IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2019.Doi Number

An Internet Water Army Detection Supernetwork Model

YING LIAN[1,#], XUEFAN DONG[1,#], YUXUE CHI[2,3], XIANYI TANG[4], YIJUN LIU[2,3]

[1]Beijing Jiaotong University, No.3 Shangyuancun Haidian District Beijing 100044 P. R. China
[2] Institutes of Science and Development, CAS, No.15 ZhongGuanCunBeiYiTiao Alley, Haidian District, Beijing 100190, PR China
[3] University of Chinese Academy of Sciences, No.19A Yuquanlu, Beijing 100049, PR China
[4] CAS Center for Interdisciplinary Studies of Social and Natural Sciences, Chinese Academy of Sciences, No.15 ZhongGuanCunBeiYiTiao Alley, Haidian District, Beijing 100090, PR China

Corresponding author: Yijun Liu. Author (e-mail: yijunliu@casipm.ac.cn).

# These authors contributed equally to this work

The present paper is supported by the National Natural Science Foundation of China under Grant Nos. 71573247 and 91746106, and by the Chinese Postdoctoral Science Foundation under Grant No. 2018M641177.

ABSTRACT The emergence of Internet water armies has strongly affected the information quality of online communication platforms, thus disrupting the order of the Internet. Accurate detection of Internet water armies is therefore of great significance. Based on the supernetwork theory, a new Internet water army detection model is proposed in this paper, in which a supernetwork with four layers is established, including social subnetwork, information subnetwork, psychological subnetwork, and negative keyword subnetwork. Then, personal information of users, dissemination process of information, transformation process of different psychologies, similarity between different keywords, and the connections between different subnetworks are considered in the model. Thus, 9 composite indexes are proposed, the majority of which are used for the first time in detecting Internet water armies. A dataset selected from the largest online communication platform in China, the Weibo website, is used to test the performance of the model. Four existing water army detection models introduced in previous studies are used to provide a comparison analysis. The results show that our proposed model has better performance in terms of accuracy and stability than the other four existing models, which thanks to the employment of the supernetwork theory. We believe that our propose model could be helpful for information researchers to further understand the complex nature of Internet water armies, as well as for the government to better manage the Internet.

INDEX TERMS Internet Water army detection; Supernetwork theory; Social media; Feature measurement; Machine learning classification

I. INTRODUCTION

With the development of Internet technologies, social media platforms are increasingly being favored by users owing to their unique advantages of timeliness, arbitrariness, and concealment. However, despite these advantages, a mysterious group has entered into public view, namely, "Internet Water Army". According to Chen *et al.* [1], the Internet water army refers to a specific group of users employed by interest organizations or individuals to post purposeful comments and articles on the Internet. Two principles of their behavior can be found: to avoid being exposed and to increase the influence of their opinions [2]. In addition, most of these specific users are college students and unemployed persons. To some extent, the increasing activity of water armies has strongly affected the development of some certain events, thus disrupting the social order and equity. For instance, according to a report by “Aju Business Daily”, a famous media in South Korea, up to hundreds of Internet water armies were employed by the National Intelligence Service to participate in the South Korea’s General Election in 2012, largely contributing to the success of Park Geun-hye. In 2017, the No. 12 report named “Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation” was published by University of Oxford Computing Program [3]. It studied the network campaign activities in 28 countries, and analyzed in detail the social media manipulation behavior, considering various kinds of Internet water armies within the Facebook and Twitter, thus providing strong evidence of the fact that Internet water armies have globally invaded several spheres. Therefore, accurate detection of Internet water armies is of great importance. However, Internet water armies are often accompanied with rigorous organizational features. When employed by interest organizations or individuals, they quickly disguise themselves as normal active users and post a large amount of information during a short period of time, in order to satisfy the interests of the employers. In particular, they are skilled in using rumors to manipulate public opinions, by applying false evidence to make their viewpoints more convincing and by interacting with other water armies to hide themselves well. Hence, it is difficult to detect Internet water armies from a large number of online users.

The supernetwork was first proposed by Nagurney and Dong in 2002 [4]. It is above and beyond existing networks. The supernetwork involves multi-layer, multi-level, multi-attribute characteristics, as well as congestion and coordination. Previous studies relevant to this point mainly focused on three aspects: Variational Inequality [5], Hypergraph [6], and System Science [7]. In addition, this theory has been widely applied in various areas, such as supply chain [8], transportation network system [9], e-commerce economic network [10], knowledge management, and communication trend [11], showing great effectiveness. The supernetwork can be used to reflect the connections and interactions between different subnetworks, thus providing a more comprehensive description of the whole system compared to other traditional approaches. With respect to the present paper, studies focusing on the application of the supernetwork theory in network public opinion area shed light on the understanding of online information analysis, especially for the information within the Internet communication platforms (see the works of Ma and Liu [12]; Tian *et al.* [13]; Liu *et al.* [14]; Wang *et al.* [15]). Based on these valuable investigations and achievements, we can better and more comprehensively identify the features of Internet water armies, and thus proposing a more effective detection model.

Therefore, in the present paper, we establish a water army detection supernetwork, in which there are four subnetworks, considering the user, information, psychology, and keywords, respectively. In addition, based on this supernetwork, 9 composite indexes are created, namely the cluster coefficient, density degree of posting time, the ratio between followed and following, information dissemination breadth and depth, psychological value, psychological transformation intensity, negative keyword proposition, and content similarity. Moreover, three commonly applied machine learning approaches and four existing detection models are used to provide a comparison analysis.

A. RESEARCH OBJECTIVES

Given the importance of Internet water army detection and the advantage of supernetwork theory, the present paper establish a new Internet water army detection model based on the supernetwork theory. In the proposed model, four subnetworks, namely, user subnetwork, information subnetwork, psychology subnetwork and keyword subnetwork are considered. From these four perspectives, we propose 9 composite indexes through comprehensively studying the features of Internet water armies, most of which are not applied in previous studies. To train and test our proposed model, we retrieve nearly 20000 comment information of three microblogs posted by Internet water armies on the Weibo website, a Chinese social media platform. In addition, four existing Internet water army detection models that have displayed effective performance in previous studies are also applied to provide a comparison analysis. We show that our model has better performance regarding the accuracy and stability in detecting Internet water armies than the other four existing models.

B. CONTRIBUTIONS

The main contributions of the present paper can be concluded as following: First, despite the importance of classifiers, effectively identifying the features of Internet water armies also plays a vital role in achieving accurate detection. However, most features recognized by existing studies are data-driven, which may decrease the universality of the proposed model [16]. In order to solve this problem, we establish a new model with a supernetwork structure to comprehensively describe the features of Internet water armies from a theoretical perspective. Therefore, this paper is helpful for filling the gap in the literature concerning the detection of Internet water armies. Second, although the data used in this paper only focuses on the Weibo Website, we also expect our findings can be helpful for the government and company managers for detecting Internet water armies on other social communication platforms, such as Twitter, Facebook and WeChat, in order to enable them to more effectively manage the Internet and avoid the viral marketing, respectively. Moreover, in addition to the application in Internet water army detection, our proposed model can be also used to deal with some other detection problems, taking the identification of online rumors and online opinion leaders as examples, which have attracted high attention of scholars in recent years.

This paper is organized as follows: In section 2, a number of previous relevant studies are analyzed and compared. In Section 3, the model is established based on the discoveries of Internet water armies' features. Section 4 includes an empirical study, in which the results of comparison experiment are presented and discussed. In Section 5, a conclusion has been drawn for the study.

