**International Journal of Digital Crime and Forensics**
Volume 14 • Issue 3

# An Unhealthy Webpage Discovery System Based on Convolutional Neural Network

Zengyu Cai, School of Computer and Communication Engineering, Zhengzhou University of Light Industry, China
Chunchen Tan, School of Computer and Communication Engineering, Zhengzhou University of Light Industry, China
https://orcid.org/0000-0001-8569-2584
Jianwei Zhang, College of Software Engineering, Zhengzhou University of Light Industry, China
https://orcid.org/0000-0002-3178-0607
Tengteng Xiao, Hundsun Technologies Inc., China
Yuan Feng, School of Computer and Communication Engineering, Zhengzhou University of Light Industry, China*

## ABSTRACT

Currently, with the popularity of the internet, people are surrounded by a large number of unhealthy pages which have a serious impact on the physical and mental health of visitors. To protect the legitimate rights and interests of internet users from infringement and maintain the harmonious and stable development of society, a new unhealthy webpage discovery system is needed. First, this paper proposed the knowledge of unhealthy webpages and web crawlers, and then the whole system's plan and design were introduced. The test results show that the unhealthy webpage discovery system can meet the needs of users. This experiment uses a CNN algorithm to classify the text and completes the collection and classification of unhealthy information through URL acquisition and URL filtering. The experimental results show that the unhealthy webpage discovery system based on a convolutional neural network can greatly improve the accuracy of unhealthy webpage discovery and reduce the omission rate, which can meet the needs of users for unhealthy webpage discovery.

## KEYWORDS

Artificial Intelligence, Deep Learning, Information Filtering, Machine Learning, Network Security, Neural Network, Text Classification, Web Crawler

## INTRODUCTION

With the rapid development of information technology, the Internet has long penetrated all aspects of people's lives. According to Hootsuite data, as of January 27, 2022, the number of social media users is 4.62 billion, an increase of 424 million (10%) over the same period last year (Social, 2022). The development of the Internet has brought people convenience, but also led to some problems. The emergence of increasingly unhealthy webpages harms social stability and people's physical and mental health. The discovery of unhealthy webpages is an important way to solve the stable development

DOI: 10.4018/IJDCF.315614 *Corresponding Author

of society, and this has been the focus of a growing research stream. Oram et al. (2021) developed a phishing web detection model based on a light gradient booster that uses phishing websites to monitor and filter phishing webpages by mimicking URLs. Namara et al. (2018) investigated how to adapt Facebook's privacy features to users' personal preferences based on user-defined privacy. Ertam (2018) used web scraping to collect classified news headlines and summaries from a news agency website and classified the test data using vector learning and deep learning methods. Yizhi Liu et al. (2021) developed an efficient mobile malicious webpage detection framework based on deep learning and edge clouds, which applies the ideas of edge computing and multi-device load optimization to MMWD, which can optimally deploy multiple device resources and detect mobile malicious web pages more effectively. Sajedi (2019) used an integrated algorithm to assign weights to weak classifiers. Then, he used a genetic algorithm to select the optimal weak classifier member and set optimal settings for optimal integration. Patil and Patil (2019) proposed using feature selection methods and machine learning to detect malicious web pages to detect unhealthy webpages.

Although the current research has achieved numerous results, the existing results are mainly the application of traditional machine learning or pattern matching methods, and there are problems such as a high false positive rate, high false negative rate, and high labor cost. Therefore, this paper constructs an unhealthy web page discovery system based on a convolutional neural network. A web crawler crawls the key text information in unhealthy web pages, and the key information crawled is filtered and screened by a convolutional neural network, which greatly improves the recognition rate of unhealthy web pages, helps to purify cyberspace, and improves the network environment.

In addition, we use some abbreviations, such as DQN (Deep Q-Network), PPO (Proximal Policy Optimization), URL (Uniform Resource Locator), CNN (Convolutional Neural Network), SMTP (Simple Mail Transfer Protocol), MD5 (Message-Dest algorithm), and SVM (Support Vector Machines).

## BACKGROUND

### Machine Learning

Machine learning is a multidisciplinary interdisciplinary discipline involving probability theory, statistics, computer science, algorithmic complexity theory, and other disciplines. Machine learning is a broad interpretation of an algorithm that mimics the various abilities of humans (Graham et al., 2022). Machine learning algorithms come in many forms of classification, such as learning strategies, knowledge acquisition, application areas, comprehensive classification, and learning classification. There are different machine learning algorithms according to different classification forms. The simplest and most commonly used classification is according to the learning form, which can divide machine learning algorithms into three categories: supervised learning, unsupervised learning, and reinforcement learning.

