblog posts.

1) USER IS VERIFIED OR NOT

If the microblog user belongs to the authenticated user, then the microblog user has greater influence and authority than other ordinary users. Because personal authentication needs to provide real identity information, so authenticated users are more intense to maintain their image, then they will publish high-quality microblogs. Downstream users prefer to forward microblogs issued by authenticated upstream users. This feature can be obtained from users basic information. This feature belongs to 0, 1 features, which can be described as Eq. (1):

$$Verified(u_1) = \begin{cases} 1, & Verified \\ 0, & Unverified \end{cases} \tag{1}$$

2) WHETHER THE GENDER OF MICROBLOG USERS IS THE SAME

Different sexuality users have different interests and preferences for the same microblog posts. Of course, different sexuality users have different interests in the content of microblog post. So different sexuality users also have an impact on the forwarding of microblog posts. Therefore, we describe a factor that affects the forwarding microblog posts through whether the sexuality of microblog users is the same.

This feature can be expressed by Eq. (2)

$$Gender(u_i, u_j) = \begin{cases} 1, & Same \\ 0, & Different \end{cases} \quad (2)$$

3) NAMED ENTITY SIMILARITY BETWEEN USERS

There are a lot of named entities, such as person names, place names and organization names. After users register, then they can post their microblog posts. Usually, if the content of the microblog posts contains familiar places to other users, such as the home and old school of the users, then the users may be more interested in the microblog posts. It is easier to forward the microblog posts. Therefore, we extract named entities from specific microblog posts, and also extract named entities from basic information of the users, and then describe their similarity through improved Jacquard similarity.

$$sim(n_u, n_w) = \sqrt[n]{\frac{|n_u \cap n_w|}{|n_u \cup n_w|}}. \quad (3)$$

where the named entity set n_u, n_w which represents the basic information of users u and w. The number n of named entity set is used to adjust the value of similarity. Because when a named entity matches the user's information once in microblog posts, the traditional Jacquard similarity may be very small, but the user's intimacy to microblog posts is relatively high, so we adjust it by squaring n. If they have the same named entities, the more similar they are, the more likely they are to forward the microblog posts.

4) GEOGRAPHIC LOCATION SIMILARITY

If the location of downstream users'(fans') lives and the related to content of microblog posts are close to each other, users may be more interested in these microblog posts. In real life, forwarding their microblogs are easier among the related users. The distance formula between downstream users and upstream users and micro-blog posts can be obtained by the Eq. (4).

$$\begin{aligned} D_{funs,idos,pos} = {} & alpha * Dis(Coor_{funs}, Coor_{idos}) \\ & + (1 - alpha) * Dis(Coor_{funs}, Coor_{post}) \end{aligned} \quad (4)$$

where the parameter α represents an adjusting factor of the distance between downstream users and upstream users. If the microblog posts do not contain geographic location $\alpha = 1$, otherwise, the initial settings $\alpha = 0.5$. $Coor_{funs}$ and $Coor_{idos}$ indicate the latitudes and longitudes of downstream and upstream users' geographical locations. $Coor_{post}$ indicates the location of upstream users' micro-blog posts. The user's city can be obtained from the user's basic personal information. Geocoding API interface provided by Baidu Map Open Platform can transform geographical location into coordinate representation [39]. $Dis(A, B)$ represents the distance between two objects A's and B's two geographic locations, where the object A represents $Coor_{funs}$ and the object B can be $Coor_{idos}$ or $Coor_{post}$. The distance is measured by the distance of the earth's surface. For a location, it includes its longitude and latitude. Then the distance of the earth surface of two objects A and B can be calculated by Eq. (5).

$$Dis(A, B) = 2 * R * \arcsin(\frac{\sqrt{2 * (1 - \eta + \theta - \tau)}}{2}). \quad (5)$$

where $\eta = cos(laB - laA)$, $\theta = cos(LaA) * cos(LaB)$, $\tau = \theta * cos(lnB - LnA)$, laA and laB indicate the latitude of objects A and B. lnA and lnB indicate the longitude of objects A and B. R indicates the radius of the earth which is 6371 km. In the above Eq. (4), if the distance between the three is larger, the closer the downstream users are to the upstream users and micro-blog posts, and vice versa. Then the similarity can be used to measure the geographic similarity between users by Eq. (6).

$$Sim_{location}(funs, idos, post) = \frac{1}{1 + D_{funs,idos,post}} \quad (6)$$

5) OTHER RELEVANT USER FEATURES

In addition to the above features, it also includes the length of user registration time, the number of users concerned, and the number of users' fans. The longer the user registering time is, the more familiar the user is with the microblog platform. At the same time, the user has more social relationships and more interests than other newly registering users. Then the user may be more likely to forward other people's microblog posts than other users. The number of users concerned, to a certain extent, reflects the user's interest in other users' microblog posts. At the same time, the more the number of users concerned is, the more users can receive their favorite microblog posts are. So it provides more possibilities for users to forward microblog posts. So the number of users concerned has a certain impact on the forwarding microblog posts. The number of users' fans, to a certain extent, shows the degree of interest of other users to themselves, and also reflects the influence of the users. The more the number of users' fans are, the more attractive the users are. To maintain their fans' continuous attention to themselves, microblog users will publish a large number of microblog posts to maintain this relationship, then the user may forward more microblog posts.

B. CONTENT FEATURES

Whether a microblog user forwards a microblog post is not only related to the microblog users, but also closely related to the microblog posts. Microblog Posts contain abundant information, including text, pictures, videos, URL links. In this subsection, we quantify the content characteristics of microblog posts that affect the forwarding of microblog users.

1) DOES A MICROBLOG POST CONTAIN A TOPIC TAG

Topic tag is a tag that users add to microblog posts when they send microblog posts. This tag outlines the theme of this microblog posts. Usually, microblog posts with tags are mostly hot topics, and the more microblog will be forwarded.

This feature of microblog post w_i can be obtained by Eq. (7).

$$Hashtag(w_i) = \begin{cases} 1, & IncludeHashtag \\ 0, & UnincludeHashtag \end{cases} \tag{7}$$

2) TOPIC SIMILARITY BASED ON INTEREST DRIFT

Users' interests will constantly change as time goes on. They will generally forget some of their previous interests and generate some new ones. This is very similar to the principle of the Ebinhaus forgetting curve. Users remember something and forget it slowly over time. Based on the Ebinhaus curve as the interest drift of microblog users, this paper uses the forgetting rule curve [40], calculates the microblog weight of target users in different time, and defines the range $[1, T]$ of time t.

$$b_t = k/((\lg t)^c + k) \qquad t \geq 1. \tag{8}$$

where b is the amount of memory preservation, t is the number of days from the current time to the research time, and c and k are the parameters. Ebinhaus proved that when $c = 1.25$, $k = 1.84$, the law of change of this function is approximately the same as that of human forgetting.

From the above Eq. (8), we can get the weight of each word corresponding to each microblog, and then add these weighted words together to get the user's interest word bag, as shown in Fig. 2.

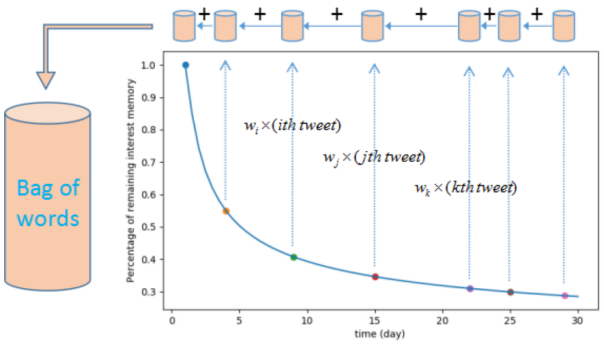

FIGURE 2. Construction of word bag model based on interest drift.

By using the word bag, we can construct the word vector with interest drift, and then use the LDA (Latent Dirichlet Allocation) topic model to get the user's topic distribution P_{user}. We can get the topic distribution P_{weibo} of Sina microblog posts, which is used to express the topic distribution of specific microblog posts. To measure the consistency of interests and topics of microblog users, the difference between the two probability distributions is described by relative entropy (Kullback-Leibler divergence) [41]:

$$D_{KL}(P_{weibo}|P_{user}) = \sum P_{weibo} \cdot \log \frac{P_{weibo}}{P_{user}}. \tag{9}$$

Because the relative entropy is asymmetric. That is to say $D_{KL}(P_{weibo}|P_{user}) \neq D_{KL}(P_{user}|P_{weibo})$, the user's interest preference distribution is different from the topic distribution of microblog posts. If the same weight is set at 0.5, the following definition can be made:

$$D_{KL} = 0.5 * D_{KL}(P_{weibo}|P_{user}) + 0.5 * D_{KL}(p_{user}|P_{weibo}). \tag{10}$$

The smaller the difference D_{KL} between user's interest distribution and microblog topic distribution is, the higher the topic similarity is. Then the microblog user may be more interested in the microblog posts, and the microblog can be forwarded more easily. The topic similarity is measured by Euclidean similarity. Similarly, it is not difficult to get the similarity between users.

$$D_{KL} = 0.5 * D_{KL}(P_{weibo}|P_{user}) + 0.5 * D_{KL}(p_{user}|P_{weibo}). \tag{11}$$

C. STRUCTURAL FEATURES

In research of microblog user forwarding, we regard microblog users as a node in the network. The relationship between users constitutes the edges of the network. Ultimately, the relationship between users and users constitutes a relationship network. The relationship structure of users has an important impact on microblog forwarding. In this subsection, we quantify the mutual characteristics of microblog network structure that affect microblog user forwarding.

