Hindawi
Security and Communication Networks
Volume 2021, Article ID 3294610, 16 pages
https://doi.org/10.1155/2021/3294610

Review Article

Survey on Astroturfing Detection and Analysis from an Information Technology Perspective

Tong Chen, Jiqiang Liu, Yalun Wu, Yunzhe Tian, Endong Tong, Wenjia Niu, Yike Li, Yingxiao Xiang, and Wei Wang

Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, Beijing Jiaotong University, Beijing 100044, China

Correspondence should be addressed to Endong Tong; edtong@bjtu.edu.cn and Wenjia Niu; niuwj@bjtu.edu.cn

Received 23 September 2021; Accepted 9 November 2021; Published 1 December 2021

Academic Editor: Zhe-Li Liu

With the development of the Internet, user comments produced an unprecedented impact on information acquisition, goods purchase, and other aspects. For example, the user comments can quickly render a topic the focus of discussion in social networks. It can promote the sales of goods in e-commerce, and it influences the ratings of books, movies, or albums. Among these network applications and services, "astroturfing," a kind of online suspicious behavior, can generate abnormal, damaging, and even illegal behaviors in cyberspace that mislead public perception and bring a bad effect on Internet users and society. Hence, the manner of detecting and combating astroturfing behavior has become highly urgent, attracting interest from researchers both from information technology and sociology. In the current paper, we restudy it mainly from the perspective of information technology, summarize the latest research findings of astroturfing detection, analyze the astroturfing feature, classify the machine learning-based detection methods and evaluation criteria, and introduce the main applications. Different from the previous surveys, we also discuss the new future directions of astroturfing detection, such as cross-domain astroturfing detection and user privacy protection.

1. Introduction

With the rapid development of the Internet, more people are communicating with each other through network applications and services. Recently, how people work, contact, acquire information, and purchase has been greatly influenced by e-commence, the emergence of social network (e.g., Facebook, Twitter, WeChat, and Weibo) in particular. Among these network applications and services, there widely exists a kind of suspicious online behavior, namely, astroturfing, mostly appearing in business and political events as others' opinions often have an important impact on an individual's impression of a certain subject [1]. In other words, the attitude and opinions of online users are very likely to be affected by other users.

Recently, breaking news has exposed the traces of astroturfing. The dairy products company Mengniu hired astroturf organizations to attack the reputation of their competitor Yili, another dairy products company, in October 2010. Astroturfing has been a widespread concern in society. It has also been exposed that coordinated attacks can be launched by a group of astroturfers. They generate positive or negative opinions to draw public attention or arouse curiosity. Such a practice is referred to as "cyber gossip," which can mislead the public and put the competing businesses at serious risk. In October 2015, Amazon sued 1114 members of the "Internet water army," accusing them of providing inveracious reviews for goods and services on Amazon.com, which was in violation of the United States laws [2]. In April 2016, a technology social networking site in the U.S. announced that the "internet water army" on Twitter became a secret weapon during the U.S. presidential election [3]. We can see that astroturfing is developing fast worldwide.

Obviously, astroturfing can lead to abnormal, damaging, and even illegal behaviors in cyberspace, misleading the public perception and causing a negative effect on society and netizens. Astroturfing is difficult for online users to be

aware of, and it poses a challenge to distinguishing truth from falsehood. Therefore, it is urgent to figure out how to detect and combat astroturfing. Researchers from IT and sociology are highly interested to study astroturfing.

In the current paper, we will study astroturfing mainly from an IT perspective rather than a sociology perspective. Generally speaking, IT covers any form of technology, i.e., any equipment or technique used by a company, institution, or any other organization that handles information. Thus, our main focus is to design algorithms and technologies to effectively detect online astroturfing and help users quickly identify potential online astroturfers. Furthermore, we give a comprehensive and practical discussion from a technical perspective, rather than from the social analysis view.

Unfortunately, as a kind of suspicious online behavior, astroturfing has a much closer relationship with and has differences that can be easily confused with other suspicious online behaviors: traditional spam [4, 5], fake review [6, 7], social spam [8–12], and link farming [13, 14]. Hence, firstly, we should attempt to find out the essential characteristics of astroturfing. Then, we find out how to design a semiautomatic or automatic computer algorithm and realize how it will be naturally carried out to suit different data sizes. To wrap up, from the IT perspective, we hope that the computer has the learning ability to effectively detect unknown astroturfing activity in big social data. In this paper, we will summarize the astroturfing detection and analysis based on the astroturfing feature, the basic method, evaluation, and application.

The structure of this article is organized as follows: section 2 will discuss the astroturfing feature. Section 3 will present the learning-based astroturfing detection methods and describes the effective detection evaluations. Section 4 shows the astroturfing detection applications. Finally, the future directions are envisaged in section 5.

2. Astroturfing Feature Characterization from IT Perspective

2.1. Feature in Definition. The group or individual involved in the action of "astroturfing" is called an "astroturfer" or "Internet water army." It is very difficult to define astroturfing in a qualified and accurate way. Therefore, several descriptive explanations of "astroturfing" are summarized as follows:

(i) The behavior of covering up the sponsors of a particular organization, such as public relations, advertisings, and politics, can make such a behavior seem like it originated from the grassroots participants [15].

(ii) The "fake" grassroots participants are established by public organizations [16]. The astroturfing situation occurs when a large number of people are employed to post some employer-mandated statements or claims via various public social channels that they do not subjectively support [17, 18].

(iii) The expensive fraud at the national level is explained as follows [19]: to convince the public, the astroturfing propagates "fake" claims among the public. Furthermore, it is costly to employ astroturfing and spread the employer-mandated claims. Hence, the influence of astroturfing is constrained by financial support [20].

From these explanations of "astroturfing," we generalized the following feature keywords:

(i) Money business: if astroturfing exists, there must be an employer who offers money indirectly to the lobbying firms or directly to the grassroots.

(ii) Content effect: any astroturfing should aim to timely achieve the effect of astroturfing by posting a large number of employer-mandated opinions to public channels or by improving the corresponding search rankings so the employer-mandated claims can obtain wider public attention. The content consists of two parts: claims and posts.

Therefore, the definition of the concept features of astroturfing should consists of two parts, namely content effect and money business. Furthermore, these two parts can be utilized to guild the astroturfing detection. This can be simply represented as follows:

$$f_{\text{Astroturfing}} \in V(\text{money business}) \cup V(\text{content effect}) \sqrt{b^2 - 4\text{ac}}. \quad (1)$$

In the aforementioned formula, V(money business) and V(content effect), respectively, represent the feature sets of money business and content effect. However, from the IT perspective, the money business feature is very difficult to mine as a lot of evidence is missed if we rely on the open Web data alone, whereas the other feature can be obtained in public Internet. Thus, all IT research studies on astroturfing belong to the category of mining and utilizing the content effect feature.

2.2. Feature in Suspicious Behavior Category. In the behavior category of suspicious online behaviors, compared with other known suspicious behaviors, astroturfing has some similarities, as well as differences from the traditional ones, such as spam [4], fake review [6], spam (social) [8, 9], and link framing [13]. We make a comparison of them in Table 1.

As observed in Table 1, we select five aspects to make a comparison. The application indicates the carrier of a suspicious online behavior. Among various suspicious behaviors, astroturfing mainly appears in business-to-consumer (B2C) and customer-to-customer (C2C) e-commerce applications, such as TaoBao, Amazon, and Netflix. In addition, astroturfing also appears in various recent social networks, such as Weibo, Facebook, and Twitter. In comparison, the spam appears in the applications of email and subscriber management system (SMS) alone.

As for the aspect of the participant, astroturfing is implemented by artificial intelligence (AI) or an individual via the online astroturfing platform or in other ways online that can achieve the purpose of performing requiring effects.

Table 1: Comparison of various suspicious online behaviors.

	Application	Participant	Time	Visual	Scale
Spam	E-mail and SMS	Artificial intelligence	No	No	Huge
Fake review	B2C and C2C e-commerce	Artificial intelligence or individual	No	Yes	No
Spam (social)	Social network	Artificial intelligence or individual	No	Yes	No
Link farming	Search engine and social contact	Artificial intelligence or individual	Long time	No	—
Astroturfing	B2C and C2C e-commerce	Artificial intelligence or individual	Short time	Yes	Huge

Similarly, fake reviews, spam (social), and link farming also utilize an AI program or an individual.

The time denotes the participating time within the particular task cycle. In an astroturfing mission, generally speaking, the duration of astroturfing behavior is not long. Usually, it takes a few hours or a few days to achieve the purpose of generating sufficient public influence through a series of astroturfing operations. However, the link farming behavior usually needs more time.

The participant visual refers to whether the suspicious behavior of participating is visible to normal social network users. To achieve the purpose of confusing the public, the participating behavior of astroturfing must be visible to the public. In addition, the scale represents the actual number of users participating in such abnormal network behaviors. In most cases, for an effective astroturfing campaign, at least a few hundred astroturfers are involved.

Hence, astroturfing's main features from the suspicious online behavior category are the ones that can cover those in the fake review and social spam behaviors. This can be simply represented as follows:

$$f_{\text{Astroturfing}} \supset f_{\text{Fake Review}}, f_{\text{Astroturfing}} \supset f_{\text{Social Spam}}. \tag{2}$$

Hence, we can exploit the detection approach to fake reviews and social spam to detect astroturfing.

