# Detecting Social Spammers in Sina Weibo Using Extreme Deep Factorization Machine

Yuhao Wu[1], Yuzhou Fang[1], Shuaikang Shang[2], Lai Wei[1], Jing Jin[2], and Haizhou Wang[1(✉)]

[1] College of Cybersecurity, Sichuan University, Chengdu 610065, People's Republic of China
{wuyuhao,fangyuzhou,weilai@scu.edu.cn}@stu.scu.edu.cn, whzh.nc@scu.edu.cn

[2] College of Computer Science, Sichuan University, Chengdu 610065, People's Republic of China
{shangshuaikang,jinjing.0306}@stu.scu.edu.cn

**Abstract.** Online social networks (OSNs) are a form of social media that allow users to obtain news and information as well as connect with others to share content. However, the emergence of social spammers disrupts the normal order of OSNs significantly. As one of the most popular Chinese OSNs in the world, Sina Weibo is seriously affected by social spammers. With the continuous evolution of social spammers in Sina Weibo, they are gradually indistinguishable from benign users. In this paper, we propose a novel approach for social spammer detection in Sina Weibo using extreme deep factorization machine (xDeepFM). Specifically, we extract thirty features from four categories, namely profile-based, interaction-based, content-based, and temporal-based features to distinguish between social spammers and benign users. Furthermore, we build a detection model based on xDeepFM to implement the effective detection of social spammers. The proposed approach is empirically validated on the real-world data collected from Sina Weibo. The experimental results show that this approach can detect social spammers in Sina Weibo more effectively than most of the existing approaches.

**Keywords:** Online social networks · Social spammers · Sina Weibo · Extreme deep factorization machine

## 1 Introduction

With the rapid development of information technology, online social networks (OSNs) have had a significant impact on public life, which enable users to conduct massive-scale and real-time communication [2]. However, OSNs have gradually emerged social spammers, namely users with malicious purposes in OSNs. As social spammers continue to evolve, their behaviors including guidance of


Z. Huang et al. (Eds.): WISE 2020, LNCS 12342, pp. 170–182, 2020.
https://doi.org/10.1007/978-3-030-62005-9_13

online public opinion, malicious commentary, defamation, and ideology infiltration have posed huge damage to normal social order and even national stability [7]. For instance, during the presidential election in the U.S. in 2016, social spammers spread a large number of fake tweets in Twitter, and the decisions of many voters were affected by such tweets [19]. Accordingly, the detection of social spammers in OSNs is of great significance.

Currently, most of the existing research of social spammer detection is carried out on Twitter [1,11,17] and Facebook [4,6,18], and there are relatively few studies based on the Chinese OSNs, such as Sina Weibo [8,13]. Sina Weibo is one of the largest Chinese microblogging services in the world, which has a significant influence on the Chinese social communication. Meanwhile, Sina Weibo has also become one of the most active OSNs of social spammers for its popularity. Spammers widely conduct social spamming, which threaten the quality of Sina Weibo's social network [8]. The research of social spammer detection in Sina Weibo needs to be undertaken more comprehensive and in-depth.

This paper proposes a novel approach for detecting social spammers in Sina Weibo using extreme deep factorization machine (xDeepFM) [12], which consists of three components: data collection component, feature extraction component, and detection component. To begin with, the data collection component is responsible for collecting user data from Sina Weibo. Next, the feature extraction component is used to extract features of social spammers and benign users. Eventually, in the detection component, a xDeepFM model is employed for detection.

**Contributions.** The main contributions of the paper are summarized as follows:

- A novel deep learning-based approach for detecting social spammers in Sina Weibo is proposed, which mainly comprises three combined components, namely, a data collection component, a feature extraction component, and a detection component.
- A total of thirty features are extracted to identify social spammers in Sina Weibo accurately. These features can be divided into four categories: profile-based features, interaction-based features, content-based features as well as temporal-based features.
- A deep learning model, xDeepFM, is employed for detection. The evaluation results show that it significantly outperforms the widely used models.

