Sensitive Data Detection of Social Network Based on Improved Random Forest Algorithm

Weiyan Tang
College of Computer Science, Chongqing University, China
twy00108@163.com

Jun Luo
College of Computer Science, Chongqing University, China
luor1990@163.com

Abstract: *The characteristics of social network sensitive data are complex, which leads to the difficulty of detecting social network sensitive data, so to study the sensitive data detection method of social network based on improved Random Forest (RF) algorithm. Simulate login to social network, and capture social network information by means of web crawler and collector. The Topology-Based Hierarchical Trait (TBHT) topology feature logic algorithm optimized by Naive Bayesian (NB) algorithm is used to extract sensitive data features of social networks from social network information. The RF algorithm is improved by adaptive node splitting, and a sensitive data detection model based on the improved RF algorithm is built by combining the characteristics of social network sensitive data. Social network information is input into the model, and relevant detection results are obtained. The experimental results show that the data acquisition mode using web crawler and collector runs stably and has a large amount of data acquisition, and the extracted data features are efficient. The accuracy of the improved RF algorithm in data classification is more than 97.5%. Therefore, this method is a powerful and practical method for detecting sensitive data of social networks.*

Keywords: *Random forest algorithm, web crawler, naive bayes, sensitive data, feature extraction, data detection.*

Received January 16, 2025; accepted October 1, 2025
https://doi.org/10.34028/iajit/23/2/4

1. Introduction

Social networking originated from networking, and the starting point of online social networking is E-mail [16]. The internet is essentially a network of computers [10] E-mail. It has solved the problem of remote mail transmission [19], and is still one of the most common applications on the internet [7], as well as the starting point of social networking. With the evolution of social networking, people's image on the network is more complete. At this time, social networking came into being [2]. Making friends is just a starting point of social networks [14], like Google its starting point of social networking [6] is just to obtain personal information and friends list [3]. Social networks have gone through different stages of development: early conceptualization-SixtDegrees represents six degree separation theory; The stage of meeting strangers-Friendster help build weak relation to improve social principle of theory; stage of entertainment MySpace create a rich multimedia personalized space and attract attention; social graphStage-Facebook replicates offline reality network, came to the theory of online low-cost management [21]; cloud social networking stage the whole process of Social Network Services (SNS) development is to gradually integrate offline life information transfer to online for low-cost management, which makes virtual social networking and real world social networking more and more cross.

However, the booming development of social network has brought convenience and opportunities, but it has also raised a series of serious data security issues. With the rapid flow and aggregation of massive information in social networks, it inevitably contains a large amount of sensitive data, such as personal identity information, privacy records, trade secrets, politically sensitive speech, etc. Once these sensitive data are leaked or maliciously exploited, it not only seriously infringes on personal privacy and rights, but may also threaten the commercial interests of enterprises, social stability, and even national security. Therefore, how to accurately and efficiently detect sensitive data in social networks has become a key issue that urgently needs to be addressed. With the development of the Internet, social networks continue to develop rapidly. In social networks, there are a lot of sensitive data generated by communication [13]. These sensitive data may include the user's gender, region, occupation, hobbies, ID card number, age and other information, involving the user's personal information and privacy. If we do not pay attention to these sensitive data [15], it will lead to the disclosure of users' personal information and privacy, which will seriously threaten the safety of users' personal property. Therefore, it is necessary to implement effective protection [4] for social network sensitive data to avoid user personal information disclosure [22] and ensure the security of social network communication. To protect these sensitive data, it is necessary to conduct effective data detection.

In social networks, many researchers have studied the detection of sensitive data. For example: Akinyelu [1] studied Machine Learning (ML) and Nature Inspired

(NI) spam detection technology of E-mail spam, network spam, social network spam and comment spam. ML our solution is one of the most effective detection solutions at present. However, most ML algorithms are computationally complex, so some researches introduce NI algorithms to further improve the speed and generalization performance of ML algorithm. Through the investigation of data detection technologies based on ML and NI, it can provide information for major companies to design more effective social network sensitive information filtering systems. Rebhi *et al*. [17] studied the stable community detection method based on time reuse graph. This method first uses a mixed sensitive data detection algorithm that considers both relationship heterogeneity and node similarity to find the best static graph partition at every moment. Then consider the time dimension to find the final stable community. Finally, users who publish sensitive data are extracted from composite graphs and real social networks. Idocin *et al*. [9] studied the community detection and social network analysis based on the Italian war in the 15th century. They modeled the social network based on human interaction and proposed the affinity function to capture the nature of local interaction between each pair of participants in the network. The new Borgia clustering algorithm is used to detect the data in the community, and a detection scheme suitable for large and complex social networks is designed, so that sensitive data can be detected for social networks of different scales. Yokotani and Takano [23] studied the prediction of online criminals and victims and their time of crime and damage through daily chat time and online social network activities. Preventing crimes and losses caused by the leakage of sensitive information through social networks has become a long-term challenge in global cyberspace. By using unsupervised and supervised ML to detect sensitive information released by people in social networks, criminal events and locations can be predicted.

The improved Random Forest (RF) algorithm has the characteristics of high precision, good accuracy, difficult to fall into over fitting, strong anti-noise ability and strong adaptability to high-dimensional data and data sets [11, 20]. Therefore, the research on sensitive data detection method of social network based on improved RF algorithm is proposed. This method first needs to collect the data of social networks, then extract the characteristics of sensitive information according to the collected data, and finally use the RF algorithm to detect and classify sensitive information.

2. Social Network Sensitive Data Detection

2.1. Social Network Data Capture Based on Web Crawler

2.1.1. Social Network Simulation Login

Social network simulation login is different from user login. No matter what method is used to collect social network data, it can only be accessed after simulated login. Through the analysis of the header information sent, it can be seen that ordinary users log in with user names and passwords in clear text. The main parameters in the header information are the name value of password and the value of vk. When the social network login is successful, the returned cookie contains a gsid field. You can access the social network by sending a request using the get method to obtain this parameter. The encoding format returned by the page is generally Unicode Transformation Format 8-bit (UTF-8). Browser login is more complicated than mobile login, and the specific steps are as follows:

1. Add (Username), the user name is calculated by base64: username=base64. Encoding (urllib. quote (uscrname)) [:1 1]; Base64 encoding takes every three 8-bit characters as a group according to the length of the string. For each group, first obtain the American Standard Code for Information Interchange (ASCII) encoding of each character, and then convert the ASCII encoding into 8-bit binary to obtain a group of 3*8=24-bit bytes. Then divide the 24-bit bytes into four 6-bit bytes, and fill two high-order zeros in front of each 6-bit byte to obtain four 8-bit bytes. Convert the four 8-bit bytes into decimal system, and compare with the base64 encoding table to get the corresponding encoded characters.
2. Request the prelogin link address. Different social networks have different connections.
3. Encrypt the password. Firstly, create an RSA public key. For the two parameters of the public key, the social network gives fixed values, but both are hexadecimal strings, which are pub key and "10001," convert the two values to decimal, and finally convert the encrypted information to hexadecimal.
4. Request a pass. Use POST to send the request and verify whether the social network login is successful. Refer to the content obtained after POST. If retcode=101, it means login failed; if retcode=0, it means login succeeded.
5. After successfully logging in to the social network, extract the URL address to be used from the content returned by the server, then use the get method to send a request to the server for the URL address, and save the cookie information of this request, which is the cookie to be logged in.