Clustering coefficient [42] is an important indicator of social networks, which describes the local cohesion around nodes. In social networks, a clustering coefficient can be used to calculate the probability that a friend of someone's friends is also a friend. If the friends of a user's friends are still friends, it shows that there is a high degree of aggregation between these users, then the relationship between the user and his friends maybe more ''intimate'' and have more exchanges, then it will be easier to forward other people's microblogs. Besides, we defined the user and local network structure influences.

Neighborhood overlap [43] represents the strength of the relationship between users and users. Among all the friends of two users, the more common friends they have, the more similar their interests are. If upstream users post a microblog post, the topic interest of the post may be close to downstream users. The downstream users may be more inclined to forward their microblog posts.

1) USER INFLUENCE

In the microblog user network, the more fans a user has and the more influence he has; the less a user pays attention to idols, the more influence he has. We can find that PageRank algorithm [44] is very suitable to measure a user's influence.

$$P(u_i) = \beta \sum_{u_j \in fans(u_i)} \frac{p(u_j)}{idols(u_j)} + \frac{1-\beta}{N}. \tag{12}$$

where $P(u_i)$ is the PageRank ranking value of users u_i. $fans(u_i)$ is the set of users' fans u_i. $idols(u_j)$ is the set of

users' idols. N is the total number of users. β is a damping coefficient. In practical applications, the empirical value of β is usually set to 0.85.

Because hundreds of millions of micro-blog users, social networks have small-world phenomenon. The relationship network with a certain user is extremely huge. And so, the PageRank iteration algorithm takes a lot of space and time. By using the idea of PageRank ranking algorithm, it can be approximately replaced by Eq. (13).

$$f(u_i) = \frac{fans(u_i)}{idols(u_i) + 1} \tag{13}$$

2) LOCAL NETWORK STRUCTURE INFLUENCE

The relationships between microblog users, such as concern and forwarding, constitute the social network of microblog users. The relationships between users and all their neighbors constitute the user's self-centered network. The user's neighbor nodes can be divided into two types. These microblog users that have forwarded others are called active users, the other users that have forwarded others are called inactive users. If the number of ring structures connected with users was more, the probability of users forwarding microblog posts would be smaller. The relationship between the number of rings and the probability of forwarding is not linear, but rather exponential decay. The law of exponential decay can be fitted to get the relationship between the number of rings and forwarding. Based on this rule, we describe the user's structural influence, which is expressed in the following form.

$$F(S_v, G_v) = e^{-d|c(s_v)|}. \tag{14}$$

where $|c(s_v)|$ is the number of rings that active user v connects with other neighbor nodes in the self-centered network. d is the delay coefficient ($d > 0$). G_v represents the number of self-centered networks that the active user v consisting of user-centered connects with neighbor nodes.

D. INTERACTION FEATURES

In the microblog user network, the closer the relationship between users is, the easier they will be recognized, and the easier their posting microblog posts will be forwarded. Microblog users can pay attention to other users. If a user pays attention to other users, the user can see microblog posts published by other users, which provides the possibility of users' forwarding behavior. Of course, users can pay attention to each other, which indicates that their relationship is more ''intimate''. When a microblog user publishes a message about other users or wants to mention it. When awakening other users to pay attention to the microblog post, they usually use @a user in the microblog post to show that the relationship between the microblog and users is important; when a user posts a microblog post, other users can forward, like and comment on the microblog, and the higher the frequency of such interaction, the closer their relationship will be. Then microblog posts will be more easily forwarded. Some users may browse microblog frequently, often forward other people's microblog, and interact with other people frequently. While some users may rarely browse microblog, almost no interaction with other people, so the user's forwarding activity has a great impact on forwarding microblog posts. we quantify the interaction characteristics of microblog users that affect the forwarding of microblog users.

1) WHETHER USERS ARE FRIENDS WITH EACH OTHER

In the Microblog user network, if users pay attention to each other, they are called friends. If one user i pays attention to another user j, it usually indicates that the user i is interested in the user j's life and interests. If the other user j also pays attention to the user i, then they are in a two-way relationship, and they are friends and relationships are more intense. They will have more identification and the same values, which they will resonate more likely. Therefore, the user i and the user j may be more willing to forward the other party's microblog posts, and the feature is expressed as follows:

$$Friend(u_i, u_j) = \begin{cases} 1 & Mutual\ attention \\ 0 & Others \end{cases} \tag{15}$$

2) WHETHER THE USER MENTIONED

Whether the Microblog user mentions the concept of a user means that the Microblog user mentions someone in the microblog content when posting someone. If a Microblog user i mentions microblog user j when posting Microblog. The microblog user j will receive a microblog message that microblog user i mentions. At this time, Microblog user j will see the one posted by microblog users i. Then it is possible for Microblog user j to forward this Microblog posts posted by Microblog user i. Furthermore, Microblog user i refer to Microblog user j. It indicates that the relationship between Microblog users i and j is relatively friendly. The strength of their friendly relationship may be stronger than that of mutual friends. Due to the existence of this ''intimate'' relationship, Microblog user posts between Microblog users are more likely to get the mutual forwarding. This feature can be expressed as follows:

$$Mention(u_i, u_j) = \begin{cases} 1 & Mutual\ @ \\ 0 & Others \end{cases} \tag{16}$$

3) USER COMMENT

When the microblog user posts a microblog post, other users can comment on the microblog post. If a user is commenting on the microblog post, this user is interested in the post who posted the microblog. To a certain extent, the comment users also have some interest in the users who post microblog post. The closer the user interacts, the easier it is for the microblog posts to be forwarded. Therefore, whether the user posts a comment posted by the other user as a factor affecting whether the microblog is forwarded or not. It can be expressed by the following formula:

$$Comment(u_i, u_j) = \begin{cases} 1 & Mutual\ Comment \\ 0 & Others \end{cases} \tag{17}$$

4) USER's FORWARDING INFLUENCE

If a microblog user has great influence and his microblog content is very attractive, then the microblog is easy to be forwarded. The microblog posts posted by official, news, and celebrities will be commented, liked, and forwarded by a large number of users. Therefore, the forwarding influence of the microblog posts can be measured to describe the forwarding influence of the users. It can be expressed by the following formula:

$$Influence(u_i) = \frac{1}{m}\sum_{j=1}^{m} RN_j \tag{18}$$

where m indicates the number of historical microblog posts of the user i, RN_j indicates the number of times the user i's of jth microblog post is forwarded.

5) FORWARDING ACTIVITY OF DOWNSTREAM USER

Different Microblog users have their characteristics when they are using microblog. Some users like to browse microblog posts for information. Some users like to post Microblog posts to increase their popularity. At the same time, some users tend to forward other people's microblog posts. The user's forwarding behavior can cause the information to spread. The intensity of the information diffusion is closely related to the enthusiasm of the user's forwarding behavior. In order to measure the enthusiasm of a user, we uses the number of microblog posts forwarded by users in a unit time to indicate the forwarding activity of downstream users. This feature can be expressed by the following formula:

$$Activity = \frac{N}{T} \tag{19}$$

where N is the total number of microblog posts forwarded by the downstream user within the total time T.

E. FEATURE NORMALIZATION

The above features have different meanings, the values responding to different features vary greatly. For example, the number of users' fans can reach millions, while the similarity of interests among users is between 0 and 1. Whether a microblog post contains a topic tag is a binary expression of 0 or 1. Because of the difference of each index, it is difficult to compare their contribution, which will also affect the forwarding prediction model. In order to reduce the difference of each index, we use the maximum and minimum normalization method to normalize the value of each index, which maps the value of each feature to the interval of [0,1].

$$x_i = \frac{x_i - \min\limits_{i\le i\le n}\{x_i\}}{\max\limits_{i\le i\le n}\{x_i\} - \min\limits_{i\le i\le n}\{x_i\}}. \tag{20}$$

where x_i is a vuale of a feature. n is the size of the sample space. In summary, we firstly classify these features affecting microblog posts forwarding into four categories according to these characteristics: user features, microblog features, network structure features, and user interaction features. 23 features and their types are obtained.