**Paper Organization.** The rest of this paper is organized as follows. In Sect. 2, related work in the field of social spammer detection is introduced. The proposed social spammer detection approach is elaborated in Sect. 3. Furthermore, Sect. 4 describes the experimental setup and evaluation results. Finally, Sect. 5 concludes the research and plans for future work.

## 2 Related Work

In this section, we summarize some important studies on social spammer detection using machine learning approaches in recent years. Machine learning approaches are effective in detecting social spammers. Generally, the

machine learning approaches can be categorized into classical machine learning approaches and deep learning approaches, which are introduced separately below.

The classical machine learning approach performs social spammer detection in OSNs by training a classical machine learning classification model. Chu et al. [9] measured and characterized the behaviors of humans, spammers, and cyborgs on Twitter. Also, an automated classification system was designed for detecting social spammers. After that, Yang et al. [21] made an analysis of the evasion tactics utilized by spammers and further designed several new features to detect more spammers. In their work, random forest (RF), decision tree (DT), and some other classical machine learning models were applied. In [1], an integrated social media content analysis platform was proposed, which leverages three aspects of features including user-generated content, social graph connections, and user profile activities to identify social spammers. Meanwhile, supervised machine learning models such as support vector machine (SVM), and RF etc. were used for detection, and RF model got the highest accuracy of 96.07%. Alghamdi et al. [2] utilized focused interest patterns of users and combined unsupervised and supervised machine learning to detect social spammers. In [3], supervised machine learning algorithms including RF, SVM, and logistic regression (LR) were employed to detect organized behaviors. Meanwhile, user-based features, temporal-based features, and features of collective behavior were used to distinguish users. Fazil et al. [10] defined six new features and redefined two features in their work. They focused on four categories features, namely metadata-based, content-based, interaction-based, and community-based features. These features were fed to RF, DT, and Bayesian network (BN) machine learning classifiers to detect social spammers.

In recent studies of social spammer detection, deep learning approaches begin to be more and more wildly used. Comparing with classical machine learning approaches, deep learning approaches have better generalization performance, especially when dealing with big data. Cai et al. [5] proposed an extreme learning machine (ELM)-based approach for social spammer detection in Sina Weibo. In their work, features were extracted from message content and behavior, and ELM is used to identify social spammers. In [11], a deep neural network based on contextual long short-term memory (LSTM) architecture that exploits both content and metadata to detect social spammers was proposed. The results showed that the approach had a high area under curve value. Moreover, a deep learning-based social spammer detection scheme was proposed in [17], i.e. DeBD, which utilizes tweet joint features, tweet metadata temporal features, and feature fusing.

## 3 The Proposed Social Spammer Detection Approach

In this section, we describe the proposed approach for detecting social spammers in Sina Weibo, the architecture of which is shown in Fig. 1. It is composed of data collection component, feature extraction component, and detection component. The details of each component of the proposed approach are described below.