2.1.2. Social Network Data Capture

After the social network simulation login is successful, the application crawler grabs and downloads the social network pages [5] from the social network according to certain logic and algorithm, which is an important part of the search engine [8]. The principle of web crawler is to regard social network as a directed graph, in which the pages on the website are the nodes distributed on the directed graph, and the relationship between pages is

based on the directed edge of the graph. The workflow of web crawlers is based on these directed edges (URL link), start crawling the page from one or more present nodes, get the content on the page, and get other URL links in the page from the content, then continue to traverse other nodes in the network according to these URL links.

This paper adopts a breadth first crawling strategy, which will first process all links on the initial social network page, and then process the next layer of pages corresponding to this page after the completion of the page processing, until the completion of traversal. This can effectively control the crawl depth of the page according to the amount of social network data required, and can avoid the situation that you cannot stop crawling when crawling deep level pages. Social network data capture process:

1. Select a seed user (that is, provide the URL of the initial social network page) as the starting point of the crawler.
2. Process the URLs in the URL collection to be accessed, then download the web pages pointed to by the URLs using multithreading or parallel technology, collect social network web page data with the help of HTTP and other web protocols, and store the collected pages.
3. Analyze the collected social network pages, extract the user ID, standardize the URL address formed by the ID, unify the format and store it for the next step.
4. Analyze the obtained URL address, eliminate duplicate and invalid URLs, and then organize and store the obtained URL. During program execution, the program will select the next URL to be processed from the queue. When the URL is empty, the crawler will terminate. In practical applications, the size of these queues is usually properly controlled. If the number of URLs in the queue is too large, it will increase the load on the server. On the contrary, if the number of URLs in the queue is small, it will reduce the pressure on the server during the execution of the crawler, and improve the speed and efficiency of the crawler. Therefore, the size of the URL queue to be downloaded needs to be set reasonably. The URL queue to be downloaded cannot be too small. If it is too small, the queue will be empty quickly. If there is no new URL, the program will stop, which will affect the efficiency of program execution.
5. Store the social network data downloaded to the local page for post-processing.

When collecting social network data, you can set collection rules according to social network data collection requirements to obtain user IDs, generate unique URL addresses according to IDs, and then use web crawlers to obtain social network data, which improves collection rate and the integrity of the collected content. To address privacy concerns and ensure compliance (such as complying with the General Data Protection Regulation (GDPR) and platform policies) when capturing social network data, it is necessary to establish a set of implementation processes that cover ethical approval, data anonymization, and full process compliance management, as follows:

1) Submit application: to ensure that data collection activities comply with ethical and legal norms and avoid infringing on user privacy, researchers need to submit an ethics review application to the institutional ethics committee (such as universities, enterprises, or third-party ethics review agencies), which should clearly state the research purpose, data sources, collection scope, anonymization methods, and data usage plan; during the approval process, it is necessary to strictly follow the principles of "legality, fairness, and transparency" in GDPR regarding data processing, and ensure compliance with the data usage terms in the target platform policy, in order to balance data utilization and privacy protection needs within the compliance framework. If the captured data involves User Generated Content (UGC) and the user has not publicly shared it, explicit consent from the user is required. Provide clear consent options through the platform Application Programming Interface (API) or user interface, and record the user's consent behavior.
2) Data preprocessing: firstly, clean the raw data, handle missing values, correct formatting errors, and delete duplicate records. Then, filter and delete fields and direct identifiers (such as user ID, name, etc.,) that are unrelated to the study, retaining only necessary information. If the data contains sensitive fields (such as phone numbers), partial anonymization can be performed (such as hiding some characters) to ensure that redundant and risky content has been removed before entering the anonymization stage.
3) Anonymization processing: select appropriate anonymization techniques based on data type and risk: process classified data (such as gender, occupation) by generalizing or replacing them with categories; for numerical data such as age and income, anonymization is achieved by generalizing to intervals or adding random noise; text data (such as addresses, comments) can be reduced in risk through generalization or direct deletion; geographic location data (such as latitude and longitude, postal code) needs to be generalized into regions or grids to protect privacy. At the same time, pseudonymization technology can be selected according to the needs to ensure that the data can still meet the analysis requirements after anonymization.
4) Verification and evaluation: after anonymization is completed, it is necessary to conduct are identification risk assessment of the data to check if there is still a risk of uniquely identifying users through quasi-identifiers. At the same time, the utility of the data should be evaluated to ensure that the

anonymized data can still be used for statistical analysis or ML tasks, and that the distribution and accuracy have not been severely compromised by anonymization. In addition, it is necessary to check whether the anonymization process complies with relevant regulations (such as GDPR, the California Consumer Privacy Act (CCPA)) and platform policy requirements, in order to avoid data being unusable due to violations.

5) Document record: version management of anonymized scripts and configuration files, recording processing timestamps to accurately restore the anonymization process during subsequent audits or repeated operations.

The flowchart for obtaining social network data based on web crawlers [12] is shown in Figure 1. The ID queue is a storage structure used to temporarily store a series of IDs, which may have been generated from the initial "SeedID" after some processing and will be further used in subsequent processes involving "Personal release information" and so on. Data storage "should be a place for storing relevant data". As shown in Figure 1, data related to different IDs (such as id1, id2, etc., corresponding to a series of data, such as id1 1, id1 2, etc.,) will be stored here for long-term preservation or for subsequent operations to call upon these social network data.

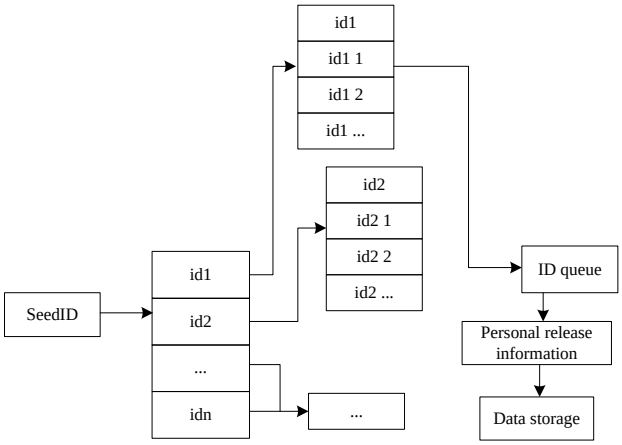

Figure 1. Flow chart of social network data acquisition based on web crawler.

