

Cross-site Prediction on Social Influence for Cold-start Users in Online Social Networks

QINGYUAN GONG, YANG CHEN, and XINLEI HE, School of Computer Science, Fudan University, China and Shanghai Key Lab of Intelligent Information Processing, Fudan University, China and Peng Cheng Laboratory, China

YU XIAO, Department of Communications and Networking, Aalto University, Finland

PAN HUI, Department of Computer Science, University of Helsinki, Finland and CSE Department, Hong Kong University of Science and Technology, Hong Kong

XIN WANG, School of Computer Science, Fudan University, China and Shanghai Key Lab of Intelligent Information Processing, Fudan University, China

XIAOMING FU, Institute of Computer Science, University of Göttingen, Germany

Online social networks (OSNs) have become a commodity in our daily life. As an important concept in sociology and viral marketing, the study of social influence has received a lot of attentions in academia. Most of the existing proposals work well on dominant OSNs, such as Twitter, since these sites are mature and many users have generated a large amount of data for the calculation of social influence. Unfortunately, cold-start users on emerging OSNs generate much less activity data, which makes it challenging to identify potential influential users among them. In this work, we propose a practical solution to predict whether a cold-start user will become an influential user on an emerging OSN, by opportunistically leveraging the user's information on dominant OSNs. A supervised machine learning-based approach is adopted, transferring the knowledge of both the descriptive information and dynamic activities on dominant OSNs. Descriptive features are extracted from the public data on a user's homepage. In particular, to extract useful information from the fine-grained dynamic activities that cannot be represented by the statistical indices, we use deep learning technologies to deal with the sequential activity data. Using the real data of millions of users collected from Twitter (a dominant OSN) and Medium (an emerging OSN), we evaluate the performance of our proposed framework to predict prospective influential users. Our system achieves a high prediction performance based on different social influence definitions.

This work is sponsored by National Natural Science Foundation of China (No. 62072115, No. 71731004, No. 61602122), CERNET Innovation Project (NGII20190105), the project "PCL Future Greater-Bay Area Network Facilities for Large-scale Experiments and Applications (LZC0019)," the Research Grants Council of Hong Kong (No.16214817), the 5GEAR project and FIT project from the Academy of Finland, and the EU H2020 COSAFE project.

Authors' addresses: Q. Gong, Y. Chen, and X. He, School of Computer Science, Fudan University, Shanghai 200438, China and Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200438, China and Peng Cheng Laboratory, Shenzhen 518066, China; emails: {gongqingyuan, chenyang, xlhe17}@fudan.edu.cn; Y. Xiao, Department of Communications and Networking, Aalto University, 02150 Espoo, Finland; email: yu.xiao@aalto.fi; P. Hui, Department of Computer Science, University of Helsinki, 00014 Helsinki, Finland and CSE Department, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong; email: panhui@cse.ust.hk; X. Wang, School of Computer Science, Fudan University, Shanghai 200438, China and Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai 200438, China; email: xinw@fudan.edu.cn; X. Fu, Institute of Computer Science, University of Göttingen, 37077 Göttingen, Germany; email: fu@cs.uni-goettingen.de.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2021 Association for Computing Machinery.

1559-1131/2021/05-ART6 \$15.00

<https://doi.org/10.1145/3409108>

CCS Concepts: • **Human-centered computing** → **Social networking sites**; • **Networks** → *Online social networks*;

Additional Key Words and Phrases: Social influence, cold-start users, cross-site linking

ACM Reference format:

Qingyuan Gong, Yang Chen, Xinlei He, Yu Xiao, Pan Hui, Xin Wang, and Xiaoming Fu. 2021. Cross-site Prediction on Social Influence for Cold-start Users in Online Social Networks. *ACM Trans. Web* 15, 2, Article 6 (May 2021), 23 pages.

<https://doi.org/10.1145/3409108>

1 INTRODUCTION

Online social networks (OSNs), such as Facebook [50], Twitter [14, 31], and Pinterest [23], have become very popular all over the world and attracted billions of users [28]. In OSNs, users are able to generate contents online, manage their own curation, and interact with their friends. Moreover, OSNs are important platforms for information diffusion [2]. A user can subscribe to another user’s updates by following her (on asymmetric social networks like Twitter) or making friends with her (on symmetric social networks like Facebook). Receiving the contents published by other users, a user can express her feedbacks by posting comments, re-posting the contents (e.g., the “retweet” function on Twitter), or using the “like” function to quickly show the appreciation for the contents.

Social proof theory [10] shows that people tend to imitate and refer to others for guidance when they need to make decisions or take actions. The concept of *social influence* [1, 4] has been introduced to quantify a user’s overall impact within the social network, which is an important concept in sociology and viral marketing [4]. In social networks, a piece of information can reach a massive number of audiences through the network via a “word-of-mouth” way of diffusion. Users may achieve different levels of social influence, because of the substantial difference in their credibility, expertise, and social connectivity. Some users, known as influential users, could quickly deliver information to a large number of audiences. Researchers have made numerous efforts in quantifying users’ social influence in OSNs [1, 4, 31, 32, 42]. Most of these studies use Twitter to apply their methods, thanks to the fact that Twitter is a dominant and mature OSN service and many users have generated a large amount of data for the judgment.

For an OSN with more than 200M monthly active users (MAUs), we call it a *dominant OSN*. Accordingly, for an OSN with less than 200M MAUs, we call it an *emerging OSN*. At the global level, there are very few dominant OSNs, for example, Facebook and Twitter. These OSNs have a wide coverage of users all over the world for more than one decade. Although dominant OSNs serve billions of users, there are still a number of emerging OSNs, normally with special focuses. For example, Foursquare [7] focuses on location-centric activities, and Medium (<https://www.medium.com/>) offers blog services. New emerging OSNs are launched from time to time, whose users usually have relatively smaller account ages.¹

Existing methods are proposed to identify the influential users on dominant OSNs, relying on their social connections or the rich contents they have generated. Because of the diversity and usefulness of emerging OSNs, a user might sign up new accounts on emerging OSNs from time to time, but remain her principal social connections and activities on dominant OSNs. This will lead to a cold-start problem. When one user just joins an emerging OSN and becomes a cold-start user, only a partial of her social connections are established, and she might have generated very

¹The account ages on emerging OSNs are in general younger than the account ages on dominant OSNs. For example, Twitter was launched in 2006 and Medium was launched in 2012. According to our dataset, the average account age of Medium users is 3.84 years, while that of Twitter users is 8.33 years.

few activities [37, 48]. It is challenging to determine whether a cold-start user will become an influential user on an emerging OSN using traditional approaches [1, 4, 31, 33, 44, 55]. To make an accurate judgment of whether one cold-start user will become an influential user, we propose the idea of “cross-site prediction” by leveraging the users’ demographic and behavioral information on dominant OSNs. Nowadays, the accounts owned by the same user on different OSNs are connected and made public in forms of *cross-site linking* or the *social hub websites*. Many emerging OSNs have provided the *cross-site linking* function [18] to authorize users to log in with their accounts on dominant OSNs, while social hub services [19] allow a user to show her accounts on multiple OSNs on a same page.

In this article, we build a cross-site prediction system to judge whether a cold-start user on an emerging OSN will become an influential user. This system leverages a user’s data on a dominant OSN to extract features to distinguish between influential users and ordinary users. Useful features are chosen according to a user’s descriptive information and dynamic activities. For the dynamic activities, we use deep learning technologies to extract dynamic features from the constructed activity sequences of each user. Aggregating the output dynamic features and the summarized descriptive features, a supervised machine learning-based classifier is used to implement the final prediction. The evaluation takes Twitter and Medium as the case study to demonstrate the prediction performance of our system. We have made the following three key contributions:

- We formulate the problem of predicting whether a cold-start user on an emerging OSN will become an influential user. We introduce the idea of studying users’ social influence across OSNs, and leverage users’ data on dominant OSNs to help the prediction.
- We build a cross-site influential user identification system to implement the prediction. The framework of the system is based on a supervised machine learning method. In special, we integrate deep learning approaches to process the users’ dynamic activity data. We extract features from the users’ descriptive information and dynamic activities on dominant OSNs to help the machine learning classifier make accurate decisions.
- The pair of emerging OSN Medium and dominant OSN Twitter is taken as a case study. We use real user data to examine the performance of our proposed cross-site prediction system to identify prospective influential users on Medium. Experimental results demonstrate that the system can achieve an accurate prediction in terms of different social influence definitions.

The rest of this article is organized as follows: We discuss the background and related work in Section 2. We present the definition of social influence and the data collection procedure in Section 3. The overall cross-site prediction framework is introduced in Section 4. We use Medium and Twitter as the emerging and dominant OSNs to demonstrate the system implementation details in Section 5 and further evaluate the prediction performance of the system using real data in Section 6. Some practical issues are discussed in Section 7. We present the conclusion and some prospective future work in Section 8.