II. RELATED WORKS

Looking from the historical perspective, detection of Internet water armies was originally studied in the context of mail service system. In this point of view, the Internet water armies refer to the word "spammers", who have attracted the attention of online users by sending spams with obvious unsolicited commercial features. In February 1999, the official announcement of the RFC2502 (Anti-Spam Recommendations for SMTO MTAs) (http: //www.faqs.org /rfcs/rfc2502.html) marked the beginning of anti-spam technology and spam identification technology research, which could be viewed as one of the most notable landmarks in this area. Following this announcement, how to accurately identify the spammers have drawn an increasing interest of scholars for several decades. In general, most existing methods focusing on the spammer detection in the mail service systems are content-feature-based, by distinguishing the differences between spam and normal emails. This trend may thank to the difficulty in identifying user information of spam posters. For instance, Sakkis *et al.* [17] used the stacked generalization to combine classifiers and created an automatically text categorization model that can improve the filtering effectiveness. Islam *et al.* [18] created a new integrated technique of e-mail classification and applied it in the spam filtering problem. In recent years, with the development of the Internet, some online social platforms, such as Twitter and Weibo, have been created. They not only provided a new form for individuals to post and exchange information, but also led to the formation of a new form of spammers, whose posting information is no longer limited to emails, but extending to various forms, such as microblogs and comments [19,20]. So far, the word "spammer" is not restricted in the mail service system but has much broader meanings. In this way the spammers could be also called "Internet water armies". Much evidence has confirmed the negative influences of Internet water armies on the order of Internet and even the society as a whole [1,2,21, 22]. For instance, as Zhang and Lu [22] stated that, due to the rumors and false information posted by Internet water armies, the normal network order and social harmony and stability have been seriously disturbed. Thus, effectively identifying the spammers or Internet water armies in social Web sites is of great significance. The work implemented by Heymann *et al*. [23] refers to the first paper regarding this point. They stated that existing methods for spammer detection in email and Web cannot be well applied in social Web sites, and accordingly proposing some tentative suggestions for future studies.

In recent years, the machine learning algorithms has been commonly applied in the detection of Internet water armies. Thus, the detection process can be generally divided into two steps: the exaction of features of Internet water armies and the selection of appropriate classifiers. With respect to the latter, Wu [24] used an enhanced BPNN (Back Propagation Neural Network) classifier with weighted learning strategies to filter Internet water armies. Results indicated that the employment of such mechanisms can improve the accuracy performance. In addition, Zhang *et al.* [25] and Behjat *et al.* [26] introduced the Particle Swarm Optimization (PSO) into the Internet water army identification problem. These evaluable achievements to a large extent increase the detection accuracy of Internet water armies. Sharaff *et al.* [27] provided an excellent comparative study focusing on the classification algorithms for spammer detection. Various algorithms, including but not limited to Decision Tree, Support Vector Machine and Bayesian models were considered in their paper. However, some scholars argued that mechanisms relying solely on machine learning methods are sometimes inadequate in detecting Internet water armies [28], and the addition of the consideration regarding the features of Internet water armies could achieve better performance [29]. Thus, how to effectively capture the features of Internet water armies has attracted much attention of researchers. Up to now, various features have been identified, covering the aspects of user, content, behavior, network, etc. Chakraborty *et al.* [30] and Wu *et al.* [31] provided a great survey presenting the state-of-the-art in the area of social spammer detection. In their works, existing relevant methodologies were well discussed and compared. In addition, to obtain the first-hand data, Chen *et al.* [1] pretended to be part of Internet water armies. This is a laudable attempt. Owing to their valuable achievements, some basic features of Internet water armies were identified, including response ratio, posting interval, and active time. Their pioneering findings shed light on the understanding of Internet water army behavior for future studies. Lee *et al.* [32] applied the "honeypots" to collect the Internet water army data from Twitter and found the features of followers/following relationship and user links by analyzing these data. Moh and Murmann [33] established a feature-based matrix to determine the trust degree of users. Las-Casas *et al.* [34] introduced a new detection methodology based on the source network, in which a supervised classification technique and network-level metrics are employed. Santos *et al.* [35] proposed a content-based spam tweets filtering approach. Ahmed and Abulaish [36] introduced an online network-based identification approach, considering 14 features of Internet water armies. Gillani *et al.* [37] introduced a novel economic metric to improve the effectiveness of spam detector. By observing the relationship between different Internet water armies, Jeong *et al.* [21] applied the Triad Significance Profile (TSP) and Social Status (SS) to detect Internet water armies. In addition, an ensemble technique was also proposed in their work. Moreover, in order to achieve a higher level of detection accuracy, some scholars also begun to discover the features of Internet water armies by comprehensively considering different aspects. Taking Wu *et al.* [38] as an example, who proposed a graph-structured approach to present the inherent connection between social spammers and spam messages.

Shen *et al.* [39] proposed a Multi-View Learning for Social Spammer Detection (MVSD) framework, in which the user and network information are taken into account together. Inuwa-Dutse *et al.* [28] used an optimized set of features independent of historical data, which offered more insights into the behavior of spammers in social Web sites.

To sum up, it can be noted that detection methodologies simultaneously considering content-based, behavior-based, and network-based features of Internet water armies are extensively studied. In addition, these methods have yielded significant results. However, there are still some limitations. For instance, the considered network-based features only focus on several simple structure parameters; the psychological aspect of Internet water armies is not taken into account, including psychological classifications and transformation process; various features are just statistically listed without any analysis of their internal relations; most existing models are data-driven and lack theoretical bases (For example, in the work of Miller *et al.* [40], up to 107 data-driven indexes were applied), thus decreasing the universality of the model [16]. Therefore, by applying the supernetwork theory, a new Internet water army detection model is established from the theoretical perspective. This model considers a four-layer supernetwork, including social, information, physiological, and negative keyword subnetworks. In addition, the connections between the nodes in the same subnetwork and different subnetworks are analyzed as well, in order to provide more effective and comprehensive theoretical supports for Internet water army detection.

III. SUPERNETWORK BASED WATER ARMY DETECTION (SWAD) MODEL

A. MODEL

Most classic methods within the social network area only focused on the problem of “who” and ignored some other parameters, such as time, place, event, and psychology. In addition, the connections between these parameters are frequently not considered. Inspired by the supernetwork theory, some scholars began to apply it into the social network field, especially for the Internet public opinion networks, to comprehensively analyze the interaction mechanism and inherent influences between different elements [12-15]. In particular, the created public opinion supernetwork contains four subnetworks: environmental subnetwork, social subnetwork, psychological subnetwork, and viewpoint subnetwork [12]. In the present paper, to better apply this theory to the water army detection problem, we reconstruct the structure of the network, establishing a supernetwork based water army detection (SWAD) model. For a certain public opinion case, this model can be used to provide a comprehensive description by considering the social, information, psychological, and viewpoint perspectives. In addition, the social subnetwork is set as the main subnetwork, and the other three subnetworks connect with the social subnetwork.

(1) **Social Subnetwork:** this subnetwork indicates the reply relationship between two users participating in a specific Internet public opinion case, in which each user is set as a node and each reply relation is set as an edge.

(2) **Information Subnetwork:** this subnetwork measures a summary of posted information within a specific Internet public opinion case, in which each information is set as a node and we not consider their interval connections.

(3) **Psychological Subnetwork:** this subnetwork represents the psychological types of users included in a public opinion event, in which each type is set as a node and the transformation process between them is set as edges. In addition, the psychological type is identified based on a psychological lexicon.

(4) **Negative Keyword Subnetwork:** this subnetwork measures the negative keywords included in the posts, in which each negative keyword is set as a node and an edge between each two nodes means that they are consisted in a similar post.

In addition, we also consider the connections between the nodes within different subnetworks. To clearly explain their internal mechanism, an example is shown in Fig.1. In Fig.1, there are five, three, three, and five nodes in the social, information, psychological, and negative keyword subnetworks. In other words, there are five users (a_1, a_2, a_3, a_4, a_5), three information (e_1, e_2, e_3), three psychological types (p_1, p_2, p_3), and five negative keywords (k_1, k_2, k_3, k_4, k_5) in a certain Internet public opinion case. Table 1 shows the relationship between them, in which user a_1 has the psychological type p_3, users a_2, a_3, a_4 have the psychological type p_1, and user a_5 as the psychological type p_2; posts of a_1, a_3, a_5 include information e_2, post of a_2 includes information e_1 and e_3, and post of a_4 includes information e_3; there no negative keywords in the posts of a_1 and a_5, negative keywords k_2, k_3, k_4 are included in the post of a_2, negative keywords k_3, k_4, k_5 are included in the post of a_3, and negative keyword k_1 is included in the post of a_4.