2.2. Feature Extraction of Social Network Sensitive Data Based on Naive Bayesian Algorithm

The Topology-Based Hierarchical Trait (TBHT) topology feature logic algorithm optimized by Naive Bayesian (NB) algorithm is used to extract sensitive data features of social networks from social network information [18].

2.2.1. Naive Bayesian Algorithm

NB algorithm belongs to the general name of a class of classification algorithms, which are based on NB theorem, so they can also be collectively called NB classification. NB quantification can reasonably solve the events that often occur in real life. If two events are separated by C and D indicates that the event C and events D the probability of occurrence is $Q(C)$, $Q(D)$ to indicates that the probability of two events occurring at the same time is $Q(CD)$ to indicates that in the event C the event D probability of occurrence $Q(D/C)$ to indicates that there are:

$$Q(D/C) = Q(CD)/Q(C) \tag{1}$$

Event C and event D are both independent events:

$$Q(CD) = Q(D)Q(C/D) = Q(C)Q(D/C) \tag{2}$$

Set sample space Ω a division of D_1, D_2, ..., D_n, can meet D_i pairs are mutually exclusive, and $D_iD_j=\varnothing(i\neq j)$, $\sum D_i=\Omega(i=1, 2, ..., n)$ then there is a full probability formula:

$$\begin{aligned} Q(C) &= Q(C \cap \Omega) = Q(C \cap \sum D_i) = Q(\sum CD_i) \\ &= Q(CD)\sum Q(CD_i) = \sum Q(D_i)Q(C/D_i) \end{aligned} \tag{3}$$

Hypothesis $Y\in\Omega$ belongs to an unknown type of social network data sample, a_j belongs to a certain type. If the social network data sample Y is for a specific type a_j then the classification problem is determined $Q(a_j|Y)$, that is, getting social network data samples Y determine social network data samples the best classification of Y. In the given dataset B each type inside a_j under the condition of prior probability, the most likely classification is the best classification. A more direct way to calculate this possibility is through Bayesian theorem. Bayesian theorem provides a method to calculate the hypothetical probability, which can be expressed as:

$$Q(a_j|Y) = Q(Y|a_j)Q(a_j)/Q(Y)Q(C) \tag{4}$$

The above formula belongs to Bayesian formula, where $Q(a_j)$, $Q(a_j|Y)$, $Q(Y|a_j)$ represent prior probability, posterior probability and joint probability respectively.

1. Prior probability $Q(a_j)$:a_j of the prior probability of is $Q(a_j)$, about a_j the background knowledge belonging to accurate classification opportunity can be reflected by a priori probability. If there is no such prior knowledge, the same prior probability can be

assigned to each type to be selected. However, in general, you can use the a_j number of samples $|a_j|$ and the total number of samples $|B|$ that the ratio of:

$$Q(a_j) = \frac{|a_j|}{|B|} \quad (5)$$

2. Joint probability $Q(Y|a_j)$: joint probability is also called conditional probability, that is, when known as a_j type, social network data samples Y the probability of occurrence. If set $Y=<c_1, c_2, ...c_m|a_j>$, then there are:

$$Q(Y|a_j) = Q(c_1, c_2, ... c_m|a_j) \quad (6)$$

3. Posterior probability $Q(Y|a_j)$:a_j the posterior probability of is $Q(a_j|Y)$, refers to when social network data samples Y when given, a_j probability of establishment. Can reflect social network data samples Y after the establishment confidence of a_j.

2.2.2. Feature Extraction of Social Network Sensitive Data

The TBHT topology feature logic algorithm optimized based on NB algorithm can capture the features of social network sensitive data according to the topological relationship between social network data. At the same time, the transformation of various social network sensitive data characteristics through probability logic can form a necessary condition for the later detection of social network sensitive data using improved RF algorithm, thus improving the accuracy of social network sensitive data detection. TBHT topological feature logic algorithm inherits the probability operation characteristics of NB algorithm, and has independent logic processing performance. Therefore, the relation of TBHT topological feature logic algorithm is a double subset, which can be expressed as:

$$Q(c_i kx) = \frac{|Q(xkc_i) \wedge Q(c_i \wedge)|}{\sum_i Q(x)^{\infty} Q(a_j|Y)} \quad (7)$$

$$\begin{gathered} Q(xkc_i) = Q(c_1) \sum_{i\in m}^{d_1} (c_i d^m) \forall \\ Q(c_2) \sum_{i\in m}^{d_2} (c_i d^m) \forall \\ Q(c_3) \sum_{i\in m}^{d_3} (c_i d^m) \forall \\ Q(c_m) \sum_{i\in m}^{d_m} (c_i d^m) \forall \end{gathered} \quad (8)$$

In Equations (7) and (8), $Q(c_i kx)$ represents the probability that social network sensitive data type c_i appears under the condition of given social network data x, $Q(xkc_i)$ represents the corresponding occurrence probability of social network data x when considering the probability constant k and sensitive data type c_i, $Q(x)$ represents the occurrence probability of social network data x, and $Q(a_j|Y)$ represents the sensitive data feature a_j under the specific data set condition Y. c is a collection of sensitive data types of social networks, i and m are the coefficient of sensitive data type and characteristic type of social network, k and d are a set of probability constants and social network data mining feature constraints. When Equation (7) can meet Equation (8) as a constant introduced, it means that the construction of the mathematical model for feature extraction of social network sensitive data has been completed. The output of the mathematical model is the result of feature extraction of social network sensitive data, as shown in Figure 2.

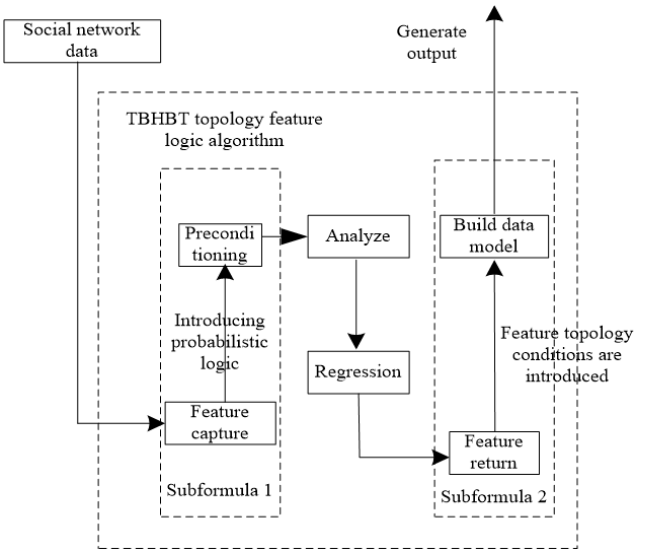

Figure 2. Structure of feature extraction for sensitive data in social networks.

