IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 April 2022, accepted 23 April 2022, date of publication 2 May 2022, date of current version 2 September 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3171846

RESEARCH ARTICLE

Social Spammer Detection via Convex Nonnegative Matrix Factorization

HUA SHEN[1,2,3], BANGYU WANG[2], XINYUE LIU[2], AND XIANCHAO ZHANG[2]
[1]Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China
[2]School of Software, Dalian University of Technology, Dalian 116620, China
[3]College of Mathematics and Information Science, Anshan Normal University, Anshan 114007, China

Corresponding author: Xianchao Zhang (xczhang@dlut.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61272374 and Grant 61300190, in part by the Key Project of Chinese Ministry of Education under Grant 313011, and in part by the Foundation of Department of Education of Liaoning Province under Grant L2015001.

ABSTRACT With the increasing popularity of social network platforms such as Twitter and Sina Weibo, a lot of malicious users, also known as social spammers, disseminate illegal information to normal users. Several approaches are proposed to detect spammers by training a classifier with optimization methods and mainly using content and social following information. Due to the development of spammers' strategies and the courtesy of some legitimate users, social following information becomes vulnerable to fake by spammers. Meanwhile, the possible social activities and behaviors vary significantly among different users, which leads to a large yet sparse feature space to be modeled by existing approaches. To address issues, in this paper, we propose a new approach named CNMFSD for spammer detection in social networks, which exploits both content information and users interaction relationships in an innovative manner. We have empirically validated the proposed method on a real-world Twitter dataset, and experimental results show that the proposed CNMFSD method improves the detection performance significantly compared with baselines.

INDEX TERMS Social spammer detection, matrix factorization, social regularization term.

I. INTRODUCTION

Social networks, such as Twitter, Facebook, and Sina Weibo, are increasingly used to disseminate and share information easily and quickly. However, it is a double-edged sword since the success of social networks also attracts more social spammers [1]. They try to seize our privacy, send us unwanted information, publish malicious content and links [2], and promote commodity information, which thoroughly impacts social stability and organizational management models [3]. According to a study by Nexgate [4], the number of social spammers grows so fast that one in two hundred social messages is spam. Meanwhile, to increase their influence and be undetected, spammers collude with each other to construct the criminal communities [5]. Thus, social spammer detection is a challenging task for researchers. Successful social spammer detection presents its significance to improve the quality of user experience, and positively impact the overall value of the social systems going forward [6].

In the past decade, researchers have tried different techniques to detect spammers, such as link analysis [7] and content analysis [8], [9]. The methods of content-based detection of spammers mainly focus on analyzing and extracting users' features and then directly applying existing classification approaches such as support vector machines (SVM) to detect spammers [9]–[11]. Recently, more advanced deep learning-based approaches have been proposed to detect social spammers only based on content [12]–[14]. However, with the development of spamming strategies, these methods could not accurately detect spammers with new strategies, only relying on the extracted features. Another category of methods is proposed to detect spammers via social network analysis [15]. These methods assume that spammers cannot establish an arbitrarily large number of social trust relations with legitimate users. The users, who have relatively low social influence or social status in social networks, will be determined as spammers. Unfortunately, only depending on

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

network information, these methods are hard to distinguish between legitimate users and spammers.

Some approaches [16]–[18] have been proposed to detect spammers via both content and network analysis, which identify spammers more accurately than the traditional approaches. The main challenge in detecting social spammers is that the possible social activities and behaviors are more varied and complex, and they constitute a much larger feature space. As a result, spammers are more challenging to detect. Therefore, it is crucial to design more effective methods for extracting users' features. Meanwhile, the reflexive reciprocity [19] indicates that many users simply follow back when they are followed by someone for the sake of courtesy. It is easier for spammers to acquire a large number of follower links in social networks. Thus, with the perceived social influence, they can avoid being detected. However, the interactions between spammers and legitimate users are usually unilateral. In most cases, spammers share a message and then mention (i.e., @) legitimate users. On the contrary, legitimate users constantly interact with legitimate users but have few interactions with spammers. Consequently, it is more reasonable to take the interactions among users into consideration when detecting spammers.

To address these challenges, we propose to take advantage of both social network interaction information and content information.

In this paper, we study the problem of social spammer detection with social interaction and content information. In essence, we investigate: (1) how to model the social interaction information and content information properly; and (2) how to seamlessly utilize both social interaction and content information for the problem we are studying. Our solutions to these two challenges result in a novel spammer detection framework name Convex-NMF based Supervised Spammer Detection with Social Interaction (CNMFSD). Based on statistical analysis, we observe that spammers and legitimate users have different characteristic distributions. Thus, we use a matrix factorization model to collaboratively induce a succinct set of latent features for spammers and legitimate users, respectively, and this latent feature learning process is guided by the label information. The generated features are then used as the input representation for a spammer classification model. Then we refine the latent features with predicted label information and social interaction information. Finally, the refined latent features are used as the input representation for the final classification. The main contributions of this paper are outlined as follows:

- We propose a three-stage optimization model that conducts feature extraction and classifier learning simultaneously. First, we use Convex Non-negative Matrix Factorization (CNMF) [20], [21] and Non-negative Matrix Factorization (NMF) to induce latent feature from content information, then train an SVM classifier and finally, refine latent features using social interaction information as the input representations of the classifier. Through iteratively learning among content information, social interaction regularization, and classification model, the proposed method can train an accurate classifier.
- We propose a novel method to induce latent features and a novel social interaction regularization term. Using CNMF, we get the latent content matrix of spammers and legitimate users, respectively, and then obtain the user feature latent matrix by NMF according to the latent content matrix. The latent feature refine process is guided by the social interaction relationship matrix and the label information.
- We evaluate our method on a large-scale real-world social network data set from Twitter, one of the largest social networks in the world. The experimental results show that the proposed framework can identify more spammers compared with baseline approaches. We conduct experiments to demonstrate the significance of using CNMF to induce latent features for spammers and legitimate users, respectively, and validate the effectiveness of the new social interaction regularization term.

The structure of this paper is organized as follows. In Section II, we review existing work in social spam detection. In Section III, we formally define the problem of social spammer detection with content and social interaction information. In Section IV, we propose a new model to integrate both content and social interaction information for spammer detection. In Section V, we report empirical results on a real-world dataset. Finally, we conclude and present the future work in Section VI.

II. RELATED WORK

Many different methods have been proposed to combat social spammers since Heymann *et al.* [22] firstly surveyed potential solutions and challenges in social spammer detection. Masood *et al.* [6] elaborated a classification of spammer detection techniques, including fake content, URL-based spam detection, detecting spam in trending topics, and fake user identification. In this paper, we only focus on the binary classification task, i.e., spammer or legitimate user identification.