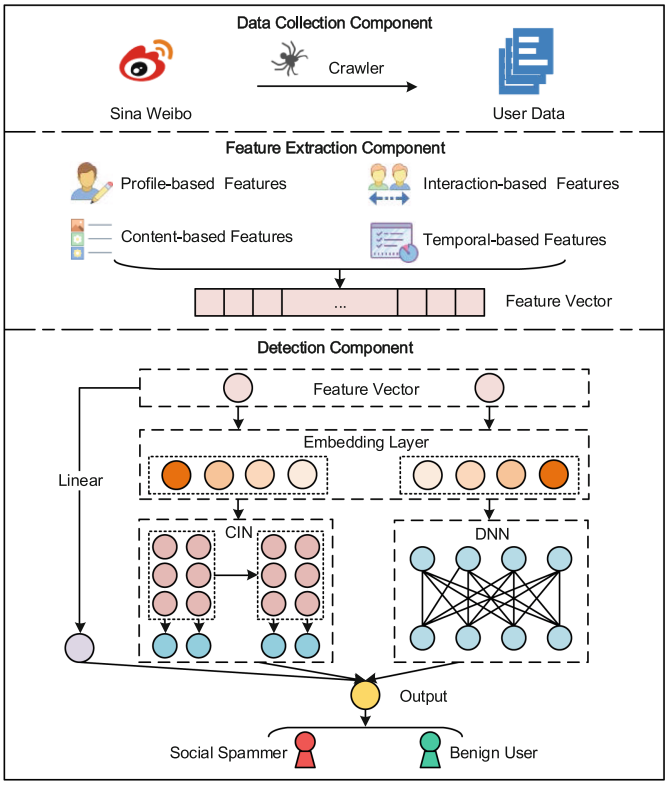


**Fig. 1.** The architecture of the proposed approach.

### 3.1 Data Collection

Sina Weibo provides developer APIs for data accessing, and these APIs are great ways for researchers and developers to collect user data from Sina Weibo at no cost. However, these APIs have some strict restrictions on data collection. To meet the needs of research, a high-performance and multi-threaded web crawler is developed. The crawler can create multi-tasks with multiple proxy IP to cycle through them and build a series of API requests to download raw HTML data from the web. Then, the valid data such as user profiles and posts of users is extracted, which is stored in a database next.

### 3.2 Feature Extraction

In this component, a total of thirty features of social spammers and benign users are extracted and divided into profile-based, interaction-based, content-based, and temporal-based features. Table 1 summarizes all the considered features.

**Table 1.** Summary of the extracted features.

| Profile-based features | Interaction-based features |
|---|---|
| Default avatar | # of likes (mean) |
| Default nickname | # of comments (mean) |
| Length of nickname | # of reposting (mean) |
| Completeness of profile | Ratio of reposted posts |
| Ratio of follower count | Diversity of posting sources |
| Active level | |
| **Content-based features** | **Temporal-based features** |
| # of mentions (mean & variance) | Time interval (mean & variance) |
| # of hashtags (mean & variance) | Time interval (maximum & minimum) |
| # of URLs (mean & variance) | Burstiness parameter of time interval |
| # of punctuations (mean & variance) | Information entropy of time interval |
| # of interjections (mean & variance) | |
| # of words (variance) | |
| # of pictures (variance) | |
| Post sentiment score (mean) | |

**Profile-Based Features.** Profile-based features are extracted from profiles of users, which can reveal the differences between benign users and social spammers.

*Default avatar and nickname (DA and DN):* Lots of social spammers use the default avatars and default nicknames [20]. Hence, $DA$ and $DN$ are considered in our work. The value of $DA(u)$ is 1 if user $u$ uses the default avatar, otherwise it is 0, and the calculation of $DN(u)$ is similar to $DA(u)$.

*Length of nickname (LN):* The $LN$ feature was used and achieved good results in [20]. Also, we adopt $LN$ in our work. Since there are strict restrictions on $LN$ in Sina Weibo, the value range of $LN(u)$ of user $u$ is $\{LN(u)|2 \leq LN(u) \leq 30\}$.

*Completeness of profile (CP):* Users of Sina Weibo can fill in or change their profiles. Benign users have real friend-making demands, so they tend to fill in their profiles carefully. However, the profiles of social spammers are usually incomplete. Thus, $CP$ is used in our work, like the way of using this feature in [21].

*Ratio of follower count (RF):* The $RF$ feature is widely used [8–10]. To compute this feature, we follow the method used in these works.

*Active level (AL):* The $AL$ feature is to measure how active a user is. In our work, we define $AL$ of user $u$ as

$$AL(u) = \sum_{i=1}^{M} \beta_i \times \varphi_i, 0 < \beta_i < 1, \tag{1}$$

where $\beta_i$ is the value of the $i^{th}$ level, $\varphi_i$ is the weight of the $i^{th}$ level, and $M$ is the number of used levels. In our work, whether verified (verification is 1, otherwise 0) and the normalized user level are weighted to calculate a user's $AL$ feature. The value range of $AL(u)$ is $\{AL(u) \,|\, 0 < AL(u) \leq 1\}$.