2 RELATED WORK

In this section, we illustrate the background and related work of this study. We first introduce the existing studies of social influence in Section 2.1, and then discuss the cross-site linking function in Section 2.2. Section 2.3 highlights our contributions over previous work.

2.1 Existing Studies of Social Influence

In a social network, users interact with and influence each other. Different users have different levels of capability to spread information on the platform. The concept of social influence [1, 4, 31, 33, 44, 55] has been proposed and widely used in social network-related studies. To our knowledge, existing research about social influence definitions can be summarized into two categories, i.e., from the social graph perspective and the user interaction perspective. For the first category, researchers referred to the social graph to define the social influence metrics, such as number of followers [1, 4, 55] and PageRank [31, 33, 44]. For the second category, researchers proposed the social influence metrics by referring to the interactions between users. Metrics within the second category include the number of retweets [4, 55], number of replies [55], and number of mentions (i.e., the “@” function in microblogs) [4]. These metrics represent the impact users generate in OSNs from different angles. The number of followers a user has indicates how popular this user is concerning social connections. The number of retweets means how many times a user’s tweets have been re-posted, showing the popularity of her tweets. The number of mentions reflects how frequently a user has been mentioned by others. Each of these metrics is meaningful and characterizes a user’s social influence from some aspect. Besides the social influence of an individual user, researchers also explore the social influence of a group of people. Li et al. [32] proposed the idea to measure the social influence of a community by aggregating the social influence of users involved. Meanwhile, instead of analyzing the social influence from a static view, Song et al. [42] studied the problem of identifying influential users in social networks from a dynamic aspect.

Metrics based on the social graph emphasize the importance of social connections in information diffusion, due to the fact that a user will receive the posts published by her followings or her friends automatically. However, Wilson et al. [50] have shown that the social graph cannot reflect the information diffusion accurately, since Facebook users tend to be inactive with most of her friends. OSNs offer a number of functions to support the interactions between users; for example, the “like,” “comment,” “repost/retweet,” and “mention.” Thanks to these functions, information can be delivered between users in a more trackable way. Accordingly, the information diffusion in this way can also be described by a graph, i.e., the interaction graph $G = (V, E)$. Each node in V represents one user. An edge e_{ij} between the user u_i and the user u_j is established if there exist some visible interactions from u_i to u_j . In this article, we study the social influence based on the definitions by considering the social interactions between users.

As a popular and mature OSN service, Twitter has been the platform widely used in studying the social influence. However, the aforementioned metrics are not very suitable for cold-start users on emerging OSNs, who are still new to the OSN and lack of activity data. For example, it takes quite some time to let the Medium users grow up to an active one and generate enough “clap” interactions between them. Using the aforementioned metrics to identify influential users in their cold-start stage might lead to some inaccurate results. To remedy this problem, we make use of the cross-site linking function to leverage a user’s information on dominant OSNs to realize the prediction.

2.2 Cross-site Linking in OSNs

Many of the emerging OSNs, such as Foursquare [7]/Swarm [8, 51] and Medium, take advantage of their users’ accounts on dominant OSNs to enhance their function-orientated services. They support a *cross-site linking function* [18], allowing users to link their accounts on dominant OSNs, e.g., Twitter and Facebook. In this way, users can log in to the emerging OSNs with their Twitter or Facebook accounts, avoiding the problem of managing multiple accounts. By using the cross-site linking function, a user is allowed to post the same piece of information to multiple OSNs simultaneously and copy social connections from dominant OSNs to emerging OSNs.

Cross-site linking is widely applied in the literature as a way to take advantage of the aggregated social footprints to study user behaviors across multiple OSNs. Zhong et al. [58] proposed the concept of social bootstrapping, i.e., copying established social connections from a dominant OSN into an emerging OSN. Their study demonstrated how users build their social connections on a new OSN by referring to the cross-site linking function. Zhang et al. [57] studied the relationship between social interactions on emerging OSNs and social ties on established ones. Farseev et al. [12] aggregated different kinds of information such as locations, texts, photos, and demographic attributes from Foursquare, Twitter, Instagram, and Facebook to construct a multi-source dataset. They applied the dataset to predict users' demographic information. Jia et al. [27] proposed a model to predict whether a user will become a volunteer based on her public information on Facebook and Twitter.

In this work, we make use of cross-site links between an emerging OSN and a dominant OSN. From this new angle, we study how to predict whether a cold-start user on an emerging OSN will become an influential user. We select the "Medium-Twitter" pair as a case study, since Medium allows a user to link her profile to her account on Twitter. It is worth noting that the cross-site linking function is not the only way to connect the accounts of one same user on multiple OSNs. Due to the prosperity of emerging OSNs, users are often active in more than one OSN. The accounts matching has been proved possible to be realized through the self-posted URLs of one user's homepages on other OSNs [18], the public online profile services such as about.me [19], or the same publicly accessible email address shown on different OSNs [46]. There also has been a series of research works studying the alignment of users' accounts according to their usernames or generated contents in different OSNs [15, 26, 30, 34]. We add more details in the discussion in Section 7.1.

2.3 Contributions over Previous Work

Summarizing previous works on social influence studies in Section 2.1, we find that, on one hand, they always focus on a single OSN; for example, Twitter, which is a well-developed social network where users have generated rich behavior data. On the other hand, the methodology adopted by many previous works is to rank the users by the value of indegree [1, 4, 55] or PageRank [31, 33, 44] referring to the social graph or the interaction graph. Under the circumstance that users are having accounts on multiple OSNs, quite a number of accounts are experiencing the cold-start phase after they register their accounts on emerging OSNs. Considering the limitations of previous methods that only count the existing published contents, social connections, and visible interactions, we formulate the problem of predicting the potential influential users across OSNs.

In the cold-start phase, users may only provide basic account information (sometimes only a username randomly generated by the OSN service provider) without publishing countable contents or building regular social connections [37, 48]. Since users' generated data is not off-the-shelf on the emerging OSN, previous approaches are not applicable. Thanks to the cross-site linking function [18], we are able to reach their Twitter accounts and collect their historical behaviors on Twitter. Taking Medium as the case study, we propose a machine learning-based cross-site prediction system to analyze the connections between users' behavior on Twitter to their potential social influence on Medium. The proposed system is able to leverage the dominant OSN (Twitter) to help detect the potential influential users on the emerging OSN (Medium).

3 SOCIAL INFLUENCE DEFINITION AND DATA COLLECTION

Medium is a blog-sharing social network launched in August 2012 by Evan Williams, the co-founder and former CEO of Twitter.² On Medium, each user has a profile page, showing her

²<https://medium.com/@ev>, accessed on May 1st, 2019.

demographic attributes such as username and profile photo, her social connectivity information including the numbers of followings and followers, and a paragraph of her biography. Medium allows a user to link her Facebook and Twitter accounts to her Medium profile page. The linked Facebook and Twitter accounts are also shown on this page. Visitors to the Medium page can further access the user’s Facebook and Twitter pages conveniently. A post on Medium is called a “story.” If a user is impressed by a story published by other users, she can click a “clap” button to express her appreciation and support. The profile page also provides links to three additional tabs, i.e., “Latest,” “Claps,” and “Responses,” showing the latest stories published by the user, the stories she has clapped for, and the stories she has commented on, respectively.³ The stories in these tabs are organized with a reverse chronological order. In our investigation, Medium serves as the emerging OSN, and Twitter acts as the dominant OSN.

3.1 Social Influence Definition

On Medium, the main form of user-generated contents is Medium story, which is similar to blogs. Medium users are able to express their opinions by publishing stories. Unlike Twitter, where the maximum number of characters in one tweet is limited to 140, there is no length limit for the stories. If one user likes the content of a published story, she can click the “clap” button to show her support.

Definition 3.1. Given the interaction graph $G = (V, E)$, where V is the node set formed by Medium users, and E is the edge set formed by the interactions of “claps” between users. We add weight to each edge in the interaction graph G to signify the total number of “claps.” If user u_i has made m claps to user u_j ’s published stories, we set w_{ij} as m . The social influence of a Medium user u_i is represented by the *total* number of “claps” made by other users to the Medium posts published by her. Formally:

$$\mathcal{I}_T(u_i) = \left\{ \sum_j w_{ij} \mid u_i, u_j \in V, e(u_j, u_i) \in E \right\}. \quad (1)$$

The stories published by a user may receive different levels of attention. It is possible that one user receives a large number of “claps” just because of a single popular story, while most of her other stories have received very few “claps.” An alternative social influence definition of a Medium user is based on the “H-index” [24] metric, which has been widely used for quantifying a scholar’s citation impact.

Definition 3.2. We use $c_{s_1}, c_{s_2}, \dots, c_{s_n}$ to denote the numbers of claps received by the published stories s_1, s_2, \dots, s_n , respectively. The social influence of a Medium user u_i is the output of the H-index operator \mathcal{H} . Formally:

$$\mathcal{I}_H(u_i) = \mathcal{H}(c_{s_1}, c_{s_2}, \dots, c_{s_n}), \quad (2)$$

where the operator \mathcal{H} will return an integer value p , indicating that for user u_i , p stories have received at least p “claps,” while each of the other $n - p$ stories has received less than p “claps.” The returned value is taken as the influence value $\mathcal{I}_H(u_i)$ of user u_i .