Many approaches employed machine learning methods to train a classifier to detect spammers. SMFSR [16] jointly modeled user activities' information and the social following information to learn a classifier. SSDM [17] incorporated users' text information and social following information into an efficient spare supervised model for spammer detection. Mateen *et al.* [23] proposed a hybrid technique that utilizes user-based, content-based, and graph-based characteristics for spammer profiles detection. Gupta *et al.* [24] presented a policy for the detection of spammers on Twitter and used the popular techniques, i.e., Naive Bayes, clustering, and decision tree.

An important line of research in spam detection relies on analyzing the tweet content, as shown in [25] and [26] where suspicious use of hashtags or URLs is traced. The main objective in [26] is to study the semantics of short texts or messages in contrast with a set of Wikipedia text pages that are modeled and used as an aggregation of entities. The work presented in [25] stresses the need for efficient URL detection schemes utilizing different features such as lexical ones and dynamic behaviors.

Other directions adopted in detecting Twitter spammers focus on discovering traits or patterns that best describe the spammer's behavioral profile. In such works like [27], the main contribution is to determine deceptive double characters for user profiles, which is done by analyzing non-verbal behavior variables as a function of time, such as follows and retweets. Also, Sumner *et al.* [28] follow a similar technique. Direct approaches to checking up the user's portfolio include, but are not limited to, the notion of having no profile photo/biography/personal tweets or a suspiciously high/low number of followers/followees. Examples of different profile-based behavior analysis activities are demonstrated in [29] and [30].

Different from discovering traits or patterns, some work considers social network information to identify spammers. Ghosh *et al.* [31] investigated link farming on Twitter and proposed a ranking scheme to deter spam. Yang *et al.* [32] proposed a criminal account inference algorithm by exploiting criminal accounts' social relationships. Cao *et al.* [33] presented the SybilRank algorithm relying on social graph properties to rank users. Cui *et al.* [34] proposed a Hybrid Factor Non-Negative Matrix Factorization method to incorporate the predictive factors for user-post specific social influence prediction.

In addition, there is other research work apply deep learning techniques to detect social spam. Wu *et al.* [35] applied a deep learning technique to identify spam on Twitter by learning the syntax of many tweets using the word vector technique to perform pre-processing and create high-dimension vectors. Selvaganapathy *et al.* [36] used a deep neural network technique and a feature selection Boltzmann machine for detecting and classifying malicious URLs. Ban *et al.* [14] extracted deep learned features from Twitter text automatically using the DL-based Bi-LSTM technique for spam detection. Alom *et al.* [13] proposed an approach based on deep learning techniques, which leveraged both tweet text as well as users' meta-data (e.g., number of followings/followers, and so on) to detect spammers. DeepSBD [12] is a state-of-the-art deep learning approach to detect social spammers using deep learning approaches via aggregating different types of features, including profile, temporal behavior, activity sequence behavior, and content behavior. However, this work ignores the importance of social relations among users, which would help the model improve prediction performance.

III. PROBLEM DEFINITION

In this section, we introduce the notations used in this paper firstly and then formally define the problem that we study.

A. USER SETS

Let $\mathcal{U} = \{u_i^l\}_{i=1}^m$ denote the labeled user set, where m is the number of labeled users, and $\mathcal{Y} = \{y_i\}_{i=1}^m$ denote the corresponding label set. Let $\mathcal{U}' = \{u_i'\}_{i=1}^{m'}$ denote the unlabeled user set, where m' is the number of unlabeled users.

B. USER CONTENT MATRICES

Each user $u_i \in \{\mathcal{U}, \mathcal{U}'\}$ posts a set of tweets, and we can extract a content vector $\mathbf{x}_i \in \mathbb{R}^n$ for each u_i, where n is the number of content features. Let $X \in \mathbb{R}^{n\times m}$ denote the content feature matrix for labeled users and $X' \in \mathbb{R}^{n\times m'}$ be the content feature matrix for unlabeled users.

C. USER-USER INTERACTION

Each user u_i can interact with other users, and we use R denote the user-user interaction matrix, where R_{ij} represents the number of mentions (i.e., @) from u_i to u_j. Note that in this paper, we aim to model the interactions among users as well as consider the users' types. Thus, we extract the interaction graph among all users, i.e., $R \in \mathbb{R}^{(m+m')\times(m+m')}$.

Based on the given notations, we formally define the problem of social spammer detection as follows:

D. SOCIAL SPAM DETECTION TASK

Given a set of labeled users $\mathcal{U}$, content information X of labeled users, identity label information $\mathcal{Y}$ of labeled users, social network interaction information R, unlabeled user set $\mathcal{U}'$ and their content feature matrix X', our task is to learn a classifier with parameter set W to automatically classify each user $u_i' \in \mathcal{U}'$ as a spammer or legitimate user, i.e., $y_i' \in \{-1, 1\}$.

IV. THE PROPOSED APPROACH

The proposed approach CNMFSD is shown in Figure 1. The basic idea of CNMFSD is taking advantage of content information and user interaction relationships to train an effective classifier to detect spammers. In particular, we take the labeled legitimate user content matrix X_N and the labeled spammer content matrix X_S into consideration, respectively, where $X = [X_N; X_S]$. We use the CNMF technique to extract latent content features $U = \{U_N, U_S\}$ and the latent user features $V = \{V_N, V_S\}$. After that, we use V to train an accurate classifier W by considering the interaction relationships (i.e., R) among labeled users. Besides, we use the learned U and the NMF technique to extract the user feature matrix V' for unlabeled users $\mathcal{U}'$ based on the user content matrix X'. Based on the trained classifier W and the learned V', each user u_i' will be assigned a label y_i'.

Next, we introduce the details of the proposed CNMFSD. We first introduce the matrix factorization technique for the induction of latent features from content information and the classifier for detecting spammers and then propose a new social regularization term to model social network interaction information. Finally, a novel three-stage spammer detection framework is proposed, which integrates

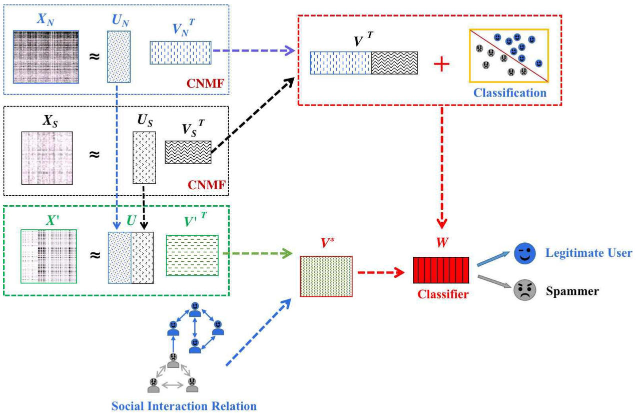

FIGURE 1. Overview of the proposed CNMFSD.

content information learning, social regularization, and classification.