**Interaction-Based Features.** The posts of users can be commented, reposted, and liked by others. These interactions often reflect the difference between benign users and social spammers. Therefore, interaction-based features are extracted.

*Mean of the number of likes, comments, and reposting (ML, MC, and MR):* The number of likes, comments, and reposting on a user's posts can quantify the popularity of the user, and most of posts of social spammers are illogical, so they have few likes, comments, or reposting. Hence, $ML$, $MC$, and $MR$ are employed, which can be computed using the method in [15].

*Diversity of posting sources (DS):* Generally, posts come with posting sources, such as computer, mobile, etc. Benign users' posts tend to have different posting sources, while social spammers' posts usually have few posting sources. Therefore, we consider the $DS$ feature and use the Margalef diversity index to calculate it, which is given by

$$DS(u) = \frac{\gamma - 1}{\ln K}, \tag{2}$$

where $\gamma$ denotes the number of types of posting sources of the user, $K$ represents the number of posts of the user $u$.

*Ratio of reposted posts (RR):* A number of posts of social spammers are reposted from other users, or generated using probabilistic methods. We thereby use $RR$ feature to distinguish users, which is defined as the ratio of the number of reposted posts to the total number of posts [10,13].

**Content-Based Features.** Generally, the content of different posts of social spammers is similar and the writing habits of social spammers are illogical. Hence, content-based features are employed to identify social spammers.

*Mean and variance of the number of mentions (MM and VM):* In Sina Weibo, users use "@" to mention other users when posting. The number of mentions in posts can significantly distinguish users [10]. Hence, $MM$ and $VM$ of user $u$ are defined as

$$MM(u) = \frac{1}{K}\sum_{i=1}^{K}\eta_i, \text{ and } VM(u) = \frac{1}{K}\sum_{i=1}^{K}(\eta_i - MM(u))^2, \tag{3}$$

where $\eta_i$ is the number of mentions in the $i^{th}$ post.

*Mean and variance of the number of hashtags (MH and VH):* In Sina Weibo, "#" is used by users to participate in the discussion of topics while posting posts [9,10]. Accordingly, $MH$ and $VH$ are taken as two features, which can be computed in the same way as $MM$ and $VM$.

*Mean and variance of the number of URLs (MU and VU):* URLs are usually used by social spammers for advertising, monetization, etc. [9]. The number of URLs can judge the quality of users' posts [10]. Thus, $MU$ and $VU$ are employed, which can be computed like $MM$ and $VM$.

*Mean and variance of the number of punctuations (MP and VP):* The use of punctuations in posts can reflects a user's writing habits. In the posts of social spammers, the use of punctuations is often unreasonable. For this reason, $MP$ and $VP$ are used, which can be calculated like $MM$ and $VM$.

*Mean and variance of the number of interjections (MI and VI):* An interjection is a word or expression that occurs as an utterance on its own and expresses a spontaneous feeling or reaction, such as "oh", "ah", "o", "ha", etc. These words can reflect a user's writing style. Thus, $MI$ and $VI$ are employed. To compute them, we follow the method used in $MM$ and $VM$.

*Variance of the number of words (VW):* The word counts of different posts of a social spammer are usually similar [20]. As such, $VW$ is given to identify social spammers in our work.

*Variance of the number of pictures (VNP):* When posting, users can add pictures to their posts. Generally, the number of pictures on each post of a social spammer is similar. Thus, $VNP$ is used.

*Mean of the post sentiment score (MS):* Features of sentiment are the features extracted through sentiment analysis of posts [14,20]. In our work, sentiment scores of posts are calculated, and the $MSS$ feature is employed. The value range of $MS$ is $\{MS(u) \mid 0 \leq MS(u) \leq 1\}$.