Given the above definitions, the level of one user u_i ’s social influence can be represented as the variable $\mathcal{I}(u_i)$. A larger value of $\mathcal{I}(u_i)$ indicates a higher level of social influence. Based on the value $\mathcal{I}(u_i)$, the users can be categorized into two groups: *influential users* and *ordinary users*.

³<https://medium.com/@chenyang03>.

With a threshold δ , the discrimination of these two user groups can be formulated as

$$u_i \text{ is } \begin{cases} \text{an influential user} & \text{if } I(u_i) > \delta, \\ \text{an ordinary user} & \text{if } I(u_i) \leq \delta. \end{cases} \quad (3)$$

The threshold variable δ can be determined according to the specific scenario. A larger δ indicates that a higher influence value is required for a user to be selected as an influential user.

3.2 Data Collection

To obtain a dataset for our study, we need the activity records of Medium users for labels and the data they generate on Twitter for features. We used Breadth First Search (BFS) to crawl a set of Medium IDs. BFS has been widely used in OSN data collection, such as in References [16, 21, 50]. The data crawling started from the user Evan Williams (<https://medium.com/@ev>), the CEO of Medium, and lasted from January 23, 2019 to February 1, 2019. To implement BFS, we maintained a queue of Medium user IDs and put Evan Williams’s Medium ID into the queue first. We did the crawling in an iterative way. In each iteration round, we picked the first Medium ID from the head of the queue and obtained the lists of the corresponding user’s followings and followers using a Python-based crawler. The Medium IDs of these followings and followers, if they have not been crawled, would be further added to the end of the queue. This process was repeated until the number of user IDs we collected reached the threshold of 2.5M. Based on these IDs, we randomly selected 350,000 of them to form the dataset. For each Medium user in this dataset, we further crawled her Medium profile and published stories according to her user ID. Among the 350,000 Medium users in our dataset, 194,850 (55.67%) of them have linked their Twitter accounts. Accordingly, we crawled these users’ profiles and published tweets on Twitter using the Twitter API. The crawled public user data from their Medium and Twitter accounts are used for our study.

As discussed in Reference [50], a user normally interacts with only a small portion of her social connections. Some of the interactions between users, for example, page views, are stored in the backend and invisible to the public. Differently, some interactions, such as “claps” or “comments,” are visible to the public. Both of the two types of interactions are useful to signify information diffusion on the platform. Since our work is based on the publicly accessible data, we take the visible interactions to calculate the social influence metrics. On Medium, the primary type of interactions between users is the “claps” received by the Medium stories.

4 OVERALL FRAMEWORK TO PREDICT PROSPECTIVE INFLUENTIAL USERS ON EMERGING OSNS

For a cold-start user on an emerging OSN, she might have generated a limited amount of activity records. Our objective is to predict whether a user will become an influential user at the very beginning of her voyage on an emerging OSN. Existing social influence metrics can not be applied directly, since they normally require a user to be mature. To remedy this problem, we refer to the user’s information on well-established dominant OSNs, and we believe these contents are more informative to judge her potential social influence on the corresponding emerging OSN. We build a cross-site prediction framework to predict whether a user will become an influential user on an emerging OSN, connecting her behavior on a dominant OSN through feature construction and prediction model formulation. The cross-site prediction problem is formulated in Section 4.1. We further discuss the feature selection in Section 4.2 and introduce the decision maker in Section 4.3.

4.1 Problem Formulation

In the cross-site prediction problem, there is an emerging OSN ($O_{Emerging}$) as the target site and a dominant OSN ($O_{Dominant}$) as the source site. To predict whether a cold-start user will become

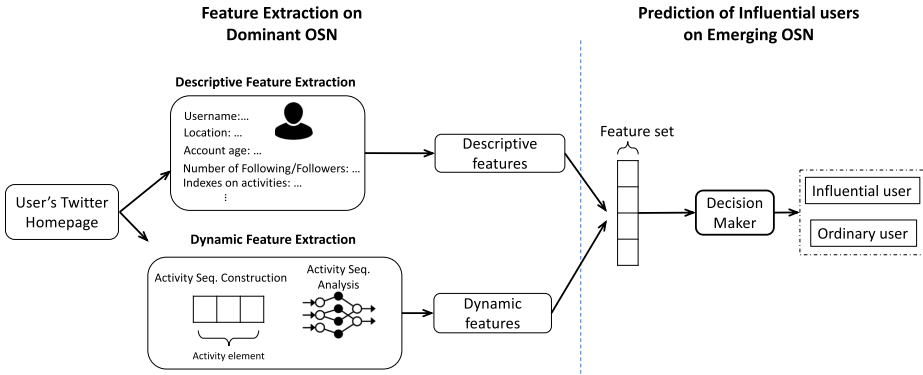


Fig. 1. Workflow of the cross-site prediction system on social influence for cold-start users in emerging OSNs.

an influential user on $O_{Emerging}$, we refer to the user's account on $O_{Dominant}$ to compensate the lack of activity data on $O_{Emerging}$.

For each user u_i , the ground-truth label y_i indicates whether she is an influential user on $O_{Emerging}$. y_i is a binary tuple with the sum of the two elements as 1. If u_i is an influential user, then we set y_i as $[1, 0]$. If u_i is not an influential user, then we set y_i as $[0, 1]$. The predicted result of u_i is denoted as \hat{y}_i . We introduce a supervised machine learning model f to do the prediction. Each user u_i is described by the same set of features, represented by a d -dimensional feature vector X_i . We aim to do the prediction for cold-start users on $O_{Emerging}$, while all features are selected from $O_{Dominant}$. We focus on both the descriptive information and dynamic activities of the users to form two categories of features, i.e., descriptive features and dynamic features. As shown in Figure 1, the values of these two types of features are fed into a decision maker to conduct the prediction. As in Equation (4), our prediction model f takes u_i 's feature vector X_i on $O_{Dominant}$ as the input and output the judgment of whether this user will be an influential user as \hat{y}_i :

$$\hat{y}_i = f(X_i). \quad (4)$$

Our task turns into finding a suitable f to obtain the \hat{y}_i close to y_i . Suppose there are N users in total in the user dataset U . We define $Y = [y_1, y_2, \dots, y_N]^T$ to represent the ground-truth of whether each user is an influential user in $O_{Emerging}$. The objective function of the prediction model can be represented as

$$\Gamma = \min |Y - f(\mathbf{X})| = \min \frac{1}{N} \sum_{i=1}^N |y_i - f(X_i)|, \quad (5)$$

where $|y_i - f(X_i)|$ measures the difference between the predicted result and the ground-truth label for user u_i .

To examine whether a user will be an influential user on Medium, a straightforward idea is to see whether she is an influential user on a selected dominant OSN. As in References [4, 55], a user's social influence on Twitter can be quantified by the total number of "retweets" received by her published original tweets. We use Spearman's rank correlation coefficient [43] to quantify the correlations between the users' social influence values on Medium and Twitter based on all the 194,850 pairs of valid user accounts in our crawled dataset. The value of this coefficient is between -1 and 1 . It reveals negative correlations between the pair of examined ranks if the value is close to -1 , positive correlations if the value is close to 1 , and almost no correlations if the value is close to 0 . The corresponding Spearman's rank correlation coefficients between the two

social influence definitions on Medium and the social influence rank on Twitter are 0.231 and 0.232, respectively. This means that users' social influence on Twitter and on Medium cannot be considered as equivalent, but are positively correlated at a weak level. Therefore, we cannot simply conclude that a user has a larger social influence value on Twitter will also have a higher level of social influence on Medium. To solve this problem, we propose to build a supervised machine learning model to leverage the Twitter data to predict whether a cold-start user will become an influential user on Medium.

4.2 Feature Selection

We summarize the public data that users are possible to generate into four types, i.e., demographic information, account information, social connections, and historical activities. To characterize the user behavior comprehensively, our system processes the user's information on dominant OSNs from both static and dynamic perspectives.

Descriptive Features: There are usually a lot of attributes in a user's profile on $O_{Dominant}$. These attributes describe the user's properties from different angles. Some attributes represent a user's demographic information, for example, name, gender, age, and home location. In addition, some attributes represent a user's statistical information, such as number of published posts, number of friends (on symmetric OSNs), number of followings/followers (on asymmetric OSNs), number of check-ins (on location-based social networks). All these attributes can be selected as descriptive features to distinguish between influential users and ordinary users on $O_{Emerging}$. The descriptive features of users are categorized as demographic, account, social, and UGC (user-generated content).