A. MODELING CONTENT INFORMATION

Since the content information is sparse, building models using the original content information may fail to predict the label of the user precisely. To solve this issue, we propose to factorize the content information matrix X of labeled users into two latent matrices, U and V, which represent the latent factors of users and content, respectively.

Intuitively, the content posted by spammers and legitimate users may be different. To capture the difference in the characteristic distributions, we propose to separately consider them when conducting the factorization. Specifically, the factorization form of CNMF is as follows:

$$X_N \simeq (X_N G_N) V_N^T, \quad U_N = X_N G_N \tag{1}$$

$$X_S \simeq (X_S G_S) V_S^T, \quad U_S = X_S G_S \tag{2}$$

$$V = \{V_N, V_S\}, \quad U = \{U_N, U_S\} \tag{3}$$

where $X_N \in \mathbb{R}^{n \times m_N}$ represents the content matrix of the legitimate users, $X_S \in \mathbb{R}^{n \times m_S}$ represents the content matrix of the spammers, and $m_N + m_S = m$. $V_N \in \mathbb{R}^{m_N \times k}$ is latent *user-feature* matrix of legitimate users, $U_N \in \mathbb{R}^{n \times k}$ is the *content latent* matrix of legitimate users, $V_S \in \mathbb{R}^{m_S \times k}$ is spammer features latent matrix, and $U_S \in \mathbb{R}^{n \times k}$ is spammer content latent matrix.

The objective function of CNMF [21] is as follow:

$$\begin{aligned}\mathcal{O}_C &= \|X_N - (X_N G_N) V_N^\top\|_F^2 + \|X_S - (X_S G_S) V_S^\top\|_F^2 \\ s.t.\ & G_N \geq 0, \quad V_N \geq 0,\ G_S \geq 0,\ V_S \geq 0\end{aligned} \tag{4}$$

The reason that we choose CNMF to factorize the content matrix X, instead of using NMF, is that CNMF imposes the constraint by defining U lie within the column space of X. This constraint restricts U to convex combinations of the columns of X so that we could interpret the columns of U as weighted sums of certain data points, which may be helpful to induce user latent features. Meanwhile, another advantage of CNMF is that it can effectively reduce the possibility of overfitting.

B. CLASSIFICATION MODEL

Using the latent features V_i for each labeled user u_i^l, we can train a supervised classifier. We choose the hinge loss used in Support Vector Machines (SVM) [37] as the loss function. To make the gradient computation and the subsequent optimization more tractable, we use the following smoothed hinge loss

$$h(z) = \begin{cases} \frac{1}{2} - z & z \leq 0 \\ \frac{1}{2}(1-z)^2 & 0 < z < 1 \\ 0 & z \geq 1. \end{cases} \tag{5}$$

We then define the classifier based on the latent features, and the objective function builds on Eq. (5) with SVM as follows:

$$\mathcal{O}_W = \sum_{i=1}^{m} h(y_i V_i^\top W) + \frac{\lambda_W}{2} \|W\|_2^2, \tag{6}$$

where λ_W is the regularization coefficient, $W \in \mathbb{R}^k$ is the classifier parameter, $y_i \in \{1, -1\}$ is the label information. If $y_i = 1$, the user u_i^l is a spammer; otherwise, u_i^l is a legitimate user. Note that, although we apply a linear classifier $V_i^\top W$ to the latent user features, non-linear classifiers can be applied using kernel trick.

C. UNLABELED USER CONTENT FACTORIZATION

To extract the latent features from the unlabeled user content matrix X', a naive approach is still applying CNMF. However, this approach may result in extracting latent user features randomly due to the lack of matrix factorization guidance. To address this issue, we propose to directly use U_N and U_S as the base content features, which are optimized under the guidance of the label information. Besides, we use NMF to conduct the matrix factorization as follows:

$$\begin{aligned} X' &\simeq UV'^\top \\ UV'^\top &= (U^\top U)^{-1} U^\top X' \end{aligned} \tag{7}$$

where $U \in \mathbb{R}^{n \times 2k}$ is the column-wise concatenation of U_N and U_S, and $V' \in \mathbb{R}^{m' \times 2k}$ is the latent feature matrix for unlabeled users. Correspondingly, we have the loss function for the unlabeled user content factorization as follows:

$$\mathcal{O}'_C = \|X' - UV'^\top\|_F^2. \tag{8}$$

D. LABEL PREDICTION FOR UNLABELED DATA

For each unlabeled user u'_i, we use both the learned latent feature $V'_i \in \mathbb{R}^{2k}$ using Eq. (8) and the classifier parameter vector $W \in \mathbb{R}^k$ via Eq. (6). Different from existing work, we cannot directly multiply them together to obtain the label. Specifically, we divide $V'_i \in \mathbb{R}^{2k}$ into two vectors $V'_{Ni} \in \mathbb{R}^k$ and $V'_{Si} \in \mathbb{R}^k$, where V'_{Ni} denotes the latent feature component learned from legitimate users, and V'_{Si} denotes the latent feature component learned from spammers. Then we use the following rules to determine the label of an unlabeled user:

$$y'_i = \begin{cases} 1 & if\ V'^{\top}_{Ni} W + V'^{\top}_{Si} W > 0; \\ -1 & otherwise. \end{cases} \tag{9}$$

E. SOCIAL REGULARIZATION

Using Eq. (9) can obtain the prediction label for each unlabeled data. However, users are connected and interact in the social networks, and the links between users reflect users' closeness. One intuitive solution is to exploit users' social relations to regularize the matrix factorization for modeling those interactions. Previous studies have shown that users' social relationships could be exploited to regularize the decomposition of the feature matrix [16], [17], which improves the performance of identifying spammers.

According to our previous research work [38], we have four kinds of following relations between users, which are shown in Figure 2(a): [spammer, spammer], [legitimate user, legitimate user], [legitimate user, spammer], and [spammer, legitimate user]. Figure 2(b) shows the social relationships used in [16]. They claimed that the latent features of spammers are different from their neighbors in the social network significantly, and conversely, legitimate users usually have similar latent features with their friends since they share similar interests and may perform similar social activities. Therefore, they only considered two kinds of relationships: [legitimate user, legitimate user] and [spammer, legitimate user]. Besides, in [17], they considered three types of relationships: [spammer, spammer], [legitimate user, legitimate user], and[legitimate user, spammer](Figure 2(c)), since the fourth relation can be easily faked by spammers.

However, with the development of spamming strategies, spammers collude with each other to construct the criminal communities for hiding themselves, and many legitimate users may follow spammers out of courtesy [19], which means that the second and third relationships are no longer reliable. To avoid this issue, in this paper, we propose to use the interaction relationships (e.g., the *mention* action "@") instead of the following relationships. This motivation is based on one observation that there is almost no [legitimate user, spammer] in interaction relationships (Figure 2(d)).