2.3. Feature in Social Application. As mentioned above, the exploitation of the detection approach to fake reviews and social spam helps detect astroturfing. We will discuss the feature in the operable actions of social applications. We try to generalize a formula to calculate the astroturfing probability by an activation function $H(s)$, which is as follows:

$$H(s) = \frac{e^s}{1+e^{-s}}, \qquad s = \sum_{i=1}^{d} \omega_i x_i, \tag{3}$$

where x_i is the i_{th} feature and ω_i is the corresponding weight. It is noted that the like action m_{like} and the comment action n_{comment} are key to generate the content effect, and hence, these two action features can be composed and assigned with bigger weights. The composed feature is calculated as follows:

$$x_{\text{compose}} = (m_{\text{like}} - M_0)^2 (n_{\text{comment}} - N_0)^2, \tag{4}$$

where M_0 and N_0 are the thresholds of the like action number m_{like} and comment number n_{comment}, respectively, for a specific poster. Then, we describe the common feature x_i in real social applications and list the corresponding notation for all features and their parameters in Table 2.

2.4. Like

2.4.1. Following Similarity. The following similarity F_{sim} of the accounts u_i and u_j can be calculated as follows:

$$F_{\text{sim}}(u_i, u_j) = \frac{\left|F(u_i) \cap F(u_j)\right|}{\left|F(u_i) \cup F(u_j)\right|}, \tag{5}$$

where $F(u)$ denotes the original tweets set for the account u. Astroturfing has low following similarities.

2.4.2. Low Followers. Astroturfing has fewer followers, while their following is more. Thus, the index of low followers F_{low} can be calculated as follows:

$$F_{\text{low}} = \frac{|\text{following}|}{|\text{followers}|}. \tag{6}$$

Generally, astroturfing has a relatively high F_{low}.

2.5. Comment

2.5.1. Percentage of Replies. Astroturfing hardly read and reply to others' comments. Posting new comments rather than reading and replying are preferred. Thus, we can calculate the ratio of the number of replies to the total number of comments from the same user as follows:

$$p = \frac{\text{the number of replies}}{\text{the number of total comments}}, \tag{7}$$

where p represents the ratio. Astroturfing has a low value of p.

2.5.2. The Ratio of Similar Comments. To minimize their own workloads and obtain the maximum benefits, astroturfing will make a number of published duplication or other people's comments when they make false comments. The ratio of similar comments is high.

2.6. Retweet

2.6.1. Retweet Similarity. The two retweets are for the same tweet and are created within a threshold time window. The retweet similarity RT_{sim} between the two accounts u_p and u_q can be computed as follows:

Table 2: Parameters and notation.

Notation	Meaning
u_i	Account
p	The percentage of replies, and astroturfing has a relatively low value.
P	$P_{p/q}$ is the probability of word $w_{p/q}$ appears, and $P_{p,q}$ is the joint probability.
s	The consecutive sentence
v	Semantic vector representation
k	The piece of a certain sentence, and K is the total number of pieces
l	The total quantity of the sentences in the k_{th} piece
F_{sim}	Following similarity, and $F_{\text{sim}}(u_i, u_j)$ denotes the following similarity of the accounts u_i and u_j
F	Original tweets set, $F(u)$ denotes the original tweets set for the account u
F_{low}	Low followers, and astroturfing has a relatively high value.
RT_{sim}	Retweet similarity, and $\text{RT}_{\text{sim}}(u_p, u_q)$ denotes the retweet similarity between the accounts u_p and u_q
RT	$\text{RT}(u_p) = (u_p, T_1, \text{tid}_1), (u_p, T_2, \text{tid}_2), \ldots, (u_p, T_n, \text{tid}_n)$, T_i is the retweeting time, and tid_i is the retweet ID
RT_{ratio}	The most dominant application's percentage
$nb_{/\text{day}}$	The posting frequency, and the astroturfing has a relatively high value
Cl_{rat}	The number of received clicks, and astroturfing has a relatively low value
$P(w_q \mid w_p)$	The transition probability from the word w_p to w_q
$\text{PTP}_{i,:}$	Sentence transition, and astroturfing has a relatively high value
$O_{p,q}$	Word cooccurrence
$\text{sco}^{\text{as}}(s_1, s_2)$	The average score of two consecutive sentences s_1 and s_2
$\text{sco}^{\text{bs}}(s_1, s_2)$	The best score of two consecutive sentences s_1 and s_2
$\text{sim}(s_i, s_{i+1})$	Pairwise sentence similarity
SD	Semantic dispersion, and the astroturfing has a relatively high value

$$\text{RT}_{\text{sim}}(u_p, u_q) = \frac{\left\|\text{RT}(u_p) \cap \text{RT}(u_q)\right\|}{\left\|\text{RT}(u_p) \cup \text{RT}(u_q)\right\|}, \tag{8}$$

where $\text{RT}(u_p) = (u_p, T_1, \text{tid}_1), (u_p, T_2, \text{tid}_2), \ldots, (u_p, T_n, \text{tid}_n)$, T_i refers to the moment for retweeting, and tid_i represents the retweeted ID of u_p.

2.6.2. Retweet Time Distribution. As the tweet is constantly exposed to astroturfers, it is continually retweeted. As a result, the average retweet time for astroturfing is much longer than that for normal internet users. The astroturfing tweets have a higher standard deviation and the lowest kurtosis. A near-zero skewness is also noticed for most astroturfing tweets, suggesting the even retweeting of these tweets.

2.6.3. The Most Dominant (MD) Application's Percentage. Twitter applications from the third party are often used to produce retweets for collusion-based astroturfing services. The percentage of the amount of retweets generated by the MD application to the total quantities of the retweets can be calculated as follows:

$$\text{RT}_{\text{ratio}} = \frac{\text{dominant application retweets}}{\text{total reweets}}. \tag{9}$$

On an average, approximately 90% of the crowdturfing retweets and nearly 40% of the normal retweets are generated by dominant applications.

2.6.4. The Number of Unreachable Retweeters. Most astroturfing users will not follow the initial tweeter who posts the tweet when retweeted since the astroturfing services randomly recommend tweets to uncertain users without focusing on the relationships between these users. Approximately 80% of the astroturfing tweets have about eighty percentage "unreachable" retweeters. As the identifiable feature, the proportion of "unreachable" retweeters for astroturfing is relatively high.

2.7. Post

2.7.1. Average Interval Time of Posts. It is considered that normal users are less aggressive when posting comments, while astroturfing is not interested in online discussion but is keen to complete the task in the least amount of time, suggesting a shorter average interval time from paid posters. Thus, for the same user, we calculate the mean interval time among different adjacent user comments within each signal episode, computing the total average overall execute episodes. The result indicates that 60% of the potential paid posters are posting at a speed of 200 seconds per interval, whereas only 40% of the normal users post comments at this speed.

2.7.2. The Posting Frequency. It refers to the average number of posts generated per day and is computed as follows:

$$nb_{/\text{day}}(\text{posts}) = \frac{nb_{\text{total}}(\text{posts})}{\text{age}_{\text{days}}(\text{account})}. \tag{10}$$

Astroturfing has a high posting frequency.

2.8. Click

2.8.1. The Number of Received Clicks. The links of retweeting are seldom clicked on as it is not within the duty of astroturfing accounts. The number of received clicks could

be small. We calculate the proportion of the total number of clicks to the retweet number per tweet as follows:

$$Cl_{\text{rat}} = \frac{\text{the number of received clicks}}{\text{the number of retweets per weet}}. \quad (11)$$

It is found that a smaller number of clicks is received for approximately 90% of the links within the astroturfing posted tweets than the total number of retweets, implying that retweets are performed by astroturfing accounts in other ways.

2.9. Active

2.9.1. Online Time Distribution. The user ID usually changes in astroturfing. The user ID is normally discarded when a mission is completed, and another ID will be adopted for a new mission. Thus, the number of active days of an online user is analyzed. The result indicates a high similarity of the percentage of potential paid users to that of common users in the groups with active days of 1, 2, 3, and 4 days. However, a few normal users participate in the discussion for 5 days or more, and there is no astroturfing active over 4 days.

2.10. Sentence

2.10.1. Sentence Transition. Given W, which denotes the corresponding words set, we calculate the transition probability from word w_p, which can form a p_{th} row matrix PTP, and it can be calculate as follows:

$$\begin{aligned} &\text{PTP}_{i,:} = \left[P\left(w_1 \mid w_p\right), \ldots, P\left(w_q \mid w_p\right), \ldots, P\left(w_n \mid w_p\right)\right], \\ &\text{s.t.} \sum_{w_q \in W} \left(w_q \mid w_p\right) = 1, \end{aligned} \quad (12)$$

where $P(w_q \mid w_p)$ denotes the transition probability from the word w_p to w_q. The PTP value of the composite comments is larger than that of the real comment.

2.10.2. Word Cooccurrence. There are cooccurrence patterns for words between the two different adjacent sentences. For the word w_p with w_q, the cooccurrence score can be defined as $O_{p,q} = \log(P_{p,q}/P_p P_q)$, in which $P_{p/qp/q}$ means the probability of the word $w_{p/q}$ appears, and $P_{p,q}$ denotes the joint probability.