Dynamic features: A user is able to conduct different types of activities on $O_{Dominant}$ from time to time. Besides the attributes in profiles, we also make use of the fine-grained dynamic user activities for the prediction. An example to show the necessity of dynamic features is the tweets published by a Twitter user. Compared with the statistical indices such as the total number of published tweets, fine-grained dynamic user activities provide a detailed view to characterize a user. One user's published tweets form an activity sequence. For each tweet, we use an *activity vector* to describe it from different angles. Taking a sequence of activity vectors as the input, we propose an *activity sequence analysis module* implemented by recurrent neural networks (RNNs) [36] to deal with the details of each tweet publishing activity and output the dynamic features.

4.3 Decision Maker

As shown in Figure 1, the prediction framework is a two-step model to determine whether a cold-start user will become an influential user on $O_{Emerging}$. After obtaining both the descriptive and dynamic features for each of the N users in the first step, we define the input matrix $X \in R^{N \times d}$ for the decision maker. The i th row in X denotes the feature values of the i th user, obtained from $O_{Dominant}$. In the second step, a decision maker is used to train a supervised machine learning model to fully utilize the features and predict whether a cold-start user will become an influential user on Medium.

For a cold-start Medium user that enables the cross-site linking function and links her profile to Twitter, we are able to access her Twitter account shown on her Medium profile page. For a Twitter user, she could click the "Log in with Twitter" button to register a Medium account and become a new user on Medium. This new Medium account will be naturally linked to her Twitter account. By visiting the URL of her Twitter homepage, our system is able to collect her Twitter data and further use the collected information to judge whether this user will become an influential user on Medium.

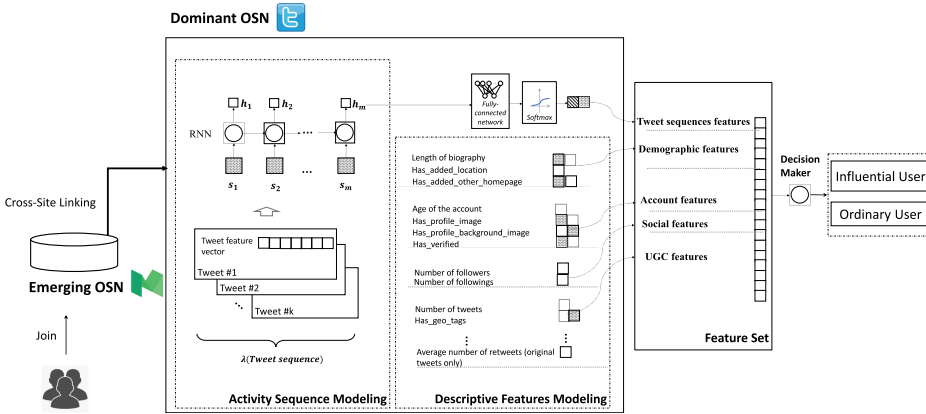


Fig. 2. Implementation of the cross-site influential user prediction system.

5 IMPLEMENTATION OF THE CROSS-SITE PREDICTION SYSTEM

In this section, we explain how we implement each important building block of our cross-site influential user prediction system using Medium and Twitter as the case study. Figure 2 shows the implementation details of this system. Since we select Twitter as $O_{Dominant}$, the system needs to retrieve the user data from Twitter to make the prediction. As mentioned in Section 4.2, we use two groups of features to describe a user, i.e., descriptive features and dynamic feature. Section 5.1 illustrates how we get the descriptive features from the profile of a user’s Twitter account. Section 5.2 illustrates how the dynamic activities are constructed as a sequence and further processed by RNNs to get the dynamic feature. After obtaining the two feature groups, a supervised machine learning-based decision maker is used to judge whether the user will become an influential user or an ordinary user on Medium. Section 5.3 explains how we implement the decision maker.

5.1 Descriptive Feature Extraction

The system introduces descriptive features based on a user’s publicly accessible information. Considering the services provided by Twitter, we list the selected four subsets of users’ descriptive features in Table 1, including the demographic information, the account information, statistics of social connections such as the numbers of followers or followings, and statistics of her UGCs, such as the number of published tweets and the average time interval between successive tweets. Details of the four feature subsets are listed as follows:

- On Twitter, a user can choose to fill the information fields in her profile page, including a biography to describe herself, her current location, and the URL of her other online personal homepages. The personal information is maintained by the user herself, which generally reveals the preferences of her. We utilize this basic information to extract the **Demographic Features**.
- Additional descriptive data in the homepage records the status of the account, including the age of the account, whether the user has changed the default profile image and the background image, and whether this account has been “verified” as an account of public interest. We take this part of information as the **Account Features**.
- Twitter supports the social networking function by endowing a user to follow anyone she is interested in. Several works show that a large number of followers is essential to generate social influence, since the larger audience will definitely accelerate information

Table 1. Features for Potential Influential User Prediction

Feature Group	Feature Category	Feature List
Descriptive Features	Demographic Features	<ul style="list-style-type: none"> · Length of biography · Has_added_location · Has_added_other_homepage
	Account Features	<ul style="list-style-type: none"> · Age of the account · Has_profile_image · Has_profile_background_image · Has_verified
	Social Features	<ul style="list-style-type: none"> · Number of followers · Number of followings
	UGC Features	<ul style="list-style-type: none"> · Number of tweets · Has_geo_tags · Number of lists subscribed to · Number of original tweets · Number of retweets · Number of tweets “liked” by the user · Total number of “likes” received (original tweets only) · Total number of tweets she posts “likes” · Average number of “likes” received (original tweets only) · Total number of retweets (original tweets only) · Average number of retweets (original tweets only)
Dynamic Feature	Dynamic Feature	<ul style="list-style-type: none"> · Probability of being an influential user (according to activity sequence)

diffusion [45, 56]. In the meanwhile, there are studies about social influence for marketing in OSNs showing that the number of followings is also an important factor, which should be considered together with the number of followers for measuring the social influence [11]. Considering about these findings, we extract both the numbers of followings and followers as the **Social Features**.

- Besides receiving tweets from her followings or lists subscribed, Twitter users can post tweets, i.e., in the form of publishing original tweets or re-tweeting other users’ tweets. The numbers of “likes,” replies, or retweets are also representative indicators of whether the audience is paying attention to the user [4, 41]. We summarize comprehensive content generating features from users’ tweeting behavior, including her preference of enabling the geography tags for tweets or not, her activeness of posting tweets, and the numbers of “likes” and retweets she has received from other Twitter users. These make up the **UGC Features**.

5.2 Dynamic Feature Extraction

The descriptive features used in our model reflect an aggregative view of each user on Twitter. For a Twitter user, the main activities include tweet publishing and retweeting. These activities lead to information propagation and diffusion on Twitter. Our system involves an *activity sequence*

Table 2. Elements of the Input Vector to RNN

Key	Type	Exemplified value
if_retweet, if this tweet is a retweet or an original tweet	list	[1,0] for retweet
Number of “likes” received	integer	10
Number of “retweets” received	integer	5
Timezone of the published time, one of the six time intervals in a day when the tweet published	list	[0,1,0,0,0,0] for 7:00 am
Number of images	integer	1
Number of words	integer	1
Number of users being “@”	integer	1
Number of URLs	integer	1
If this tweet is in English	Boolean	1

analysis module to process the historical activity data of one user. The module is implemented with deep learning techniques, which have been widely used in user behavior analytics in OSNs [17, 20, 35, 54]. For each user, we construct a sequence s of the tweets she has published, including the original tweets and retweets. Each element s_t in the sequence corresponds to a tweet, describing the tweet from the perspectives of the publishing behavioral characteristics such as the publishing timezone, the number of users being “@” in this tweet, and the attentions this tweet attracts, which are represented as the number of “likes” received and the number of “retweets” received. Each tweet can be described as a vector of 15 elements. The detailed dimensions are listed in Table 2.

We introduce an RNN to process the tweet sequence s with m elements, i.e., (s_1, s_2, \dots, s_m) . RNN is a representative deep neural network based on a chain-like architecture. It consists of an array of cells, each with a number of neurons. The cells in the neural network deal with the input vector of the tweet sequence successively. At each time slot t , one tweet vector is fed into the neural network. One cell processes the input element at a time, generating a cell hidden state variable h_t . The hidden state h_{t-1} generated in the previous cell will be fed into the current cell together with s_t at time slot t . Processing these two inputs represented with a function g , the cell generates the current hidden state h_t ,

$$h_t = g(h_{t-1}, s_t). \quad (6)$$

One sparkling feature of RNNs is that the cells are able to memorize the intermediate processing results of previous cells. The neural network is often trained to minimize an optimal objective function through gradient descent. Unfortunately, due to their iterative nature, standard RNNs suffer from both vanishing and exploding gradients [29]. The long short-term memory (LSTM) neural network [25] solves this problem by incorporating the “gate” control in the cells.