Based on the above analysis, we all know that it is helpful to take advantage of the interaction relationships, and correspondingly, we propose the following social regularization term for predicted label calibration based on the learned latent user features as follows:

$$\mathcal{R}_S = \frac{1}{2}\sum_{i=1}^{m+m'}\sum_{j=1}^{m+m'} \hat{y}_i\hat{y}_j a_{ij}||\hat{V}_i - \hat{V}_j||_2^2. \tag{10}$$

where $\hat{V}_i$ and $\hat{V}_j \in \{V, V'\}$ are the learned user latent features, and $\hat{y}_i$ and $\hat{y}_j$ are the corresponding labels. a_{ij} is a weight factor to measure the relative interaction frequency between u_i and u_j, which is defined as follows:

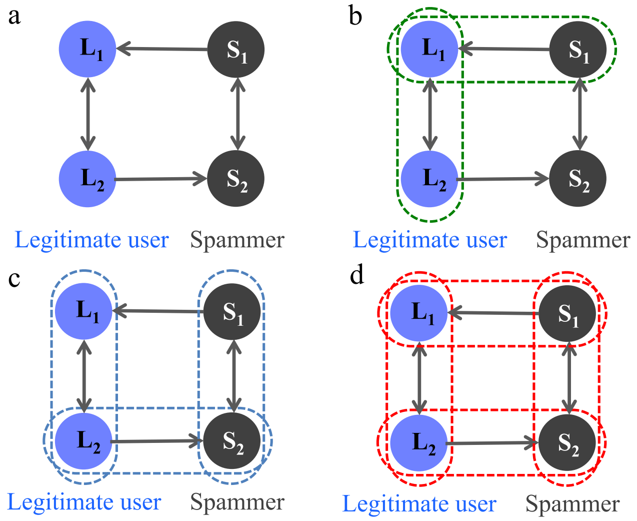

FIGURE 2. Social relationships among users [38].

$$a_{ij} = \begin{cases} 0 & if\ \min(R_{ij}, R_{ji}) = 0 \\ \dfrac{2 * \min(R_{ij}, R_{ji})}{R_{ij} + R_{ji}} & if\ \min(R_{ij}, R_{ji}) > 0 \end{cases} \tag{11}$$

where R_{ij} is the number of mentions from u_i to u_j. We can see that if both user u_i and u_j belong to the same class and $a_{ij} \neq 0$, the latent feature distance of these two users would be reduced; otherwise, the distance will be enlarged, which is in accord with our intuition. This social regularization term is used for calibrating the predicted labels and helping to learn better user latent features V^* with the following objective function:

$$\mathcal{O}_S = \frac{\lambda_{\mathrm{NMF}}}{2}\mathcal{O}'_C + \frac{\lambda_R}{2}\mathcal{R}_s + \frac{\lambda_{V'}}{2}||V'||_F^2. \tag{12}$$

Using the learned V^*, we can predict the labels again using Eq. (9).

F. OPTIMIZATION

To optimize the proposed CNMFSD, we first minimize $\mathcal{O}_C$ to get U and V based on the labeled data. Then, using the gradient descent method to minimize $\mathcal{O}_W$ to update W by fixing V. Given U, we get V' using the least square method. Finally, we minimize $\mathcal{O}_S$ to update V' to get V^* using R. The details are introduced in the following.

1) COMPUTING U AND V

Since $U = XG$, we need to get G firstly. Although the latent matrix of spammers and legitimate users are induced, respectively, their optimization methods are the same. Following the alternating updating rules introduced in [21], we can use the following updating rules to solve G and V.

$$V_{ij} \leftarrow V_{ij}\sqrt{\frac{(X^T XG)_{ij}}{(VG^T X^T XG)_{ij}}} = V_{ij}\sqrt{\frac{(X^T U)_{ij}}{(VU^T U)_{ij}}} \tag{13}$$

$$G_{ij} \leftarrow G_{ij}\sqrt{\frac{(X^TX)_{ij}}{(X^TXGV^TV)_{ij}}} \tag{14}$$

2) COMPUTING W

Since W is only related to V and $\mathcal{Y}$, it is easy to solve W using stochastic gradient descent (SGD). We derive the gradients of W as follows.

$$\frac{\partial \mathcal{O}_W}{\partial W} = \sum\nolimits_{i=1}^{l} y_i V_i^T h'(y_i V_i W) + \lambda_W W \tag{15}$$

where the gradient of the smoothing hinge loss $h(z)$ is

$$h'(z) = \begin{cases} -1 & z \leq 0 \\ z-1 & 0 < z < 1 \\ 0 & z \geq 1 \end{cases} \tag{16}$$

The solutions of W lead to the following update rules:

$$W \leftarrow W - \eta\frac{\partial \mathcal{O}_W}{\partial W} \tag{17}$$

where η is the learning rate in the procedure of stochastic gradient descent.

3) COMPUTING V'

Since $X \simeq UV'$ and we have gotten U, we can use least squares to update V'. The updating rule is as follow.

$$V'_{ij} \leftarrow V'_{ij}\sqrt{\frac{(X'^\top U)_{ij}}{(V'U^\top U)_{ij}}} \tag{18}$$

4) COMPUTING V^*

We minimize $\mathcal{O}_S$ to update V' by using user interaction information. After the convergence of objective function, we get the latent features V^*, which equals to V' at the last iteration. Firstly, we show another form of R_S.

$$\begin{aligned} R_S &= \frac{1}{2}\sum_i\sum_j \hat{y}_i\hat{y}_j a_{ij}||\hat{V}_i - \hat{V}_j||_2^2 = Tr(V'\Delta V'^\top) \\ \Delta &= D_Y(D_S - S)D_Y \end{aligned} \tag{19}$$

where $S = (a_{ij}) \in \mathbb{R}^{(m+m')\times(m+m')}$, D_Y is diagonal matrix whose diagonal element is y_i, D_S is the diagonal matrix whose diagonal element is the sum of the corresponding row of S. Note that this term considers both labeled and unlabeled data.

Then we derive the gradients of V^* as follows.

$$\frac{\partial \mathcal{O}_S}{\partial V'} = \lambda_{\text{NMF}}(-2X'^\top U + 2V'U^\top U) + \lambda_R V'\Delta + \lambda_{V'}V' \tag{20}$$

The solutions of V' leads to the following update rules:

$$V'_{ij} \leftarrow V'_{ij}\sqrt{\frac{(\lambda_{\text{NMF}}X'^\top U)_{ij} + (\lambda_R V'\Delta)^-_{ij}}{(\lambda_{\text{NMF}}V'U^\top U)_{ij} + (\lambda_R V'\Delta)^+_{ij} + \lambda_{V'}V'}} \tag{21}$$

where A^+ and A^- are negative matrix, and $A = A^+ - A^-$.