TABLE 1. Set of all initial features.

UF[1]	Feat1: user is verified or not.
	Feat2: whether the gender of Microblog users is the same.
	Feat3: named entity similarity between users.
	Feat4: geographic location similarity.
	Feat5: user registration time.
	Feat6: number of idols.
	Feat7: number of fans.
MF[2]	Feat8: does a microblog post contains topic tag.
	Feat9: does a microblog post contains URL.
	Feat10: topic similarity based on interest drift.
	Feat11: TF-IDF spatial vector similarity.
	Feat12: length of microblog posts.
	Feat13: does the microblog post contain picture.
	Feat14: does the microblog post contain video.
NSF[3]	Feat15: clustering coefficient.
	Feat16: neighborhood overlap.
	Feat17: user influence.
	Feat18: local network structure influence.
UIF[4]	Feat19: whether users are friends with each other.
	Feat20: whether the user mentioned.
	Feat21: user comment.
	Feat22: user's forwarding influence.
	Feat23: forwarding activity of downstream user.

[1]UF = User features
[2]MF = Microblog features
[3]NSF= Network structure features
[4]UIF= User interaction features

IV. FEATURE SCREENING MODEL BASED ON FILTER AND WRAPPER

The factors affecting the accuracy of the algorithm may include thousands of features, but most of them are redundant or unrelated to the construction of the model. And so, the feature selection to improve the performance of the algorithm not only reduces the time complexity of the algorithm, but also reduces the space complexity of the algorithm.

A. FILTER SINGLE FACTOR TEST

When the data is preprocessed, it is necessary to select valuable features for microblog forwarding prediction and input them into a machine learning algorithm for training. Generally speaking, feature selection is considered from two aspects:

- **Whether a feature diverges or not**: If a feature does not diverge, the less information it provides. In this paper, we use different analyses to screen the features that affect microblog post forwarding, and remove the feature that the variance value is too small.
- **Relevance between features and targets:** The greater the correlation between independent variables and dependent variables, it shows that the change of independent variables has a greater impact on whether the microblog posts forwards or not. In this paper, the features are divided into discrete and continuous. χ^2 test is used for discrete variables, and point-two-column correlation analysis is used for continuous features.

For the analysis of single feature factor of Filter, we first need to preliminarily screen the complete set $S = \{X^{(1)}, \cdots, X^{(23)}\}$ of features, screen out the features with small variance, and then get the feature set S_1 ($S_1 \subseteq S$). Then, we analyze the correlation of feature sets S_1, divide them into different sets according to the type of feature values, divide them into discrete ones S_{11}, and continuous ones S_{12}. In feature sets S_{11}, we use the χ^2 test to get feature sets S_{21}. In feature sets S_{12}, we use point-two-column correlation analysis to get feature sets S_{22}, and finally, we get feature sets S_2. The feature set obtained by merging the final feature set S_2 obtained by a single feature test of Filter selection (algorithm 1).

Algorithm 1 Features Filler Selection

```
Input:
  S = {X^(1), ··· , X^(n)};
  Given Threshold φ1, φ2, φ3;
Output:
  S2;
Begin
  S1 ⇐ φ,S11 ⇐ φ,S12 ⇐ φ;
  S2 ⇐ φ,S21 ⇐ φ,S22 ⇐ φ;
  For i in |S| do;//variance analysis
    If f_anovo(X^(i)) > φ1;
     S1 = S1 + X^(i);
    End If;
  End For;
  //Partition Discrete/Continuous Value Sets
  For j in |S1| do
   If X^(j) is discrete value;
     S11 = S11 + X^(j);
   Else;
     S12 = S12 + X^(j);
   End If;
  End For;
  For k in |S11| do;
   IF f_χ²(X^(k)) > φ2;
     S21 = S21 + X^(k);
   End If;
  End For;
  For l in |S12| do;//point-two-column analysis
   IF f_point−biserial(X^(l)) > φ3;
     S22 = S22 + X^(l);
   End If;
  End For;
  S2 = S21 + S22;
End.
```

1) VARIANCE ANALYSIS

This method belongs to a preliminary pretreatment of all feature sets. We select the feature that the values of selected independent variables change little.

All features $S = \{X^{(1)}, \cdots, X^{(23)}\}$ are normalized by the previous values, and the final data are mapped to the [0,1] interval. The more the distribution of the independent variable data centralizes, the smaller the change of the value of the independent variable is, the smaller the influence of the change of the independent variable on the dependent variable is. Then the information provided by the independent variable to the prediction model is less. The features should be deleted. On the contrary, if the larger the change of the independent variable is, the more the information brought by the independent variable is, and it should be preliminary. The independent variable is selected as a feature of the model.

We use variance to measure the radian of the change of the independent variable, and then, measure the amount of information that the independent variable brings.

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}. \tag{21}$$

where x_i represents the value of the ith independent variable under a certain type of feature. N is the total number of instances of this type features, μ is the average value of all independent variables under this type of feature.

Through the variance analysis of all the values corresponding to each feature in the feature set $S = \{X^{(1)}, \cdots, X^{(23)}\}$, the features whose variance value is less than the given threshold are deleted. After the variance analysis, the feature set S_1 is preliminarily screened.

2) χ^2 TEST

χ^2 test is a very classic method in statistics. It is a very widely used hypothesis test method. The theoretical basis of χ^2 test is distribution. This hypothesis test method is used to test the distribution of actual data and theory. Its null hypothesis H_0: there is no difference between the observed frequency and the expected frequency.

The basic idea of the χ^2 test is to first propose a null hypothesis H_0, then assume that H_0 is true, then, we calculate the value based on the assumption H_0. And then, we compare the calculated χ^2 value with the significance level. According to the set significance level, and finally, decide whether to accept or reject the null hypothesis H_0. In predicting microblog post forwarding problem, it can be obtained χ^2 according to the actual forwarding distribution and the theoretical forwarding distribution. If the χ^2 value is larger, the correlation between the two features is larger. If the χ^2 value is smaller, the correlation between the two distributions is smaller. If the χ^2 value is equal to 0, then the two distributions are completely irrelevant. But, to arrive at the final relevance conclusion (ie, related or unrelated), you need to set a threshold α, which is also a significance level. If the calculated χ^2 value is greater than the value α, it means that the actual observed value deviates from the theoretical value. The original hypothesis should be rejected. It indicates that the independent variable and the dependent variable are related. If the χ^2 value is less than α, the null hypothesis H_0 cannot be rejected. It indicates that

the actual observation value is less deviated from the theoretical value, and the original hypothesis should be accepted. It indicates that there is no significant difference between the comparative data. that is, the two distributions are not related.

The feature set S_1 is obtained by filtering selection-variance analysis. According to the feature value types in the feature set S_1, the feature set can be divided into S_{11} whose feature values are discrete and S_{12} whose feature values are continuous. Since the χ^2 test requires the independent variable to be discrete, we will perform the χ^2 test on the feature set S_{11}. χ^2 is computed by Eq. (22):

$$\chi^2 = \sum \frac{(A-E)^2}{E} = \sum_{i=1}^{k} \frac{(A_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{(A_i - np_i)^2}{np_i}. \tag{22}$$

where A_i is the observation frequency of the ith feature, and E_i indicates the expected frequency of the ith feature. n indicates the total frequency, p_i indicates the expected frequency rate of the ith feature.

The χ^2 value of the feature is calculated by the above formula, then the set threshold value is compared with the critical value corresponding to the χ^2 value, and the feature set S_{21} is finally obtained by deleting the feature less than the critical value in the feature set S_{11}.

3) POINT-TWO COLUMN CORRELATION ANALYSIS

The point-two column correlation analysis has high requirements for the application scenario. The requirement variable is continuous, and the dependent variable is a discrete Boolean value. After variance analysis of the complete set $S = \{X^{(1)}, \cdots, X^{(23)}\}$, the discrete feature set S_{11} and the continuous feature set S_{12} are obtained. We will perform point two-column correlation analysis on the feature set S_{12}. The point two columns are related to the linear correlation between the independent variable and the dependent variable. The type of the dependent variable value is a Boolean value. The corresponding independent variables can be obtained by dividing the values of the dependent variable. The correlation coefficient of the two columns (independent variables and independent variables) is expressed by Eq (23).

$$R_{pq} = \frac{\bar{X_p} - \bar{X_q}}{S_X} \cdot \sqrt{pq} \tag{23}$$

In the microblog post forwarding process, if the microblog posts is forwarded by the users, then it is indicated by 1, otherwise, it is represented by 0. If n instances are studied, there are u instance of the microblog post forwarding, and there are v instances which the microblog posts v are not forwarded. Then, $p = \frac{u}{n}$, $q = \frac{v}{n} = 1 - p$. $\bar{X_p}$ represents the average value of the independent variables corresponding to the forwarding microblog posts. $\bar{X_q}$ represents the average value of the independent variables corresponding to the not forwarding microblog posts. S_X indicates the standard deviation of all instance dependent variables. After analyzing the correlation of the point-two columns of the feature set, the feature whose correlation coefficient is smaller than the set threshold is deleted, and finally, the feature set S_{22} is obtained.