Fed with the two parts of information, each LSTM cell involves three gates to control the processing of the input data, i.e., input gate, forget gate, and output gate. The three gates are realized with the sigmoid function σ as

$$f_t = \sigma(W_f[h_{t-1}, s_t] + b_f), \quad (7)$$

$$i_t = \sigma(W_i[h_{t-1}, s_t] + b_i), \quad (8)$$

$$o_t = \sigma(W_o[h_{t-1}, s_t] + b_o), \quad (9)$$

where W_f, W_i, W_o are the weight variables and b_f, b_i, b_o are the bias variables.

Each cell has a cell state variable c , which is updated based on the information filtered by the three gates. c_t is updated with an intermediate variable \tilde{c}_t

$$\tilde{c}_t = \tanh(W_c[h_{t-1}, x_t] + b_c). \quad (10)$$

Combining with the information produced by the input and forget gates, the cell state variable is updated as

$$c_t = f_t c_{t-1} + i_t \tilde{c}_t. \quad (11)$$

The hidden state of each cell is generated based on the cell state variable and the information produced by the output gate

$$h_t = o_t \tanh(c_t). \quad (12)$$

The above two equations are the updating operations for standard LSTM networks. There are several advanced versions of LSTM structures. Here, we introduce the Phased LSTM (PLSTM) network [38], which considers the irregularity in the sampling of the activity sequences of different users. The advantage introduced by PLSTM lies in the additional control variable k_t to the update of the cell states. The Equations (11) and (12) are changed into

$$c_t = k_t (f_t c_{t-1} + i_t \tilde{c}_t) + (1 - k_t) c_{t-1}, \quad (13)$$

and

$$h_t = k_t (o_t \tanh(c_t)) + (1 - k_t) h_{t-1}. \quad (14)$$

This new variable k_t works as an additional gate, whose openness controls the update of the variables c_t and h_t . The two state variables will keep the same as in the previous time slot if the input element at the current time slot is not that important, different from the standard LSTM cells that update the state variables for each input element. The value of k_t for each time slot is determined along with other PLSTM parameters in the training process.

Taking the hidden state of the last cell as the output of the neural network for user u_i , our framework further applies a fully connected network to transform it into a 2-dimensional vector ϕ_i . Applying a softmax function, we obtain a new 2-dimensional vector Φ_i , with each dimension $\Phi_{i,q}$ ($q \in \{1, 2\}$) representing the probability of whether u_i will be an influential user or an ordinary user on Medium, considering the tweet publishing behaviors of her.

$$\Phi_{i,q} = \text{softmax}(\phi_i) = \frac{\exp(\phi_{i,q})}{\exp(\phi_{i,1}) + \exp(\phi_{i,2})}, (q \in \{1, 2\}). \quad (15)$$

We list $\Phi_{i,1}$, the probability of being an influential user (according to activity sequence), as the dynamic feature of user u_i in Table 1.

To determine the parameters used in the RNN network, the activity sequence analysis module takes the output binary probabilities Φ_i as its judgment about the user u_i according to her activity sequence on Twitter. The parameters used in the model are tuned in the training process with a loss function as

$$L = -[y_{i,1} \cdot \log(\Phi_{i,1}) + y_{i,2} \cdot \log(\Phi_{i,2})], u_i \in U. \quad (16)$$

Compensating to the descriptive features extracted from the static user information, the above operations by the PLSTM network realize an end-to-end analysis over user's tweet publishing behaviors, capturing the subtle features of users' sequential activities.

Table 3. Construction of the Experimental Datasets

Dataset	Influence Metric	No. of Influential Users	No. of Ordinary Users
Dataset1	Total number of “claps”	1,867	157,698
Dataset2	H-index of “claps”	1,908	157,657

5.3 Classifiers to Implement the Decision Maker

Obtaining the features depicting a user’s behavior on Twitter, our system uses a decision maker to make the final decision whether the cold-start user will be an influential user on Medium. Supervised machine learning classifiers are introduced to build the connections between users’ behavior characteristics on Twitter and the possibility of being influential users on Medium. To train a selected classifier to do an accurate prediction, we need to construct a training and validation subset with the labeled user instance, telling whether each of them will become an influential user or an ordinary user. We use the data of mature users on Medium that have linked their Twitter accounts to obtain the ground-truth information of influential users and ordinary users. Classifiers can be implemented by using classic algorithms including Random Forest [3] and Decision Tree [40], as well as new algorithms including CatBoost [39] and XGBoost [6]. CatBoost and XGBoost are representative tree boosting systems, which have been widely adopted in recent machine learning competitions such as Kaggle.

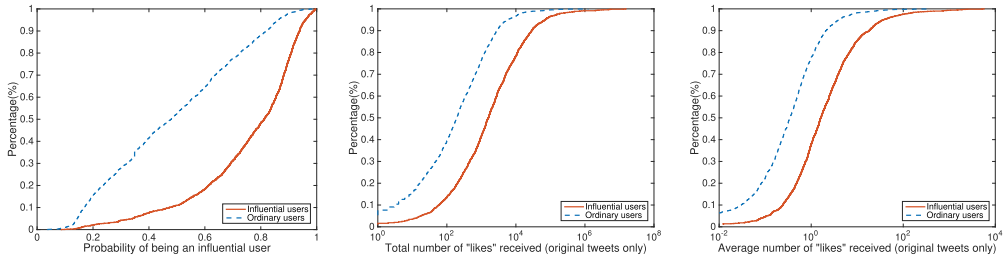
6 PERFORMANCE EVALUATION

To show the usefulness of the whole system in predicting influential users, we build a series of prediction tasks relating to two definitions of users’ social influence explained in Section 2.1. For each influence definition, we consider the following prediction tasks: After analyzing the public data on Twitter generated by the cold-start users on Medium, (a) can we accurately predict whether she will become an influential user on Medium, (b) which features in the descriptive and dynamic feature groups have the higher discriminative power, and (c) which recent recurrent neural network design is the most suitable one to conduct the activity sequence analysis. We introduce the dataset used and the metrics to evaluate the prediction performance in Section 6.1. A feature-based comparison between influential users and ordinary users is conducted in Section 6.2. We explore the prediction performance of different neural networks for implementing the activity sequence analysis module in Section 6.3 and compare among different algorithms for the decision maker in Section 6.4. We study the discriminative power of individual features in Section 6.5.

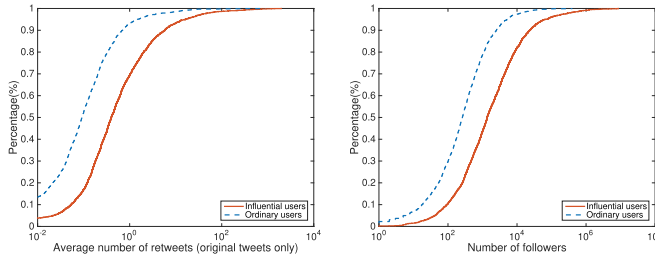
6.1 Experimental Datasets and Performance Metrics

As discussed in Section 3.1, we use both the total number of “claps” received by one user and the H-index value of the numbers of “claps” received by the user’s published Medium stories as social influence metrics. Two labeled datasets are formed accordingly. In the following study, we take two types of influence metrics, i.e., the total number of “claps” and the H-index of the “claps” received to define the ground-truth datasets as Dataset1 and Dataset2. The construction of the experimental datasets is listed in Table 3.

- Dataset1: We first check the total numbers of “claps” received by the contents generated by one user. We found that the top 1% of users concerning the total number of claps they have received published about 13.33% Medium stories. In the meanwhile, 83.86% of the claps are received by these stories. Therefore, we adopt 1% as the threshold to label the influential users concerning the social influence defined by the number of claps received. Excluding



(a) Probability of being an influential user according to activity sequence (b) Total number of “likes” received by original tweets (c) Average number of “likes” received by original tweets



(d) Average number of retweets received by original tweets (e) Number of followers

Fig. 3. Behavioral difference between influential and ordinary users.

the users whose Twitter homepages are inaccessible, we take 1,867 influential users and all the other 157,698 ordinary users to construct the training and validation subset, also the test subset. The proportion between the numbers of user instances in the training and validation subset and the test subset is 4:1.

- Dataset2: For each user, we also compute the H-index of the numbers of “claps” received by her published Medium stories. Considering this metric of social influence, we take the top 1% users as the influential users. The resulting H-index threshold of the top 1% is 10. We take the users whose H-index is more than 10 as the influential users. After excluding the users whose Twitter homepages are inaccessible, we obtain 1,908 Medium users labeled as the influential users and all the other 157,657 labeled as ordinary users to construct the training and validation subset, and the test subset. The proportion between the numbers of user instances in the training and validation subset and the test subset is also 4:1.

Following the workflow described in the previous section, we construct the dynamic and descriptive features for the given training/validation and test subsets. The prediction performance of the system is evaluated by feeding the user instances in the test subset to the trained classifier. There are several metrics to study and quantify the prediction performance. In this work, we apply the metric AUC (Area Under Curve) [13]. It denotes the probability that this classifier would rank a randomly chosen influential user higher than a randomly chosen ordinary user.