**Temporal-Based Features.** Temporal-based features are extracted from the time of posts. In our work, the series of time intervals between each post of a user is defined as $\theta = [\chi_1, \chi_2, \ldots, \chi_{K-1}]$, where $K$ is the number of posts of the user. The following temporal-based features are used to distinguish users.

*Mean and variance of the time interval (MT and VT):* In [8], the regularity of the time of user's posts is considered, and the variance of the post time was employed. In our work, $MT$ and $VT$ of user $u$ are defined by

$$MT(u) = \frac{1}{K-1} \sum_{i=1}^{K-1} \chi_i, \text{ and } VT(u) = \frac{1}{K-1} \sum_{i=1}^{K-1} (\chi_i - MT(u))^2, \quad (4)$$

where $\chi_i$ is the time interval of two consecutive posts.

*Maximum and minimum of the time interval (MAT and MIT):* A number of social spammers do not post for a long time after posting a large number of posts in a short time. Thus, the features $MAT$ and $MIT$ are used. We sort the series of time interval to get a new series: $\theta' = [\chi'_1, \chi'_2, \ldots\ldots, \chi'_{K-1}]$ ($\chi'_i \leq \chi'_{i+1}, 1 \leq i \leq K-1$). Then, $MAT$ and $MIT$ of the user $u$ are computed by

$$MAT(u) = \frac{1}{\mu}\sum_{i=1}^{\mu}\chi'_i, \text{ and } MIT(u) = \frac{1}{\mu}\sum_{\mathrm{i}=K-\mu}^{\mathrm{K}-1}\chi'_i. \tag{5}$$

After analysis, when $\mu = 5$, this pair of features can distinguish users well.

*Burstiness parameters of time interval (BT):* $BT$ can distinguish users as well. There are three special values for $BT$: $\varepsilon - 1$, $\varepsilon$, $\varepsilon + 1$, which can be understood as a completely regular behavior, a completely Poisson behavior, and the most bursty behavior, respectively [16]. Generally, the values of $BT$ of social spammers are close to $\varepsilon - 1$ and $\varepsilon + 1$.

*Information entropy of time interval (IT):* The Shannon entropy can be applied to quantify the regularity of the posting time interval series of users [16]. Thus, $IT$ is employed in our work, and the smaller the $IT$, the greater the probability that the user is a social spammer.

### 3.3 xDeepFM for Detection

The detection component is based on the xDeepFM model [12], which combines a compressed interaction network (CIN) and a deep neural network (DNN) into a unified model. xDeepFM can learn certain bounded-degree feature interactions explicitly, and learn arbitrary low-order and high-order feature interactions implicitly as well. The characteristics of xDeepFM, being able to learn patterns of combinatorial features automatically and generalize to unseen features, have a great effect on the detection of social spammers.

After completing extracting features of users and forming feature vectors, each feature vector is divided into two types of features: continues and categorical features. Continuous features are those that have numerical values, while categorical features are features that can take on one of a limited number of possible values. This component transforms categorical features and continuous features by one-hot encoding and normalization, respectively. A user's feature vector can be represented as $x_i$ ($i$ denotes the $i^{th}$ user among all users), the output of xDeepFM is the probability that the user is a social spammer, which is given by

$$\hat{y_i} = \sigma(w_1^T x_i + w_2^T d_i + w_3^T c_i + b), \tag{6}$$

where $\sigma$ is the sigmoid function, $d_i$ and $c_i$ are the outputs of the DNN and CIN, respectively. $w_1$, $w_2$, $w_3$, and $b$ denote learnable parameters. Moreover, the cross-entropy cost function is used for overcoming the learning slowdown problem of quadratic cost function, which is defined by

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n} y_i log\hat{y}_i + (1-y_i)log(1-\hat{y}_i), \tag{7}$$

where $n$ is the total number of training feature vectors of users, and the value range of $i$ is $\{i|1 \leq i \leq n\}$. In addition, $y_i$ is the true category of the $i^{th}$ user (social spammer or benign user). Furthermore, the optimization process is to minimize a objective function, it is defined as

$$\mathcal{J} = \mathcal{L} + \lambda\|\Theta\|, \tag{8}$$

where $\lambda$ is the regularization term, $\Theta$ represents the parameter set of the model.