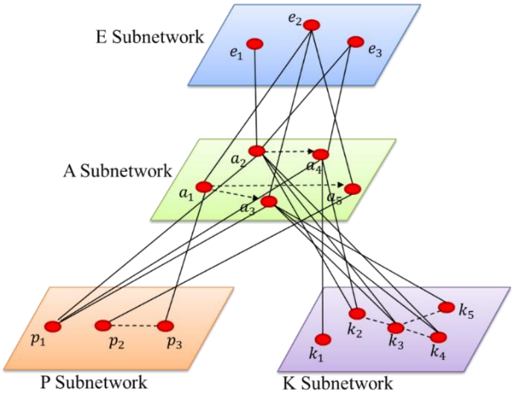

FIGURE 1. An example of the SWAD Model. Note that A Subnetwork is the social subnetwork, E Subnetwork is the information subnetwork, P subnetwork is the psychological subnetwork, and K Subnetwork is the negative keyword subnetwork; the dotted lines represent the intra-subnetwork connections, named edges; the solid lines represent the inter-subnetwork connections, named super edges; the social subnetwork is a directed graph, and the psychological subnetwork, information subnetwork and negative keyword subnetwork are undirected graphs.

TABLE 1. Relationship between two nodes within the supernetwork

Social Subnetwork	Information Subnetwork	Psychological Subnetwork	Negative Keyword Subnetwork
a_1	e_2	p_3	——
a_2	e_1, e_3	p_1	k_2, k_3, k_4
a_3	e_2	p_1	k_3, k_4, k_5
a_4	e_3	p_1	k_1
a_5	e_2	p_2, p_3	——

B. FEATURES

Based on the established detection model, we propose 9 composite features, shown in Table 2. In addition, the instance shown in Fig.1 is also applied to explain each composite index.

TABLE 2. Internet water army detection index system

Subnetworks	Features
Social Subnetwork	Cluster Coefficient C_i Density degree of posting time D_i The ratio between followed and following δ_i
Information Subnetwork	Information dissemination breadth $R(e_i)$ Information dissemination depth $O(e_i)$
Psychological Subnetwork	Psychological value p_i Psychological persistent intensity P_{ij}
Negative Keyword Subnetwork	Negative keyword proportion β_i Content Similarity SIM_{ij}

a. Features in the social subnetwork

(1) Cluster Coefficient C_i

$$C_i = \frac{2E_i}{h_i(h_i - 1)}. \tag{1}$$

According to the graph theory, the cluster coefficient is an indicator measuring the aggregation degree of the nodes within a network, which can be calculated through formula (1). For a certain node *i*, the number of edges connecting to this node h_i ($h_i \geqq 1$) is also the number of *i*'s neighbors. The maximum number *N* of edges between these h_i nodes is equal to h_i*(h_i-1)/2. In addition, E_i measures the actual number of edges between these h_i nodes. The cluster coefficient is the ratio between the maximum number of edges and the actual number of edges. C_i value closer to 1 means that the actual number of edges is closer to the number of maximum number of edges, indicating a stronger degree of aggregation, while C_i value closer to 0 represents a weaker degree.

In online social networks, there is a common phenomenon: *my friend's friend is my friend*. This triangular relationship reflects a strong degree of aggregation. Within the cluster of normal users, such a relation commonly exists, while for the cluster of Internet water armies, their respective friends are relatively independent. In other words, the cluster coefficient of normal users is larger than that of water armies. Therefore, the cluster coefficient could be cited as an effective indicator to distinguish between normal users and water armies.

(2) Density degree of posting time D_i

$$D_i = -\sum q_{ij} \log q_{ij}, \tag{2}$$

$$q_{ij} = \frac{n_{ij}}{N_i}. \tag{3}$$

The density degree of posting time D_i is proposed according to the entropy form, which can be calculated through formula (2). For a specific public opinion case, the duration is represented by *t*. We equally divide *t* into *m* aspects. The value of *m* is set according to the basis that maximizes the entire posts equally distributed in each aspect. The q_{ij} is a ratio parameter that can be calculated based on formula (3), where n_{ij} is the number of posts of user *i* in the *j*th division of time period (j = 1,2,…,m), and N_i is the entire number of posts of user *i*.

Based on the discovery of water armies, we notice that in order to achieve the purpose of maximum publicity effects, water armies always post a large number of microblogs during a short period of time. While for normal users, the posted information contains their own viewpoints, which will require some time to be thought out. In other words, users with larger density degree are more likely to be water armies.

(3) Ratio between followed and following $\boldsymbol{\delta_i}$

$$\delta_i = \frac{f_i}{g_i}. \tag{4}$$

In formula (4), f_i and g_i measure the number of followers and the number of following, respectively, which are two important personal features of every user in online communication platforms. Normal users prefer to follow their families, friends, and colleges, and these people will also be their followers. However, for water armies, as the task requirement and concealment demand, most have a high following and few followers, and the gap is significant. Therefore, the user with smaller value of follower /following ratio are more likely to be water armies.

b. Features in the information subnetwork

(1) Information dissemination breadth $\boldsymbol{R(e_i)}$

The dissemination breadth of specific information can be measured by the ratio between the number of super edges connecting this information and the number of whole super edges between the nodes within a social subnetwork and information subnetwork. The super edges are defined as the edge connection the nodes in different subnetworks (see the solid lines in Figure 1). This indicator can be calculated by formula (5).

$$R(e_i) = \frac{F(e_i)}{N}, \tag{5}$$

where $R(e_i)$ is the dissemination breadth of information e_i, $F(e_i)$ measures the number of super edges involving e_i, and N is the number of all super edges between the nodes within the social subnetwork and information subnetwork.

(2) Information dissemination depth $\boldsymbol{O(e_i)}$

The dissemination depth O (e_i) of information e_i can be obtained using formula (6), where $F(e_i)$ measures the number of super edges containing e_i, $A(e_i)$ is the number of super edges connecting to the user whose posts contain information e_i within the social subnetwork, N_a is the total number of users within the social subnetwork, and N is the number of all super edges between a social subnetwork and information subnetwork. In addition, if information e_i has not been replied, O (e_i)=1.

$$O(e_i) = \frac{F(e_i)/A(e_i)}{N/N_a}. \tag{6}$$

The information impacting degree has been confirmed to be an important indicator for detecting water armies [41]. However, most previous studies only use the node degree to measure this indicator, which is so simple and cannot provide an effective and comprehensive expression. For example, in reality, the "Big V" users (the influential users) are always wrongly classified into the Internet water army category because they have a large node degree. Therefore, in the present paper, in accordance with the features of information posted by water armies, only users with larger dissemination breadth and smaller dissemination depth are seen as water armies. To clearly explain these two indicators, an example is provided, as shown as Fig. 2. There are six super edges between the nodes within the social subnetwork and information subnetwork, which are e_1a_2, e_2a$_1$, e_2a_1, e_2a_5, e_3a_2, and e_3a_4. Based on formulas (5) and (6), the dissemination breadth and depth of information e_1, e_2, and e_3 can be calculated, as displayed in Table 3.

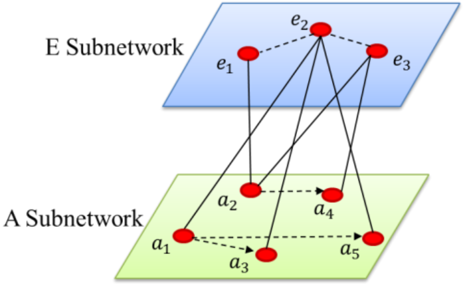

FIGURE 2. Social subnetwork and information subnetwork graphs

TABLE 3. Results of dissemination breadth and depth of information e_1, e_2, and e_3

Information	$F(e_i)$	$F(e_i)$	N	N_a	$R(e_i)$	$O(e_i)$
e_1	1	——	6	5	0.17	1
e_2	3	2	6	5	0.5	0.13
e_3	2	1	6	5	0.33	1.65

c. Features in the psychological subnetwork

(1) Psychological type $\boldsymbol{p_i}$

In the present paper, the psychological type is represented as a real value, which is similar to that proposed by Ma and Liu [12]. There are two aspects within a certain psychological type, namely the tendency and strength. In particular, three kinds of psychological types are considered: p_i < 0 represents the psychological type with negative psychological tendency and the psychological strength of $|p_i|$; p_i = 0 represents the psychological type with neural psychological tendency and the psychological strength of 0; p_i > 0 represents the psychological type with positive psychological tendency and the psychological strength of $|p_i|$. In addition, the interval of p_i is set as (+∞, -∞). It is a fact that, within the development of public opinion events, different individuals hold different psychological types and strengths. In addition, the psychological type of a specific post is labeled based on the number of keywords involved in a psychology lexicon. In specific, the text of each post is segmented by using the ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) [42], thus we can extract the nouns, adjectives and verbs of each post. Then, we match these words to different psychological tendencies and strengths by employing the HowNet sentiment lexicon in order to calculate the psychological value of each post.