After the χ^2 test of the feature set S_{11}, the feature set S_{21} is obtained. After the feature set S_{12} is analyzed by the point two-column correlation analysis, the feature set S_{22} is obtained. The final feature set $S_2 = S_{21} \bigwedge S_{22}$ is obtained by analyzing the single factor feature of the filter selection (Algorithm 1).

B. WRAPPER MULTIVARIATE TEST

Through the analysis of the previous single feature factors, the features that can not improve the efficiency of the microblog forwarding prediction are deleted, and the feature set S_2 is obtained. In order to delete the redundant features in the feature set S_2, the feature set S_2 is processed by Wrapper multi-factor analysis. We use the Wrapper multi-factor analysis to find the optimal combination of features under given particular feature set. Because each feature in the feature set S_2 may have a strong correlation with whether the microblog posts are forwarded or not. If the correlation or similarity between the features is very higher, then these features will cause feature redundancies. That is, these Features may be linearly related after they are mapped to the feature space. Therefore, redundant features need to be removed.

In the complete search strategy [45], it is divided into two categories: exhaustive search and non-exhaustive search. The algorithm represented in the exhaustive search is Breadth-First Search, which performs all possible combinations. The time complexity of the algorithm is $o(2^n)$. The most representative algorithm in non-exhaustive search is the branch and bound search algorithm (Branch and Bound). The algorithm adds a branch and bound based on the exhaustive search algorithm. It is determined that some branches are impossible to search for a feature set better than the optimal solution. Then the branch is cut off. Using different search algorithms in the search process can improve the search efficiency of the algorithm, but the time complexity is still $o(2^n)$. The feature set obtained by using the complete search method is optimal. However, it may take a lot of time. Especially in the process of microblog post forwarding prediction, some features take a lot of time to calculate, If a complete search algorithm is used to find the best feature set, then the time spent exponentially increases. In the heuristic search algorithm, the representative algorithm is sequential forward selection and sequential backward selection. The algorithm utilizes the idea of a greedy algorithm, and each selected feature must make the evaluation function. The shortcoming of the algorithm is that only the feature can be added. The added feature cannot be deleted. This will cause the algorithm to fall into the local solution. The time complexity of the heuristic search algorithm is very low,

but its optimal feature set is not necessarily optimal. It is difficult to find the optimal feature set in the microblog forwarding prediction. The random search method obtains a feature set by randomly selecting different features, each time. If the next feature set is more excellent than the previous feature set, then the last feature is taken. The set is replaced with the current feature set to obtain a new feature set. The number of trainings can be manually set. Since the number of training times is flexible, a superior feature set can be obtained in the controlled time. We adopt the LVW (Las Vegas Wrapper) [46] algorithm to looks for the optimal combination of features for forwarding prediction for microblog posts.

The LVW algorithm belongs to a random algorithm of the Wrapper class. It is a typical stochastic optimization algorithm. It also has the characteristics of the probabilistic algorithm. It allows to randomly select the next step in the process of performing algorithm. In most cases, there are many alternative features in the process. The randomness choice is usually less time-consuming than the optimal selection. Therefore, the LVW algorithm can greatly reduce the complexity of the algorithm. The Las Vegas Wrapper algorithm has very good features, and the solution obtained by the Las Vegas algorithm will not be wrong. As long as the Las Vegas algorithm gets a solution, the solution must be correct. For a given feature set, the algorithm randomly selects a feature subset and then cross-validates whether the performance of the indices of the feature subset is higher than the previous feature subset. If the feature subset is better than the previous feature subset, then the current feature subset is taken as the optimal feature subset. And then, the above steps are repeated until the end condition. The current feature subset is used as the final selected feature subset. The algorithm 2 demonstrates the Las Vegas Wrapper process our feature set S_2. Because each feature in feature set S_2 may have a strong correlation, the correlation between features is very large, then these features with high similarity will cause feature redundancy. That is, these features may be linearly correlated after mapping to feature space. So redundant features need to be removed.

In the above algorithm, the LVW algorithm randomly selects one of its feature subsets from the feature set in step 13. It uses a given learning algorithm ψ to calculate the error E' of the feature subset C by cross-validation method in step 16. In steps 17∼21, if the error E' is less than the previous minimum error E, or is not much different from the previous error and the feature subset C contains fewer features, then the feature subset is preserved. In steps 22∼24, if the error E' of the selected feature subset calculated under cross-validation is not lower than the previous one E, or the number of features in the subset is not reduced, then LVW continues to select a new feature subset. The above operations are performed repeatedly until the given threshold T is satisfied. Finally, the feature subset S_2^* selected by the LVW algorithm is output.

Algorithm 2 Las Vegas Wrapper

```
01 Input:
02   Data set D; Feature set S2; Error E
03   Given Learning algorithm ψ
03   Given control parameter T //end program
04 Output:
05   S2*;
06 Begin
07   E ⇐ ∞;
08   d ⇐ |S2|//the Number of Elements in Feature Set;
09   S2* ⇐ S2;
10   t ⇐ 0;
11   C ⇐ φ//randomly select feature set C;
12   While t < T
     //Randomly selection of a feature subset S2'
13     S2' = rand(), (S2' ⊆ S2)⋀(S2' ⊈ C);
14     C = C + S2';
15     d' = |C|;
     //the error of feature set on data set by cross-validation
16     E'=CrossValidatior(ψ(D^C));
17     If (E' < E)⋁((E' = E)⋀(d' < d))
18       t = 0;
19       E = E';
20       d = d';
21       S2* = C;
22     Else;
23       t = t + 1;
24     End If;
25   End While;
26    Return S2*;
27 End;
```

V. AN ENSEMBLE LEARNING MODEL FOR FORWARDING PREDICTION OF MICROBLOG USERS

In this section, we adopt the AdaBoost algorithm and build an ensemble learning (multi-feature based classification) model to predict the forwarding microblog posts for microblog users. The parameters of the ensemble learning model are the best feature combinations S_2^*.

The AdaBoost algorithm continuously changes the weights of the sample data and the weak classifier during the training process, and finally combines all the weak classifiers in a certain way to obtain the desired strong classifier. In the first round of training, the AdaBoost algorithm first assigns a weight to each training sample. The initial values of all features are $W_i = 1/N$, and then use a weak classifier to train these samples to get the classification error rate of the model. If the classification error of the weak classifier becomes greater, then we decrease the weight of the weak classifier. If the samples with the trained process is misclassified, we increase their weights. We decrease the weights for the correctly classified samples with the trained process. We use the first training round to get new weight samples,

and then use the second weak classifiers to train these samples again, and get the weights of the second classifiers and the weights of the new training round of every sample. The training continuously repeats the above process until the given loop number or the given error is reached. Finally, the weak classifiers learned in each round are combined in some combination to get the final strong classifier.

The weak classifier indicates that the error rate of the two classification is lower than 50%. That is, the weak classifier is better than the random guess. Any classification algorithm can be used as a weak classifier, such as a classical learning algorithm, decision trees, logistic regression, naive Bayes and SVM. The weak classifier used in this paper is a single-layer decision tree. The number of nodes is only one, also known as stump. It is the most commonly used weak classifier of AdaBoost algorithm. A strong classifier indicates that a polynomial learning algorithm can learn it and the correct rate is high. Popularly speaking, it is a classifier that combines weak classifiers. The classification accuracy of this classifier is very high.

The AdaBoost algorithm briefly is described as follows:

- First, the training sample weight is initialized, and the first weak classifier is trained by using the sample data set.
- Secondly, the weight of the weak classifier is determined according to the classification results of the weak classifier. If the error classification rate is larger, the weight of the corresponding weak classifier is smaller. At the same time, the weight of the wrong samples in the data set is increased, and finally the updated weight of the training samples is obtained; then the new samples are used to train the second weak classifier.
- Repeat the second step until the training number or classification error rate set in the training is obtained. Finally, we obtain the number of weak classifiers equal to the number of training times.
- The weak classifiers are combined to obtain the final strong classifier.

In order to better understand the specific running process of the AdaBoost method, the algorithm will be described in detail in the following.

Given a training data set $\{(x_1, y_1), \cdots, (x_n, y_n)\}$. N indicates the size of the training data set. y_i indicates whether the microblog post x_i is forwarded or not. In the training sample, y_i is labeled as 1 or -1. $y_i = 1$ indicates that the user forwards the microblog post, and $y_i = -1$ indicates that the user does not forward the microblog post. Some parameters of the AdaBoost algorithm indicate the meaning of certain quantities in Table 2.