## 4 Experiments and Evaluation

In this section, we evaluate the performance of the proposed approach for detecting social spammers. Our experiments are conducted on a Ubuntu 18.04.3 LTS platform with an Intel Xeon E5-2618L v3 CPU and a NVIDIA GeForce RTX 2080TI GPU (64 GB RAM). In the numerical results analysis, four metrics are taken into consideration to evaluate the performance of detection models, namely accuracy, precision, recall, and F-score. Each experiment is repeated ten times independently, and the average results are shown.

### 4.1 Dataset

In order to train our detection model, we collect a mass of user data. Then, five researchers in related fields are invited to manually annotate the collected user data. To be specific, the five researchers first independently annotate user data, and then mutually verify the results of annotation. Eventually, a dataset with 10,000 social spammers and 10,000 benign users is constructed after completing annotation, data balance processing, and feature extraction. Further, the dataset is divided into 60% for training, 20% for validation, and 20% for testing. The brief description of the dataset is shown in Table 2.

**Table 2.** Overview of the dataset.

| Category | Training | Validation | Testing | Posts |
|---|---|---|---|---|
| Benign user | 6,000 | 2,000 | 2,000 | 118,199 |
| Social spammer | 6,000 | 2,000 | 2,000 | 96,307 |
| Total | 12,000 | 4,000 | 4,000 | 214,506 |

### 4.2 Performance Evaluation

To evaluate the effectiveness of the proposed approach, first, we complete model training on the constructed dataset, and the training process of our detection model is shown in Fig. 2. Notably, the xDeepFM model that we used is stable and convergent. Moreover, overfitting is clearly well suppressed by using dropout technology. This yields better performance for training and testing. Further, we compare the xDeepFM model with baseline models that are widely used in this field including SVM [1,3], LR [3], DT [10,21], and RF [1,10]. Table 3 and Fig. 3 show the experimental results. It can be seen that the classical machine learning models of LR, SVM, and DT have almost no advantages over the ensemble machine learning models of RF in this case. Among all detection models, our detection model xDeepFM has the highest accuracy, precision, recall, and F-score. Such results indicate that the characteristics of xDeepFM, the ability to learn and generalize patterns of combinatorial features, have obvious advantages in detecting social spammers.

**Table 3.** Performance of different detection models.

| Model | Accuracy | Precision | Recall | F-score |
|---|---|---|---|---|
| SVM [1,3] | 0.9544 | 0.9365 | 0.9752 | 0.9554 |
| LR [3] | 0.9673 | 0.9564 | 0.9791 | 0.9676 |
| DT [10,21] | 0.9724 | 0.9693 | 0.9757 | 0.9725 |
| RF [1,10] | 0.9794 | 0.9790 | 0.9801 | 0.9795 |
| **xDeepFM** | **0.9816** | **0.9829** | **0.9802** | **0.9815** |

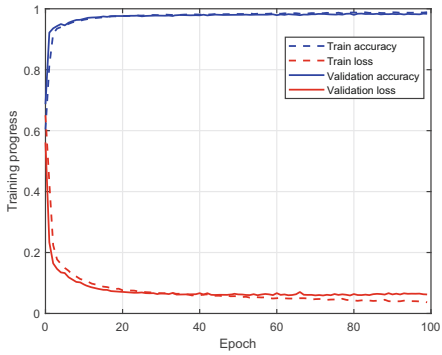


**Fig. 2.** Training progress of our detection model xDeepFM.

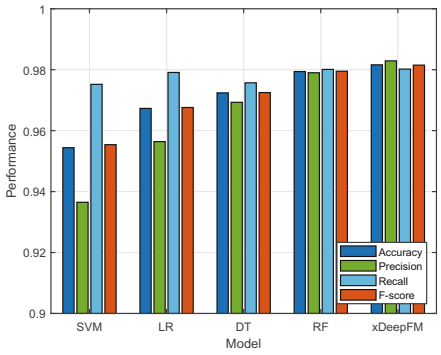


**Fig. 3.** Performance comparison between xDeepFM and other detection models.

### 4.3 Feature Ablation Study

In order to evaluate the validity of each category of feature, we conduct feature ablation test on the full feature set and four subsets of the full feature set using xDeepFM model. The subsets of the feature set can be represented by the set-difference function as