(2) Psychological transformation intensity *P_{ij}*

Despite various psychological types, there is also a dynamical transformation process that should be considered. For example, assuming that the initial psychology value of user *i* is p_i, p_i may change to p_j when user *i* receives information from its neighbors. The intensity of this transformation connection between p_i and p_j is measured by the parameter p_{ij} that can be calculated through formula (7).

$$p_{ij} = \begin{cases} sign(p_i \times p_j)/\mid p_i - p_j \mid, \ p_i \neq p_j \\ \quad\quad 1 \quad\quad\quad\quad\quad\quad , p_i \neq p_j \end{cases}, \tag{7}$$

where *sign* (*x*) is a symbol function: when $x \geq 0$, $sign(x)=1$; when $x<0$, $\text{sign}(x)=-1$. For Internet water armies, owing to their common strong purposeful posting behavior, the psychological value is hard to be changed by other users. Therefore, to some extent, the users with higher p_{ij} are more likely to be Internet water armies. To provide a clear explanation of this composite index, the calculation results based on above example (see Fig.1) are presented. In Fig.1, there are three different kinds of psychological types: p_1 is negative with the strength of 1, p_2 is neural with the strength of 0, and p_3 is positive with the strength of 1. Based on the connections (edges) between them, their psychological persistent intensity can be obtained, which is formed as a matrix, as shown in Table 4.

TABLE 4. Psychological persistent intensity matrix based on an example

P_{ij}	p_1	p_2	p_3
p_1	1.00	1.00	-0.50
p_2	1.00	1.00	1.00
p_3	-0.50	1.00	1.00

d. Features in the negative keyword subnetwork

(1) Negative keyword proposition β_i

The proportion accounted by negative keywords can be calculated by formula (8):

$$\beta_i = \frac{n_{negi}}{N_{neg}}, \tag{8}$$

where β_i measures the negative keyword proposition of the *i*th post, n_{negi} is the number of negative keywords within the post, and N_{neg} is the total number of negative keywords within the collected posts. In the present paper, a negative lexicon is established based on the posts of more than 100 collected public opinion cases. In addition, the employed word segmentation software is the ICTCLAS [42]. In general, water armies are frequently accompanied with larger β_i.

(2) Content Similarity *SIM_{ij}*

In most cases, we find that the contents posted by water armies are relatively similar, containing various negative keywords. This phenomenon is mainly due to their requirement of the largest number of posts in the shortest amount of time and their organizational features. In accordance with these valuable discoveries, we employed a method introduced in previous studies [43] to calculate the similarity between two different posts. This method is based on the term frequency–inverse document frequency (TF-IDF) model. Assuming that there are *l* posts and *g* negative keywords, a connection matrix between posts and negative keywords can be obtained, as follows:

$$\begin{bmatrix} w_{11} & w_{11} & \cdots & w_{1g} \\ w_{21} & w_{22} & \cdots & w_{2g} \\ \cdots & \cdots & \cdots & \cdots \\ w_{l1} & w_{l2} & \cdots & w_{lg} \end{bmatrix}, \tag{9}$$

where w_{ij} measures the weight of the *j*th negative keyword for the *i*th post ($i=1,2,\ldots,l;j=1,2,\ldots,g$). In addition, w_{ij} can be obtained using formulas (10)–(12).

$$w_{ij} = TF_{ij} \times IDF_j, \tag{10}$$

$$TF_{ij} = \frac{u_{ij}}{\sum_{j=1}^{g_i} u_{ij}}, \tag{11}$$

$$IDF_j = \log\frac{O}{o_j}, \tag{12}$$

where u_{ij} measures the occurrences of negative keyword *j* in post *i*, g_i is the number of negative keywords within post *i*, $\sum_{j=1}^{g_i} u_{ij}$ represents the occurrences of all negative keywords in post *i*, *O* is the number of posts with negative keywords, and o_j measures the occurrences of negative keyword *j* in all posts. In addition, in the present paper, the similarity between two posts are measured by the cosine form, as shown by formula (13).

$$SIM_{i1,i2} = \cos\theta_{i1,i2} = \frac{\sum_{j=1}^{g} w_{i1j} \times w_{i2j}}{\sqrt{(\sum_{j=1}^{g} w_{i1j}^2) \times (\sum_{j=1}^{m} w_{i2j}^2)}}. \tag{13}$$

In order to show an explanation, the similarity matrix of the above instance (see the Fig.1) is calculated, as shown by formula (14).

$$SIM_{ij} = \cos\theta = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0.5 & 0.28 & 0 \\ 0 & 0.5 & 1 & 0 & 0 \\ 0 & 0.28 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \cdot \quad (14)$$

Then, by applying the formula (15), the similarity of a certain post can be calculated.

$$SS_i = \frac{\sum_{o=1, o\neq m}^{O} SIM_{ij}}{O}, \quad (15)$$

where SS_i measures the similarity degree of post *i*.

C. FRAMEWORK

The framework of our proposed model is shown in Fig.3, in which the input is the post data selected from the Weibo website, and the output is the classification result. In particular, there are three steps:

Step 1-Data collection and preprocessing: data of Internet water armies is bought from the Taobao online shopping platform, and the normal data are obtained by using the clawer software named "Octopus". In addition, the data are relevant to several specific cases. Each post involves two aspects: the user information and the content information, in which the former contains personal description, posting time, the number of followers, and the number of following, etc. For the latter, the psychological value and negative keywords can be obtained based on the psychological lexicon and negative keyword lexicon, respectively. Then, we categorized these messages based on their posing users.

Step 2-Model establishment: based on the supernetwork theory, our proposed detection model with four subnetworks is established. Then, the features can be calculated. It should be noted that for the users who posted different messages and had different psychologies or information, the cumulative calculation method is implemented for the calculation process.

Step 3-Classfication: based on the calculated indexes, we employed three commonly applied machine learning methods to train the classifiers, namely the Neural network, Naive Bayes, and Support Vector Machine. These three approaches were chosen mainly because of their great performance in previous studies.

The overview of the proposed model is briefly outlined in Algorithm 1.

```
Algorithm 1 SWAD model
Input: Selected messages
Output: List of messages posted by non-water armies and water armies
for each selected message do
  Extract the elements of each message, including user, information, psychological type, and negative keywords.
end for

Identify the relationship between the extracted elements.

Construct the four-layer supernetwork based on the execrated elements and identified relationships by considering social, information, psychological and negative viewpoint aspects.

for each message in training set do
    Calculate the values of nine proposed features of each message through the constructed supernetwork and formulas of (1)-(15).
end for

Design the structure of each machining learning algorithm.

Use the values of nine features of training samples as the inputs to train the machining learning algorithms.

while Algorithm is running do
  if a message contained in the test set is received then
    Characterize the message by calculating the values of nine features.
    Use trained machine learning algorithms to predict the label of message.
  end if
end while
Return the list of messages posted by non-water armies and water armies.
```