2.10.3. Average Score. The average score of two consecutive sentences s_1 and s_2 can be represented as $\text{sco}^{\text{as}}(s_1, s_2)$, which can be calculated as follows:

$$\text{sco}^{\text{as}}\left(s_1, s_2\right) = \frac{1}{|s_1||s_2|} \sum_{w_p \in s_1, w_q \in s_2} O_{p,q}. \quad (13)$$

2.10.4. Best Score. The best score of two consecutive sentences $s_1$and s_2 can be represented as $\text{sco}^{\text{bs}}(s_1, s_2)$, which can be calculated as follows:

$$\text{sco}^{\text{bs}}\left(s_1, s_2\right) = \max_{w_i \in s_1, w_j \in s_2} O_{i,j}. \quad (14)$$

For a reviewer, with a sentence sequence $\{s_1, s_2, \ldots, s_n\}$, the coherence measure is defined based on the total average cooccurrence scores within the sentence, which can be calculated as $\text{sco}_r(r) = 1/n\sum_{i=1}^{n-1} \text{sco}(s_i, s_{i+1})$. For the synthetic comments, the SCO value is more possibly produced compared with the normal users' comments.

2.10.5. Pairwise Sentence Similarity. Such feature focuses on the similar signal word or topic that both represent by two adjacent sentences. For the user who gives the comment, the coherence score can be represented as the mean pairwise sentence similarity among all pairwise sentences, which can be calculated as $1/n - 1\sum_{i=1^{n-1}} \text{sim}(s_i, s_{i+1})$, and $\text{sim}(s_1, s_2) = 2|s_1 \cap s_2|/|s_1| + |s_2|$.

2.10.6. Semantic Dispersion. Given each sentence's vectorized semantic representation, the semantic dispersion enables a quantified representation of the review content's dispersion. Let $v_1, v_2, \ldots, v_n$ be the semantic vector representation corresponding to a review's each sentence; the semantic dispersion is defined as follows:

$$\text{SD} = \frac{1}{n} \sum_{i=1}^{n} \left\| v_i - \text{centroid} \right\|, \quad (15)$$

where centroid $= 1/n \sum_{i=1}^{n} v_i$. It is expected that the synthetic reviews usually have a larger SD than the truthful reviews.

2.10.7. Running length. We count the total quantity of sentences in the k_{th} piece and denoted as l_k, and the overall compactness of the review is measured by $\sum_{k=1}^{K} l_k/K$, in which K is the total quantity of pieces. This measure is denoted as the running length.

3. Learning-Based Astroturfing Detection Approach

To completely portray the character of spams, many researchers started to use ensemble models, using many kinds of information to train a classifier for spams detecting. It is mainly summarized into three parts, which are supervised, semisupervised, and unsupervised learning.

3.1. Supervised Learning-Based Detection Approach. Supervised learning-based method can be employed to realize the detection of a review spam. The underlying mechanism is to consider the task of review spam detection as a categorization task of separating the reviews into spam and nonspam comments.

3.2. Content Feature Learning

3.2.1. Expectation-Maximization (EM) model. The underlying idea of this method is to utilize the EM algorithm to construct a label prediction approach to give the prediction

for the untagged tweets. Based on the tweets set that contains of labelled and unlabelled tweets, the EM method is trained for a text classification task. Firstly, this model employs the Naïve Bayes classifier using the labelled tweets. For the E step of this model, this well-trained classifier can give the label prediction for the unlabelled tweets. For the M step, this model will give a reestimate for the word features' probabilities on the basis of the learned labels' prediction. Reasonably, the classifier within E can be updated, and the new classifier can be utilized to give a new prediction to the tweets in M. The overall interaction processing will continue until the number of changes in the predicted labels is below 0.01% of the total unlabelled tweets.

3.2.2. Support Vector Machines (SVM) and Naive Bayes. The semantic flow difference between the synthetic and truthful reviews can be leveraged to identify the review spam. Thus, the SVM and Naive Bayes classifiers can be used to distinguish if a review is truthful or synthetic.

3.3. Behavior Feature Learning

3.3.1. Adaboost. This famous method can be regarded as an iterative approach, and the underlying idea can be concluded as a merging task. Firstly, it involves training various "weak" classifiers based on the same training set, and for the following processing, it involves combining different "weak" classifiers into a "stronger" one. Such processing is realized by changing the data distribution. According to the classification accuracy of each sample, we determine the weights of each sample according to the classification accuracy and the overall classification precision. Moreover, the lower classifier will be retrained with a new date set equipped with updated weight. Thus, the final decision classifier can be obtained by merging the well-trained lower classifiers.

3.3.2. K-Nearest. This algorithm can be utilized to solve tasks, such as regression and classification, which can be regarded as a nonparametric method. The input for this algorithm consists of k samples that closets to each other within the feature space. The constant k is defined by the user, and in the phase of classification, the unlabelled sample is assigned to the most frequency label tagged to the k-closest samples. For the detection of astroturfing based on behaviors, the aim is to differentiate between the two types of tweets: one receiving retweets from astroturfing accounts and the other receiving retweets from normal accounts.

Four novel retweet-based features are found by Song and Lee et al. [21]. It enables us to discriminate astroturfing tweets from others. The four features can be concluded as: (1) the distribution of retweet time, (2) the most dominant application proportion, (3) the total quantity of unreachable retweeters, and (4) the received clicks amount. Based on the features of retweet, the authors constructed the K-Nearest Neighbors and evaluated the accuracy of this model according to the ground truth.

3.4. Structure Feature Learning

3.4.1. Logistic Regression. According to the logistic function-estimating probability, such an algorithm measures the relationship between the dependent classification variable and one variable or various independent variables. In this context, logistic regression can be regarded as the distribution of cumulated logistic. Therefore, this method can utilize a similar technology to process the same group of problems of probit-regression, in which probit-regression utilizes cumulated normal distribution.

3.4.2. Support Vector Machines (SVM). Chen et al. [22] investigated a web forum and discovered that the comment spams and their senders have some common features, such as the proportion of spam publication, publisher ID for spams, first poster, reply comment, publish timing, and tweet activity. The relationship between different spam publishers is obtained in their research. They also built a classifier SVM with an RBF kernel to detect the spammer. There are two important hyperparameters in the radial basis function (RBF) kernel-based SVM that needs to be optimized, namely, C and γ. During the learning of the model, employing multiple five-folds cross-validation to the training set and taking the F measure as the optimization measure to realize the grid searching of C and γ can be concluded as follows:

$$(C, \gamma) \in \{10^x | -3 \le x \le 3, x \in Z\} \times \{10^y | -5 \le y \le 2, y \in Z\}. \tag{16}$$

3.5. Multiple Feature Learning

3.5.1. Random Forests. This model can be regarded as a kind of ensemble learning approach designed to realize classification, regression, and other machine learning tasks. A random forest is realized by constructing a huge number of decision trees during training. It outputs the classification result or the average prediction score of the single tree.

Lee et al. [23] proposed a comprehensive analysis of astroturfing and trained a random forest-based classifier to detect astroturfers. They identified a few valuable features such as profile features, content features, and social network features. Accordingly, the authors calculated the feature values of each user in the training and validation set. The authors select the popular classification algorithm: random forest. Lee et al. developed a classifier based on a random forest that can predict whether the user is normal.

3.5.2. Neural Autoencoder Decision Forest. It has been proven that autoencoder is a kind of robustness algorithm, and it can give an unsupervised description within the feature mode. Random forest can be regarded as the set of multiple decision trees that can solve the problem of overfitting effectiveness. Extensive experiments show that this method performs well in real world practice.

Dong et al. [24] proposed a unified model which is trainable and end-to-end based on the interesting characteristic of the autoencoder and random forest. In this model, Dong et al. utilize the hidden representation of the feature generated by the autoencoder. The whole model is jointly trained by two models, namely, the stochastic decision tree model and the differentiable one. The final prediction is generated by the decision forest.

4. Semisupervised Learning-Based Detection Approach

Evidence from other areas suggests that the learner accuracy can be considerably enhanced by combining the unlabelled data with a small amount of labelled data compared to methods that are completely supervised.

4.1. Content Feature Learning

4.1.1. Two-View Method. Li et al. [25] proposed a semisupervised method with two views for utilizing a large quantity of the available unlabelled comments to ensure spam detection established by the cooperative training algorithm framework. Their dataset was manually labelled. Out of the 6000 reviews from the Internet (https://www.Epinions.com), 1394 reviews are marked as spam.

4.1.2. PU-Learning Model. Liu et al. [26] designed a kind of semisupervised learning method that is trained based on the set that consist of a few good samples and a quantity of unlabelled samples. Liu et al. named the model the PU-learning model. In experimental evaluation, such a model can achieve the accuracy of 83.7% under the F measure.

4.1.3. FakeGAN Model. The lacking of substantial ground truth poses a major challenge to the classification methods for deceptive review detection. Hence, Aghakhani et al. [27] proposed the system FakeGAN, where the learning approach based on semisupervised neural network is firstly employed to detect a deceptive spam. Unlike the standard GAN models, the FakeGAN adopts three models in total, which consists of two discriminators and one generator. In reinforcement learning (RL), the generator is modelled as a stochastic policy agent, and Monte Carlo search algorithm is utilized in the discriminator to estimate and pass the intermediate action value, which can act as the RL reward for generator.

4.2. Behavior Feature Learning

4.2.1. C4.5. For the given data set, each tuple can be represented as a group of attributed vales, and each tuple belongs to one of the mutually exclusive classes. The purpose of C4.5 is to find the mapping from the attribution to the classification through learning, and this mapping can be utilized to classify the unlabelled new objects.