2.3. Sensitive Data Detection Based on Improved Random Forest Algorithm

The RF algorithm is improved by adaptive node splitting, and a sensitive data detection model based on the improved RF algorithm is built by combining the characteristics of social network sensitive data. Social network information is input into the model, and relevant detection results are obtained.

2.3.1. Random Forest Algorithm

RF algorithm is an integrated learning algorithm, which is a set of regression trees, and outputs by averaging the detection values of all regression trees. The RF algorithm uses the automatic resampling technology to overcome the over fitting problem of the regression tree and greatly improve the performance of the model. In addition, it can process high-dimensional data, suitable for numerical variables and category variables, and can be parallelized to adapt to large data sets. For this reason, this paper applies the RF algorithm to complete the detection of social network sensitive data.

The main factors that affect the performance of the RF model in the detection of sensitive data in social networks include the detection intensity of a single tree and the correlation between trees. If the detection intensity of a single tree is better, the detection

performance of social network sensitive data of the overall RF model is better. The smaller the correlation between trees, the better the social network sensitive data detection performance of the RF model will be. These two factors can be controlled by the number of variables preselected for each tree and the number of trees during the detection of sensitive data of social networks.

In the training of regression tree, Classification and Regression Tree (CART) algorithm. It is a dichotomous recursive segmentation technique, which divides the training set currently constructed from the above extracted social network sensitive data features into two sub training sets, so that each non leaf node of the spanning tree has two branches. Non leaf nodes represent the characteristics of social network sensitive data, while leaf nodes represent the detection values of social network sensitive data given by the tree model. CART algorithm steps are as follows:

1. Select a social network sensitive data feature according to certain conditions, and divide the node of the tree into two branches according to this feature.
2. Repeat the above steps recursively on each branch until one of the following conditions is met: the decrease of deviation is less than the given limit value; the number of samples in the node is less than the given limit value; the depth of the tree is greater than a given threshold value.

The regression tree is constructed from top to bottom. The characteristics of sensitive data of social networks are selected by calculating the best partition points, described by the node's impurity indicator *GINI*, which is defined as follows:

$$GINI = 1 - \sum_{i=1}^{M} p_i^2 Q(c_i kx) \tag{9}$$

Where, p_i is the sample in the node belongs to the class probability of i; M is the number of classes in the node.

In order to avoid the problem that the regression tree is too large and the resulting social network sensitive data detection is over fitting, the regression tree needs to be pruned to cut the branches that do not contribute much to the model and improve the detection efficiency of social network sensitive data. Pruning complexity parameter c_p value, c_p value is a measure of how well the tree with new nodes improves the goodness of fit of the model. In addition, the important parameters that affect the performance of the regression tree are the minimum sample number of nodes, the minimum sample number of leaf nodes, and the depth of the tree. The flow of RF algorithm detecting sensitive data of social network is shown in Figure 3.

The steps of RF algorithm to detect sensitive data of social networks are as follows:

1. Set the characteristics of social network sensitive data in the training set as $X=\{x_1, x_2, ..., x_n\}$, the sensitive data detection result is $Y=\{y_1, y_2, ..., y_n\}$.
2. From X, Y randomly select a sub sample set in X_b, Y_b, as a training set.
3. To X_b, Y_b training a regression tree model r_{fb}.
4. To b=1, 2, ..., B_s repeat steps 3 and 4 for continuous training.

After the training, a new sample x the RF model gives the sample attribute detection value by averaging the detection values of all regression trees:

$$p = \frac{1}{B_S} \sum_{b=1}^{B_S} r_{fb}(x) \tag{10}$$

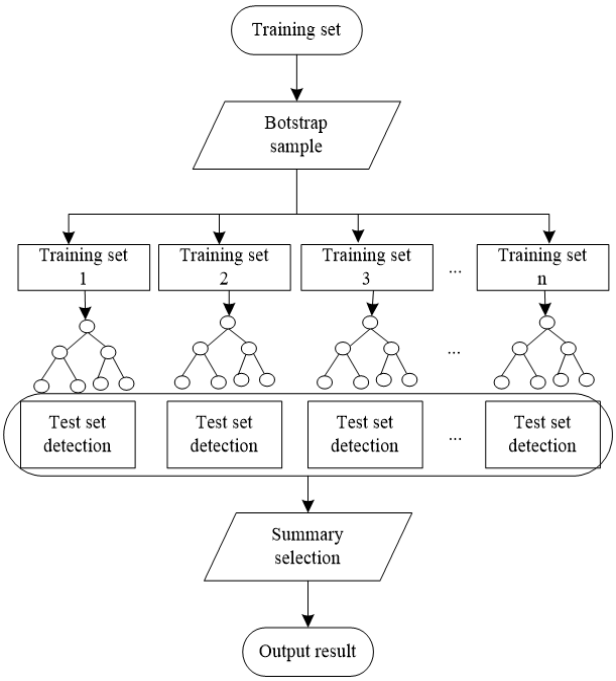

Figure 3. Sensitive data detection process.

2.3.2. Sensitive Data Detection of Social Networks Based on Improved Random Forest Algorithm

The RF algorithm has a high classification accuracy in detecting sensitive data of social networks, has a good tolerance for outliers and noise, and is not prone to over fitting. SANS-RF algorithm proposed in this paper (Self-Adaptive Node Split Random Forest, adaptive node splitting RF algorithm) the node splitting algorithm of the decision tree in the algorithm is optimized through the adaptive selection process of parameters, so as to improve the classification accuracy of the algorithm and further improve the detection accuracy of sensitive data of social networks.

Based on algorithm principles and task characteristics, preliminarily determine the range of hyperparameters. For example:

Tree depth: insufficient depth may lead to underfitting, while excessive depth may lead to overfitting.

Splitting coefficient (α, β): in the splitting of composite nodes, α and β control the weight or threshold of the splitting. Adjustments should be made based on data distribution (such as class imbalance), with *a*

controlling for feature importance and β controlling for sample distribution weights.

Grid search is a method of finding the optimal parameters by exhaustively searching for preset hyperparameter combinations. The specific implementation steps are as follows:

1. For the RF model, hyperparameters that need to be tuned include the depth of the tree [5, 10, 15, 20]; the number of trees [50, 100, 200]; minimum sample size for node splitting [2, 5, 10]; the values of α are [0, 0.2, 0.4, 0.6, 0.8, 1], and the values of β are [0, 0.2, 0.4, 0.6, 0.8, 1].
2. Evaluate the performance of each hyperparameter combination through cross validation to avoid overfitting. Divide the data into k subsets, train with k-1 subset each time, validate with 1 subset, repeat k times, and calculate the average performance. For each hyperparameter combination in the search space, train the model on the training set, evaluate the model performance on the validation set, save the validation set performance indicators for each combination, and select the optimal combination based on the validation set performance indicators.