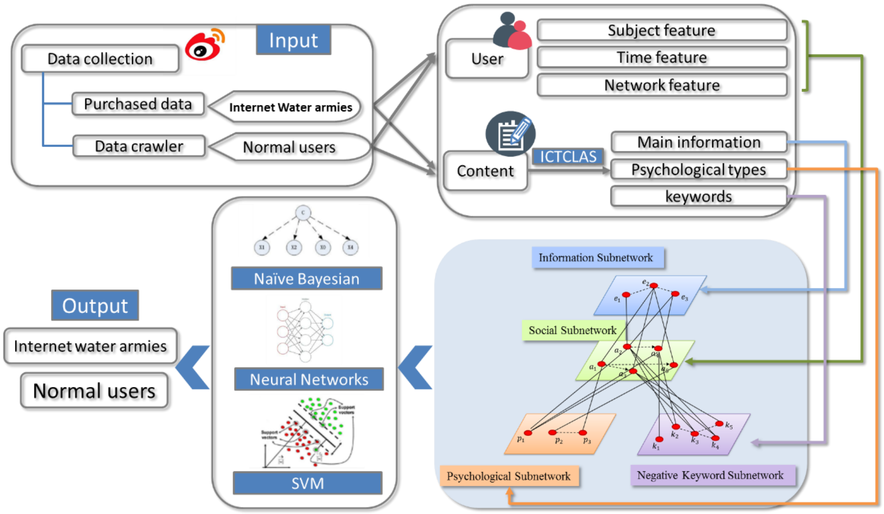

FIGURE 3. Framework of SWAD model

IV. EMPIRICAL STUDIES

A. DATA

In general, it is difficult to correctly identify the water armies from a large number of Internet users with the accuracy rate of 100%, unless the water army promotion companies provide real data. However, achieving accurate data is extremely important for the classifier training process within the machine learning algorithms. To deal with this problem, the artificial judgement mechanism was frequently applied in previous studies [21,40]. The main reason for using this method is that it is very difficult to establish communications with water army promotion companies. However, to a large extent, it will inevitably decrease the accuracy of the classification between water armies and normal users. Therefore, in order to provide a more convincing dataset, we tried to purchase Internet water army data from water army promotion companies. After long searches and difficult negotiations, we found a promotion company in the Taobao website (one of the largest online shopping platforms in China), and finally obtained the comment information of three microblogs posted by water armies on the Weibo website. Then, by using the web crawler software (Octopus), the comments to these three microblogs posted by normal users were also obtained, thus yielding the dataset. Moreover, to adopt the machine learning algorithms, the dataset is divided into two aspects: training set and test set. In this paper, we discussed two cases by respectively setting the ratio between these two sets 70/30 and 80/20. The data information is presented in Table 6.

TABLE 6. Information of the dataset

	Value	
Number of water armies	1253	
Number of non-water armies	9776	
Number of messages	19607	
Case 1		
	Training	**Test**
Number of messages posted by water armies	2968	1272
Number of messages posted by non-water armies	10757	4610
Total	13725	5882
Case 2		
	Training	**Test**
Number of messages posted by water armies	3392	848
Number of messages posted by non-water armies	12294	3073
Total	15686	3921

B. RESULTS

In the present paper, three commonly applied machine learning methods are used to train the classifiers, namely the Neural network (NN), Naive Bayes (NB), and Support Vector Machine (SVM). The reason for employing these four classical methods is that they have shown great performance in many recent papers focusing on the online water army detection [22, 36]. In addition, as the main contributions of this paper were to introduce the supernetwork theory into water army detection problem and evaluate its effects on detection performance, using classic classification algorithms could provide a more fairly assessment, to a large extent. Previous studies have comprehensively explained the mechanisms of these three approaches in detail (see [44-47]), so these contents will not be described here. In this paper, the kernel function of SVM is set as the RBF function and a 10-fold cross-validation method is applied. With respect to the NN, a three-layer Back Propagation Neural Network (BPNN) was employed, and the activation functions of hidden layer and output layer within the neural network method were both set as the sigmoid function. By following Ruan and Tan [46], the training function and the performance function were respectively set as the Levenberg-Marquardt and the Mean Squared Error (MSE). Additionally, the number of hidden layers and output layers was both set as 1, and the number of neurons in each hidden layer and output layer was respectively set as 4 and 1. In specific, output 1 indicates the message posted by non-water armies, while output o is for those posed by water armies.

Moreover, to provide a comparison analysis, we also selected four existing Internet water army detection models as the baseline, i.e., those proposed by Tian *et al.* [48], Zheng *et al.* [49], Dai and Wang [50], and Zhang and Lu [22]. These four models were chosen because the datasets employed by them are all from the Weibo website, and thus the negative effects caused by language differences can be avoided. To accurately reconstruct the features proposed by these papers, we strictly followed the steps shown in their papers and employed one of the most popular word segmentation software, ICTCLAS, and its embedded lexicon. It should be noted that, due to Zhang and Lu [22] did not describe the clustering method in their work, the K-means algorithm is applied to calculate the similarity of posts in this paper. Moreover, according to previous practice [21,40,51], the performance of the Internet water army detection model was evaluated based on four metrics, namely *Accuracy*, *Precision*, *Recall*, and *F1-score*, which are obtained through a confusion matrix. A general form of the confusion matrix is shown in Table 7.

TABLE 7. General form of Confusion Matrix

	Predicted	
True	Water Army	Non-Water Army
Water Army	True Positive (TP)	False Negative (FN)
Non-Water Army	False Positive (FP)	True Negative (TN)

In Table 7, the True Positive is the number of water armies that are correctly detected, the False Negative is the number of water armies that are incorrectly detected, the False Positive is the number of non-water armies that are incorrectly detected, and the True Negative is the number of non-water armies that are correctly detected. Based on these parameters, four assessment metrics, namely *accuracy*, *precision*, *recall*, and *F*1-*score*, can be calculated, as shown by formula (16) to (19).

$$Accuracy = \frac{|TP+TN|}{|TP+TN+FP+FN|}, \tag{16}$$

$$Precision = \frac{|TP|}{|TP+FP|}, \tag{17}$$

$$Recall = \frac{|TP|}{|TP+FN|}, \tag{18}$$

$$F1\text{-score} = \frac{2\times Precision\times Recall}{Precision+Recall}. \tag{19}$$

With each method for each model run for 100 times, and the average values were taken as the final results. The confusion matrixes of each model for case 1 and case 2 are shown in Table 8 and Table 9. The comparison results of five detection models for each case are displayed in Table 10 and Table 11. In addition, the bar graphs of case 1 and case 2 are respectively displayed in Figure 4 and Figure 5.

By considering four selected assessment metrics, our proposed model achieved better performance in Internet water army detection compared to other four previous models under the ratio between training and test sets of both 70/30 and 80/20. In specific, for case 1, the Accuracy of the proposed model by using NN and SVM is 88.98% and 88.13, respectively. With respect to the NB, the model proposed by Zheng *et al.* [49] had the highest Accuracy, which is 89.49%. The Precision of the proposed model by using NN is 74.11% in case 1, which is respectively 3%, 1.33%, 3.07%, and 1.68% higher than the models proposed by Tian *et al.* [48], Zheng *et al.* [49], Dai and Wang [50], and Zhang and Lu [22]. In addition, the proposed model performed best regarding the Recall by using all three machining learning algorithms in case 1, which is 75.86%, 75.39% and 47.53%, respectively. In terms of the F1-score, our proposed model performed better than the comparison models through the use of NN and SVM, which is 74.75% and 73.09%, respectively. With respect to NB, the model proposed by Zheng *et al.* [49] achieved the highest F1-socre, which is 75.24%. For the case with the ratio between training and test sets of 80/20, the results of our proposed model were relatively better than those of other four comparison models regarding the assessment metrics of Accuracy, Recall and F1-score as well. In terms of the Precision, the proposed model only achieved the highest value through the application of NN, which is 77.12%. The models proposed by Zhang and Lu [22] performed better than other four models using NB and SVM in the F1-score, which is 78.57% and 75.78%, respectively.

In addition, we found that the performance of these four existing models in the present paper is worse than that displayed in their respective works, which is mainly due to differences in the dataset. Moreover, consistent with Liang *et al.* [52], we also found that different learning algorithms will mostly achieve relatively similar results.