In the detection of astroturfing, Xu et al. [28] presented an analysis of the whole system of spam attack from multiple perspectives. They used the profile attributes, QA attributes, and SN attributes of the users as features to train a classifier (C4.5) for detecting the spammers.

4.2.2. Unsupervised Learning-Based and Other Detection Approach. Facing the challenge of implementing an accuracy label for a review spam dataset, it is not inapplicable to use supervised learning, occasionally. A new unsupervised text mining model is proposed.

4.3. Content Feature Learning

4.3.1. Latent Dirichlet Allocation (LDA)-Based Model. LDA is a document theme generation model used to analyze the topics studied by each user's cluster. This model takes a text corpus as the input, and it outputs K themes. Each theme is a list of words and is ranked according to the relevance to this theme.

Dong et al. [29] realized the LDA model-based unsupervised topic-sentiment joint (UTSJ) probability model. To obtain the topic-sentiment joint probability distribution vector for each comment, UTSJ makes the first attempt to employ the Gibbs sampling algorithm to realize the estimation of parameters for the maximum likelihood function with the offline way. Furthermore, the UTSJ model takes a kind of offline training to separately obtain the random forest-classifier and the SVM-classifier. Extensive experiments show that the UTSJ model is obviously better than other baseline models.

When detecting astroturfing with LDA, Yang et al. [30] used OpenCLAS and the words toolkit developed by Sogou to divide each message into a single word. Moreover, Yang et al. further improved the accuracy of LDA by combining each tweet and the corresponding review as an individual document. They also filtered out the most frequent words that appeared in the top 10. Yang et al. employed LDA to the subdocument of the corpus, which is randomly sampled. Extensive experiments show that when compared to the state-of-the-art methods in this field, LDA outperforms the others.

4.3.2. MF Model. Ma et al. [31] employed a matrix factorization model based on the orthogonal nonnegative matrix trifactorization model (ONMTF) to learn the lexicon knowledge from the spam. They suggest learning external information on the level of topic instead of studying on the level of word. The underlying idea of this model is to realize the clustering of data samples based on the feature distribution. Furthermore, we cluster data instances based on the distribution of features. ONMTF can be realized by optimizing the follow formula:

$$\begin{aligned} &\min_{U,H,V\geq 0} \left\| X - \mathrm{UHV}^T \right\|_F^2, \\ &\text{s.t.} \quad U^T U = I, \quad V^T V = I, \end{aligned} \tag{17}$$

where X denotes the context matrix. U is the low dimension representation of the word and V indicates the low dimension representation of the user. They are nonnegative matrices. By adding a least squares penalty to the level of topic in the ONMTF model, Ma et al. project the initial context information to the topic space.

4.3.3. Semantic Language Model (SLM). Lau et al. [32] proposed the SLM that can be regarded as a unsupervised model to solve the problem of text mining. SLM can be integrated to a semantic model to detect the fake reviews published by astroturfing. The underlying idea of this model is to establish a kind of approximate computing method that can imply the fake degree of reviews. More specifically, Raymond et al. utilized SLM to evaluate the semantic overlap among different reviews. Instead of the unsupervised detecting of review spams, Chen et al. also proposed an idea about the high-order associate mining to capture the concept associate knowledge, which is context-sensitive. Assuming that the semantic of one review is extremely similar to the other review, these two reviews are likely to be spam reviews with high probability. Based on the Amazon reviews collected, Raymond et al. constructed a dataset whose AUC can achieve 99.87%.

4.4. Structure Feature Learning

4.4.1. DetectVC. Liu et al. [33] employed a detection method named DetectVC that can realize the robustness and deficiency detection for volowers and customers. DetectVC utilizes the inherent motivation and purpose of the volowers and the customers. This method combines the graph structure of the relationship within the tweet users' followers and the prior knowledge collected from the follower context.

4.4.2. Markov. The Markov model can be regarded as a kind of stochastic model that can be utilized to construct a randomly changing system. The Markov model assumes that the future state depends on the current state alone, and it has no relation to the event that happens before it. Generally speaking, such an assumption is used only in the reasoning and computing process of the model. Thus, it is reasonable that a given model has the property of Markov under the model and the probability prediction fields.

In the structure-based detection of astroturfing, Fakhraei et al. [34] proposed a method to detect abnormal users under the scenario of multiple relational social networks. Social network can be regarded as a multiple relational graph with a time stamp where users are represented as vertices and different activities between them are edges. The authors use the mixture of Markov models. Fakhraei et al. assumed that the operating of each user is generated by the mixture of Markov model. Specifically, each cluster in the spam sender or nonspam sender is associated to a mixture component y. Taking the cluster y for a user as a condition, the authors assume that the action sequence for a user is generated by the corresponding Markov. The joint probability can be calculated as follows: y is the user's cluster and $x_i, \ldots, x_n$ is the action sequence, $P(x_i|y)$ denotes the probability of x_i when the cluster is y.

$$P(y, x) = P(y)P(x_1 \mid y)\prod_{i=2}^{n} P(x_i \mid x_{i-1}, y). \tag{18}$$

4.4.3. Recurrent Convolutional Neural Network (RCNN) Model. Zhang et al. [35] made the first attempt to propose a fake review detection method based on deceptive review identification by RCNN, namely DRI-RCNN. This method makes use of the context and deep learning technology to identify fake reviews. Zhang et al. utilized the RCNN vector to represent each word in a review based on the deceptive context and truthful context property of reviews and word embedding. Furthermore, the authors also developed a deep neural network that combines max pooling and ReLU filter to detect fake reviews.

4.5. Multiple-Feature Learning

4.5.1. CrossSpot Model. Jiang et al. [6] proposed the CrossSpot, which can be regarded as a scalable searching approach. CrossSpot finds dense suspicious areas under multiple mode data and realizes sorting based on the degree of suspiciousness. This method starts from a potentially suspicious module, and it takes iterative for updating to determine the optimal setting value for mode j. Meanwhile, CrossSpot makes sure that the included value for all other modes will keep the same. The aforementioned process of updating will continue until they converge.

4.6. Graph-Model Based. Numerous methods are based on graph models, especially the structure-based ways in previous papers [8, 13, 34]. The graph model-based technology can be widely used in the detection of fake reviews, social spam [8, 9], link farming [13], etc.

Ratkiewicz et al. [36] proposed a machine learning framework that combines the features of information spreading network on Twitter, including the topological feature, context-based feature, and crowdsourced feature. This framework is designed to detect the early stage of the policy fake information spreading. To describe the information flow of the Twitter community, Ratkiewicz and Conover et al. constructed a directed graph in which the "node" denotes the individual user account and the "edge" indicates the operations of retweets and following.

Hu et al. [9] made the first attempt to analyze the sentiment differences of spam sender and the normal user, and designd an optimization formula that can incorporate sentiment information into the social spam sender detection framework. They conduct the modelling of content information with the two constrains of the learned factor matrix U and learn those two constrains from the social network information in addition from the sentiment information. More specially, Hu and Tang et al. constructed user emotion

information-based undirected graph in the processing of emotion modelling.

Shehnepoor et al. [37] realized a new framework named NetSpam. These frameworks use spam features to model the review data set as a heterogeneous information network and map the spam detection processing to the category problem within such networks.

Liu et al. [38] realized a classification method based on the complex probability graph that can solve the abnormal users detection problem. To obtain an initial effective estimation for the nodes (reviews, authors, and products) in the graph, Liu et al. made use of the attention machine to train a neural network and study the embedding representation of multiple modes using the text and various other features. On the basis of the prior calculation of the node, this classification method captures the relationship among different types of nodes based on the construction of a heterogeneous graph.

The existing works in the astroturfing detection field only take one or two types of astroturfing objects into account, e.g., text comment, reviewer, reviewer group, and product. Noekhah et al. [39] proposed a multi-iterative graph-based opinion spam detection (MGSD), which can be regarded as a graph-based model. MGSD utilizes a multiple iterative algorithm that takes various factors into consideration to update the entities' score. Furthermore, to improve the detection accuracy of MGSD, Noekhah et al. combined the feature fusion technology and machine learning algorithm to select more weighted features and new features from various categories. Extensive experiments show that the feature selection technology and the feature fusion technology can improve the performance of astroturfing detection.

To summarize, we compare the learning mode, feature, and based model of different literature in Table 3 to show "how" each work is executed.

5. Evaluation Criterion

There are many metrics that can be utilized to evaluate the astroturfing detection algorithm's performance, for example, accuracy, F1 score, precision, recall, AUROC, FPR,. All these metrics are used for classification models. When there is a low spam post ratio, the accuracy cannot be taken as a strong metric because of the domination by the majority nonspam class. Since it is not expected that a normal user's review is regarded as a spam, the spam detection approach should have a relatively high accuracy. In addition, to identify as much spam as possible, the spam detection approach should have high recall. For example, when the detection system utilizes manual classification to categorize the spam under the original filter, select the top misclassification with high recall to reduce the possibility of identifying entities completely. Meanwhile, the misclassification for detecting the normal review as a spam can be corrected later. Neither precision nor recall is the priority. Thus, the evaluation metric will take the weighted mean of precision and recall, which can be named the F measure.