6.2 Feature-based Comparison between Influential Users and Ordinary Users

Before feeding the feature vectors into the classifier for training, we take Dataset1 as an example and plot the difference between influential users and ordinary users in Figure 3 according to five selected features. Based on the users in the training and validation subset, Figure 3(a) plots the distributions of the tweeting sequence analysis module output, i.e., the dynamic feature, for both

the influential and ordinary users. Influential users achieve higher values of the probability being an influential user considering their cross-site dynamic activities. Figure 3(b) and Figure 3(c) show that the influential users receive more “likes” than ordinary users for the tweets they published for both the metrics of the total number of “likes” received and the average number of “likes” received for all published tweets. We can also find the obvious difference between these two types of users concerning the number of retweets received and followers in Figure 3(d) and Figure 3(e), respectively. For every pair of compared sequences, we calculate the variances of each of the two sequences and find that these two variances are unequal. Accordingly, we do Welch’s t-test [49] and compute the corresponding p-value to study the significance of difference between influential and ordinary user instances. For each selected feature, the p-value of the Welch’s t-test is smaller than 0.001. This verifies that the selected features are significant enough to differentiate between influential users and ordinary users. We also conduct similar comparative analyses for Dataset2 and obtain similar results.

6.3 Comparison between Different Algorithms to Implement the Activity Sequence Analysis Module

We first examine the prediction performance of the activity sequence analysis module. We construct the sequences of the tweets published by users and process with different recurrent neural network designs, including the Dilated RNN (DRNN) [5], Bi-LSTM [22], Bi-GRU [9], and PLSTM [38]. DRNN constructs the “skip” connections over the tweet sequence to form a new subsequence and utilizes standard RNNs to deal with the subsequences. Bi-LSTM is a combination of a forward and a backward LSTM networks, concatenating the current hidden states generated by each LSTM network as the hidden state for the current input. Bi-GRU is a simplified gating approach similar to Bi-LSTM, but with fewer gates and higher computational efficiency. PLSTM is a variation of LSTM, which is able to deal with the irregular sampled elements of the input sequences. It takes the irregularity of the generation time of tweets into consideration. To ensure a fair comparison, we implement backward and forward PLSTM/DRNN networks to form a Bi-PLSTM and a Bi-DRNN. The number of neurons used in each neural network cell is set as 32. The learning rate is 0.001, and dropout rate is configured as 0.5. In particular, the layer of the DRNN is set as 2, meaning that only one layer of “skip” connection is used over the tweet sequence. The batch size is 100 and one epoch is used for the training.

The experiments are conducted for both the datasets with different social influence definitions. Figure 4 shows the AUC values of the prediction performance taking the total number of claps and the H-index of the numbers of claps as the definitions of the social influence, respectively. The performance of the system for both the datasets are shown together in one figure. Concerning the AUC values in the figure for Dataset1, the Bi-PLSTM network performs the best, reaching 0.791, while the lowest of 0.757 is output by the Bi-LSTM network. Similar trends are shown for Dataset2. The AUC value 0.812 is still achieved by the Bi-PLSTM network, while the lowest one is output by the Bi-LSTM network, with an AUC value of 0.752. We further conduct statistical tests on the results produced by the neural network models. Each pair of the neural network models are significantly different (p-value < 0.001, McNemar’s test). Therefore, we choose the Bi-PLSTM network to implement the activity sequence analysis module in the following experiments.

6.4 Prediction Performance of Different Classifiers for the Decision Maker

In this subsection, we evaluate the prediction performance of different classifiers for the decision maker. From the previous evaluation, we find that the activity sequence analysis module implemented with Bi-PLSTM network achieves the best AUC value. As a result, we utilize such a framework to generate the dynamic feature of users. Aggregating the descriptive features, we evaluate

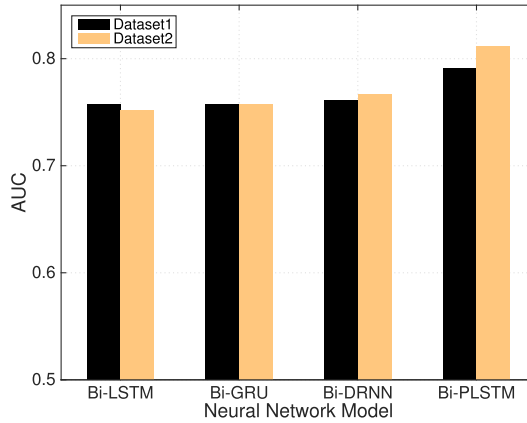


Fig. 4. Predictability of different implementations of the activity sequence analysis module.

the performance of the decision maker using different classifiers, including CatBoost [39], XG-Boost [6], Random Forest [3], and Decision Tree [40]. The parameters in the decision maker are determined through the parameter tuning phase, based on the user instances in the training and validation subset. We aim to obtain a set of parameters that could help the classifier achieve the highest AUC value. We apply a grid search to sweep the parameters space of the given classifier. Once a set of parameters is given, we can obtain an AUC value using 5-fold cross-validation. Afterwards, the trained decision maker is able to make the judgment for each user instance in the test subset, i.e., whether she will become an influential user on Medium, by referring to her information on Twitter. We evaluate the prediction performance of a trained decision maker using the test subset. Parameter settings of the classifiers and the prediction performance metrics are shown in Table 4. We can see that CatBoost achieves the highest AUC value for both datasets. We further conduct the McNemar’s test for each pair of the classifiers and find that every two classifiers are significantly different (p -value < 0.001 , McNemar’s test). In the following experiments, we use CatBoost as the decision maker to implement the prediction system.

6.5 The Discriminative Power of the Features

As in References [17, 18, 47], we conduct the χ^2 analysis to measure the discriminative power of each feature. A larger χ^2 value means that the corresponding feature has a stronger discriminative power [53]. Table 5 and Table 6 show the important features in the implementation for Dataset1 and Dataset2, respectively. For each of the datasets, the top eight features cover the four feature categories. In both tables, the dynamic feature ranks in the leading position, indicating the usefulness of the activity sequence analysis module. Considering the χ^2 values, there does not exist a determining feature that is able to identify influential users directly. The number of followers only shows the popularity of the account, while is not able to reflect whether its published content will be paid attention to by other users. We can also see that features including average number of “likes” received (original tweets only), total number of “likes” received (original tweets only), average number of retweets (original tweets only), and total number of retweets (original tweets only) are also very discriminative. The features depicting multiple aspects of users’ behaviors are showing their importance for the influential user prediction. The results of χ^2 values show that the existing conclusions on the users’ influence in one OSN are still meaningful when we consider the users’ social influence across OSNs [11, 45, 56].

Table 4. Prediction Performance of Different Classifiers for the Decision Maker

DataSet	Classifier	Parameters	AUC
Dataset1	CatBoost	learning_rate = 0.1, depth = 5, l2_leaf_reg = 15, iterations=150, border_count=100	0.848
	XGBoost	learning_rate=0.005, min_child_weight=5, max_depth=6, gamma=0, subsample=0.6, colsample_bytree=1, booster="gbtree," objective="binary:logistic"	0.842
	Random Forest	criterion="entropy," max_depth=5, n_estimators = 100	0.821
	Decision Tree	criterion="entropy," max_depth=11, min_samples_split=0.1, min_samples_leaf=1, random_state = 1	0.791
Dataset2	CatBoost	learning_rate = 0.1, depth = 5, l2_leaf_reg = 10, iterations=130, border_count=120	0.852
	XGBoost	learning_rate=0.005, min_child_weight=5, max_depth=6, gamma=0.0, subsample=0.6, colsample_bytree=0.7, booster="gbtree," objective="binary:logistic"	0.848
	Random Forest	criterion="entropy," max_depth=5, n_estimators = 100	0.843
	Decision Tree	criterion="gini," max_depth= 13, min_samples_split=0.1, min_samples_leaf=1, random_state = 1	0.807

Table 5. χ^2 Statistic for Dataset1: Influential User Definition Based on the Total Number of "Claps" Received

Rank	χ^2	Feature	Feature Category
1	3018.963	Average number of "likes" received (original tweets only)	UGC
2	2705.380	Total number of "likes" received (original tweets only)	UGC
3	2402.741	Probability of being an influential user	Dynamic
4	2328.664	Number of followers	Social
5	2280.995	Number of lists subscribed to	UGC
6	1931.263	Total number of retweets (original tweets only)	UGC
7	1617.441	Average number of retweets (original tweets only)	UGC
8	680.689	If the account has been verified	Account

Table 6. χ^2 Statistic for Dataset2: Influential User Definition Based on the H-index of the Numbers of “Claps” Received

Rank	χ^2 (Dataset2)	Feature	Feature Category
1	2307.307	Probability of being an influential user	Dynamic
2	2258.269	Number of followers	Social
3	2184.304	Number of lists subscribed to	UGC
4	2109.919	Total number of “likes” received (original tweets only)	UGC
5	1939.184	Average number of “likes” received (original tweets only)	UGC
6	1686.578	Total number of retweets (original tweets only)	UGC
7	1481.443	Average number of retweets (original tweets only)	UGC
8	681.169	If the account has been verified	Account

Utilizing the trained models in the previous subsection, we examine the power of typical individual feature, which is often used as the social influence metric on Twitter. By inputting one feature into the CatBoost classifier at one time, we obtain the AUC value of the prediction performance. The total number of “likes” received by original tweets makes the classifier reaching the AUC values as 0.743 for Dataset1 and 0.728 for Dataset2, while total number of “retweets” received by the original tweets reaches the AUC values as 0.768 for Dataset1 and 0.756 for Dataset2. The results are good, but still clearly inferior than the prediction performance of using the entire feature set. This verifies that the typical social influence metrics on Twitter are correlated with that on Medium, but the influential users on Twitter cannot be directly taken as the potential influential users on Medium.