TABLE 8. Confusion matrix results of each model obtained through NB, NN, and SVM under the ratio between training and test sets of 70/30

		SWAD	Tian *et al.* [48]	Zheng *et al.* [49]	Dai and Wang [50]	Zhang and Lu [22]
NB	**TP**	965	864	939	881	875
	FN	307	408	333	391	397
	FP	343	362	285	368	303
	TN	4267	4248	4325	4242	4307
NN	**TP**	959	921	934	957	922
	FN	313	351	338	315	350
	FP	335	359	343	374	343
	TN	4275	4251	4267	4236	4267
SVM	**TP**	948	882	886	879	856
	FN	324	390	386	393	416
	FP	374	381	359	373	301
	TN	4236	4229	4251	4237	4309

TABLE 9. Confusion matrix results of each model obtained through NB, NN, and SVM under the ratio between training and test sets of 80/20

		SWAD	Tian *et al.* [48]	Zheng *et al.* [49]	Dai and Wang [50]	Zhang and Lu [22]
NB	**TP**	622	426	613	519	605
	FN	226	422	235	329	243
	FP	191	205	183	329	165
	TN	2882	2868	2890	2744	2908
NN	**TP**	637	499	618	525	617
	FN	211	349	230	323	231
	FP	189	222	201	348	201
	TN	2884	2851	2872	2725	2872
SVM	**TP**	633	465	647	507	585
	FN	215	383	201	341	263
	FP	208	229	231	345	187
	TN	2865	2844	2842	2728	2886

TABLE 10. Comparison results of five Internet water army detection models under the ratio between training and test sets of 70/30

		SWAD	Tian *et al.* [48]	Zheng *et al.* [49]	Dai and Wang [50]	Zhang and Lu [22]
Accuracy	NB	88.95%	86.91%	**89.49%**	87.10%	88.10%
	NN	**88.98%**	87.93%	88.42%	88.29%	88.22%
	SVM	**88.13%**	86.89%	87.33%	86.98%	87.81%
Precision	NB	73.78%	70.47%	**76.72%**	70.54%	74.28%
	NN	**74.11%**	71.95%	73.14%	71.90%	72.89%
	SVM	71.71%	69.83%	71.16%	70.21%	**73.98%**
Recall	NB	**75.86%**	67.92%	73.82%	69.26%	68.79%
	NN	**75.39%**	72.41%	73.43%	75.24%	72.48%
	SVM	**74.53%**	69.34%	69.65%	69.10%	67.30%
F1-score	NB	74.81%	69.18%	**75.24%**	69.89%	71.43%
	NN	**74.75%**	72.18%	73.28%	73.53%	72.68%
	SVM	**73.09%**	69.59%	70.40%	69.65%	70.48%

TABLE 11. Comparison results of five Internet water army detection models under the ratio between training and test sets of 80/20

		SWAD	Tian *et al.* [48]	Zheng *et al.* [49]	Dai and Wang [50]	Zhang and Lu [22]
Accuracy	NB	**89.36%**	84.01%	89.34%	83.22%	89.59%
	NN	**89.80%**	85.44%	89.01%	82.89%	88.98%
	SVM	**89.21%**	84.39%	88.98%	82.50%	88.52%
Precision	NB	76.51%	67.51%	77.01%	61.20%	**78.57%**
	NN	**77.12%**	69.21%	75.46%	60.14%	75.43%
	SVM	75.27%	67.00%	73.69%	59.51%	**75.78%**
Recall	NB	**73.35%**	50.24%	72.29%	61.20%	71.34%
	NN	**75.12%**	58.84%	72.88%	61.91%	72.76%
	SVM	74.65%	54.83%	**76.30%**	59.79%	68.99%
F1-score	NB	**74.89%**	57.61%	74.57%	61.20%	74.78%
	NN	**76.11%**	63.61%	74.15%	61.01%	74.07%
	SVM	74.96%	60.31%	**74.97%**	59.65%	72.22%

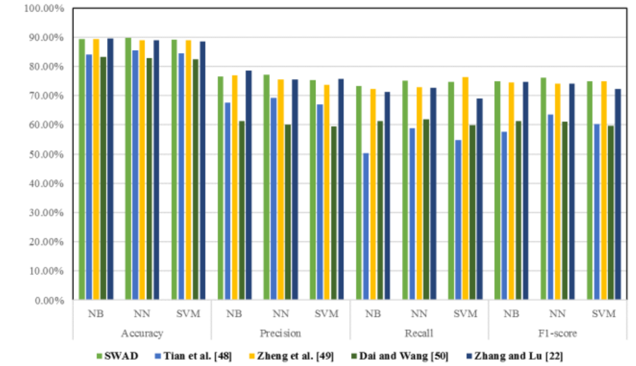

FIGURE 4. Bar graph of comparison results of five Internet water army detection models under the ratio between training and test sets of 70/30

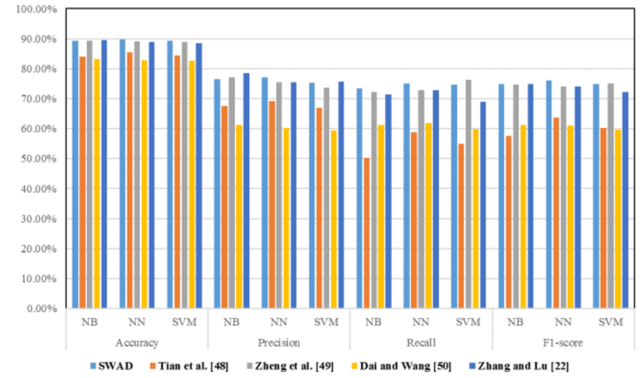

FIGURE 5. Bar graph of comparison results of five Internet water army detection models under the ratio between training and test sets of 80/20

Furthermore, we also carried out the sensitivity analysis to measure the contribution of each group of features on the overall performance of the proposed model. There are five models in total. Details are shown in Table 12.

TABLE 12. Details of five models in the sensitivity analysis

	Excluded group of features
Model 1	None
Model 2	Features in the social subnetwork
Model 3	Features in the information subnetwork
Model 4	Features in the psychological subnetwork
Model 5	Features in the negative viewpoint subnetwork

Three machine learning methods with similar designed structures and parameter settings demonstrated in above contents were used in the sensitivity analysis, including NN, NB, and SVM. The ratio between training and test sets was set as 80/20. The evaluation metrics of Accuracy, Precision, Recall and F1-score were applied. With each algorithm for each model run for 100 times, and the average values were taken as the final results. The results with three algorithms are respectively shown in Figure 6, Figure 7 and Figure 8.

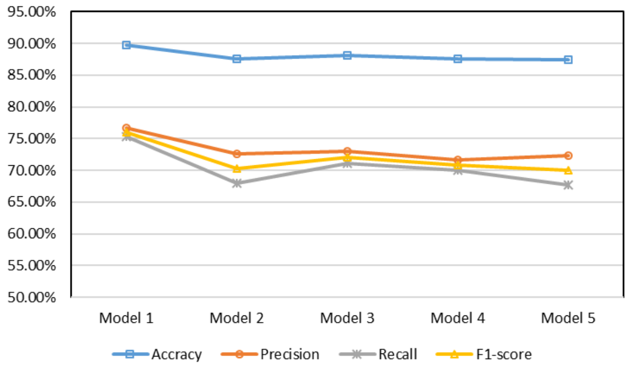

FIGURE 6. Comparison results of five models with NN

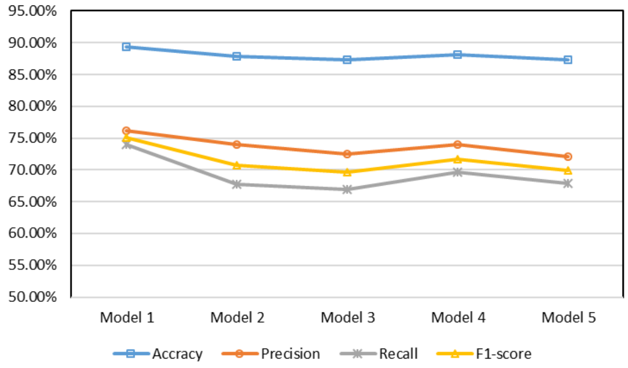

FIGURE 7. Comparison results of five models with NB

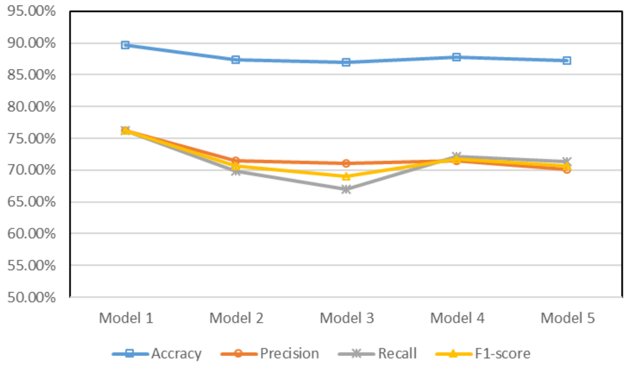

FIGURE 8. Comparison results of five models with SVM

It can be noticed that, the Model 1 that contained all features performed best in Accuracy, Precision, Recall and F1-score by using three machine learning algorithms. In addition, by comparing the other four models, we found that the performance of the Model 5, which excluded the features contained in the negative viewpoint subnetwork, seems to be relatively worse. This finding indicates that the posted contents play a more important role in the detection of Internet water armies compared to the other groups of features, to a large extent. In this view, Chen *et al.* [1] proposed that, in most cases, Internet water armies are paid to post messages on online communication platforms. Unlike normal users, they always post comments by simply following the template provided by their employers without carefully checking or changing the contents. Consequently, the messages posted by Internet water armies tends to be similar [49]. Moreover, Internet water armies are frequently required to post fake or negative information [1], thus leading to a higher proportion of negative keywords than normal users.