The ensemble learning algorithm based on AdaBoost is described as follows:

- Step 1. The training data set is initialized. In the first round, all training samples are given the same weight $w_{1i} = 1/N$: then the initial weight distribution D_1 of the training data set can be obtained:

$$D_1 = (\omega_{11}, \omega_{12}, \cdots, \omega_{1N}). \tag{24}$$

TABLE 2. Symbol define for ensemble learning.

1	N	The size of samples in training data set
2	D_t	The weight distribution of samples in training data set
3	w_{ti}	The weight of the ith sample in training data set
4	H	Weak classifier
5	H_{final}	The final stronger classifier
6	e	classification error
7	α_t	The weight of weak classifier
8	T	the number of base learners

where w_{1i} indicates weight of the ith sample and the first round in the training sample data set. Each sample is given the same weight during the initialization process because it is not known which sample instance is the most similar to the law whether the microblog post is forwarding or not at the very beginning.

- Step 2. Select the current weak classifier h which its error is the lowest as the basic classifier H_t for the tth iteration. The classification error e_t of the tth iteration of the weak classifier on the training set is computed by using Eq. (25):

$$e_t = \sum_{i=1}^{N} \omega_{ti} Loss(H_t(x_i) \neq y_i) \tag{25}$$

The tth classifier H_t calculates the error (loss) of forwarding and predicting N instances in the training set, and then compares with the actual forwarding situation. The prediction result of the tth classifier H_t is -1 and 1. -1 indicates no forwarding, 1 means forwarding. The values learned from the above training data set are also -1 and 1. The prediction result of the classifier is compared with the real result of the training set to obtain by the loss function. If the prediction result is the same as the real result, the value of the loss function is 0, otherwise, the value of the loss function is 1. It is not difficult to find that the difference between the larger the prediction result of the tth classifier H_t and the actual situation is, the larger the loss function is. The final classification error is also inseparable with the weight of each sample value. If the weight is larger and the classification is wrong, the classification error of the weak classifier will become larger. Therefore, the classification error of the weak classifier is not only related to the classification result of the classifier itself but also closely related to the weight of each sample instance in the training data set.

- Step 3. Calculate a proportion of the weak classifier in all classifiers:

$$\alpha_t = \frac{1}{2} ln(\frac{1 - e_t}{e_t}) \tag{26}$$

In ensemble learning, a plurality of weak classifiers in the training process are combined into a strong classifier in some way. Each weak classifier is different.

How to measure the difference of the weak classifier of the AdaBoost? It mainly sets the proportion of the weak classifier in the entire weak classifiers. The proportion is closely related to the classification error of the weak classifier in the training sample. In the above equation, it is found that the smaller the classification error is, the smaller the weak classifier weight is, and vice versa.

- Step 4. Update the weight distribution $D_{t+1,i}$ of the training samples:

$$\omega_{t+1,i} = \frac{\omega_{ti} exp(-\alpha_t y_t H_t(x_i))}{Z_t}, \quad i = 1, \cdots, N. \quad (27)$$

The expression $\omega_{t+1,i}$ represents the weight of the ith instance data in the $t+1$ times training data after training the t times of the weak classifier. That is, the sample weight value of training data of the $t+1$th times is correlated with the sample weight of the tth times. In the Eq. (27), $y_t H_t(x_i)$ represents the product of the actual forwarding situation of the microblog posts and the prediction result. Because y_t and $H_t(x_i)$ take on the range of -1 and 1. The prediction result is the same as the actual forwarding result, then the product is 1, otherwise -1. Since the exponential function is a monotonically increasing function. ω_{ti} and α_t are positive. The weight of the sample value becomes smaller when the forwarding prediction result is the same as the actual result. When the actual results are different from the forwarding prediction result, the weight value of the sample is increased. It will cause the misclassified sample to be amplified, and the weaker classifier of the latter round is more strict. It is more conducive to improving the accuracy of the model prediction. Z_t in Eq. (27) is used to normalize the weight of each sample. Its value represents the sum of the weights of all sample instances.

$$Z_t = \sum_{i=1}^{N} \omega_{ti} exp(-\alpha_t y_i H_t(x_i)) \quad (28)$$

By calculating the weight of each sample in each training data set, the weight distribution of the entire sample data set can be finally obtained, and its distribution can be expressed by Eq. (29):

$$D_{t+1} = (\omega_{t+1,1}, \omega_{t+1,2}, \cdots, \omega_{t+1,N}) \quad (29)$$

- Finally, by calculating the weight of the weak classifier for each time, the weak classifiers are weighted and summed to obtain the final classifier, and a strong classifier can be obtained by the symbol function:

$$H_{final} = sign(\sum_{t=1}^{T} \alpha_t H_t(x)) \quad (30)$$

The algorithm first calculates the weight of the data sample, then calculates the classification error of the current classifier, and then characterizes the importance of the classifier. So that, the loop is finally repeated to obtain T classifiers. Finally, by combining each weak classifier, it gets a strong classifier. The principle of the algorithm is shown in Fig. 3. The three ellipses in each column represent the result of a weak classifier. Firstly, the training error of the training data set is obtained by using the weak classifier. Then the weight α_t of the classifier is determined by the training error, and finally, the strong classifier is obtained by combining the weak classifiers.

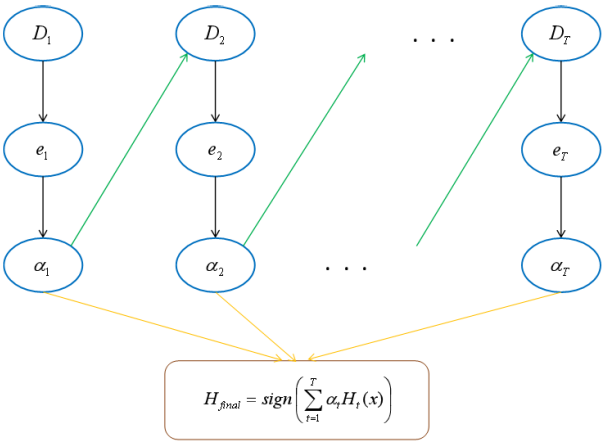

FIGURE 3. AdaBoost Algorithm.

VI. EXPERIMENT

A. EXPERIMENTAL ENVIRONMENT

The specific experimental environment includes hardware and soft parts. Some parameters are listed as follows: CPU: Intel(R) Pentium(R CPU G4500@3.5GHZ, Memory: 4.00GB, Hard disk: ITSolid StateDisk; Development language: Python; OS: Windows 10 Professional 64bit; Database: SQL Server 2008.

We extract data from the Sina Microblog platform (the largest Microblog platform in China). First, we search hot news topics in Baidu News as background hot topics. Then, by starting from these hot topics and using keywords and topic features, we search a user related to the hot topic as the seed node of the data. We use this seed node to crawl his friends' list, and then his friends as seed nodes to crawl his friends' friends. And so, we can crawl a more intimate Microblog user relationship network. Because some users provide too little research information, in order to solve the sparsity of data, the information will be filtered through the following rules to obtain the final research target users, whose screening rules are as follows:

- The number of fans and idols of users should be more than 30;
- Users' forwarding more than 3 microblog posts per month;
- Users' registered accounts exceed one year.

In the end, 5148 user data and 895621 microblog posts were obtained. We divide the data set into a training data set and test data set. The training data set is from January 1, 2017, to October 31, 2017, and the test data set from November 1, 2017, to December 31, 2017.

B. INDICES

In microblog post forwarding prediction, the Precision, Recall, F1 value are used to measure the performance indicators. Their definitions are as follows:

Precision: For a given data set, the accuracy rate represents the proportion of the number of the forwarded microblog posts which they are correctly distinguished to the predicted number of the forwarded microblog posts.

$$Precision = \frac{TP}{TP + FP} \tag{31}$$

Recall rate: For a given data set, the accuracy rate represents the proportion of the number of the forwarded microblog posts which they are correctly distinguished to the actual number of the forwarded microblog posts.

$$Recall = \frac{TP}{TP + FN} \tag{32}$$

F1 score: It is the harmonic average value of the precision and the recall. The relationship between the precision and the recall rate is one by one. When the pursuit of precision increases, the recall rate will be lower; when the pursuit of recall rate increases, the precision will be lower. So the F1 value is a comprehensive evaluation index which combines the precision and recall rate.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{33}$$

where TP represents the number of the microblog posts which we predict them to be forwarded and they are forwarded actually in the test data. FP represents the number of the microblog posts which we predict them to be forwarded and they are not forwarded actually in the test data. FN represents the number of the microblog posts which we do not predict them to be forwarded and they are forwarded actually in the test data. The confusion matrix below can better understand the meaning of each parameter.