- **Supervised learning:** As the name implies, there is a supervised sample as a supervisor, the classifier is adjusted to the weight range of the sample; the sample is used as a comprehensive template for learning and training, and the data information waiting for training from the environment is obtained for training modification. A process in which the knowledge base finally achieves the optimal feature weight. This kind of learning effectively avoids the errors caused by itself with machine learning and is mainly used in classification and prediction. Regression analysis and statistical classification are commonly used in supervised learning algorithms.
- **Unsupervised learning:** Unsupervised learning means there is no reference sample template, and machine learning needs to be used to summarize the scattered features to their feature weights. The method of outward diffusion collects data information for analyzing feature weight values. It is a kind of blind learning that mainly uses K-means clustering to establish a data center. The error is reduced by looping and decrementing operations to achieve the classification purpose.

- **Reinforcement learning:** Reinforcement learning is a reinforcement signal dominated by the environment to control the ability of reinforcement learning. The goal is to achieve agents interacting with the environment to maximize the returns algorithm. Reinforcement learning problems do not have a clear "correct action" monitoring signal, which differs from supervised and unsupervised learning. They mainly manifested reinforcement learning in the reinforcement signal. Therefore, the algorithm cannot optimize the network by calculating the error between the action and the "correct action." Since the algorithm requires real-time interaction with the environment, it leads to the acquisition of the lagging reward signal of environmental feedback. Common reinforcement learning algorithms include DQN and PPO. The Markov decision process is a common model.

## Unhealthy Webpage

Unhealthy webpages that profit by illegal means or violate public order. Good customs do not conform to social values and are not conducive to the healthy development of the Internet.

In the early days of the Internet, webpages were mainly knowledge-based information. However, because of the Internet's continuous iteration and rapid development, the network has more characteristics, such as entertainment and functionality, which brings more unhealthy webpages. In recent years, unhealthy webpages have changed from knowledge-based to profit-making, and many unhealthy webpages violate the law and morality.

- **Violation of the law:** Such unhealthy webpages involve a wide range of types, including obscenity, pornography, violence; gambling, criminal skills instigation; drugs, illegal drugs, controlled knives, firearms, fake documents, trading information; fake stocks, credit cards, lottery fraud, among which obscene, pornographic, and vulgar webpages are the most prominent. Obscene and pornography are currently the largest types of illegal webpages. Most of them use vulgar titles to attract clicks, including articles, videos, animations, and pictures of human body parts (Wang et al., 2017).
- **Violation of morality:** Unhealthy webpages that violate morality refer to all kinds of webpages that violate social order and good customs and do not conform to the mainstream values of society, including video, audio, images, and text.
- **Destruction of information security:** The destruction of information security webpages refers to high-risk information containing viruses, Trojan horses, and backdoors, which pose a security threat to the computers and data of visitors. Because of the networkization of viruses, web browsing has become the most important channel for virus transmission, and web page hacking accounts for over 90% of the total virus transmission. Moreover, because of the frequent occurrence of application software vulnerabilities and browser plug-in vulnerabilities, it is difficult to manage such high-risk information only by relying on the security awareness of netizens.

## Web Crawler

### *Definition and Classification of Web Crawlers*

A web crawler, also known as a web spider, captures data on the network according to certain rules. We usually divide them into the following four categories: general web crawlers, focused web crawlers, incremental web crawlers, and deep web crawlers.

- **General Network Crawler:** The General Network Crawler, similar to its name, is suitable for the whole network and uses web information crawling, web information analysis, URL filtering, and initial URL collection. Because of its wide crawling range, it is often used for businesses to provide data for portal sites and large Web service providers. A parallel working mode is

generally adopted to solve the problem of slow crawling speed. Because the application value of this method is relatively large, efficiency can be improved using depth priority and breadth priority strategies.

- **Focused web crawler:** The focused web crawler, as its name implies, is a highly targeted way of crawling, only crawling content related to the set theme. Due to the strong pertinence, it saves the cost of storing webpage information and improves efficiency.
- **Incremental web crawler:** This crawler algorithm is difficult to implement. Since the working method is to crawl new or changed pages, time and space consumption is relatively small. It includes an updating module, local URL set, and crawl URL set.
- **Deep web crawler:** The main content of a deep web crawler is deep web content, such as those that can only be obtained after logging into the website. Because of the huge amount of data and information stored in deep webpages, there are two data structures in the web scraping mode for storing different web page contents.