V. CONCLUSIONS

In the present paper, a water army detection supernetwork model is proposed based the supernetwork theory, in order to detect water armies with the purpose of affecting politics. According to the unique features and the supernetwork theory, we established a supernetwork with four layers, namely social subnetwork, information subnetwork, psychological subnetwork, and negative-keyword subnetwork. Hence, a system consisting 9 indexes was created, most of which are used for the first time in water army detection problem, such as information dissemination breadth and depth, psychological persistent intensity, and *content similarity*. In addition, by using the supernetwork theory, our proposed model considers the information covered by the nodes and edges within similar subnetwork and different subnetworks, rather than the features that can be easily discovered statically. Moreover, to verify the effectiveness of our proposed model, we selected four commonly applied assessment metrics and four recent existing models. The results show that our proposed model exhibits better performance in identifying water armies with political purpose, and also provides high stability. In addition, the sensitivity analysis was also carried out in the present paper in order to measure the contribution of each group of features on the overall performance of the proposed model. We found that the features contained in the negative viewpoint subnetwork play a more important role compared to the other features. This finding to a large extent is consistent with some relevant studies [1, 49].

However, along with the improvement of detection technologies, the evolution of water armies never stops. They have become more organizational and concealed and are therefore more difficult to be identified. Nevertheless, their strong purposeful interest is fixed, and the structure of online communication platforms is unchanged, which can be effectively described and measured by classic theories.

Thus, the models combined with theory-driven and data-driven principles should be the main direction of future studies focusing on this area. For example, despite the contribution of the supernetwork theory to describe network structure features of water armies, applications of several psychological classical theories may produce some surprising results.

ACKNOWLEDGMENT

Thankfulness shall be expressed to the reviewers for their useful discussions and comments on this manuscript.

REFERENCES

[1] C. Chen, K. Wu, V. Srinivasan, and X. Zhang," Battling the Internet Water Army: Detection of Hidden Paid Posters," in *Proc. Int. Conf. Adv. Soc. Netw. Anal. Min,* pp.116-120. ACM, 2013.

[2] K. Zeng, X. Wang, Q. Zhang, X. Zhang, and F. Y. Wang, "Behavior modeling of internet water army in online forums," *IFAC Proc. Vol.,* vol.47, no.3, pp. 9858-9863, Aug. 2014.

[3] M. Bradshaw, and P. N. Howard, "Troops, Trolls and Troublemakers: A Global Inventory of Organized Social Media Manipulation," *Working paper. The University of Oxford,* 2017.

[4] A. Nagurney, and J. Dong,"Supernetworks: Decision-Making for the Information Age," *Cheotenham: Edward Elgar Publishers,* 2002.

[5] A. Nagurney, Z. Liu, M. G. Cojocaru, and P. Daniele, "Dynamic electric power supply chains and transportation networks: an evolutionary variational inequality formulation," *Transp. Res. Pt. e-Logist. Transp. Rev.,* vol.43, no.5, pp. 624-646, Sept. 2007.

[6] N. Wang, W. Xu, Z. Xu, and W. Shao, "A survey on supernetwork research: Theory and applications," *35th Chin. Control. Conf. (CCC)* (pp.1202-1206). IEEE, 2016.

[7] F. Liao, T. Arentze, and H. Timmermans, "Supernetwork approach for multimodal and multiactivity travel planning," *Transp. Res. Rec. J. Transp. Res. Board,* vol. 2175, no.1, pp. 38-46, 2010.

[8] T. Wakolbinger, and A. Nagurney, "Dynamic supernetworks for the integration of social networks and supply chains with electronic commerce: modeling and analysis of buyer-seller relationships with computations," *NETNOMICS: Economic Research and Electronic Networking*, vol.6, no.2, pp. 153-185, 2004.

[9] A. Nagurney, " On the relationship between supply chain and transportation network equilibria: a supernetwork equivalence with computations," *Transp. Res. Pt. e-Logist. Transp. Rev.,* vol.42, no.4, pp. 293-316, Jul. 2006.

[10] A. Nagurney, T. Wakolbinger, and L. Zhao, "The evolution and emergence of integrated social and financial networks with electronic transactions: a dynamic supernetwork theory for the modeling, analysis, and computation of financial flows and relationship levels," *Comput. Econ.,* vol. 27, no.2-3, pp. 353-393, 2006.

[11] X. I. Yun-Jiang, and Y. Z. Dang, "Method to analyze robustness of knowledge network based on weighted supernetwork model and its application," *Syst. Eng. Theory Pract. Online*, vol.27, no.4, pp.134-140, 2007.

[12] N. Ma, and Y. Liu, "Superedgerank algorithm and its application in identifying opinion leader of online public opinion supernetwork," *Expert Syst. Appl.,* vol.41, no.4, pp.1357-1368, Mar. 2014.

[13] R. Y. Tian, and Y. J. Liu, "Isolation, insertion, and reconstruction: three strategies to intervene in rumor spread based on supernetwork model," *Decis. Support Syst.,* vol.67, no.2, pp.121-130, Nov. 2014.

[14] Y. Liu, Q. Li, X. Tang, N. Ma, and R. Tian, "Superedge prediction: what opinions will be mined based on an opinion supernetwork model?" *Decis. Support Syst.,* vol.64, no.3, pp.118-129, Aug. 2014.

[15] G. Wang, Y. Liu, J. Li, X. Tang, and H. Wang, "Superedge coupling algorithm and its application in coupling mechanism analysis of online public opinion supernetwork," *Expert Syst. Appl.,* vol.42, no.5, pp.2808-2823, Apr. 2015.

[16] S. Kwon, M.Cha, and K. Jung, " Rumor Detection over Varying Time Windows," *Plos One*, vol.12, no.1, e0168344. 2017.

[17] G. Sakkis, I. Androutsopoulos, G. Paliouras, V. Karkaletsis, C. D. Spyropoulos, and P. Stamatopoulos, "Stacking classifiers for anti-spam filtering of e-mail," *Int. J. Prod. Res.,* vol.52, no.19, pp.5857-5879, 2001.

[18] M. R. Islam, W. Zhou, M. Guo, and Y. Xiang, "An innovative analyser for multi-classifier e-mail classification based on grey list analysis," *J. Netw.*

Comput. Appl., vol.32, no.2, pp.357-366. Mar. 2009.

[19] Y. Lv, J. Liu, H. Chen, J. Mi, M. Liu, and Q. Zheng, "Opinioned post detection in Sina Weibo,"*IEEE Access*, vol.5, pp.7263-7271, Mar. 2017.

[20] X. Liu, and C. Liu, "Information diffusion and opinion leader mathematical modeling based on microblog," *IEEE Access*, vol.6, pp.34736-34745, Jun. 2018.

[21] S. Jeong, G. Noh, H. Oh, and C. K. Kim, "Follow spam detection based on cascaded social information, " *Inf. Sci.*, vol.369, pp.481-499, Nov. 2016.

[22] W. Zhang, and J. Lu, "An Online Water Army Detection Method Based on Network Hot Events," in *Proc. Int. Conf. Meas. Technol. Mechatron. Autom.*, (pp.191-193). IEEE Computer Society, 2018.

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

[24] C. H. Wu, "Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks," *Expert Syst. Appl.*, vol.36, no.3, pp.4321-4330, Apr. 2009.

[25] Y. Zhang, S. Wang, P. Phillips, and G. Ji, "Binary PSO with mutation operator for feature selection using decision tree applied to spam detection," *Knowledge-Based Syst.*, vol.64, no.1, pp.22-31, Jul. 2014.

[26] A. R. Behjat, A. Mustapha, H. Nezamabadi-Pour, M. N. Sulaiman, and N. Mustapha, "A PSO-Based Feature Subset Selection for Application of Spam/Non-spam Detection," in *Proc. Int. Conf. Comput. Commun. Eng.*, Vol.378, pp.675-679, IEEE. 2012.