In Y. R. Chen and H. H. Chen [22], the baseline for detecting the confusing forum spams is generated by leveraging the spreadsheets. Yang et al. [40] developed a Sybil detector based on measuring, and the ground truth is provided by Renren company. The Sybil is utilized widely on Renren, and it detects over 100 thousand Sybil accounts. Sedhai and Sun [41] made use of 2104 manually labelled classes as the experimental ground truth. Moreover, based on the KNN classification, the labels will increase with a more efficient way.

5.1. Precision, Recall, and F1 Score. The precision and recall are the commonly used classification metrics. Precision assesses the true positive part under the samples that are classified into positive. Recall assesses the proportion of positive samples that are labelled correctly. The F1 score is the weighted mean of precision and recall, which trades off these two metrics.

$$\begin{aligned} \text{Precision} &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ \text{Recall} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ F1 &= \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \end{aligned} \tag{19}$$

where TP denotes true positive, which means that a sample is classified to be positive as the sample is actually positive, FP indicates false positive, which means that a sample classified as positive but is actually negative, while FN denotes false negative, which means that a sample is classified as negative but is actually positive.

The weights of the two parts in F1 score are equal. When a detection algorithm can maximize the precision and recall at the same time, the algorithm will perform well. Thus, if one algorithm performs well in both sides appropriately, it will be better that it performs extremely well in one side and poorly in the other [42].

In the area of spammer detection, many approaches used precision, recall, or F1 score to act as the evaluation metric and assessed the models' performance, such as Hu et al. [43], Hu et al. [44], Hu et al. [45], Sedhai and Sun [41], Liu et al. [33], Dong et al. [29], Liu Pang [46], Liu et al. [38], Dong et al. [24], and You et al. [47].

5.2. AUROC and AUPR. AUROC means the area under the ROC, and it draws how TPR (True Positive Rate) changes according to the changes of FPR (false positive rate), in which TPR denotes the proportion of a sample classified as correctly positive, and FPR indicates a sample that is classified as positive but is actually negative. AUPR is the area under precision-recall. In the work of Fakhraei et al.[34], in order to avoid overoptimistic estimates of the PR curve and the ROC, Fakhraei used AUROC and AUPR to estimate the performance of their method. However, in the work of Song et al. [21] and Wang et al. [48], they used AUROC alone as the evaluate metric.

Table 3: Comparison of the learning mode, feature, and based model among different literature.

Literature	Learning mode			Feature	Based model
	Supervised learning	Semi-supervised learning	Unsupervised learning		
Lee et al. [21]	√	—	—	Behavior	K-nearest
Chen et al. [22]	√	—	—	Structure	SVM
Lee et al. [23]	√	—	—	Multiple	Random forest
Dong et al. [24]	√	—	—	Multiple	Decision forest
Li et al. [25]	—	√	—	Content	Two view
Liu et al. [26]	—	√	—	Content	PU-learning
Aghakhani et al. [27]	—	√	—	Content	GAN
Xu et al. [28]	—	√	—	Behavior	C4.5
Dong et al. [29]	—	—	√	Content	LDA
Yang et al. [30]	—	—	√	Content	LDA
Ma et al. [31]	—	—	√	Content	MF
Raymond et al. [32]	—	—	√	Content	SLM
Liu et al. [33]	—	—	√	Structure	DetectVC
Fakhraei et al. [34]	—	—	√	Structure	Markov
Zhang et al. [35]	—	—	√	Structure	RCNN
Jiang et al. [6]	—	—	√	Structure	CrossSpot

5.3. TPR, FPR, and FNR. TPR and FPR have been explained above. FNR (False Negative Rate) indicates the proportion of a sample that is classified as negative but is actually positive. In the actual experiment, we hope that the TPR is as large as possible, while the FPR and FNR are as small as possible.

$$\begin{aligned} \text{TPR} &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \\ \text{FPR} &= \frac{\text{FP}}{\text{FP} + \text{TN}}, \\ \text{FNR} &= \frac{\text{FN}}{\text{TP} + \text{FN}}. \end{aligned} \tag{20}$$

In their works, Lee et al. [23], Barushka and Hajek [49], and Lee et al. [50] used the FPR and FNR as metrics to evaluate the classifier for spammer detection. In the work of Morales et al. [51] and Xu et al. [28], both of them used the TPR and FPR as measures to evaluate the detection performance of the classifier.

5.4. Accuracy and ER. Accuracy assesses the proportion of samples that are classified correctly for all examples. Meanwhile, the ER (error rate) measures the fraction of misclassified examples over all examples. They are also the most frequently used evaluation metric.

$$\begin{aligned} \text{Accuracy} &= \frac{\text{TP} + \text{TN}}{P + N}, \\ \text{ER} &= \frac{\text{FP} + \text{FN}}{P + N}. \end{aligned} \tag{21}$$

Morales et al. [51] and Wang et al. [52] used the ER as the experimental metric for evaluation. Meanwhile, Lee et al. [23], Dong et al. [24], Aghakhani et al. [27], Yang et al. [30], Zhang et al. [35], Li et al. [53], Dhingra and Yadav [54], and You et al. [47] used the accuracy as a measure to evaluate the classifier they built. Furthermore, as shown in Table 4, we compared several literature under different evaluation criterions to analyze the evaluation stand in the field of astroturfing detection, in which the best performance under different criterions are bolded.

6. Applications

Astroturfing can disrupt the normal order of the network and bring many negative effects to society and people's life. Hence, it is essential to design schemes for the assistance of normal users, administrators, law enforcers, etc. Astroturfing detection can help users distinguish truth from falsehood and obtain the information that they really need. In social networks, astroturfing detection is of importance for multiple applications. This section concerns with several typical applications, for instance, the detection of astroturfing in social networks, astroturfing account recognition, and deceptive reviews identification.

7. Single Astroturfing Detection

7.1. Detecting Astroturfing in Twitter. Lee et al. [23] proposed a comprehensive analysis for the identification of astroturfing on Fiverr and Twitter. They found valuable features such as profile features, activity features, content features, social network features, personality features, and temporal features. For the training set and the testing set, Lee and Webb et al. computed each user's feature values and trained a corresponding classifier (SVM-based classifier and Random Forest-based classifier) to detect astroturfers on Fiverr.com and Twitter.

7.2. Automatic Review Synthesis Model. Morales et al. [51] leveraged the difference of semantic flows between the semantic and truthful reviews to identify the review spam, and they used SVM and Naive Bayes as classifiers to identify if one review is truthful or synthetic. Positive reviews are automatically generated in their model by mixing up existing reviews.

Table 4: Comparison of several literature under different evaluation criterions.

Literature	Evaluation criterion									
	Precision	Recall	F1	AUROC	AUPR	TPR	FPR	FNR	Accuracy	ER
Lee et al. [21]	—	—	—	—	—	**0.98**	0.010	—	—	—
Lee et al. [23]	—	—	**0.974**	—	—	—	**0.008**	0.248	**97.35%**	—
Dong et al. [24]	**96.08%**	94.15%	0.951	—	—	—	—	—	95.85%	—
Li et al. [25]	61.40%	62.10%	0.631	—	—	—	—	—	—	—
Liu et al. [26]	—	—	0.856	—	—	—	—	—	—	—
Aghakhani et al. [27]	—	—	—	—	—	—	—	—	89.10%	—
Xu et al. [28]	—	—	—	—	—	0.94	0.096	—	—	—
Dong et al. [29]	87.15%	83.02%	0.850	—	—	—	—	—	—	—
Fakhraei et al. [34]	>50.00%	80.00%	—	**0.914 ± 0.001**	**0.543 ± 0.005**	—	—	—	—	—
Zhang et al. [35]	—	—	0.866	—	—	—	—	—	88.15%	—
Ratkiewicz et al. [36]	—	—	—	—	—	—	—	—	96.40%	—
Liu et al. [38]	82.00%	80.00%	0.810	—	—	—	—	—	—	—
Noekhah et al. [39]	—	—	—	—	—	0.94	0.018	**0.058**	96.17%	—
Hu et al. [43]	86.50%	93.90%	0.901	—	—	—	—	—	—	—
Hu et al. [44]	91.30%	94.40%	0.928	—	—	—	—	—	—	—
Liu et al. [46]	79.48%	79.49%	0.793	—	—	—	—	—	78.62%	—
You et al. [47]	72.80%	73.30%	0.730	—	—	—	—	—	73.00%	—
Barushka et al. [49]	—	—	0.959	—	—	—	0.012	0.127	96.16%	—
Lee et al. [50]	—	—	0.966	—	—	—	0.036	0.174	93.26%	—
Sun et al. [51]	—	—	—	—	—	0.86	0.250	—	—	**21.60%**
Li et al. [53]	84.40%	**95.50%**	0.861	—	—	—	—	—	83.50%	—

Bold indicates the maximum value under each evaluation criterion.

8. Group Astroturfing Detection

8.1. Real-Time Detection System. Detecting astroturfing by establishing the classifier is a typical application for astroturfing detection. In Chen et al. [55], the fundamental architecture and the design of a detection system that identifies malicious behaviors and potential paid posters in real time are discussed. The purpose of their system is to identify potential paid posters and locate their user IDs during the information collection process. This system can not only automatically collect data from different resources/websites and report the behavior of potential paid posters but also provide valuable information for the analysts and online users. Four major components are involved: data crawler, scheduler, data analyzer, and database system.