C. FEATURE SECTION ANALYSIS ON DATA SET

1) BOOLEAN TYPE FEATURE SECTION

On training data set, we compute the variances (Fig.4) of 23 features (Table 1). Among Boolean features, we find that the variance value of "whether the sexuality of Microblog users is the same" is the largest, because the proportion of male users is about 46% and that of female users is about 54%. So the distribution of the features is more divergent. The variance value of "whether the user mentioned" is the least in Boolean types, because when microblog posts are published, there are very few users to mention others, so this leads to a very small variance of this feature.

Among the features of continuous types, we find that the variance value of "user registration time" is the largest, because new users are more casual in registering microblog account, new users may register at any time, so the feature value is more divergent. The variance value of

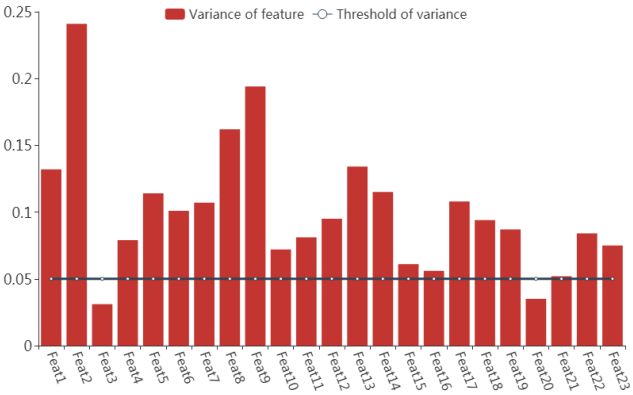

FIGURE 4. The variances of the features.

"named entity similarity between users" have the smallest, because the named entities contained in user microblog posts are very small. However, the differences in hobbies and habits of each microblog users are very large. The named entity including their microblog posts are very different. The variance values of "named entity similarity between users" are very small. By using the above variance analysis, we set the feature selection threshold at 0.05. The feature whose variance is greater than or equal to 0.05 is retained, otherwise, it is deleted. From the above analysis, we can delete "named entity similarity between users" and "whether the user mentioned".

We use χ^2-test to check the Boolean type features. We choose "User is verified or not" feature to demonstrate our experiments. We randomly select 50,000 microblog posts. Whether the user of these microblog posts and their friends forward these microblog posts or not do some statistics. We can build a fourfold table (Table 3) of χ^2-test for the feature "User is verified or not". We use SPSS software with the χ^2 test for every Boolean type feature. Some following results can be obtained. Pearson $\chi^2 = 1.769$, $p = 0.183$. We set significance level $\alpha = 0.05$. Because $\alpha < p$, we should accept the hypothesis H_0. The difference has no statistical significance. It can be considered that "User is verified or not" has not directly correlation with "whether microblog posts are forwarded or not". Then the feature "User is verified or not" should be removed from the feature set. The χ^2 values of all features are listed as follows:

TABLE 3. Fourfold table of χ^2 Test.

features	forward	not forward	sum
User is verified	18278	3184158	3202436
User is not verified	12175	2154327	2166502
sum	30453	5338485	5368938

The critical value of χ^2 for all features is 3.841 in the Table 4. We delete the feature whose χ^2 value is less than 3.841, otherwise we can keep it.

TABLE 4. χ^2 values of all Boolean type features.

Feature	Feat1	Feat2	Feat8	Feat9
χ^2 of feature	1.77	0.25	1357.12	25.78
Feature	Feat13	Feat14	Feat19	Feat21
χ^2 of feature	592.45	685.21	487.48	128.25

2) CONTINUOUS TYPE FEATURE SECTION

We use point-two column correlation analysis to select the continuous type features. The correlation coefficients of 13 continuous type features among 23 features are shown in Fig 5.

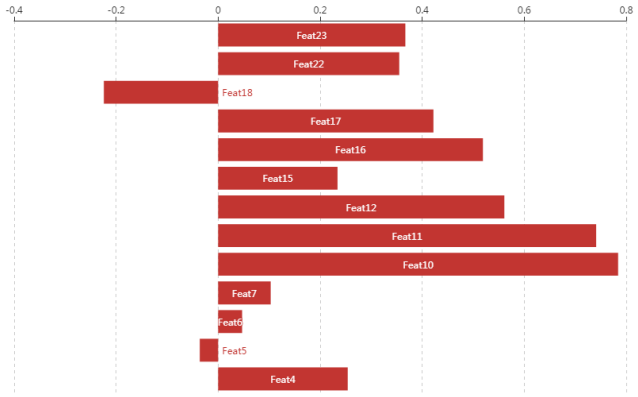

FIGURE 5. Point-biserial correlation coefficient.

We find that the minimum correlation coefficient of point two-column analysis is -0.036. The feature "user registration time" indicates that there is a very weak negative correlation between the feature and whether the microblog post is forwarded or not. Because the absolute value of the correlation coefficient of the feature is less than 0.05, we should delete the feature. In real life, users' forwarding habits are very different. Some users register for microblog very early, but they do not always forward other people's microblog posts. On the contrary, some newly registered users often forward other people's microblog posts. Because the correlation coefficient of "number of idols" is 0.047, which is less than the significance level 0.05, the feature should also be deleted. The correlation coefficient of "local network structure influence" is -0.224, which indicates that the feature is negatively correlated with whether the microblog posts are forwarded or not. Its' absolute value of the feature is slightly smaller, which indicates that the feature has a certain impact on forwarding microblog posts.

3) LVW FEATURE COMBINATION ANALYSIS

In this section, we combine the features to find the optimal features. Logistic regression model is selected for cross-validation learning algorithm ψ. Optimal feature combinations are obtained by Python programming on experimental data sets.

- **User features**: (4).geographic location similarity.
- **Microblog features**: (8).does a microblog post contain a topic tag, (9).does a microblog post contain a url, (10).topic similarity based on interest drift, (12).length of microblog posts.
- **Network structure features**: (15).Clustering coefficient, (16).neighborhood overlap, (17).user influence, (18).local network structure influence.
- **User interaction features**: (21).user comment, (22).user's forwarding influence, (23).forwarding activity of downstream user.

The number of deleted features is (7), (11), (13), (14) and (19) by using LVW feature combination analysis respectively. The "number of fans" feature is correlated with the "user influence" feature and should be deleted. However, the "user influence" feature not only shows the number of users' fans but also shows the influence of users. In the LVW feature combination analysis, the TF-IDF spatial vector similarity feature is removed, while the microblog similarity feature based on microblog drift is retained. We can find that these two features reflect users' interests better. It shows that the similarity feature of microblog posts based on interest drift model is very effective in predicting microblog forwarding. The features (13) and (14) are deleted because they are related to the length of microblog posts. Usually, the longer the length of microblog posts, the more information it contains. It is easier for long microblog posts to include pictures and videos. Feature (19) is removed, because it is related to the number of neighborhood overlaps between users. However, the feature of neighborhood overlap degree is stronger than whether the users are friends or not. And so it is removed.

D. ADABOOST PREDICTION MODEL BASED ON MULTI-FEATURE

We extract 12 features from 13 initial features by using the variance analysis, χ^2 test, point-two-column correlation analysis, and LVW feature combination analysis. These 12 features become the parameters of our proposed AdaBoost prediction model for the user's behavior of forwarding microblog posts. In order to illustrate the validity of the model, we choose four classical classification models (Naive Bayes, Logistic regression, Rand forest, SVM) in microblog forwarding prediction, and compare them with our proposed AdaBoost prediction model. We give their precision, recall and F1 score for these four model and our proposed AdaBoost prediction model in Table 5. Among the five models, the precision of our proposed AdaBoost prediction model is the highest, which reaches 0.858. The recall of naive bayes model is the highest, which achieves 0.928. However, because the precision of the naive Bayes model is very low, its' F1 score is not high. The F1 score of our proposed AdaBoost prediction model is the highest, which achieves 0.885. It indicates that our proposed AdaBoost prediction model has a good overall result, and are very effective for the forwarding behavior prediction.

TABLE 5. Performance comparison of our proposed AdaBoost prediction model with four classical classification models.

Models	Precision	Recall	F1 Score
Naive Bayes	0.725	0.928	0.816
Logistic regression	0.782	0.857	0.818
Rand forest	0.841	0.856	0.848
SVM	0.794	0.887	0.838
our proposed AdaBoost	0.858	0.913	0.885

To study the users' forwarding situation of microblog posts with different topics, we classifies microblog posts published by users by different topics. Because the microblog posts are issued by the upstream users, and the downstream users are forwarded by the upstream users. Usually, the forwarding process brings the different topic microblog posts. Among the popular microblog events, the most popular microblogs are **star gossip**, **politics** and **military affairs**, **people's livelihood** and **emergencies**. And so, we chooses these four types of microblog topics to check our proposed AdaBoost prediction model. The filtered feature set and AdaBoost prediction model are used to forecast the forwarding behavior of these four type topics. The experimental results are shown in Fig. 6.