The the proposed algorithm for social spammer detection can be found in Algorithm 1.

Algorithm 1 CNMFSD Algorithm

Input: Training content information $X = X_N, X_S$; unlabeled content information X'; user interaction relationship matrix R; hyperparameters $\lambda_{\text{NMF}}, \lambda_R, \lambda_{V'}, \lambda_W$; labels $\mathcal{Y}$; size of latent features k; learning rate η

```
Initialize $U = \{U_N, U_S\}$, $V = \{V_N, V_S\}$, $W$;
while not converged do
    Update $V_N$ according to Eq.(13);
    Update $G_N$ according to Eq.(14);
end while
$U_N = XG_N$;
while not converged do
    Update $V_S$ according to Eq.(13);
    Update $G_S$ according to Eq.(14);
end while
$U_S = XG_S$;
$U = \{U_N, U_S\}$;
$V = \{V_N, V_S\}$;
while not converged do
    Update $W$ according to Eq.(17);
end while
while not converged do
    Update $V'$ according to Eq.(18);
end while
while not converged do
    Update $V'$ according to Eq.(20);
end while
$V^* = V'$
```

Output: Content latent matrix in spammer and legitimate user space $U = \{U_N, U_S\}$; Unlabeled user latent features matrix V^*; Classification weight matrix W

V. EXPERIMENTS

In this section, we conduct extensive experiments to illustrate the effectiveness and efficiency of the proposed framework. We begin by introducing the dataset and experimental setup. Next, experimental results on spammer detection are shown that the performance of CNMFSD is better than that of baselines. Finally, we qualitatively analyze the importance of CNMF and social regularization.

In particular, we want to answer the following three research questions in our experiments:

- **RQ1**: Does the proposed CNMFSD outperform state-of-the-art baselines for the social spam detection task?
- **RQ2**: Is CNMF better than NMF for extracting latent user features?
- **RQ3**: Is the designed social regularization term effective for social spam detection?

A. DATASET

The goal of this paper is to detect social spammers using both user tweet content and interactions among users. To achieve this goal, we use a real-world Twitter dataset in our experiments. The dataset is constructed from two publicly available datasets, which are Twitter social honeypot dataset [39] and the Kwak's dataset [40]. The Twitter social honeypot dataset provides the ground truth label for each user, and Kwak's dataset contains user tweets. In particular, we extract common users from both datasets and the corresponding user tweets and then filter out the non-English tweets and the users who posted less than one tweet. Finally, there are 7,450 users (3,012 spammers and 4,438 legitimate users)in our dataset. By analyzing the mentioned actions among users' tweets, we extract the social interaction relationships. Since each user posted more than one tweet, we aggregate all the posted tweets and use tf-idf to extract content features, and the size of the content feature is 4,144.

B. BASELINES AND EXPERIMENTAL SETUP

To fairly compare the proposed CNMFSD with baselines (RQ1), we use the following traditional and state-of-the-art approaches as baselines:

- **SVM** [37]. We directly train a classifier on the content matrix X of the labeled users and then apply it to the unlabeled data X' to make predictions.
- **SMFSR** [16]. We perform a joint optimization model that simultaneously uses the matrix factorization technique on the concatenated X and X', which is guided by the social relationship graph and the label information. The learned classifier is used to predict the unlabeled users.
- **SSDM** [17]. We employ a directed Laplacian formulation to model the social network and then integrate the network information into a sparse supervised formulation for the modeling of content information.
- **LSTM$_{avg}$**. Long short-term memory network (LSTM) [41] is a commonly used deep learning approach to model text data. In particular, we use pre-trained Twitter word embeddings [42] on each user's original tweet to learn a hidden state and then average all the hidden states to learn an aggregated user representation, which is used to train the model or make a prediction.
- **LSTM$_{att}$**. Since each tweet may contribute unequally to the user representation learning, we use the attention mechanism [43] to distinguish the importance of tweets. After LSTM outputs the hidden state for each tweet, then we use a linear layer to learn the weight using the hidden state. Finally, the weighted sum of hidden states is taken as the user representation.
- **Transformer$_{avg}$**. We use a pre-trained language model [44] to learn the tweet representation (i.e., the output from *[cls]*). Then the averaged tweet representation from the same user is used as the user feature, which is further used to make predictions.
- **Transformer$_{att}$**. Similar to LSTM$_{att}$, we use the attention mechanism to assign weight to each tweet representation outputted from *[cls]*. Then the weighted representation is used as the user feature.
- **Deep-learnt** [14]. Deep-learnt uses Word2Vec [45] and Bi-LSTM (Bi-directional long short term memory [41]) to learn user embeddings, which are further used to predict the labels.
- **TextCNN** [13]. Following [46], A convolutional neural network (CNN) is used to learn tweet embeddings. Then the average of tweet embeddings from the same user represents the user's feature, which is used to make a prediction.
- **DeepSBD** [12]. This approach also employs a CNN model to learn user representations. However, it uses a low-level attention mechanism to learn user representations.

For RQ2, we aim to validate the effectiveness of using CNMF and design the following baselines:

- **NMF$_{svm}$**. We first concatenate X and X' and then apply NMF to extract latent user features. The latent features from the labeled data are used for training the SVM classifier, which is then used on the unlabeled user features to estimate the labels.
- **CNMF$_{svm}$**. The setting of CNMF$_{svm}$ is similar to that of NMF$_{svm}$, and the only difference is to use CNMF to extract the latent features.
- **NMF$_{svm}$+SI**. We use NMF$_{svm}$ method to get the latent feature matrix V' and the classifier W, and then employ the social interaction relationship matrix to refine V'.

In this work, we propose a new social regularization term (RQ3) for calibrating the predicted labels using **CNMF$_{svm}$** and the following baselines:

- **CNMF$_{svm}$+SR**. The method performs the same process of CNMF$_{svm}$ to train the latent features V' and the SVM classifier firstly and then uses the social following relationships (i.e., the links among users but without considering interaction frequency) to induce the latent features V^*.
- **CNMF$_{2-norm}$+SI**. The processes of factorizing the content information matrix X of labeled users into U and V and extracting the latent features V' from X' are similar to those of CNMFSD. Unlike CNMFSD, we use the two norms of latent features vector as the classifier instead of SVM. If the two norm of latent features vector in the legitimate user space is bigger than that in the spammer space, the user is predicted as a legitimate user. The refining process of V' is also guided by the social interaction relationship matrix.

Precision, recall, and F1-measure are used as the performance metrics. Also, we empirically set $\lambda_{\text{NMF}} = \lambda_R = \lambda_{V'} = \lambda_W = 0.1, \alpha = 0.1, k = 400$, and $\eta = 0.001$ in the following experiments. For deep learning-based baselines, we set the size of hidden state as 256 for LSTM.