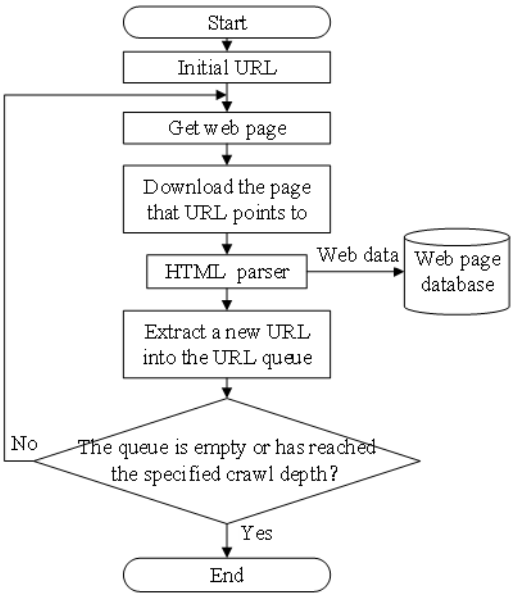
### *Web Crawler Workflow*

The first step of the basic work of a web crawler is to put the website address to be crawled into the queue to be crawled; the second step is to take out the URL to be crawled, parse the webpage and download the corresponding webpage content, and then store it in a database middle. Mark the addresses that have been crawled; the third step is to analyze whether the information has been updated on the crawled web pages. If it is updated, it needs to be reupdated for crawling; the fourth step is to see if the required conditions are met, and if not, return to step 2 and continue until the condition is met and ends. We show the workflow in Figure 1.

## Neural Networks

A convolutional neural network is a feed-forward neural network with deep structures, including convolution calculations, and represents deep learning algorithms (Goodfellow et al., 2016). Convolutional neural networks can represent learning and perform translation-invariance classification

**Figure 1. Web Crawler Workflow Diagram**



of input information according to their hierarchical structure and are also known as "translation-invariance artificial neural networks." Convolutional neural networks currently have many applications in medicine, agriculture, and the military. For example, Camara et al. (2022) invented a convolutional neural network to detect renal abdominal aortic aneurysms. In addition, Park and Park (2020) used convolutional neural networks to study the effect of class purity of training patches on crop classification performance.

### Text Classification

Text classification techniques refer to the supervised learning process of specific texts based on fixed rules. Text classification can be explained as passing unknown text categories through some rule criteria and finally attributing them to a specific category. Assuming that there is an objective function, the classification method is trained through a large amount of corpus to obtain the trained model, and then, according to the learned classifier, each feature item set is classified into the initially defined number of categories. Text classification refers to specifying the final classification label of the text in pre-classification and then associating the specified text with the classification number based on the text content. With the rise of big data and artificial intelligence, the combination of text classification and deep learning technology is increasing, and many scholars have also conducted extensive research in this field. Jindal et al. (2019) proposed a method for training deep networks using robustness to mark noise, introduced a nonlinear processing layer, and modeled the statistical data of label noise as a CNN structure to prevent the model from over-fitting wrong labels. Parvathi and Jyothis (2018) adopted the method of convolutional neural networks and used deep learning to make accurate categories predictions.

## DESIGN OF AN UNHEALTHY WEBPAGE DISCOVERY SYSTEM BASED ON A CONVOLUTIONAL NEURAL NETWORK

### Overall Design

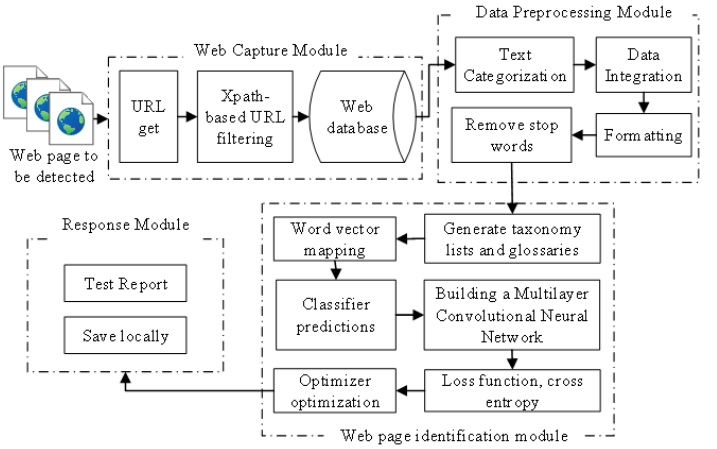
According to the unhealthy webpage discovery system's design goals, research on the unhealthy webpage discovery system, and the propagation speed of unhealthy information, this paper designs an unhealthy webpage discovery system model based on a convolutional neural network. First, enter the URL to be detected, crawl the data through the web crawler, filter the crawled data, upload the data to the webpage database, filter out the unhealthy webpages through the convolutional neural network, and finally report the identified unhealthy webpages and analyze the results. Finally, we show the system architecture in Figure 2.