8.2. Multiagent System. Analyzing the distribution and behavior characteristics of astroturfing is also helpful to better understand and monitor the astroturfing accounts. Zeng et al. [56] surveyed the behavior mode and policy of astroturfing on the Internet forums. They constructed a multiple-agent system [57], and utilized the ground truth data set of the online forums to conduct extensive experiments. Furthermore, they took the research of the factors that can impact the astroturfers' influence. Zeng et al. found that astroturfing maximized their influence by adjusting their behavior policy dynamically, and the effectiveness of their policy was highly related to the users' features.

Zeng et al. developed a multiple-agents technology-based social network environment. This model has two different agents, namely astroturfers and users. These two types of agents utilize the theme features and user features to evaluate the complex dynamics of cooperative coevolution within the multiple agent system, which takes the accumulated polarities as the basis.

8.3. IWA Social Network. Liu et al. [58] developed a new social network named as the IWA social network. IWA consists of two types of nodes, which are nodes that take IWAs as the core and the normal expanded nodes that communicate with IWAs. IWA is a kind of unnatural network, and the core nodes communicate with others because of their own economic interests or other interactions.

Considering of the features of the IWA, Liu et al. think that there are a group of members consist in the IWA social network, and each member controls several good accounts that own huge number of followers. In order to keep the property of transmission, each member confusing as the normal user. There are main account that astroturfing member can contact to others, which include both astroturfing users and normal users. In this way, IWA social networks can build connection with other normal users.

8.4. User Preference Graph. Astroturfing users may provide deceptive reviews to interfere with the judgment of normal users. Identify and filter out the deceptive reviews is also important.

Li et al. [59] extracted both textual features and contextual features. They proposed a new user preference graph to measure the user relationship. They incorporated both these features and the user preference relationship into the supervised learning framework and obtained more precision results for predicting the deceptive answer. They propose a new user graph to describe the relationship among users. Figure 1(a) shows the general process in a question-answering thread. Accordingly, other users will give a few

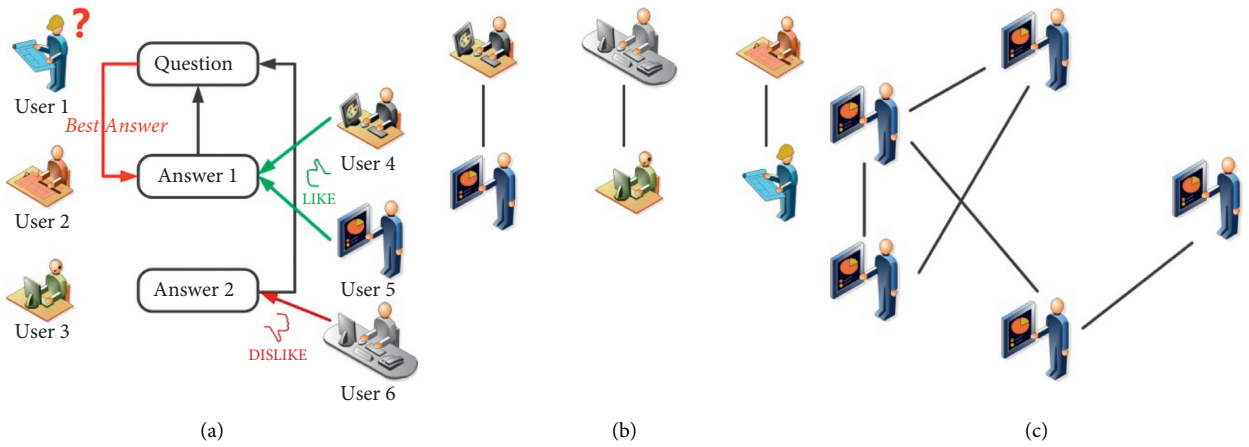

Figure 1: User Preference Graph Construction: (a) question answering, (b) user preference relation, and (c) user preference graph.

answers. Then, each user will give voting to each answer as "useful" and "unuseful" to indicate the evaluation of each user. Furthermore, the questioner will choose an optimal answer. If the two users show the same preference toward the target answer, then these two users have the same user preference, and they will have the user preference relationship. Li et al. extracted all the relationships of user preferences (shown in Figure 1(c), and each node denotes a user), in which there will be an edge exist if there is a user preference relationship between these two users.

8.5. FakeGAN System. The existing efforts in the field of spam detecting mainly focus on constructing a supervised classifier according to the review sentiment and lexical mode. Aghakhani et al. [27], inspired by the application of neural network in classification, developed the FakeGAN system. FakeGAN takes the first attempt to employ the generative adversarial network (GAN) in the task of text classification. Compared with the standard GAN model, FakeGAN model has two discriminators and a generator. Aghakhani et al. modelled the generator as a stochastic policy agent in RL and utilized the Monte Carlo search algorithm to estimate. In addition, they also took the intermediate action value as the RL reward, which will be transferred to the generator.

8.6. Weakly Supervised Fake News Detection Framework. Wang et al. [60] made the first attempt to propose a framework to detect fake news, which can be regarded as a kind of reinforced weakly-supervised learning, weakly-supervised fake news detection framework (WeFEND). WeFEND can get labelled the samples with high quality, facing the main challenge of detecting fake news with a deep learning model. WeFEND has three components, including the reinforced selector, fake news detector, and the annotator. Figure 2 demonstrates the framework of WeFEND.

9. Open Source or Prototype System

More applications have endorsed online reviews. To detect the spammer in the network, the computer science department of California proposed the system of spammer detection that utilizes a single real review as the template and to further replace the sentence with other review sentences in the storing dataset. They tested the performance of system using hotel reviews in the city of New York. The detection accuracy of this system is approximately 78%.

Fakhraei et al. [34] utilized k-gram features and probability modelling with the mixture of Markov model to obtain the relationship sequence. In addition, to improve the reasoning and prediction performance of the noise-reporting system, Fakhraei et al. proposed an analysis relationship model based on the Hinge Loss Markov Random Fields (HL-MRFs) and a kind of probability graph model that can be expanded to a high extent. The authors use GraphLab construction and Probability Soft Logic (PSL) as the experimental prototype and employ extensive experiments to evaluate their solution. Extensive experiments show that their model is effective, and the method proposed in this work can improve the prediction performance by integrating the multiple relationships feature of social network.

There are three components to support this framework. For the first component, extracting the graph feature within each relationship and experimental results show that considering the property of multiple relationships of the graph can improve the performance. For the second component, Fakhraei et al. took the action sequence of each user within these relationships into account. Furthermore, they extracted the k-gram feature and utilized the mixture of Markov model to label the spammer. Finally, for the third component, the authors of this work developed a HL-MRFs-based analysis relationship model to realize reasoning, which took the basis of the signal of the reporting system from the social network.

10. Future Directions from IT Perspective

10.1. Crossing Domain. The detection of astroturfing is an interdisciplinary research topic, and the main challenge of this topic is to apply the well-trained model to other target fields. Li et al. [61] trained their original model in the field of hotel and tested it in a restaurant field and a hospital field. However, comparing to the performance in the original

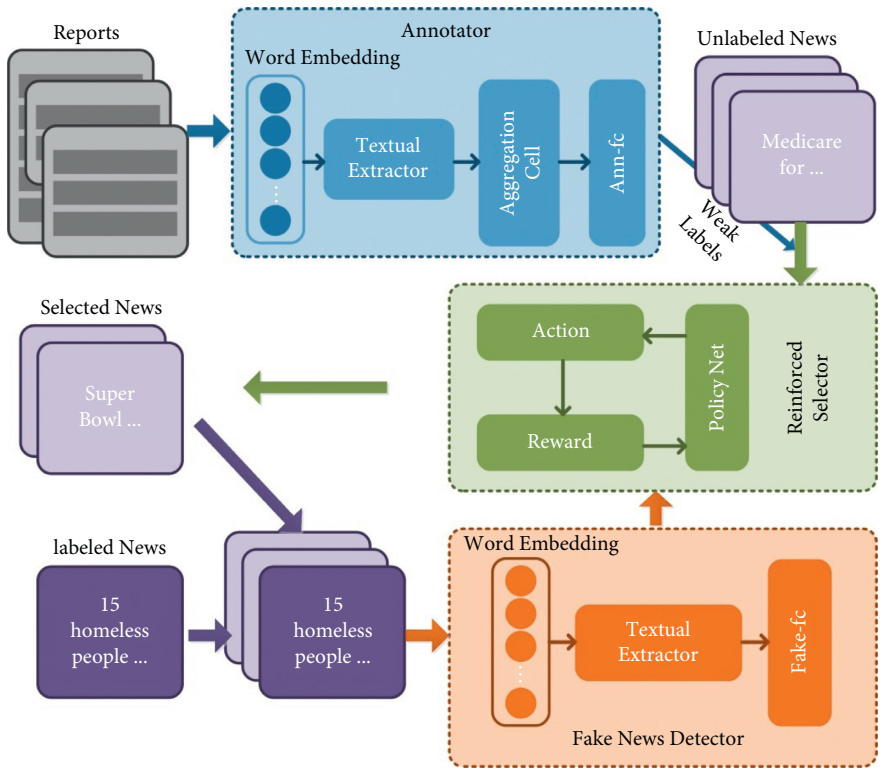

Figure 2: The architecture of proposed framework WeFEND.

field, the performance of their model dramatically decreased. Thus, the crossing domain detection of opinion astroturfing needed to be studied more deeply [62].