C. PERFORMANCE EVALUATION (RQ1)

The experimental results of the proposed CNMFSD and baseline methods on the real-world Twitter dataset are presented in Table 1. To avoid effects brought by the size of the training

TABLE 1. Social spammer detection results.

Method	Training data (30%)			Training data (50%)			Training data (70%)		
	Precision	Recall	F_1-score	Precision	Recall	F_1-score	Precision	Recall	F_1-score
SVM [37]	0.609	0.786	0.686	0.627	0.794	0.701	0.644	0.802	0.714
SSDM [17]	0.623	0.864	0.724	0.644	0.885	0.746	0.684	0.883	0.771
SMFSR [16]	0.648	0.851	0.736	0.715	0.882	0.790	0.741	0.891	0.809
LSTM$_{avg}$ [41]	0.653	0.856	0.741	0.693	0.887	0.778	0.752	0.893	0.816
LSTM$_{att}$ [41]	0.696	0.863	0.771	0.722	0.872	0.790	0.769	0.886	0.823
Transformer$_{avg}$ [44]	0.702	0.856	0.771	0.725	0.871	0.791	0.774	0.882	0.824
Transformer$_{att}$ [44]	0.744	0.854	0.795	0.752	0.885	0.813	0.791	0.885	0.835
Deep-learnt [14]	0.740	0.849	0.791	0.753	0.876	0.810	0.798	0.852	0.824
TextCNN [13]	0.745	0.827	0.784	0.774	0.829	0.801	0.809	0.832	0.820
DeepSBD [12]	0.803	0.834	0.818	0.812	0.848	0.830	0.845	0.863	0.854
CNMFSD	**0.827**	**0.857**	**0.842**	**0.831**	**0.889**	**0.859**	**0.846**	**0.892**	**0.868**

data, we vary the size of training dataset to observe the performance's trend, where "Training Data (30%)" means that we randomly select 30% data from the whole dataset as the training set and the remaining 70% data as the testing set. In the experiments, each result denotes an average of 10 runs. From the results in Table 1, we can observe that our proposed framework CNMFSD consistently outperforms other baseline methods using all metrics with different sizes of training data, especially on "Training Data (70%)".

The basic baseline is SVM, which only uses the original extracted feature X as the input to train a classifier. However, X is very sparse and noisy, which leads to the worst performance among all the baselines.

The time complexity of the proposed CNMFSD is mainly calculating V, G, and V', which is $O(nmk + 2nm'k)$. For SSDM and SMFSR, the time complexity is the same, which is $O(n(m+m')k)$. The time complexity of these three algorithms is at the same level. Compared with SSDM and SMFSR, the precision of CNMFSD is greater over 10% than theirs, and the F1-score of CNMFSD is better than theirs (over 6%) on different training sets. The recall of CNMFSD is also better than these methods' on the larger size of the training set. Though SSDM and SMFSR methods consider the social following relationships and can identify spammers effectively, they ignore that spammers collude with each other to construct the criminal communities. Besides, many legitimate users may follow spammers out of courtesy, which means that the following relationships are no longer reliable. Therefore, SSDM and SMFSR cannot achieve better performance.

However, CNMFSD takes advantage of social interaction information instead of the social following information, which reflects users' intimate relationships more objectively and is hard to fake by spammers. Meanwhile, CNMF contributes to inducing the user's latent features and reduces the possibility of overfitting. Those make that the proposed CNMFSD outperforms other baseline methods in the performance.

Moreover, we observe that the performance of SMFSR is better than SSDM. Since the original content information is sparse and the original features space is large, building models using the original features like SSDM may fail to predict the label of one user precisely. It really shows that it is helpful to induce latent features from the original features by matrix factorization for spammer detection.

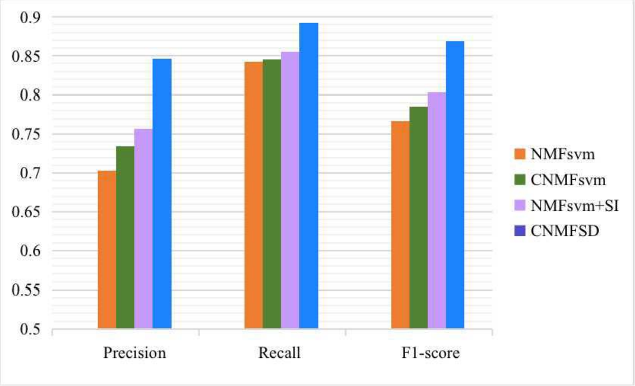

FIGURE 3. Performance of CNMF validation.

The writing styles of tweets are significantly different from general text data, and they are relatively noisy. Even using these noise data, LSTM, Transformer, Deep-learnt, TextCNN, and DeepSBD still outperform other baselines, including SVM, SSDM, and SMFSR. These results show that deep learning-based models are effective for detecting social spammers. We can also observe that attention-based approaches are better than average-based ones. This is reasonable because spammers usually post multiple tweets, and only a few of them are spam, which can avoid being detected. Thus, assigning different weights to different tweets can help the model learn better user representations and further improve its performance.

Compared with deep learning-based approaches, the proposed CNMFSD still achieves better performance. In fact, our approach uses another way to learn user representations to train a classifier, and with the help of the social regularization term, the detection performance is further boosted. Next, we use two ablation studies to demonstrate the benefits of using CNMF and designing a social regularization term in our proposed method.

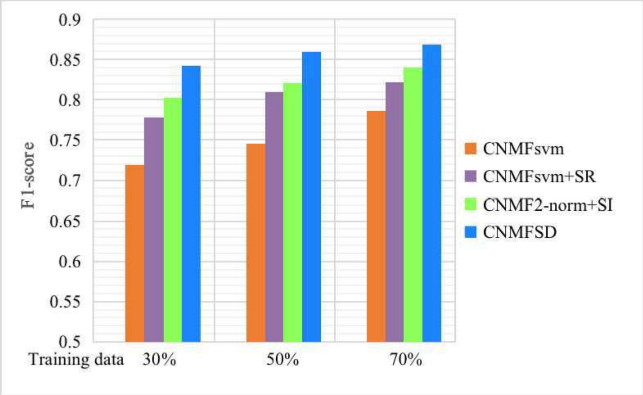

FIGURE 4. Performance of social regularization term validation.

D. CNMF EVALUATION (RQ2)

In Section III-A, we analyzed the reason of choosing CNMF in the paper. To quantitatively verify our motivation, we conduct the following experiments and select NMF$_{svm}$, CNMF$_{svm}$, and NMF$_{svm}$+SI as baselines. NMF$_{svm}$ and NMF$_{svm}$+SI induce the latent features by NMF instead of CNMF. Both CNMFSD and NMF$_{svm}$+SI use the proposed social interaction relationships and SVM to train the corresponding classifiers. The experimental results on Training data 70% are shown in Figure 3. Note that we can observe similar result patterns on both Training data 30% and Training data 50%.