[27] A. Sharaff, N. K. Nagwani, and A. Dhadse, "Comparative study of classification algorithms for spam email detection,". from Book *Efficient Identification of Users and User Sessions from Web Log Repository Using Dimensionality Reduction Techniques and Combined Methodologies* (pp.237-244) 2016.

[28] I. Inuwa-Dutse, M. Liptrott, and I. Korkontzelos, "Detection of spam-posting accounts on Twitter," *Neurocomputing*, vol. 315, pp.496-511, Nov. 2018.

[29] N. Chavoshi, H. Hamooni, and A. Mueen,"Temporal patterns in bot activities," in *Proc. 26th Int. Conf. World Wide Web Companion*, pp. 1601-1606, 2017.

[30] M. Chakraborty, S. Pal, R. Pramanik, and C. R. Chowdary, "Recent developments in social spam detection and combating techniques: a survey," *Inf. Process. Manage.*, vol.52, no.6, pp. 1053-1073, Nov. 2016.

[31] T. Wu, S. Wen, Y. Xiang, and W. Zhou, "Twitter spam detection: survey of new approaches and comparative study," *Comput. Secur.*, vol.76, pp.265-284, 2018.

[32] K. Lee, B. D. Eoff, and J. Caverlee, "Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter," in *Proc. Int. AAAI Conf. Weblogs and Soc. Media (ICWSM)*, 2011, pp.185-192, 2011.

[33] T. S. Moh, and A. J. Murmann, "Can You Judge a Man by His Friends?-Enhancing Spammer Detection on the Twitter Microblogging Platform Using Friends and Followers,". *Inf. Syst. Technol. Manage*, Springer Berlin Heidelberg, 2010.

[34] P. H. B. Las-Casas, D. Guedes, J. M. Almeida, A. Ziviani, and H. T. Marques-Neto, "Spades: detecting spammers at the source network," *Comput. Netw.*, vol.57, no.2, pp.526-539, Feb. 2013.

[35] I. Santos, I. Mi ñambres-Marcos, C. Laorden, P. Gal án-Garc ía, A. Santamar ía-Ibirika, and P. G. Bringas, "Twitter Content-Based Spam Filtering," in *Proc. Int. Joint Conf. SOCO’13-CISIS’13-ICEUTE’13*, 2014.

[36] F. Ahmed, and M. Abulaish, "A generic statistical approach for spam detection in online social networks," *Comput. Commun.*, vol.36, no.10-11, pp.1120-1129, Jun. 2013.

[37] F. Gillani, E. Al-Shaer, and B. Assadhan, "Economic metric to improve spam detectors," *J. Netw. Comput. Appl.*, vol.65, no.C, pp.131-143, Apr. 2016.

[38] F. Wu, J. Shu, Y. Huang, and Z. Yuan, "Co-detecting social spammers and spam messages in microblogging via exploiting social contexts," *Neurocomputing*, vol.201, pp.51-65, Aug. 2016.

[39] H. Shen, F. Ma, X. Zhang, L. Zong, X. Liu, and W. Liang, "Discovering social spammers from multiple views," *Neurocomputing*, vol.225, pp.49-57, Feb. 2016.

[40] Z. Miller, B. Dickinson, W. Deitrick, *et al.* "Twitter spammer detection using data stream clustering," *Inf. Sci.*, vol.260, no.1, pp.64-73, Mar. 2014.

[41] C. Fan, C. Liu, C. Zhang, and H. Wu, "Analysis of the Time Characteristics of Network Water Army Based on BBS Information,". in *Proc. Int. Conf. Intell. on Intelligent Sci. and Big Data Eng*, pp.20-28. Springer, Cham, 2015

[42] H. P. Zhang, H. K Yu, D. Y. Xiong, and Q. Liu, "HHMM-based Chinese lexical analyzer ICTCLAS," *Sighan Workshop on Chinese Language Processing* (Vol.17, pp.758-759). Association for Computational Linguistics, 2003.

[43] G. Salton, A. Wong, and C. S. Yang, "A vector space model for automatic indexing," *Commun. ACM*, vol.18, no.11, pp.613-620, Nov. 1974.

[44] Z. Yang, X. Nie, W. Xu, and J. Guo, "An Approach to Spam Detection by Naive Bayes Ensemble Based on Decision Induction," in *Proc. Int. Conf. Intell. Sci. Des. and Appl.*, Vol.2, pp.861-866, IEEE Computer Society, 2006.

[45] G. Stafford, and L. L. Yu, " An Evaluation of the Effect of Spam on Twitter Trending Topics," in *Proc. Int.*

Conf. Soc. Comput, Vol.10, pp.373-378, IEEE, 2013.

[46] G. Ruan, and Y. Tan, "A three-layer back-propagation neural network for spam detection using artificial immune concentration," *Soft Comput.*, vol.14, no.2, pp.139-150, Jan. 2010.

[47] X Qin, N. X. Xia, and S. U. Yi-Dan, "SVM-based social spam detection model," *Appl. Res. Comput.*, vol.2010, pp.765-767, 2010.

[48] X. Y. Tian, G. Yu, and P. Y. Li, " Spammer detection on Sina Micro-Blog," in *Proc. Int. Conf. Manag. Sci. Emg.* (pp.82-87). IEEE, 2014.

[49] X. Zheng, X. Zhang, Y. Yu, T. Kechadi, and C. Rong, "ELM-based spammer detection in social networks," *J. Supercomput.*, vol.72, no.8, pp.2991-3005, Aug. 2016.

[50] J. Dai, and X. Wang, "An Identification Model of Water Army Based on Data Analysis," *J. Phys. Conf. Series*, vol.1069, 012085, 2018.

[51] Y. Ma, N. Yan, R. Yan, and Y. Xue, "Detecting spam on Sina Weibo," *Int. Workshop Cloud Comput. Inf. Secur.*, pp. 404-407. 2013.

[52] G. Liang, W. He, C. Xu, L. Chen, and J. Zeng, "Rumor identification in microblogging systems based on users' behavior," *IEEE Trans. Comput. Soc. Syst.*, vol.2, no.3, pp.99-108, Sept. 2015.

YING LIAN received the B.S. degree the M.S. degree in automation from Beijing University of Chemical Technology, China, in 2012 and 2015, and the Ph.D. degree in Management Science and Engineering from Institutes of Science and Development, CAS, China, in 2018. She is currently a postdoctoral fellow in Beijing Jiaotong University, China. Her main research interests include opinion dynamics, social stability early-warning system, and sustainable development strategy.

XUEFAN DONG received the B.S. degree in information system and management from Beijing University of Chemical Technology, China, in 2012, and the M.S. degree in finance and investment from the University of Nottingham, Ningbo, China, in 2014, and the Ph.D. degree in Management Science and Engineering from Institutes of Science and Development, CAS, China, in 2018. He is currently a postdoctoral fellow in Beijing Jiaotong University, China. His main research interests include social network analysis, complex network, opinion dynamics, and computer application.

YUXUE CHI received the B.S. degree in software engineering from Xiamen University, China, and the M.S. degree in Management Science and Engineering from Institutes of Science and Development, CAS, China. She is currently a doctor candidate in Institutes of Science and Development, CAS, China. Her main research interests include neural network algorithms, opinion dynamics and complex network.

XIANYI TANG received the B.S. degree and the M.S. degree from Fudan University. He is currently a guest scholar in the CAS Center for Interdisciplinary Studies of Social and Natural Sciences, CAS, China. His main research interests include social physics, opinion dynamics, philosophy of science and nonequilibrium thermodynamics.

YIJUN LIU received Ph.D. degree from the Academy of Mathematics and Systems Science, CAS, China. She is the researcher and master tutor at Institute of science and development, CAS. Currently, she also serves as assistant director of Intercross-Science Research Center for Natural Science and Social Science, deputy executive director of Center for Social Governance and Risk Research, Institute of science and development of CAS. As a member of Sustainable Development group in the CAS, she has been studying the sustainable development, new-urbanization, and other related issues since 2004. She has hosted and undertaken a number of important research tasks from National Development and Reform Commission and National Natural Science Foundation of China, published more than 20 reports and papers. Her main research interests include opinion dynamics, social stability early-warning system, and sustainable development strategy.