10.1.1. Missing Data. Public astroturfing generally works underground, and the available data are not enough for researching their behavior. Thus, the research work for the astroturfers behavior is difficult to be carried out.

10.1.2. Optimization Method. Recently, the semisupervised approach has mainly paid attention to the cooperative training and PU learning. However, the classification performance of these methods is not good enough. Though the constructing of astroturfing spam opinion dataset is difficult, the development of unsupervised learning approaches to identify the spam opinion is very important [62]. In addition, at present, there are excellent works for applying neural network model in studying the sentiment representation of the Natural Language Processing (NLP) task. Thus, the effectiveness design for neural network algorithms is the focus of the future research.

10.1.3. Complexity of the Internet. Astroturfing work is always cross-platform, cross-channel, and it has numerous sock puppets. This makes their behavior information fragmented and difficult to link to each other.

10.1.4. Evolutionary Astroturfers. Online astroturfers not only imitate legitimate users' behaviors, such as the number and the frequency of posting tweets, to avoid being identified [63], but also dynamically adjust their behavior policy to maximize their influence[56]; thus, the learned astroturfers detection systems can quickly go stale.

There are extensive works in the field of astroturfing detection, and there are also many potential challenges in this field. Many open tasks need to be researched in-depth. Thus, we discuss the possible challenges for future researching in the domain of astroturfing detection.

The essence of astroturfing is the actors' profitable purpose and behavior patterns rather than the content. As a result, a behavior-driven suspicious pattern analysis will be a significant driving force of the detection method. Taking combination of a wide domain of behavior information will better understand and distinguish suspicious behavior from normal behavior.

In short, different practical applications present different detection requirements of suspicious behavior. Astroturfing detection can be realized by optimizing the value of suspicion degree and finding the most suspicious part in the large-scale data of behavior. The principle is that "we would rather kill all just for one" (let the user complaints), instead of excessively ensuring the accuracy.

11. Cross-Domain Astroturfing Detection

Through finding the relationship among astroturfing behaviors across different domains, this correlation among astroturfing can be utilized to combine the traditional e-mail and the spam together into the detection algorithm of astroturfing, to further improve the detection effectiveness of social spam. For instance, the web spam detection method

has studied the web link structure widely [64, 65]. Moreover, utilizing such link-based method into the online community social corrections, can improve the detecting effectiveness of social spam obviously [56].

11.1. Utilized Temporal Information. The temporal information is very significantly in the detection of astroturfing, since the astroturfing users usually post huge messages within a few minutes, or posts a large number of reviews with various accounts within a very short period of time. Thus, studying the correlated behaviors with astroturfing on the level of time features, and predicting the future astroturfing behavior is very meaningful. Meanwhile, this indicates that we need to adjust model based on the temporal features accordingly, and make the model is more comprehensive.

11.2. Protection of Privacy of Individual Online Posters. In the Internet era, the user's privacy is more easily compromised. The crowd workers' on Amazon Mechanical Turk (MTurk) used to call for "Our Privacy Needs to be protected at All Costs" [66]. How to prevent users' information from being leaked and to give protection of the individual privacy of online posters will be a future challenge.

At present, several famous websites have taken measures in the aspect of user privacy protection. For example, Twitter selectively provides some data to the public (It only provides 1% of real user data for data mining [67].) in order to keep user sensitive data within tolerable limits. Moreover, JD Mall and Taobao.com have preprocessed consumers' comments. The method is as follows: Before reviewing the product, users need to confirm their comments are set to "public" or "anonymous." If the user chooses the anonymous comment, the system will preprocess the user information, replace the user's real avatar with the default image, and replace the user's nickname with ∗∗∗. For example, this system will turn a user's name "Hellosunshine" into "H∗∗∗e". At the same time, we cannot see the user's buyer show (the user's purchase lists and comments). In this way, the privacy of users can be protected on the premise that the water army detection is normal.

In addition, we find that the user who transfers astroturfing messages successfully may be a normal user, and they participate in astroturfing messages posting with no idea. In other words, they have been confused by the original astroturfing account. Therefore, further studies are required to design more generalized methods for solving the privacy protection problem.

Data Availability

All data generated or analyzed during this study are owned by all the authors and will be used for our future research. The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant nos. 61972025, 61802389, 61672092, U1811264, and 61966009 and the National Key R&D Program of China under Grant nos. 2020YFB1005604 and 2020YFB2103802.

References

[1] H. C. Kelman, "Compliance, identification, and internalization three processes of attitude change," *Journal of Conflict Resolution*, vol. 2, no. 1, pp. 51–60, 1958.

[2] E. Choo, T. Yu, and M. Chi, "Detecting opinion spammer groups through community discovery and sentiment analysis," in *Proceedings of the Conference on Data and Applications Security and Privacy*, pp. 170–187, Fairfax,VA, USA, July 2015.

[3] A. Heydari, M. A. Tavakoli, N. Salim, and Z. Heydari, "Detection of review spam: a survey," *Expert Systems with Applications*, vol. 42, no. 7, pp. 3634–3642, 2015.

[4] Z. Alom, B. Carminati, and E. Ferrari, "Detecting spammers on twitter," in *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 1191–1198, Barcelona, Spain, August 2018.

[5] A. Sundararaj and G. Kul, "Impact analysis of training data characteristics for phishing email classification," *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, vol. 12, pp. 85–98, 2021.

[6] M. Jiang, A. Beutel, P. Cui, B. Hooi, S. Yang, and C. Faloutsos, "A general suspiciousness metric for dense blocks in multimodal data," in *Proceedings of the IEEE International Conference on Data Mining*, pp. 781–786, Atlantic City, NJ, USA, November 2015.

[7] D. J. Lemay, R. B. Basnet, and T. Doleck, "Examining the relationship between threat and coping appraisal in phishing detection among college students," *Journal of Internet Services and Information Security*, vol. 10, pp. 38–49, 2020.

[8] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: social honeypots + machine learning," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 435–442, Geneva, Switzerland, July 2010.

[9] X. Hu, J. Tang, and H. Gao, "Social spammer detection with sentiment information," in *Proceedings of the IEEE International Conference on Data Mining*, pp. 180–189, Shenzhen, China, December 2014.

[10] M. Kolomeets, A. Chechulin, and I. Kotenko, "Bot detection by friends graph in social networks," *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, vol. 12, pp. 141–159, 2021.

[11] C. Johnson, B. Khadka, and R. B. Basnet, "Towards detecting and classifying malicious URLs using deep learning," *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, vol. 11, pp. 31–48, 2020.

[12] S. Narteni, I. Vaccari, and M. Mongelli, "Evaluating the possibility to perpetrate tunnelling attacks exploiting short-message-service," *Journal of Internet Services and Information Security*, vol. 11, pp. 30–46, 2021.

[13] S. Ghosh, B. Viswanath, and F. Kooti, "Understanding and combating link farming in the twitter social network," in *Proceedings of the International Conference on World Wide Web*, pp. 56–61, Lyon, France, April 2012.

[14] A. Kitana, I. Traore, and I. Woungang, "Towards an epidemic SMS-based cellular botnet," *Journal of Internet Services and Information Security*, vol. 10, pp. 38–58, 2020.

[15] L. Akoglu, R. Chandy, and C. Faloutsos, "Opinion fraud detection in online reviews by network effects," in *Proceedings of the International AAAI Conference on Web and Social Media,*, Santa Clara, CA, United States, July 2013.

[16] J. C. Stauber and S. Rampton, "Toxic sludge is good for you: lies, damn lies, and the public relations industry," *Journalism and Mass Communication Educator*, vol. 52, no. 3, pp. 314–317, 1995.

[17] J. G. McNutt, "Researching advocacy groups: internet sources for research about public interest groups and social movement organizations," *Journal of Policy Practice*, vol. 9, no. 3, pp. 308–312, 2010.

[18] C. H. Cho, M. L. Martens, H. Kim, and M. Rodrigue, "Astroturfing global warming: it isn't always greener on the other side of the fence," *Journal of Business Ethics*, vol. 104, no. 4, pp. 571–587, 2011.

[19] T. P. Lyon and J. W. Maxwell, "Astroturf: interest group lobbying and corporate strategy," *Journal of Economics and Management Strategy*, vol. 13, no. 4, pp. 561–597, 2004.

[20] J. Hoggan and R. Littlemore, "Climate cover-up: the crusade to deny global warming," *Energy & Environment*, vol. 21, no. 3, pp. 363-364, 2010.

[21] J. Song, S. Lee, and J. Kim, "Crowdtarget: target-based detection of crowdturfing in online social networks," in *Proceedings of the ACM Sigsac Conference on Computer and Communications Security*, pp. 111–114, Denver Colorado, USA, October 2015.

[22] Y. R. Chen and H. H. Chen, "Opinion spam detection in web forum: a real case study," in *Proceedings of the International Conference on World Wide Web*, pp. 173–183, Florence, Italy, May 2015.

[23] K. Lee, S. Webb, and H. Ge, "Characterizing and automatically detecting crowdturfing in fiverr and twitter," *Social Network Analysis and Mining*, vol. 5, no. 2, pp. 1–16, 2015.

[24] M. Dong, L. Yao, X. Wang, B. Benatallah, C. Huang, and X. Ning, "Opinion fraud detection via neural autoencoder decision forest," *Pattern Recognition Letters*, vol. 132, pp. 21–29, 2020.