7 DISCUSSION

In this section, we discuss the practicability of our proposed framework. Section 7.1 discusses the methodologies for obtaining a user’s accounts on dominant OSNs. In Section 7.2, we show the potential beneficiaries of our system, i.e., emerging OSN service providers, third-party OSN application providers, and users on emerging OSNs. We discuss the user privacy and ethical issues in Section 7.3, and the limitations of our work in Section 7.4.

7.1 Obtaining a User’s Accounts on Dominant OSNs

Our design needs to refer to the data generated by the user on dominant OSNs. There are several ways to obtain one user’s accounts on multiple OSNs. Quite a number of users choose to use the cross-site linking function or social hub services. Emerging OSNs often offer the cross-site linking function [18], allowing a user to link her profile to her accounts on dominant OSNs such as Facebook and Twitter. Such cross-site linking information is publicly accessible on a user’s profile page on an emerging OSN. This function is quite popular, considering the convenience it brings to users, i.e., to login to many OSNs with their Facebook/Twitter accounts, to share the contents they generate on emerging OSNs to Facebook/Twitter, and to find their Facebook/Twitter friends on a newly registered OSN to speed up the cold-start. According to our previous study [18], about 57.10% of the Foursquare users have linked to their Facebook or Twitter accounts. Similarly, according to the measurement results of this article, 55.67% of Medium users have linked their accounts to Twitter.

Meanwhile, social hub services [19] can also be used to retrieve a user's accounts on dominant OSNs. A social hub service allows a user to manage and display her accounts on multiple OSNs on a single webpage. [About.me](#) is a representative social hub service serving millions of users, allowing a user to add her accounts on tens of OSNs to her profile page [19]. By visiting a user's profile page on a social hub, we can link between her accounts on dominant OSNs and emerging OSNs.

There also exist account-matching algorithms to identify the accounts of the same user from different OSNs by referring to the publicly visible information of users, including the current profile attributes [15], revision histories of the attributes [26], social connections, and published posts [30, 34]. These approaches can match a user's accounts on different OSNs with a good accuracy on selected pairs of OSNs. We could also make use of these approaches to identify the accounts belonging to the same person on emerging and dominant OSNs.

7.2 Potential Beneficiaries

Our system can be utilized by three kinds of entities, i.e., the service providers of emerging OSNs, the third-party OSN application providers, and users on emerging OSNs. For emerging OSN service providers, they can adopt our system to predict whether a newly registered user will become an influential user and pay attention to the prospective influential users when necessary. In addition, since our system only needs to refer to the publicly visible information to conduct the prediction, third-party OSN application providers can also utilize our system to conduct the prediction and uncover the potential influential users on emerging OSNs. Last but not least, users on emerging OSNs can refer to our solution to follow the potential influential users.

7.3 User Privacy and Ethical Issues

We respect the privacy of users. In our work, we only used the publicly accessible information on Medium and Twitter. In addition, we anonymized the user IDs, and the dataset was stored in an offline environment. The ethical assessment of this study was reviewed and approved by the Research Department of Fudan University.

7.4 Limitations of Our Design

There are two main limitations in our study. First, our approach is only able to serve the users who have enabled the cross-site linking function or used the social hub services. Second, we have only validated our approach using the dataset collected from Medium and Twitter. The further validation of our approach to more emerging OSNs would be our future work.

8 CONCLUSION AND FUTURE WORK

In this article, we study the problem of predicting whether a cold-start user will become an influential user on emerging OSNs. By introducing a user's activities and profile information on dominant OSNs, we design and implement a machine learning-based system for the prediction. We examine the performance of the proposed framework by considering two social influence definitions on Medium, i.e., the total number of "claps" and the H-index value of the numbers of "claps" received by the stories. Based on the real data crawled from Medium and Twitter, we demonstrate that our system achieves a good performance in distinguishing between influential users and ordinary users on Medium. Note that our system can be used by not only OSN operators but also third-party application providers, as the system only needs to use the publicly accessible information.

There are a number of prospective future directions to explore further. We list them as follows:

- We plan to expand our approach to other emerging OSNs. Although Medium is a representative emerging OSN, it still provides functions of content generation and propagation, which is a bit similar to Twitter. We will further examine the prediction performance of our system on different types of OSNs, which offers platforms for particular purposes such as food sharing or tourism guiding. We will study how the users' social influence is correlated on the dominant and the emerging OSNs with a particular service.
- We will further study the evolvement of a user's social influence on emerging OSNs and explore different factors that play a role in driving a cold-start user to become an influential user.
- We wish to study different types of user classification problems from the perspective of cross-site user behavior analysis, such as the detection of malicious users [17, 20] and structural hole spanners [52], instead of merely focusing on the prediction of potential influential users.

ACKNOWLEDGMENTS

A preliminary version of this article has been published in Proceedings of the 13th CCF Conference on Computer Supported Cooperative Work and Social Computing (ChineseCSCW'18). Yang Chen is the corresponding author.

REFERENCES

- [1] Eytan Bakshy, Winter A. Mason, Jake M. Hofman, and Duncan J. Watts. 2011. Everyone is an influencer: Quantifying influence on Twitter. In *Proceedings of the ACM WSDM*.
- [2] Eytan Bakshy, Itamar Rosenn, Cameron Marlow, and Lada Adamic. 2012. The role of social networks in information diffusion. In *Proceedings of the WWW*.
- [3] Leo Breiman. 2001. Random forests. *Mach. Learn.* 45, 1 (2001), 5–32.
- [4] Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P. Krishna Gummadi. 2010. Measuring user influence in Twitter: The million follower fallacy. In *Proceedings of the AAAI ICWSM*.
- [5] Shiyu Chang, Yang Zhang, Wei Han, Mo Yu, Xiaoxiao Guo, Wei Tan, Xiaodong Cui, Michael J. Witbrock, Mark A. Hasegawa-Johnson, and Thomas S. Huang. 2017. Dilated recurrent neural networks. In *Proceedings of the NIPS*. 76–86.
- [6] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the ACM KDD*.
- [7] Yang Chen, Jiyao Hu, Yu Xiao, Xiang Li, and Pan Hui. 2020. Understanding the user behavior of Foursquare: A data-driven study on a global scale. *IEEE Trans. Comput. Soc. Syst.* 7, 4 (2020), 1019–1032.
- [8] Y. Chen, J. Hu, H. Zhao, Y. Xiao, and P. Hui. 2018. Measurement and analysis of the swarm social network with tens of millions of nodes. *IEEE Access* 6 (2018), 4547–4559.
- [9] Kyunghyun Cho, Bart van Merriënboer, Çağlar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the EMNLP*.
- [10] Robert B. Cialdini. 2008. *Influence: Science and Practice* (5th ed.). Prentice Hall.
- [11] Marijke De Veirman, Veroline Cauberghe, and Liselot Hudders. 2017. Marketing through Instagram influencers: the impact of number of followers and product divergence on brand attitude. *Int. J. Advert.* 36, 5 (2017), 798–828.
- [12] A. Farseev, L. Nie, M. Akbari, and T. Chua. 2015. Harvesting multiple sources for user profile learning: A big data study. In *Proceedings of the ACM ICMR*.
- [13] Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern Recog. Lett.* 27, 8 (2006), 861–874.
- [14] Maksym Gabielkov, Ashwin Rao, and Arnaud Legout. 2014. Studying social networks at scale: Macroscopic anatomy of the Twitter social graph. In *Proceedings of the ACM SIGMETRICS*.
- [15] Oana Goga, Patrick Loiseau, Robin Sommer, Renata Teixeira, and Krishna P. Gummadi. 2015. On the reliability of profile matching across large online social networks. In *Proceedings of the ACM KDD*.
- [16] Neil Zhenqiang Gong, Wenchang Xu, Ling Huang, et al. 2012. Evolution of social-attribute networks: Measurements, modeling, and implications using Google+. In *Proceedings of the ACM IMC*.