From Figure 3, we observe that the precision, recall, and F1-score of CNMF$_{svm}$ are better than those of NMF$_{svm}$. Since using NMF to induce the latent features may cause overfitting, the performance of methods using NMF is worse than that of the method using CNMF. Although NMF$_{svm}$+SI applies synthetically the social interaction relationships and SVM to learn user latent features, the method using NMF simply factorizes the original feature matrix X into two latent matrices without any constraints on the latent matrix U. Different from NMF$_{svm}$+SI, CNMFSD imposes a constraint by defining the vectors U lie within the column space of X, which is helpful to induce user latent features. Thus, the proposed method can achieve better performance.

E. SOCIAL REGULARIZATION EVALUATION (RQ3)

To answer RQ3, we conduct the following experiments to validate the effectiveness of the proposed social regularization term for detecting spammers. Here, we use three baselines, which are CNMF$_{svm}$, CNMF$_{svm}$+SR, and CNMF2_{-norm}+SI.

Figure 4 shows the performance of the above methods on the different sizes of the training data. The performance of CNMF$_{svm}$ is worse than those methods of using social regularization, which indicates that the social regularization term can help spammer detection approaches to identity social spammers effectively. Compared with CNMF$_{svm}$+SR, CNMF2_{-norm}+SI has better performance. The results justify the theoretical analysis. That is, spammers can collude with other accomplices to build the social following relationships easily, but the social interaction relationships are hard to create. Thus, modeling users' interaction relationships in social regularize terms can improve detection performance more effectively than other regularization terms considering users following relationships.

Among these methods, CNMFSD shows the best performance, which indicates that taking full advantage of convex decomposition matrices and training user latent features with the social regularization term iteratively can improve spammer detection performance.

VI. CONCLUSION

In this paper, we propose a new framework by taking advantage of content and social interaction information for social spammer detection. Different from existing methods that utilize users' the following information, the proposed method CNMFSD integrates users' interaction information based on the trained classification model. In addition, we introduce a new strategy to induce latent features using CNMF in spammers and legitimate users space for improving the performance of detecting spammers. Experimental results on a real dataset show that CNMFSD obtains better detection performance compared with existing methods.

In this work, we employ Convex-NMF to learn latent user features for legitimate users and spammers, respectively. Such a fine-grained learning strategy makes the proposed model obtain accurate latent user representations, which further helps the model to achieve better performance. Besides, introducing social interaction into this task can also improve prediction performance.

Although the proposed model outperforms baselines, it also has some disadvantages. First, in the classifier training stage, we do not consider the social interaction graph, which is trained solely based on the outputs from CNMF. Second, we use tf-idf to extract the user content matrices. However, a spammer always posts some normal tweets to imitate the behavior of legitimate users. Thus, it is essential to distinguish the importance of tweets when we extract the user content matrix.

In future work, we will directly use raw tweets as the model input to learn user representations by distinguishing the importance of each tweet via deep learning techniques. After that, we plan to use graph neural networks to model social interactions among users.

REFERENCES

[1] A. Barushka and P. Hajek, "Spam detection on social networks using cost-sensitive feature selection and ensemble-based regularized deep neural networks," *Neural Comput. Appl.*, vol. 32, no. 9, pp. 4239–4257, May 2020.

[2] Q. Fu, B. Feng, D. Guo, and Q. Li, "Combating the evolving spammers in online social networks," *Comput. Secur.*, vol. 72, pp. 60–73, Jan. 2018.

[3] Z. Zhang, R. Hou, and J. Yang, "Detection of social network spam based on improved extreme learning machine," *IEEE Access*, vol. 8, pp. 112003–112014, 2020.

[4] Nexgate2013. *2013 State of Social Media Spam*. Accessed: Sep. 2013. [Online]. Available: http://nexgate.com/wp-content/uploads/2013/09/Nexgate-2013-State-of-Social-Media-Spam-Research-Report.pdf

[5] D. Liu, B. Mei, J. Chen, Z. Lu, and X. Du, "Community based spammer detection in social networks," in *Proc. Int. Conf. Web-Age Inf. Manage.*, Cham, Switzerland: Springer, 2015, pp. 554–558.

[6] F. Masood, G. Ammad, A. Almogren, A. Abbas, H. A. Khattak, I. U. Din, M. Guizani, and M. Zuair, "Spammer detection and fake user identification on social networks," *IEEE Access*, vol. 7, pp. 68140–68152, 2019.

[7] S. Rao, A. K. Verma, and T. Bhatia, "A review on social spam detection: Challenges, open issues, and future directions," *Expert Syst. Appl.*, vol. 186, Dec. 2021, Art. no. 115742.

[8] C. Chen, J. Zhang, Y. Xie, Y. Xiang, W. Zhou, M. M. Hassan, A. AlElaiwi, and M. Alrubaian, "A performance evaluation of machine learning-based streaming spam tweets detection," *IEEE Trans. Computat. Social Syst.*, vol. 2, no. 3, pp. 65–76, Sep. 2015.

[9] X. Zheng, Z. Zeng, Z. Chen, Y. Yu, and C. Rong, "Detecting spammers on social networks," *Neurocomputing*, vol. 159, pp. 27–34, Jul. 2015.

[10] C. Yang, R. Harkreader, and G. Gu, "Empirical evaluation and new design for fighting evolving Twitter spammers," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 8, pp. 1280–1293, Aug. 2013.

[11] Z. Chu, I. Widjaja, and H. Wang, "Detecting social spam campaigns on Twitter," in *Proc. Int. Conf. Appl. Cryptogr. Netw. Secur.*, Cham, Switzerland: Springer, 2012, pp. 455–472.

[12] M. Fazil, A. K. Sah, and M. Abulaish, "DeepSBD: A deep neural network model with attention mechanism for SocialBot detection," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4211–4223, 2021.

[13] Z. Alom, B. Carminati, and E. Ferrari, "A deep learning model for Twitter spam detection," *Online Social Netw. Media*, vol. 18, Jul. 2020, Art. no. 100079.

[14] X. Ban, C. Chen, S. Liu, Y. Wang, and J. Zhang, "Deep-learnt features for Twitter spam detection," in *Proc. Int. Symp. Secur. Privacy Social Netw. Big Data (SocialSec)*, Dec. 2018, pp. 208–212.

[15] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummadi, "Understanding and combating link farming in the Twitter social network," in *Proc. 21st Int. Conf. World Wide Web (WWW)*, 2012, pp. 61–70.

[16] Y. Zhu, X. Wang, E. Zhong, N. Liu, H. Li, and Q. Yang, "Discovering spammers in social networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 26, 2012, pp. 171–177.