For the same social network dataset, if different node splitting algorithms are selected, different decision trees will be obtained due to different selected attributes, and the classification accuracy of the resulting RF will also be different. Therefore, it is proposed to select the optimal attribute for node splitting when generating the decision tree, that is, the node splitting algorithm is linearly combined to form a new splitting rule, which is applied to the selection of node attributes. Since only ID3 and CART are the integrated node splitting algorithms in the RF algorithm, the consideration of node splitting optimization is based on these two algorithms. Its node splitting formula is expressed by attributes a for sample set D the information gain and Gini index obtained from the division are as follows:

$$Gain(D,a) = Ent(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} Ent(D^v) GINI \quad (11)$$

$$Gini(D,a) = p \sum_{v=1}^{V} \frac{|D^v|}{D} Gini(D^v) \quad (12)$$

Where, D^v is v contained by branch nodes D all in attribute a the upper value is sample of a^v:

$$Ent(D^v) = -\sum_{k=1}^{|y|} p_k \log_2 p_k \quad (13)$$

$$Gini(D^v) = 1 - \sum_{k=1}^{|y|} p_k^2 \quad (14)$$

Equations (13) and (14) respectively represent the information entropy and *Gini* value of social network dataset D, p_k the sample in the node belongs to the class probability of k.

In SANS-RF, the information gain of ID3 or the Gini index of CART is not simply used as the sole node splitting criterion to select splitting attributes under certain levels or conditions, in order to pursue the maximization of sample purity; in other cases, use the Gini index to focus on the differences in probability distributions of different categories. This can comprehensively utilize the advantages of both and avoid the limitations of a single standard. The node splitting indicators and rules of ID3 and CART are shown in Table 1:

Table 1. Comparison of node splitting algorithms.

Algorithm	Node splitting criterion	Criterion index
ID3	Maximum information gain	Partition of the data set by sample purity
CART	The Gini index is the smallest	Two different probabilities are randomly selected from the data set

In combination with the contents in Table 1, the node splitting criteria should aim at higher purity of the social network sensitive data set after division, so the splitting formula of composite node is:

$$H = \min_{\alpha,\beta,\in R} F\{D,a\} Gain(D,a) Gini(D,a)$$
$$s.t \begin{cases} \alpha + \beta = 1 \\ 0 \le \alpha, \beta \le 1 \end{cases} \quad (15)$$

Where, parameter α, β are two algorithms coefficient in $H(x)$, where H take the minimum value, that is, both ID3 and CART are optimal, can be used as a node division criterion to improve the classification effect of sensitive data on social networks.

Because the sensitive data characteristics of data in different data sets are different, the parameter selection in SANS-RF algorithm is difficult to fix. Therefore, the adaptive parameter selection process is adopted to obtain the optimal combination parameters, improve the detection effect of social network sensitive data α, β the constraints in the above formula shall be met.

For samples D, sensitive data detection error rate is defined as:

$$E(f;D) = \frac{H}{m} \sum_{i=1}^{m} \text{II}(f(x_i) \ne y_i) \quad (16)$$

A sensitive data detection model based on improved RF algorithm is built by combining the characteristics of social network sensitive data. Social network information is input into the model to obtain relevant detection results. The sensitive data detection model of social network based on improved RF algorithm is:

$$acc(f;D) = \frac{1}{m} \sum_{i=1}^{m} \text{II}\left(f(x_i) \ne y_i\right) = 1 - E(f;D) \quad (17)$$

Where, m is the number of samples.

3. Experimental Analysis

3.1. Experimental Objects

SinaWeibo is a social media platform based on user relations. Users can access it through a variety of mobile

terminals, such as Personal Computers (PCs) and mobile phones, and realize instant sharing, communication and interaction of information in the form of text, pictures, videos and other multimedia forms. Weibo is based on the open platform architecture, enabling users to simply publish content publicly, and let users interact with others through fission communication, so as to closely connect with the world. As one of the new entrances to the internet, microblog has changed the way of information dissemination and realized the instant sharing of information. SinaWeibo has multiple functional advantages such as publishing, forwarding, following, commenting, searching, and private messaging. But at the same time, because SinaWeibo's information release threshold is extremely low, the content of Weibo is not limited, the communication is rapid, the communication mode is fission, and the information interaction is simple and fast, it will lead to the production of a large number of sensitive data, which is not conducive to the positive, healthy, orderly and benign development of social networks.

Therefore, SinaWeibo is taken as the social network studied in this paper to detect sensitive data and verify the detection effect of this method on sensitive data of social networks.

3.2. Experimental Data

In order to verify the efficiency of the social network data collection method in this paper, the 1-hour data collection volume is compared with some common methods, and the results are shown in Table 2.

Table 2. Number of microblogs collected in one hour.

Acquisition method	Collection quantity/piece
API	8127
Web crawler	8296
Harvester	8762
APIs and web crawlers	10022
Collectors and web crawlers	13279

It can be seen from Table 2 that the method in this paper is used to collect social network data, and the number of data collected within one hour is the largest. The reason is that the method in this paper effectively combines the collector and the web crawler to obtain social network data more comprehensively.

To detect sensitive data of social networks, fast and complete access to a large amount of data is a prerequisite. When selecting the experimental method, we must consider the time consumption in the data acquisition process, which may be related to the network bandwidth and the amount of data processing when the program is running. In the experiment, the number of user data collected by five methods, namely API, web crawler, collector, API and web crawler integration, collector and web crawler integration, was recorded for several hours continuously. Through comparative analysis, it is concluded that in the case of long-term data acquisition, the integration of web crawler and collector is the most stable and efficient acquisition scheme. The specific experimental results are shown in Figure 4, where the horizontal axis is the time of continuous collection and the vertical axis is the amount of user information collected and processed in unit time.

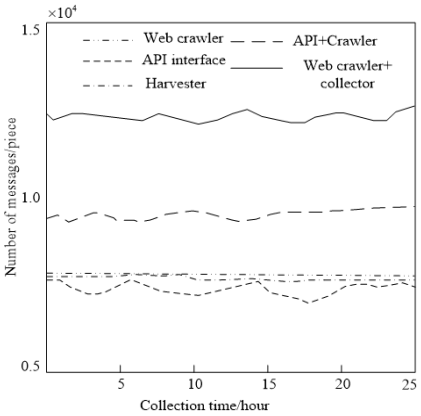

Figure 4. Comparison of continuous acquisition performance.

The overall performance analysis of the five data acquisition methods is shown in Table 3.

Table 3. Overall comparison of collection schemes.

Index	API	Web crawler	Harvester	APIs and web crawlers	Collectors and web crawlers
Speed	Low	Mid	High	Little high	High
Integrality	Mid	High	Mid	Little high	High

It can be seen from Figure 4 and Table 3 that only the data collection method of web crawler+collector adopted in this paper can maintain a stable and large amount of data collection in a long-time continuous collection, and rate fast, high integrity.