[25] F. Li, M. Huang, and Y. Yang, "Learning to identify review spam," in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 2488–2493, Barcelona Catalonia, Spain, July 2011.

[26] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu, "Building text classifiers using positive and unlabeled examples," in *Proceedings of the International Conference on Data Mining*, p. 179, Melbourn Florida, USA, November 2003.

[27] H. Aghakhani, A. Machiry, S. Nilizadeh, C. Kruegel, and G. Vigna, "Detecting deceptive reviews using generative adversarial networks," in *Proceedings of the IEEE Security and Privacy Workshops*, pp. 89–95, San Francisco, CA, USA, May 2018.

[28] A. Xu, X. Feng, and Y. Tian, "Revealing, characterizing, and detecting crowdsourcing spammers: a case study in community Q&A," in *Proceedings of the Conference on Computer Communications*, pp. 2533–2541, Hong Kong, China, May 2015.

[29] L. Y. Dong, S. J. Ji, C. J. Zhang et al., "An unsupervised topic-sentiment joint probabilistic model for detecting deceptive reviews," *Expert Systems with Applications*, vol. 114, pp. 210–223, 2018.

[30] X. Yang, Q. Yang, and C. Wilson, "Penny for your thoughts: searching for the 50 cent party on sina weibo," in *Proceedings of the International Conference on Web and Social Media*, pp. 694–697, Palo Alto, CA, USA, May 2015.

[31] H. Ma, W. Zhao, and Q. Shi, "Orthogonal nonnegative matrix tri-factorization for semi-supervised document co-clustering," in *Proceedings of the Advances in Knowledge Discovery and Data Mining*, pp. 189–200, Hyderbad, India, June 2010.

[32] R. Y. K. Lau, S. Y. Liao, R. C. Kwok, K. Xu, Y. Xia, and Y. Li, "Text mining and probabilistic language modeling for online review spam detection," *ACM Transactions on Management Information Systems*, vol. 2, no. 25, pp. 1–30, 2011.

[33] Y. Liu, Y. Liu, M. Zhang, and S. Ma, "Pay me and I'll follow you: detection of crowdturfing following activities in microblog environment," in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 3789-3790, Palo Alto, California, USA, July 2016.

[34] S. Fakhraei, J. Foulds, M. Shashanka, and L. Getoor, "Collective spammer detection in evolving multi-relational social networks," in *Proceedings of the International Conference on Knowledge Discovery and Data Mining*, pp. 1769–1778, Sydney NSW, Australia, Auguest 2015.

[35] W. Zhang, Y. Du, T. Yoshida, and Q. Wang, "Dri-rcnn: an approach to deceptive review identification using recurrent convolutional neural network," *Information Processing & Management*, vol. 54, no. 4, pp. 576–592, 2018.

[36] J. Ratkiewicz, M. Conover, M. Meiss, A. Flammini, M. Menczer, and B. Goncalves, "Detecting and tracking political abuse in social media," in *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pp. 297–304, Barcelona, Catalonia, Spain, July 2011.

[37] S. Shehnepoor, M. Salehi, R. Farahbakhsh, and N. Crespi, "Netspam: a network-based spam detection framework for reviews in online social media," *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 7, pp. 1585–1595, 2017.

[38] Y. Liu, B. Pang, and X. Wang, "Opinion spam detection by incorporating multimodal embedded representation into a probabilistic review graph," *Neurocomputing*, vol. 366, pp. 276–283, 2019.

[39] S. Noekhah, N. B. Salim, and N. H. Zakaria, "Opinion spam detection: using multi-iterative graph-based model," *Information Processing & Management*, vol. 57, no. 1, Article ID 102140, 2020.

[40] Z. Yang, C. Wilson, X. Wang, T. Gao, B. Y. Zhao, and Y. Dai, "Uncovering social network Sybils in the wild," *ACM Transactions on Knowledge Discovery from Data*, vol. 8, no. 2, pp. 5–33, 2014.

[41] S. Sedhai and A. Sun, "HSpam14: a collection of 14 million tweets for hashtag-oriented spam research," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 223–232, Santiago Chile, Auguest 2015.

[42] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[43] X. Hu, J. Tang, Y. Zhang, and H. Liu, "Social spammer detection in microblogging," in *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 1709–1714, Menlo Park, CA, USA, Auguest 2013.

[44] X. Hu, J. Tang, and H. Liu, "Leveraging knowledge across media for spammer detection in microblogging," in *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 547–556, Gold Coast Queensland, Australia, July 2014.

[45] X. Hu, J. Tang, and H. Liu, "Online social spammer detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 59–65, Québec City Québec Canada, July 2014.

[46] Y. Liu and B. Pang, "A unified framework for detecting author spamicity by modeling review deviation," *Expert Systems with Applications*, vol. 112, pp. 148–155, 2018.

[47] L. You, Q. Peng, Z. Xiong, D. He, M. Qiu, and X. Zhang, "Integrating aspect analysis and local outlier factor for intelligent review spam detection," *Future Generation Computer Systems*, vol. 102, pp. 163–172, 2020.

[48] X. Wang, B. Zhou, Y. Jia, and S. Li, "Detecting internet hidden paid posters based on group and individual characteristics," *Lecture Notes in Computer Science*, in *Proceedings of the International Conference on Web Information Systems Engineering*, pp. 109–123, Miami, FL, USA, November 2015.

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

[50] K. Lee, P. Tamilarasan, and J. Caverlee, "Crowdturfers, campaigns, and social media: tracking and revealing crowdsourced manipulation of social media," in *Proceedings of the International AAAI Conference on Weblogs and Social Media*, pp. 331–340, Cambridge, MA, USA, July 2013.

[51] A. Morales, H. Sun, and X. Yan, "Synthetic review spamming and defense," in *Proceedings of the International Conference on World Wide Web*, pp. 155-156, Rio de Janeiro Brazil, May 2013.

[52] G. Wang, T. Wang, H. Zheng, and B. Y. Zhao, "Man vs. machine: practical adversarial detection of malicious crowdsourcing workers," in *Proceedings of the USENIX Security Symposium*, pp. 239–254, San Diego, CA, USA, Auguest 2014.

[53] L. Li, B. Qin, W. Ren, and T. Liu, "Document representation and feature combination for deceptive spam review detection," *Neurocomputing*, vol. 254, pp. 33–41, 2017.

[54] K. Dhingra and S. K. Yadav, "Spam analysis of big reviews dataset using fuzzy ranking evaluation algorithm and hadoop," *International Journal of Machine Learning and Cybernetics*, vol. 10, no. 8, pp. 2143–2162, 2019.

[55] C. Chen, K. Wu, S. Venkatesh, and X. Zhang, "Battling the internet water army: detection of hidden paid posters," in *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining*, pp. 116–120, Niagara Ontario, Canada, Auguest 2013.

[56] K. Zeng, X. Wang, Q. Zhang, X. Zhang, and F. Y. Wang, "Behavior modeling of internet water army in online forums," *IFAC Proceedings Volumes*, vol. 47, no. 3, pp. 9858–9863, 2014.

[57] M. Niu, B. Cheng, Y. Feng, and J. Chen, "GMTA: a geo-aware multi-agent task allocation approach for scientific workflows in container-based cloud," *IEEE Transactions on Network and Service Management*, vol. 17, no. 3, pp. 1568–1581, 2020.

[58] W. Liu, Y. Cao, D. Li et al., "Structural analysis of IWA social network," *Applications and Techniques in Information Security*, vol. 557, pp. 141–152, 2015.

[59] F. Li, Y. Gao, S. Zhou, X. Si, and D. Dai, "Deceptive answer prediction with user preference graph," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 1723–1732, Sofia, Bulgaria, Auguest 2013.

[60] Y. Wang, W. Yang, F. Ma et al., "Weak supervision for fake news detection via reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 516–523, New york NY, USA, February 2020.

[61] J. Li, M. Ott, C. Cardie, and E. Hovy, "Towards a general rule for identifying deceptive opinion spam," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pp. 1566–1576, Baltimore, Maryland, June 2014.

[62] Y. Ren and D. Ji, "Learning to detect deceptive opinion spam: a survey," *IEEE Access*, vol. 7, Article ID 42945, 2019.

[63] S. Liu, J. Zhang, and Y. Xiang, "Statistical detection of online drifting twitter spam," in *Proceedings of the ACM on Asia Conference on Computer and Communications Security*, pp. 1–10, Xi'an China, May 2016.

[64] L. Becchetti, C. Castillo, D. Donato, R. Y. Baeza, and R. B. Eito, "Link-based characterization and detection of web spam," in *Proceedings of the International Workshop on Adversarial Information Retrieval on the Web*, pp. 1–8, Seattle, WA, August 2006.

[65] A. A. Benczur, K. Csalogany, and T. Sarlos, "Link-based similarity search to fight web spam," in *Proceedings of the International Workshop on Adversarial Information Retrieval on the Web*, pp. 9–16, Seattle, WA, August 2006.

[66] H. Xia, Y. Wang, Y. Huang, and A. Shah, "Our privacy needs to be protected at all Costs," in *Proceedings of the ACM on Human-Computer Interaction*, pp. 1–22, NY, United States, November 2017.

[67] T. Wu, S. Wen, Y. Xiang, and W. Zhou, "Twitter spam detection: survey of new approaches and comparative study," *Computers & Security*, vol. 76, pp. 265–284, 2018.