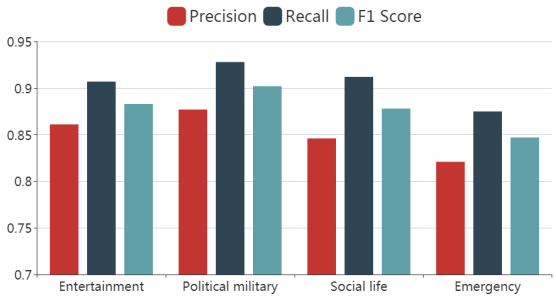

FIGURE 6. Microblogs forwarding with different Topics.

We can see that the precision, recall rate, and F1 score are the highest when predicting the microblog post forwarding behavior with the topics of politics and military. This shows that the regularity of microblog users focusing on politics and military forwarding is stronger. and the characteristic indicators depicted are relatively stable in this application scenario. The precision, recall rate and F1 score of **emergencies** are the lowest. Because of the randomness of emergencies, it exists the difference between three indices and the actual situation of emergencies. When predicting the forwarding behavior of **star gossip** and **people's livelihood**, the precision of **star gossip** is higher than **people's livelihood**. However, the recall and F1 score are opposite with the precision. Generally speaking, the recall of forwarding prediction model about these four topics are the highest, and the precisions are the lowest. It shows that our proposed AdaBoost prediction model is relatively stable.

We divide Sina microblog user into three categories: official users, well-known users, and ordinary users. To verify our proposed AdaBoost prediction model, we select the most active top 100 posting and forwarding users from our prediction result set, respectively. From Fig. 7, the proportion of the three categories under posting users and forwarding users are consistent with the actual situation. In Fig. 7 (left), there are 11 official users, 62 well-known users, and 27 ordinary users. It shows that the well-known users are the main force of posting microblog posts, because of their high attention and great influence on social networking. Especially, the famous stars often publish microblog posts interacting with fans. The official users will also publish them. They bring many news and notifications. The ordinary users publish relatively few microblog posts. In Fig. 7(Right), It shows that 7 well-known users and 93 ordinary users are the most active users of forwarding microblog posts, but none of them are official users. The official users and well-known users are very cautious about forwarding, because they may be responsible for their forwarding behavior, especially official users. The ordinary users are not more cautious about forwarding behavior.

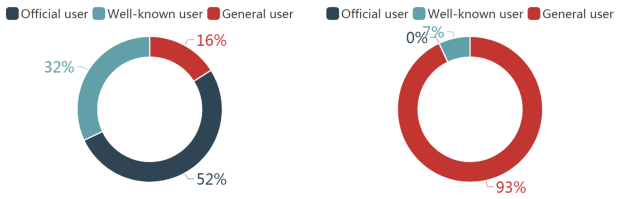

FIGURE 7. The top 100 most active posting and forwarding users.

In order to prove that the feature selection model can improve the efficiency of the algorithm, all the features proposed at the beginning (i.e.,Initial feature set (23 in total)) and the optimal features (i.e.,Optimal feature set (12 in total)) obtained and deleted are used for The Adboost algorithm to perform a comparison test. And its indexes include the time complexity in addition to precision, recall, and F1 mentioned earlier. Because the running time of the algorithm has a greater relationship with the computer hardware part, the relative time is used to represent the time complexity of the algorithm.The experimental results of performance comparison of different feature sets are shown in Table 6.

TABLE 6. Performance comparison of different feature sets.

Feature set	Precision	Recall	F1 score	T(n)
Initial feature set	0.861	0.894	0.877	58.7
Optimal feature set	0.858	0.913	0.885	1

From the experimental results, it is found that the precision of the initial feature set is higher than that of the optimal feature set. In addition, the other indicators of the initial feature set are worse than the performance of the optimal feature set. Although the number of features in the initial feature set is larger, the overall performance is not so good, which indicates that there may be redundancy between features. The experimental results show that the feature selection algorithm can greatly promote the performance of the microblog forwarding prediction model.

To demonstrate the performance of the 12 best features on forwarding microblog posts, we respectively delete one of the 12 best features and use our proposed AdaBoost prediction model to predict forwarding behavior. The precisions, recalls, F1 scores are listed in Fig. 8. In the x-axis direction, the first optimal feature set(New) represents 12 features obtained from the feature analysis, and the "4" represents the remaining other features of deletion feature (4) in the optimal feature set, etc. From Fig. 8, we can see that the precisions, recalls, F1 scores of the optimal feature set are higher than those of other sets that delete a feature. It shows that the 12 best features have a positive effect on forwarding prediction.

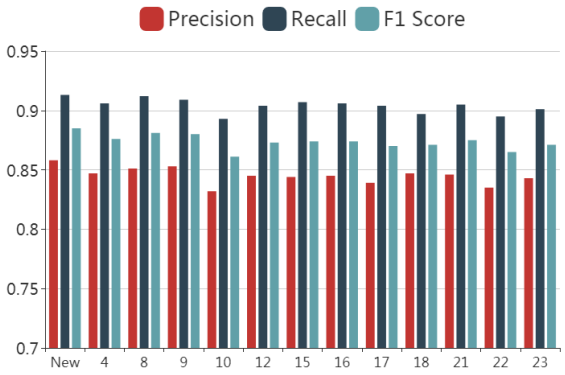

FIGURE 8. The Effect of the best features set.

VII. CONCLUSION

There are many factors affecting microblog forwarding, we comprehensively consider the factors, and put forward four types of microblog features: user characteristics, microblog characteristics, network structure characteristics, and interactive behavior characteristics. Among them, we propose many novel features, such as topic similarity based on interest drift, geographic location similarity, user aggregation coefficient, neighborhood overlap degree among users and user's forwarding influence. And the indexes are normalized. We propose a set of effective methods for feature screening of microblog, including variance analysis, χ^2 test, point-two-column correlation analysis, and LVW feature analysis. By series of feature screening, we select the best 12 features from the initial 23 features. These 12 features have a high correlation to whether microblog post is forwarded or not. Moreover, these features are relatively independent, which greatly reduces the time and space complexity of the model. Finally, we compare the classical methods with our proposed AdaBoost prediction model. The experiment proves that our proposed AdaBoost prediction model is very effective.