To verify the accuracy of the method in this paper for feature extraction of sensitive information in social networks, compare it with the feature extraction method based on Genetic Algorithm (GA) and the feature extraction method based on Convolutional Neural Network (CNN). The feature types and their examples are described as follows:

1. Vocabulary features refer to keywords or phrases directly extracted from text content that are related to sensitive information. In social media posts involving personal privacy, vocabulary features include "ID number," "bank card number," "home address," etc. The following are examples of vocabulary features:

   a) The content of the post: "My ID number number is 123456789012345678, please don't disclose it."
   b) Extracted lexical features: ID number, 123456789012345678.

2. Metadata features: Meta refers to information related to text content but not directly present in the text, such as posting time, posting location, user ID, etc. Although these pieces of information do not directly contain sensitive content, they may reveal sensitive

information through their association with other information. In a social network post about a secret meeting, Metadata features may include posting time, posting location, etc. These informations combined with text content suggest the sensitive nature of the meeting.

a) Post content: "I met celebrity XXX at XX mall today, so excited!"
b) Extracted metadata features: posting time, posting location.

3. Contextual features refer to the context, background, or related topics in which the text content is located, which helps to more accurately determine whether the text contains sensitive information. In a seemingly ordinary social network post, if the context is discussing a sensitive event or topic, then the post contains sensitive information related to the sensitive event.

- Post content: recently, there have been internal adjustments within the company, so everyone should be careful in their actions.
- Context: this post was posted on the company's internal forum, and the company has indeed been undergoing personnel adjustments recently.
- Extracted contextual features: internal company forums, personnel adjustments.

The accuracy of the three methods for feature extraction of sensitive information is shown in Figure 5.

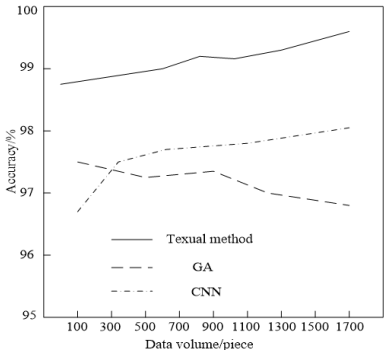

Figure 5. Accuracy of data feature extraction.

It can be seen from Figure 6 that the accuracy of data feature extraction of the three methods exceeds 95%. When the amount of data is less than 300, the accuracy of data feature extraction of GA method is higher than that of CNN method. As the amount of data increases, the accuracy of data feature extraction of GA method declines all the way, while that of CNN method gradually increases. However, the accuracy of the method in this paper is higher than that of the GA method and CNN method from the initial data feature accuracy, and the accuracy gradually increases from 98.8% with the increase of data volume. It can be seen that the feature extraction method of sensitive data adopted in this paper has a high accuracy of feature extraction.

To verify the performance of the improved RF algorithm in classifying sensitive data (text and images) in social networks, this study selected RF, Support Vector Machine (SVM), and Bidirectional Encoder Representations from Transformers (BERT) model as comparison methods, mainly based on the following considerations: Firstly, using the original RF as the baseline, directly evaluate the effectiveness of the improved strategies (such as feature selection or parameter optimization). Secondly, SVM is introduced as a representative of traditional ML, utilizing its stable performance in small and medium-sized data and high-dimensional features to form a complementary contrast with RF; finally, BERT is chosen as the benchmark model for deep learning, especially for text classification tasks, to test whether the improved RF can approach or surpass cutting-edge deep learning methods. This comparative design covers typical methods from traditional ML to modern deep learning, which can not only verify the generalization ability of improved RF, but also reveal the advantages and disadvantages of different algorithms in sensitive data classification tasks, such as the robustness of RF, the small sample performance of SVM, and the semantic understanding ability of BERT, thus providing a basis for method selection for different application scenarios. The results are shown in Figures 6 and 7.

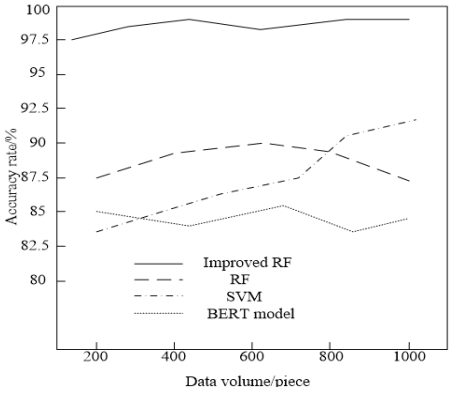

Figure 6. Comparison of classification accuracy of text information.

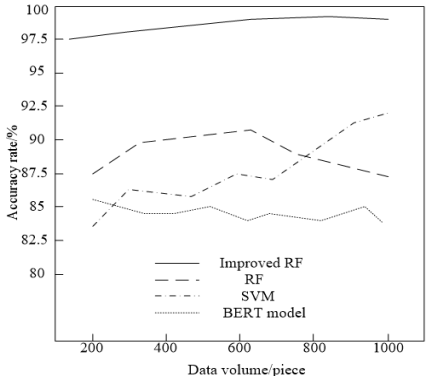

Figure 7. Comparison of classification accuracy of picture information.

Explain the classification error rates of different methods for text and image information in the form of a table, and the results are shown in Table 4.

Table 4. Information classification error rate.

Data volume/Piece	Improved RF		RF		SVM		BERT model	
	Classification error rate of text information/%	Classification error rate of picture information/%	Classification error rate of text information/%	Classification error rate of picture information/%	Classification error rate of text information/%	Classification error rate of picture information/%	Classification error rate of text information/%	Classification error rate of picture information/%
200	2.5	2.5	12.5	12.5	16.8	16.7	15.1	14.1
400	1.2	1.7	11.2	10.3	15.1	13.2	16.2	15.8
600	2.2	1.2	10.3	9.8	10.9	12.9	15.4	16.3
800	1.3	1.1	11.4	11.9	11.5	11.8	16.3	16.1
1000	1.2	1.2	13.1	13.2	8.7	8.4	16.4	16.77

To determine whether the difference in accuracy between different methods is statistically significant, the Analysis of Variance (ANOVA) method is used. The specific assumptions are as follows:

- Zero Hypothesis (H0): there is no significant difference in the average accuracy among the four methods of improved RF, RF, SVM, and BERT model.
- Alternative Hypothesis (H1): there is a significant difference in the average accuracy between at least two methods.