- [17] Qingyuan Gong, Yang Chen, Xinlei He, Zhou Zhuang, Tianyi Wang, Hong Huang, Xin Wang, and Xiaoming Fu. 2018. DeepScan: Exploiting deep learning for malicious account detection in location-based social networks. *IEEE Commun. Mag.* 56, 11 (2018), 21–27.
- [18] Qingyuan Gong, Yang Chen, Jiyao Hu, Qiang Cao, Pan Hui, and Xin Wang. 2018. Understanding cross-site linking in online social networks. *ACM Trans. Web* 12, 4 (2018), 25:1–25:29.
- [19] Qingyuan Gong, Yang Chen, Xiaolong Yu, Chao Xu, Zhichun Guo, Yu Xiao, Fehmi Ben Abdesslem, Xin Wang, and Pan Hui. 2019. Exploring the power of social hub services. *World Wide Web: Internet Web Inf. Syst.* 22, 6 (2019), 2825–2852.
- [20] Qingyuan Gong, Jiayun Zhang, Yang Chen, Qi Li, Yu Xiao, Xin Wang, and Pan Hui. 2019. Detecting malicious accounts in online developer communities using deep learning. In *Proceedings of the ACM CIKM*.
- [21] Roberto Gonzalez, Ruben Cuevas, Reza Motamedi, Reza Rejaie, and Angel Cuevas. 2013. Google+ or Google-?: Dissecting the evolution of the new OSN in its first year. In *Proceedings of the WWW*.
- [22] A. Graves and J. Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM networks. In *Proceedings of the IEEE International Joint Conference on Neural Networks*.
- [23] Jinyoung Han, Daejin Choi, Byung-Gon Chun, et al. 2014. Collecting, organizing, and sharing pins in Pinterest: Interest-driven or social-driven? In *Proceedings of the ACM SIGMETRICS*.
- [24] Jorge E. Hirsch. 2005. An index to quantify an individual’s scientific research output. *Proc. Nat. Acad. Sci. United States Amer.* 102, 46 (2005), 16569–16572.
- [25] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.* 9, 8 (Nov. 1997), 1735–1780.
- [26] Paridhi Jain, Ponnurangam Kumaraguru, and Anupam Joshi. 2016. Other times, other values: Leveraging attribute history to link user profiles across online social networks. *Soc. Netw. Anal. Mining* 6, 1 (2016), 85.
- [27] Yongpo Jia, Xuemeng Song, Jingbo Zhou, Li Liu, Liqiang Nie, and David S. Rosenblum. 2016. Fusing social networks with deep learning for volunteerism tendency prediction. In *Proceedings of the AAAI*.
- [28] Long Jin, Yang Chen, Tianyi Wang, Pan Hui, and Athanasios V. Vasilakos. 2013. Understanding user behavior in online social networks: A survey. *IEEE Commun. Mag.* 51, 9 (2013), 144–150.
- [29] Rafal Jozefowicz, Wojciech Zaremba, and Ilya Sutskever. 2015. An empirical exploration of recurrent network architectures. In *Proceedings of the ICML*.
- [30] X. Kong, J. Zhang, and P. Yu. 2013. Inferring anchor links across multiple heterogeneous social networks. In *Proceedings of the ACM CIKM*.
- [31] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media? In *Proceedings of the WWW*.
- [32] Jianxin Li, Xinjue Wang, Ke Deng, Xiaochun Yang, Timos Sellis, and Jeffrey Xu Yu. 2017. Most influential community search over large social networks. In *Proceedings of the IEEE ICDE*.
- [33] Qi Liu, Biao Xiang, Nicholas Jing Yuan, Enhong Chen, Hui Xiong, Yi Zheng, and Yu Yang. 2017. An influence propagation view of pagerank. *ACM Trans. Knowl. Discov. Data* 11, 3 (Aug. 2017), 30:1–30:30.
- [34] Siyuan Liu, Shuhui Wang, Feida Zhu, Jinbo Zhang, and Ramayya Krishnan. 2014. HYDRA: Large-scale social identity linkage via heterogeneous behavior modeling. In *Proceedings of the ACM SIGMOD*.
- [35] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J. Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. In *Proceedings of the IJCAI*.
- [36] Larry Medsker and Lakhmi C. Jain. 1999. *Recurrent Neural Networks: Design and Applications* (1st ed.). CRC Press, Inc., Boca Raton, FL.
- [37] P. Meo, E. Ferrara, F. Abel, et al. 2014. Analyzing user behavior across social sharing environments. *ACM Trans. Intell. Syst. Technol.* 5, 1 (2014), 14:1–14:31.
- [38] Daniel Neil, Michael Pfeiffer, and Shih-Chii Liu. 2016. Phased LSTM: Accelerating recurrent network training for long or event-based sequences. In *Proceedings of the NIPS*.
- [39] Liudmila Ostroumova Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: Unbiased boosting with categorical features. In *Proceedings of the NeurIPS*.
- [40] J. Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA.
- [41] Daniel M. Romero, Wojciech Galuba, Sitaram Asur, and Bernardo A. Huberman. 2011. Influence and passivity in social media. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer.
- [42] Guojie Song, Yuanhao Li, Xiaodong Chen, Xinran He, and Jie Tang. 2017. Influential node tracking on dynamic social network: An interchange greedy approach. *IEEE Trans. Knowl. Data Eng.* 29, 2 (2017), 359–372.
- [43] Charles Spearman. 1904. The proof and measurement of association between two things. *Amer. J. Psychol.* 15, 1 (1904), 72–101.
- [44] Jie Tang, Tiancheng Lou, and Jon Kleinberg. 2012. Inferring social ties across heterogenous networks. In *Proceedings of the ACM WSDM*.

- [45] Sonja Utz. 2010. Show me your friends and I will tell you what type of person you are: How one's profile, number of friends, and type of friends influence impression formation on social network sites. *J. Comput.-mediat. Commun.* 15, 2 (2010), 314–335.
- [46] Giridhari Venkatadri, Oana Goga, Changtao Zhong, Bimal Viswanath, Krishna P. Gummadi, and Nishanth Sastry. 2016. Strengthening weak identities through inter-domain trust transfer. In *Proceedings of the WWW*.
- [47] Gang Wang, Tianyi Wang, Haitao Zheng, and Ben Y. Zhao. 2014. Man vs. machine: Practical adversarial detection of malicious crowdsourcing workers. In *Proceedings of the USENIX Security*.
- [48] Xinghua Wang, Zhaohui Peng, Senzhang Wang, Philip S. Yu, Wenjing Fu, and Xiaoguang Hong. 2018. Cross-domain recommendation for cold-start users via neighborhood based feature mapping. In *Proceedings of the DASFAA*.
- [49] Bernard Lewis Welch. 1951. On the comparison of several mean values: An alternative approach. *Biometrika* 38, 3/4 (1951), 330–336.
- [50] Christo Wilson, Bryce Boe, Alessandra Sala, et al. 2009. User interactions in social networks and their implications. In *Proceedings of the ACM EuroSys*.
- [51] Rong Xie, Yang Chen, Qinge Xie, Yu Xiao, and Xin Wang. 2018. We know your preferences in new cities: Mining and modeling the behavior of travelers. *IEEE Commun. Mag.* 56, 11 (2018), 28–35.
- [52] Wenzheng Xu, Mojtaba Rezvani, Weifa Liang, Jeffrey Xu Yu, and Chengfei Liu. 2017. Efficient algorithms for the identification of top- k structural hole spanners in large social networks. *IEEE Trans. Knowl. Data Eng.* 29, 5 (2017), 1017–1030.
- [53] Yiming Yang and Jan O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proceedings of the ICML*.
- [54] Yuanshun Yao, Bimal Viswanath, Jenna Cryan, Haitao Zheng, and Ben Y. Zhao. 2017. Automated crowdturfing attacks and defenses in online review systems. In *Proceedings of the ACM CCS*.
- [55] Shaozhi Ye and Felix Wu. 2013. Measuring message propagation and social influence on twitter.com. *Int. J. Commun. Netw. Distrib. Syst.* 11, 1 (2013), 59–76.
- [56] Hema Yoganarasimhan. 2012. Impact of social network structure on content propagation: A study using YouTube data. *Quant. Market. Econ.* 10, 1 (2012), 111–150.
- [57] Peng Zhang, Haiyi Zhu, Tun Lu, Hansu Gu, Wenjian Huang, and Ning Gu. 2017. Understanding relationship overlapping on social network sites: A case study of Weibo and Douban. *Proc. ACM Hum.-comput. Interact.* 1, CSCW (2017), 120:1–120:18.
- [58] Changtao Zhong, Mostafa Salehi, Sunil Shah, Marius Cobzarenco, Nishanth Sastry, and Meeyoung Cha. 2014. Social bootstrapping: How Pinterest and Last.fm social communities benefit by borrowing links from Facebook. In *Proceedings of the WWW*.

Received May 2019; revised February 2020; accepted June 2020