[17] X. Hu, J. Tang, Y. Zhang, and H. Liu, "Social spammer detection in microblogging," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, Princeton, NJ, USA: Citeseer, 2013, pp. 2633–2639.

[18] D. M. Beskow and K. M. Carley, "Bot conversations are different: Leveraging network metrics for bot detection in Twitter," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 825–832.

[19] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "TwitterRank: Finding topic-sensitive influential twitterers," in *Proc. 3rd ACM Int. Conf. Web Search Data Mining (WSDM)*, 2010, pp. 261–270.

[20] C. Thurau, K. Kersting, M. Wahabzada, and C. Bauckhage, "Convex non-negative matrix factorization for massive datasets," *Knowl. Inf. Syst.*, vol. 29, no. 2, pp. 457–478, 2011.

[21] C. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 1, pp. 45–55, Jan. 2010.

[22] P. Heymann, G. Koutrika, and H. Garcia-Molina, "Fighting spam on social web sites: A survey of approaches and future challenges," *IEEE Internet Comput.*, vol. 11, no. 6, pp. 36–45, Nov. 2007.

[23] M. Mateen, M. A. Iqbal, M. Aleem, and M. A. Islam, "A hybrid approach for spam detection for Twitter," in *Proc. 14th Int. Bhurban Conf. Appl. Sci. Technol. (IBCAST)*, Jan. 2017, pp. 466–471.

[24] A. Gupta and R. Kaushal, "Improving spam detection in online social networks," in *Proc. Int. Conf. Cognit. Comput. Inf. Process. (CCIP)*, Mar. 2015, pp. 1–6.

[25] S. Lee and J. Kim, "WarningBird: A near real-time detection system for suspicious URLs in Twitter stream," *IEEE Trans. Depend. Sec. Comput.*, vol. 10, no. 3, pp. 183–195, May 2013.

[26] M. Shirakawa, K. Nakayama, T. Hara, and S. Nishio, "Wikipedia-based semantic similarity measurements for noisy short texts using extended naive Bayes," *IEEE Trans. Emerg. Topics Comput.*, vol. 3, no. 2, pp. 205–219, Jun. 2015.

[27] M. Tsikerdekis and S. Zeadally, "Multiple account identity deception detection in social media using nonverbal behavior," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 8, pp. 1311–1321, Aug. 2014.

[28] C. Sumner, A. Byers, R. Boochever, and G. J. Park, "Predicting dark triad personality traits from Twitter usage and a linguistic analysis of tweets," in *Proc. 11th Int. Conf. Mach. Learn. Appl. (ICMLA)*, vol. 2, Dec. 2012, pp. 386–393.

[29] B. Alghamdi, J. Watson, and Y. Xu, "Toward detecting malicious links in online social networks through user behavior," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. Workshops (WIW)*, Oct. 2016, pp. 5–8.

[30] H. Shen and X. Liu, "Detecting spammers on Twitter based on content and social interaction," in *Proc. Int. Conf. Netw. Inf. Syst. Comput.*, Jan. 2015, pp. 413–417.

[31] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummadi, "Understanding and combating link farming in the Twitter social network," in *Proc. 21st Int. Conf. World Wide Web (WWW)*, 2012, pp. 61–70.

[32] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu, "Analyzing spammers' social networks for fun and profit: A case study of cyber criminal ecosystem on Twitter," in *Proc. 21st Int. Conf. World Wide Web (WWW)*, 2012, pp. 71–80.

[33] Q. Cao, M. Sirivianos, X. Yang, and T. Pregueiro, "Aiding the detection of fake accounts in large scale social online services," in *Proc. 9th USENIX Symp. Networked Syst. Design Implement. (NSDI)*, 2012, pp. 197–210.

[34] P. Cui, F. Wang, S. Liu, M. Ou, S. Yang, and L. Sun, "Who should share what?: Item-level social influence prediction for users and posts ranking," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. (SIGIR)*, 2011, pp. 185–194.

[35] T. Wu, S. Liu, J. Zhang, and Y. Xiang, "Twitter spam detection based on deep learning," in *Proc. Australas. Comput. Sci. Week Multiconference*, Jan. 2017, pp. 1–8.

[36] S. G. Selvaganapathy, M. Nivaashini, and H. P. Natarajan, "Deep belief network based detection and categorization of malicious URLs," *Inf. Secur. J., Global Perspective*, vol. 27, no. 3, pp. 145–161, Apr. 2018.

[37] V. Vapnik, *The Nature of Statistical Learning Theory*. Cham, Switzerland: Springer, 1999.

[38] H. Shen, F. Ma, X. Zhang, L. Zong, X. Liu, and W. Liang, "Discovering social spammers from multiple views," *Neurocomputing*, vol. 225, pp. 49–57, Feb. 2017.

[39] K. Lee, B. Eoff, and J. Caverlee, "Seven months with the devils: A long-term study of content polluters on Twitter," in *Proc. Int. AAAI Conf. Web Social Media*, vol. 5, 2011, pp. 185–192.

[40] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" in *Proc. 19th Int. Conf. World Wide Web (WWW)*, 2010, pp. 591–600.

[41] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[42] F. Godin, *Improving and Interpreting Neural Networks for Word-Level Prediction Tasks in Natural Language Processing*. Ghent, Belgium: Ghent Univ., 2019.

[43] D. Bahdanau, K. H. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proc. 3rd Int. Conf. Learn. Represent., (ICLR)*, 2015, pp. 1–15.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[45] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.

[46] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Doha, Qatar, Aug. 2014, pp. 1746–1751.

HUA SHEN received the master's degree in computer applied technology from Dalian Maritime University. She is currently pursuing the Ph.D. degree in computer science with the Dalian University of Technology. She is also an Associate Professor with the College of Mathematics and Information Science, Anshan Normal University, China. Her research interests include data mining and machine learning.

BANGYU WANG received the Scholar degree in software engineering from the Dalian University of Technology, where he is currently pursuing the Graduate degree in computer science. His research interests include data mining, machine learning, and information retrieval.

XINYUE LIU received the Ph.D. degree from the Dalian University of Technology, China. She is currently an Associate Professor with the School of Software, Dalian University of Technology. Her research interests include data mining, machine learning, and information retrieval.

XIANCHAO ZHANG received the Scholar and master's degrees in mathematics from the National University of Defense Technology, China, in 1994 and 1998, respectively, and the Ph.D. degree in computer science from the University of Science and Technology of China, in 2000. From 2000 to 2003, he worked as a Research and Development Manager in some international companies. He joined the Dalian University of Technology, in 2003. He is currently a Full Professor with the Dalian University of Technology. His research interests include design and analysis of algorithms, machine learning, data mining, and information retrieval.