Perform variance analysis on the accuracy data of different methods at various data sizes as samples. Calculate the inter group variance (variance between different methods) and intra group variance (variance within the same method) to obtain the F-statistic. If the p-value obtained is less than the pre-set significance level, the null hypothesis H0 is rejected, indicating that there is a significant difference in the average accuracy of at least two methods. From the trend of the chart, the accuracy of the improved RF method is significantly higher than that of the RF, SVM, and BERT model methods. Through ANOVA, it may be found that the difference in accuracy between improved RF and the RF, SVM, and BERT models is statistically significant. From Figures 6 and 7, it can be seen that sensitive data types do not affect the classification accuracy of the three methods. Among them, when the data volume reaches 800, the SVM method achieves the same classification accuracy as the RF method for text/image sensitive data. Furthermore, as the amount of social network data that needs to be classified increases, the SVM method's classification accuracy continues to increase. The method proposed in this article can accurately classify sensitive social network data regardless of the amount of data, and its classification accuracy for sensitive data exceeds 97.5%. According to the analysis of error data, compared with the experimental comparison method, the maximum error rate of improved RF is 2.5%, the maximum error rate of RF is 13.2%, and the maximum error rate of SVM is 16.8%, the maximum error rate of the BERT model is 16.7%, and the fluctuation range of the experimental comparison method is relatively large, indicating that improved RF has lower errors and higher classification accuracy.

On this basis, in order to further verify the comprehensive performance of the method, relevant tests were conducted on the detection performance of social network sensitive data on a large-scale dataset. The test results are shown in Table 5.

Table 5. Performance test of social network sensitive data detection.

Large scale dataset name	Detection accuracy/%			
	Improved RF	RF	SVM	BERT model
HateSpeech	98.76	86.55	90.32	80.33
Sensitive information subset of SQuAD	98.57	87.49	91.47	81.47
Sensitive subset of COCO	99.12	86.35	90.33	82.63
NSFW dataset	98.31	87.89	89.66	85.21
Twitter's image and text posting dataset	97.63	85.26	86.32	84.76
Instagram's image dataset with text description	97.89	86.37	87.45	81.69

According to the analysis of the data in Table 5, for various types of large-scale datasets, the maximum accuracy of social network sensitive data detection for improved RF is 99.12%, and the minimum is 97.63%; the maximum accuracy of RF's social network sensitive data detection is 87.89%, and the minimum is 85.26%; the maximum accuracy of SVM for detecting sensitive data in social networks is 91.47%, and the minimum is 86.32%; the maximum accuracy of the BERT model for detecting sensitive social network data is 85.21%, and the minimum is 80.33%. After comparison, it can be seen that improved RF has a higher accuracy in detecting sensitive data in social networks, significantly better than the comparison method. However, its computational cost is higher, and it relies on parallelization technology to achieve efficient training and inference. This is mainly due to the increase in model complexity and the processing pressure brought by large-scale datasets themselves. To alleviate this problem, parallelization work can be considered, including feature parallelization (dividing the feature set into different computing nodes for independent processing), sample parallelization (dividing the sample set into batches for parallel training of decision trees), ensemble method parallelization (predicting results and subsequent ensemble operations of parallel computing base learners), and using distributed computing frameworks such as Apache Spark and Hadoop to achieve efficient parallel training, thereby effectively improving the processing efficiency of the model on large-scale datasets while maintaining high accuracy.

The sensitive data features are divided into five categories using this method, and the sensitive data

published by users are detected according to these sensitive data feature categories. Sensitive data are divided into five categories: gender, occupation, interest, social situation and region. The detection and classification of sensitive data in social networks are shown in Table 6.

Table 6. Detection and classification of text sensitive information.

Data number	User ID	Content of speech	Whether the information is sensitive	Sensitive information type
001	The brightest star in the night sky	That's a nice haircut	No	/
002	Cat	Chongqing also has this kind of food	Yes	location
003	Diet soda	Another day of hard work	Yes	occupation
004	Bread slicing	UEFA champions league winner!!!!	No	/
005	0276851	Oktoberfest! Begin to drink!	No	/
006	Not listening to music	I often go to KTV	Yes	location
007	Idol Messi	Don't want to go to work	No	/
008	Odontoloxia	The movie is really good.	No	/
009	Eating without gaining weight	New nails	Yes	sex
010	Fantasy deep sky	Circle of friends	Yes	Social situation

From Table 6, it can be seen that using the method proposed in this article to detect sensitive data on the information posted by these 10 users, the detection results are detailed, including the data number, user ID, and content of the posted information. When the information is sensitive, it will also display what kind of sensitive information it belongs to. Judging 'another day of hard work' as containing occupational sensitive information may result in a misjudgement. This sentence itself is quite vague and general, and does not clearly indicate the specific profession. It is more like a daily expression of emotions, belonging to a general description of work status, rather than sensitive information specific to a particular profession. This is because the classification model is overly sensitive when dealing with such broad expressions, equating work-related expressions with occupational sensitive information. In social networks, ambiguous slang and satirical expressions often pose challenges to the classification of sensitive information. Slang such as "lit" has diverse meanings, and if the model lacks sufficient training data and contextual understanding ability, it is easy to mistake it for sensitive information; satirical expression depends on context, for example, 'Great, now I have even more work to do' actually expresses dissatisfaction. If the model cannot recognize the ironic tone, it will also make a literal judgment. Although such misclassifications are not clearly reflected in the table data, identifying slang and satire is an important challenge in actual social network text processing. Therefore, in order to reduce these failure modes and misclassification situations, it is necessary to further optimize the classification model, increase the understanding and processing ability of various information such as semantics, context, slang, tone, etc., and use richer and more accurate training data to improve the performance of the model.

4. Conclusions

It can be seen from the experimental results that this method adopts the data acquisition method, which has the following characteristics: large amount of data acquisition, long time data acquisition, stable operation and no large fluctuations. At the same time, the data feature extraction method adopted in this method has a high accuracy rate, which provides a reliable basis for detecting the sensitive information of users' social networks. The improved RF algorithm has high data classification accuracy and strong practicability. In addition, this method is used to detect 10 pieces of randomly collected data, which can clearly determine the issuer ID, information content, whether it is sensitive information, and the classification of sensitive information. In conclusion, this method is a sensitive information detection method with strong detection ability and high classification accuracy for social networks.