$$F \backslash F' = \{x | x \in F \wedge x \notin F'\}, \tag{9}$$

where $F$ is the set with all features, $F'$ is the subset of $F$ with a particular category of features, and $x$ is all user data of a feature. Therefore, $F \backslash Profile$, $F \backslash Interaction$, $F \backslash Content$, and $F \backslash Temporal$ represent the feature set with profile-based, interaction-based, content-based, and temporal-based features removed from $F$, respectively.

The results of the feature ablation test are shown in Fig. 4. Firstly, the detection model xDeepFM performs much better on the feature set $F$ that contains all the features than on other feature sets, which proves that each category of extracted features has the distinguishability between social spammers and benign users. Moreover, xDeepFM performs worst using feature set of $F \backslash Content$, which indicates that the distinguishability of content-based features is the greatest. Whereas the performance of xDeepFM using feature set of $F \backslash Temporal$ is second only to that of $F$, which means that the distinguishability of temporal-based features is the worst.

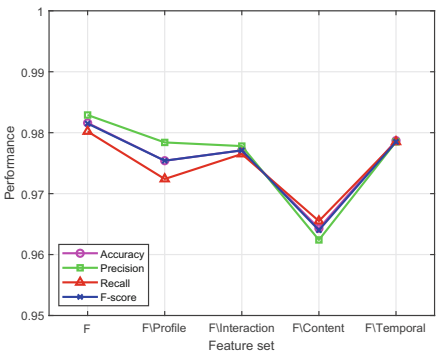


**Fig. 4.** Performance comparison on different feature sets using xDeepFM.

## 5 Conclusion and Future Work

In this paper, we have proposed a novel approach for detecting social spammers in Sina Weibo. The proposed approach mainly includes three components: data collection component, feature extraction component, and detection component. Specifically, we extracted thirty features to distinguish between social

spammers and benign users, which can be divided into profile-based, interaction-based, content-based, and temporal-based features. Furthermore, this paper built a xDeepFM model to detect social spammers. Extensive experiments on the real-world data collected from Sina Weibo showed that, compared with the widely used detection models, the proposed xDeepFM-based detection model is superior in terms of the accuracy, precision, recall, and F-score.

Future work will focus on more online social networks to further verify and improve the proposed social spammer detection approach.

**Acknowledgements.** This work is supported by the National Natural Science Foundation of China (NSFC) under grant nos. 61802270, 61802271, 81602935, and 81773548. Haizhou Wang is the corresponding author. The authors thank anonymous reviewers for their helpful comments to improve the paper.