REFERENCES

[1] P. Y. Fan, H. Wang, Z. H. Jiang, and P. Li, "Measurement of microblogging network," *J. Comput. Res. Develop.*, vol. 49, no. 4, pp. 691–699, 2012.
[2] Z. Ma, A. Sun, and G. Cong, "On predicting the popularity of newly emerging hashtags in Twitter," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 64, no. 7, pp. 1399–1410, Jul. 2013.
[3] 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 *Proc. IEEE 2nd Int. Conf. Social Comput.*, Aug. 2010, pp. 177–184.
[4] Y. L. Li, H. T. Yu, and L. X. Liu, "Predict algorithm of micro-blog retweet scale based on SVM," *Appl. Res. Comput.*, vol. 30, no. 9, pp. 2594–2597, 2013.
[5] S. B. Zhang and W. D. Cai, "Influence analysis of user characteristics to microblogging retweet behavior," *Comput. Eng. Appl.*, vol. 11, pp. 11–16, 2014.
[6] J. Chen, W. Liu, W. H. Chao, and L. H. Wang, "Microblog forwarding prediction based on hot topics," in *J. Chin. Inf. Process.*, vol. 29, pp. 150–158, 2015.
[7] J. X. Cao, J. L. Wu, W. Shi, B. Liu, X. Zheng, and J. Z. Luo, "Sina microblog information diffusion analysis and prediction," *Chin. J. Comput.*, vol. 37, no. 4, pp. 779–790, 2014.
[8] Z. F. Wang, K. L. Liu, Z. Y. Zheng, and D. Li, "Prediction retweeting of microblog based on logistic regression model," *J. Chin. Comput. Syst.*, vol. 37, no. 8, pp. 1651–1655, 2016.
[9] X. Tang, Q. Miao, Y. Quan, J. Tang, and K. Deng, "Predicting individual retweet behavior by user similarity: A multi-task learning approach," *Knowl.-Based Syst.*, vol. 89, pp. 681–688, Nov. 2015.
[10] E. F. Can, H. Oktay, and R. Manmatha, "Predicting retweet count using visual cues," in *Proc. 22nd ACM Int. Conf. Conf. Inf. Knowl. Manage. (CIKM)*, 2013, pp. 1481–1484.
[11] W. Liu, M. He, and L. H. Wang, "Research on microblog retweeting prediction based on user behavior features," *Chin. J. Comput.*, vol. 39, no. 10, pp. 1992–2006, 2016.
[12] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on Twitter," in *Proc. 43rd Hawaii Int. Conf. Syst. Sci.*, Jan. 2010, pp. 1–10.
[13] J. Y. Lee and S. S. Sundar, "To tweet or to retweet? That is the question for health professionals on Twitter," *Health Commun.*, vol. 28, no. 5, pp. 509–524, Jul. 2013.
[14] Z. Zhong, Y. Guan, Y. Hu, and C. Li, "Mining user interests on microblog based on profile and content," *J. Softw.*, vol. 28, no. 2, pp. 278–291, 2017.
[15] Y. Xiao, N. Li, M. Xu, and Y. Liu, "A user behavior influence model of social hotspot under implicit link," *Inf. Sci.*, vol. 396, pp. 114–126, Aug. 2017.
[16] M. Michelson and S. A. Macskassy, "Discovering users' topics of interest on Twitter: A first look," in *Proc. Workshop Anal. Noisy Unstructured Text Data*, 2010, pp. 73–80.
[17] Y. Guo, Y. Y. Gong, Q. Zhang, and X. J. Huang, "Retweet behavior prediction using topic model," *J. Chin. Inf. Process.*, vol. 32, pp. 134–140, 2018.
[18] J. Zhang, J. Tang, J. Li, Y. Liu, and C. Xing, "Who influenced you? Predicting retweet via social influence locality," *ACM Trans. Knowl. Discovery Data*, vol. 9, no. 3, pp. 1–26, Apr. 2015.
[19] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer, "Outtweeting the twitterers-predicting information cascades in microblogs," in *Proc. 3rd Conf. Online Social Netw.*, 2010, pp. 1–9.
[20] Z. W. Tian, L. Wang, and C. Liu, "Information dissemination mechanism analysis and model construction of micro-blog based on complex network," *Inf. Sci.*, vol. 9, pp. 15–21, 2015.
[21] M. Bagdouri and D. W. Oard, "On predicting deletions of microblog posts," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2015, pp. 1707–1710.
[22] P. Fan, P. Li, Z. Jiang, W. Li, and H. Wang, "Measurement and analysis of topology and information propagation on Sina-Microblog," in *Proc. IEEE Int. Conf. Intell. Secur. Inform.*, Jul. 2011, pp. 396–401.
[23] Q. Yan, L. Wu, and L. Zheng, "Social network based microblog user behavior analysis," *Phys. A, Stat. Mech. Appl.*, vol. 392, no. 7, pp. 1712–1723, Apr. 2013.
[24] X. Tang, Q. Miao, Y. Quan, J. Tang, and K. Deng, "Predicting individual retweet behavior by user similarity: A multi-task learning approach," *Knowl.-Based Syst.*, vol. 89, pp. 681–688, Nov. 2015.
[25] W. Yao, P. Jiao, W. Wang, and Y. Sun, "Understanding human reposting patterns on sina weibo from a global perspective," *Phys. A, Stat. Mech. Appl.*, vol. 518, pp. 374–383, Mar. 2019.
[26] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Phys. Rev. Lett.*, vol. 86, no. 14, pp. 3200–3203, Apr. 2001.
[27] Y. Xiao, C. Song, and Y. Liu, "Social hotspot propagation dynamics model based on multidimensional attributes and evolutionary games," *Commun. Nonlinear Sci. Numer. Simul.*, vol. 67, pp. 13–25, Feb. 2019.

[28] J. Liu, K. Niu, Z. He, and J. Lin, "Analysis of rumor spreading in communities based on modified SIR model in Microblog," in *Proc. Int. Conf. Artif. Intell., Methodol., Syst., Appl.*, 2014, pp. 69–79.
[29] A. Kanavos, I. Perikos, P. Vikatos, I. Hatzilygeroudis, C. Makris, and A. Tsakalidis, "Modeling retweet diffusion using emotional content," in *Artificial Intelligence Applications and Innovations* 2014.
[30] H. Ma, W. Qian, F. Xia, X. He, J. Xu, and A. Zhou, "Towards modeling popularity of microblogs," *Frontiers Comput. Sci.*, vol. 7, no. 2, pp. 171–184, Apr. 2013.
[31] O. Tsur and A. Rappoport, "What's in a hashtag?: Content based prediction of the spread of ideas in microblogging communities," in *Proc. 5th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2012, pp. 643–652.
[32] M. Jenders, G. Kasneci, and F. Naumann, "Analyzing and predicting viral tweets," in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 657–664.
[33] S. Gao, J. Ma, and Z. Chen, "Modeling and predicting retweeting dynamics on microblogging platforms," in *Proc. 8th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2015, pp. 107–116.
[34] A. Kupavskii, L. Ostroumova, A. Umnov, S. Usachev, P. Serdyukov, G. Gusev, and A. Kustarev, "Prediction of retweet cascade size over time," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2012, pp. 2335–2338.
[35] D. R. Bild, Y. Liu, R. P. Dick, Z. M. Mao, and D. S. Wallach, "Aggregate characterization of user behavior in Twitter and analysis of the retweet graph," *ACM Trans. Internet Technol.*, vol. 15, no. 1, pp. 1–24, Mar. 2015.
[36] D. Li, Z. M. Xu, S. Li, T. Liu, and X. W. Wang, "A survey on information diffusion in online social networks," *Chin. J. Comput.*, vol. 37, no. 1, pp. 189–206, 2014.
[37] H. Ogura, H. Amano, and M. Kondo, "Comparison of metrics for feature selection in imbalanced text classification," *Expert Syst. Appl.*, vol. 38, no. 5, pp. 4978–4989, May 2011.
[38] N. C. Oza, "Online ensemble learning," in *Proc. 17th Nat. Conf. Artif. Intell. 25th Conf. Innov. Appl. Artif. Intell.*, 2000.
[39] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Oct. 2007.
[40] R. Cai, Z. Hao, X. Yang, and W. Wen, "An efficient gene selection algorithm based on mutual information," *Neurocomputing*, vol. 72, nos. 4–6, pp. 991–999, Jan. 2009.
[41] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Elect. Eng.*, vol. 40, no. 1, pp. 16–28, Jan. 2014.
[42] M. Gentner, I. Heinrich, S. Jäger, and D. Rautenbach, "Large values of the clustering coefficient," *Discrete Math.*, vol. 341, no. 1, pp. 119–125, Jan. 2018.
[43] D. L. Hansen, B. Shneiderman, M. A. Smith, and I. Himelboim, "Social network analysis: Measuring, mapping, and modeling collections of connections," in *Analyzing Social Media Networks with NodeXL*, 2nd ed., 2020, pp. 31–51.
[44] J. Koo, D.-K. Chae, D.-J. Kim, and S.-W. Kim, "Incremental C-rank: An effective and efficient ranking algorithm for dynamic Web environments," *Knowl.-Based Syst.*, vol. 176, pp. 147–158, Jul. 2019.
[45] S. Ou-Yang and W. Y. Hu, "Research and application of several classical search algorithms," *Comput. Syst. Appl.*, vol. 20, no. 5, pp. 243–247, 2011.
[46] H. Liu and R. Setiono, "Feature selection and classification—A probabilistic wrapper approach," in *Industrial and Engineering Applications of Artificial Intelligence and Expert Systems*. 1997.

CHUNLONG FU received the B.S. degree in information and computational science from Neijiang Normal University, China, and the M.S. degree in computer technology from Xihua University. He is currently a Teacher of computer science with the Sichuan University Jinjiang College. His research areas include social networks and machine learning.

YAJUN DU received the D.Sc. degree in traffic information engine and control from Southwest Jiaotong University, Chengdu, China, in 2005. He is currently a Professor with the School of Computer Science and Technology, Xihua University, Chengdu, China. He has published several articles and served on program committees of both China and international conferences. His experience and research work focus on information retrieve, software engineering, search engine, Web mining, and computer networks.

BINYAN LYU received the B.S. degree in computer science and technology from Shanxi Datong University, China. She is currently pursuing the master's degree with Xihua University, China. Her research areas include information retrieval and computer networks.

QIAOYU ZHOU received the B.S. degree in building electrical and intelligent from the Engineering and Technical College, Chengdu University of Technology, China. She is currently pursuing the master's degree with Xihua University, China. Her research area includes data mining.

RUILIN HU received the B.S. degree in measurement and control technology and instrumentation program from the Shanghai University of Electric Power, China. He is currently pursuing the master's degree with Xihua University, China. His research areas include software engineering and data mining.

PENG JIA received the B.S. degree in software engineering from the Sichuan University Jinjiang College, China. He is currently pursuing the master's degree with Xihua University, China. His research areas include software engineering and computer networks.

YUJIAN ZHOU received the B.S. degree in electronics science and technology from the Chengdu College, University of Electronic Science and Technology of China. He is currently pursuing the master's degree with Xihua University, China. His research areas include software engineering and computer networks.