References

[1] Akinyelu A., "Advances in Spam Detection for Email Spam, Web Spam, Social Network Spam, and Review Spam: ML-Based and Nature-Inspired-Based Techniques," *Journal of Computer Security*, vol. 29, no. 5, pp. 473-529, 2021. https://doi.org/10.3233/JCS-210022

[2] Atroszko P., Abiddine F., Malik S., Mamun M., and et al., "Lack of Measurement Invariance in a Widely Used Facebook Addiction Scale May Thwart Progress in Research on Social-Network-Use Disorder: A Cross-Cultural Study," *Computers in Human Behavior*, vol. 128, no. 3, pp. 107132, 2022. https://psycnet.apa.org/doi/10.1016/j.chb.2021.107132

[3] Bhari P., "Use of Machine Learning and Detect Fake Profiles in a Social Media Network," *ECS Transactions*, vol. 107, no. 1, pp. 11905-11911, 2022. https://doi.org/10.1149/10701.11905ecst

[4] Cho S. and Kim H., "Privacy Preserving Authenticated Key Agreement Based on Bilinear Pairing for uHealthcare," *The International Arab Journal of Information Technology*, vol. 18, no. 4, pp. 523-530, 2021. DOI:10.34028/18/4/4

[5] Chuanxing S., Rong Z., and Lixiu S., "Research on Keyword Matching Retrieval of Web Crawler Based on Python Language," *Computer Simulation*, vol. 40, no. 3, pp. 504-507, 2023. https://doi.org/10.3969/j.issn.1006-9348.2023.03.095

[6] Fleming Z., "Using Virtual Outcrop Models and Google Earth to Teach Structural Geology Concepts," *Journal of Structural Geology*, vol. 156, no. 3, pp. 104537, 2022. https://doi.org/10.1016/j.jsg.2022.104537

[7] Ghaleb S., Mohamad M., Fadzli S., and Ghanem W., "E-mail Spam Classification Using Grasshopper Optimization Algorithm and Neural Networks," *Computers, Materials, Continua*, vol. 71, no. 3, pp. 4749-4766, 2022. https://doi.org/10.32604/cmc.2022.020472

[8] Hatcher W., Qian C., Gao W., Liang F., and et al., "Towards Efficient and Intelligent Internet of Things Search Engine," *IEEE Access*, vol. 9, pp. 15778-15795, 2021. https://doi.org/10.1109/ACCESS.2021.3052759

[9] Idocin J., Betanzos A., Cordon O., Bustince H., and Minarova M., "Community Detection and Social Network Analysis based on the Italian Wars of the 15th Century," *Future Generation Computer Systems*, vol. 113, pp. 25-40, 2020. https://doi.org/10.1016/j.future.2020.06.030

[10] Khan N., Ray R., Zhang S., Osabuohien E., and Ihtisham M., "Influence of Mobile Phone and Internet Technology on Income of Rural Farmers: Evidence from Khyber Pakhtunkhwa Province, Pakistan," *Technology in Society*, vol. 68, no. 2, pp. 101866, 2022. https://doi.org/10.1016/j.techsoc.2022.101866

[11] Liu D., Dai Q., Tang X., Zhang R., and et al., "An Improved Random Forest-Based Operation Duration Prediction of Long-Distance Tunnel Construction Considering Geological Uncertainty," *Journal of Computing in Civil Engineering*, vol. 39, no. 2, pp. 1-15, 2025. https://doi.org/10.1061/JCCEE5.CPENG-6041

[12] Liu J., Li X., Zhang Q., and Zhong G., "A Novel Focused Crawler Combining Web Space Evolution and Domain Ontology," *Knowledge-Based Systems*, vol. 243, pp. 108495, 2022. https://doi.org/10.1016/j.knosys.2022.108495

[13] Martindale N., Stewart S., Mcgirl N., Adams M., and et al., "Enabling Computation on Sensitive Data in International Safeguards with Privacy-Preserving Encryption Techniques," *Journal of Nuclear Materials Management*, vol. 49, no. 2, pp. 16-25, 2021. https://www.osti.gov/biblio/1827049

[14] Meissa M., Benharzallah S., Kahloul L., and Kazar O., "A Personalized Recommendation for Web API Discovery in Social Web of Things," *The International Arab Journal of Information Technology*, vol. 18, no. 3A, pp. 438-445, 2021. DOI:10.34028/iajit/18/3A/7

[15] Norman J., "Duplicate Sensitive Data Aggregation in Heterogeneous WSN," *International Journal of Computational Physical Sciences*, vol. 15, no. 2, pp. 131-146, 2024.

[16] Rajaraman P. and Prakash M., "Intelligent Deep Learning Based Bidirectional Long Short Term Memory Model for Automated Reply of E-mail Client Prototype," *Pattern Recognition Letters*, vol. 152, no. 12, pp. 340-347, 2021. https://doi.org/10.1016/j.patrec.2021.10.021

[17] Rebhi W., Yahia N., and Saoud N., "Stable Communities Detection Method for Temporal Multiplex Graphs: Heterogeneous Social Network Case Study," *The Computer Journal*, vol. 64, no. 3, pp. 418-431, 2020. https://doi.org/10.1093/comjnl/bxaa162

[18] Rodrigues A., Villela M., and Feitosa E., "A Systematic Mapping Study on Social Network Privacy: Threats and Solutions," *ACM Computing Surveys*, vol. 56, no. 7, pp. 1-29, 2024. https://doi.org/10.1145/3645086

[19] Sun J. and Gloor P., "E-mail Network Patterns and Body Language Predict Risk-Taking Attitude," *Future Internet*, vol. 13, no. 1, pp. 17-29, 2021. https://doi.org/10.3390/fi13010017

[20] Tao L. and Xue X., "An Improved Random Forest Model to Predict Bond Strength of FRP-to-Concrete," *Journal of Civil Engineering and Management*, vol. 30, no. 6, pp. 250-535, 2024. https://doi.org/10.3846/jcem.2024.21636

[21] Thriveni M., Rao M., and Giribabu S., "Combining the Behaviour of User and Relationships to Predicting the Links in Social Networks," *AIP Conference Proceedings*, vol. 2512, no. 1, pp. 5-16, 2024. https://doi.org/10.1063/5.0140368

[22] Wood J. and Schalkwyk I., "Reproducibility in Transportation Research: Importance, Best Practices, and Dealing with Protected and Sensitive Data," *Journal of Transportation Technologies*, vol. 15, no. 1, pp. 179-202, 2025. https://doi.org/10.4236/jtts.2025.151010

[23] Yokotani K. and Takano M., "Predicting Cyber Offenders and Victims and their Offense and Damage Time from Routine Chat Times and Online Social Network Activities," *Computers in Human Behavior*, vol. 128, no. 3, pp. 10709, 2022. https://doi.org/10.1016/j.chb.2021.107099

Weiyan Tang obtained his Bachelor's degree in Computer Science and Technology from Changchun University of Science and Technology in 2018. From 2018 to now, he is currently pursuing a Master's degree in Computer Technology at Chongqing University. His research interests include Computational Intelligence, Information Security, and Big Data Analysis.

Jun Luo obtained his Bachelor's degree in Computer Science and Engineering from Xi'an Jiaotong University in 1987. And he obtained a Master's degree in Computer Science and Engineering from Xi'an Jiaotong University in 1990. His research interests include Computational Intelligence, Information Security, and Big Data Analysis. From 1990 to 1997, he worked as a researcher at Xi'an Jiaotong University; from 1997 to now, he holds the position of Associate Professor and also serves as a Supervisor for Master's students at Chongqing University. He has published 23 academic papers, and one academic work textbook. He participated in 25 scientific research projects, and has 5 patents and two academic awards.