## References

1. Al-Qurishi, M., Hossain, M.S., Alrubaian, M., Rahman, S.M.M., Alamri, A.: Leveraging analysis of user behavior to identify malicious activities in large-scale social networks. IEEE Trans. Ind. Inform. **14**(2), 799–813 (2017)
2. Alghamdi, B., Xu, Y., Watson, J.: A hybrid approach for detecting spammers in online social networks. In: Hacid, H., Cellary, W., Wang, H., Paik, H.-Y., Zhou, R. (eds.) WISE 2018. LNCS, vol. 11233, pp. 189–198. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-02922-7_13
3. Beğenilmiş, E., Uskudarli, S.: Organized behavior classification of tweet sets using supervised learning methods. In: Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics, pp. 1–9. ACM (2018)
4. Boshmaf, Y., et al.: Íntegro: leveraging victim prediction for robust fake account detection in large scale OSNs. Comput. Secur. **61**, 142–168 (2016)
5. Cai, C., Li, L., Zeng, D.: Detecting social bots by jointly modeling deep behavior and content information. In: Proceedings of the 26th ACM Conference on Information and Knowledge Management, pp. 1995–1998. ACM (2017)
6. Cao, Q., Yang, X., Yu, J., Palow, C.: Uncovering large groups of active malicious accounts in online social networks. In: Proceedings of the 21st ACM Conference on Computer and Communications Security, pp. 477–488. ACM (2014)
7. Chakraborty, M., Pal, S., Pramanik, R., Ravindranath Chowdary, C.: Recent developments in social spam detection and combating techniques: a survey. Inf. Process. Manag. **52**(6), 1053–1073 (2016)
8. Chen, H., Liu, J., Lv, Y., Li, M.H., Liu, M., Zheng, Q.: Semi-supervised clue fusion for spammer detection in Sina Weibo. Inf. Fusion **44**, 22–32 (2018)
9. Chu, Z., Gianvecchio, S., Wang, H., Jajodia, S.: Detecting automation of Twitter accounts: are you a human, bot, or cyborg? IEEE Trans. Dependable Secur. Comput. **9**(6), 811–824 (2012)
10. Fazil, M., Abulaish, M.: A hybrid approach for detecting automated spammers in Twitter. IEEE Trans. Inf. Forensics Secur. **13**(11), 2707–2719 (2018)
11. Kudugunta, S., Ferrara, E.: Deep neural networks for bot detection. Inf. Sci. **467**, 312–322 (2018)

12. Lian, J., Zhou, X., Zhang, F., Chen, Z., Xie, X., Sun, G.: XDeepFM: combining explicit and implicit feature interactions for recommender systems. In: Proceedings of the 24th ACM Conference on Knowledge Discovery and Data Mining, pp. 1754–1763. ACM (2018)
13. Lian, Y., Dong, X., Chi, Y., Tang, X., Liu, Y.: An internet water army detection supernetwork model. IEEE Access **7**, 55108–55120 (2019)
14. Loyola-González, O., López-Cuevas, A., Medina-Pérez, M.A., Camiña, B., Ramírez-Márquez, J.E., Monroy, R.: Fusing pattern discovery and visual analytics approaches in tweet propagation. Inf. Fusion **46**, 91–101 (2019)
15. Mohammad, S., Khan, M.U., Ali, M., Liu, L., Shardlow, M., Nawaz, R.: Bot detection using a single post on social media. In: Proceedings of the 3rd World Conference on Smart Trends in Systems, Security and Sustainability, pp. 215–220. IEEE (2019)
16. Pan, J., Liu, Y., Liu, X., Hu, H.: Discriminating bot accounts based solely on temporal features of microblog behavior. Phys. A **450**, 193–204 (2016)
17. Ping, H., Qin, S.: A social bots detection model based on deep learning algorithm. In: Proceedings of the 18th IEEE International Conference on Communication Technology, pp. 1435–1439. IEEE (2018)
18. Santia, G.C., Mujib, M.I., Williams, J.R.: Detecting social bots on Facebook in an information veracity context. In: Proceedings of the 13th International AAAI Conference on Web and Social Media, pp. 463–472. AAAI (2019)
19. Shao, C., Ciampaglia, G.L., Varol, O., Yang, K.C., Flammini, A., Menczer, F.: The spread of low-credibility content by social bots. Nat. Commun. **9**(1), 1–9 (2018)
20. Varol, O., Ferrara, E., Davis, C.A., Menczer, F., Flammini, A.: Online human-bot interactions: detection, estimation, and characterization. In: Proceedings of the 11th International AAAI Conference on Web and Social Media, pp. 280–289. AAAI (2017)
21. Yang, C., Harkreader, R., Gu, G.: Empirical evaluation and new design for fighting evolving Twitter spammers. IEEE Trans. Inf. Forensics Secur. **8**(8), 1280–1293 (2013)