### Functional Design

To meet the needs of unhealthy webpage discovery, the unhealthy webpage discovery system based on a convolutional neural network mainly includes functions such as webpage collection, unhealthy webpage preprocessing, unhealthy webpage identification, unhealthy webpage response, and user management.

- **Webpage collection:** The collection of unhealthy information crawls the information of each webpage on the Internet through a web crawler, and the content of the crawled webpage is stored locally. This step includes the deduplication of web pages.
- **Unhealthy webpage identification:** This module mainly includes information processing and unhealthy webpage extraction. It uses webpage denoising, stopping word removal, text classification, topic extraction, and unhealthy information judgment.
- **Unhealthy webpage response:** Save the detected unhealthy webpage classification as a report or send the detected unhealthy webpage information to the user's mailbox. According to the

**Figure 2. System Architecture Composition Diagram**




unhealthy webpage collection and analysis module, the unhealthy webpage category is judged. Then, the user can save the data information obtained by this detection to the local in txt format for the user to analyze and view in the future or send the data information obtained by this scan to the mailbox. The SMTP mail-sending protocol is used when sending mail.

- **User management:** It includes the login of the user registered user, the user registration uses the mailbox to register to obtain the authorization code, and the user login can use the username set during registration to log in to the system. The email address is used to obtain the user's real-name authentication authorization code when registering. When registering, the user needs to set the username, enter the email address to obtain the authorization code, set the login password, and finally register. MD5 is used to encrypt the password when registering and logging in.

## Design of Unhealthy Webpage Information Collection Based on XPath Filtering

The basis of collecting and analyzing unhealthy information is the acquisition of information. This paper mainly conducts information collection and analysis on several websites prone to unhealthy information. This research's first point is how to obtain unhealthy information on the network. Web crawlers (web spiders) are used to crawl unhealthy information on Internet pages according to the user's wishes or specified rules and save the captured information for further analysis and processing. The collection process of unhealthy webpage information is: collection data configuration, URL acquisition, URL filtering, and data preservation.

- **Collection data configuration:** These are the initial collection settings for website URLs, the content settings to be collected, and the settings for the termination conditions of information collection. Information is used to maintain and manage collection sites.
- **URL acquisition:** Analyze the URL link of the website that the main user wants to monitor, observe whether the URL of the website changes regularly, construct the URL list according to the law obtained by observation and analysis, and then crawl the URL list data on this basis.
- **URL filtering:** To avoid revisiting webpages that have already crawled unhealthy webpage information when crawling unhealthy webpage information, it is necessary to mention the URL deduplication function at this time. We divide the URL deduplication steps based on a database query, hash table lookup, and MD5 compression mapping. Then, the system filters out the URLs that need to be collected from unhealthy webpages through the XPath expression filtering rules

set by the configuration module. By writing the filtering XPath expression, it is possible not only to collect unhealthy webpage information for the webpages of some related websites but also to remove the pages of irrelevant websites. This step can avoid the redundancy of grabbing data and save the system's overhead.

- **Data preservation:** Data preservation temporarily stores the data information collected from the webpage in txt format. The data information collected by the unhealthy webpage information collection module is all unstructured webpage data, and the unhealthy webpage information analysis module can read, clean, filter, and structure the information.

## Design of Unhealthy Webpage Recognition Based on a Convolutional Neural Network

### *Unhealthy Webpage Information Analysis Module System Design*

The unhealthy webpage information analysis module is mainly composed of two functions: web page information preprocessing and unhealthy webpage information processing.

- **Webpage information preprocessing:** The data captured from the webpage cannot be used directly as unhealthy webpage analysis data and needs to be preprocessed by text. It classifies the text, integrates the data, adjusts the format, and removes stop words for the data crawled by the crawler and the data collected on the Internet to prepare for the classification of unhealthy webpages using the convolutional neural network.
- **Unhealthy webpage identification:** The processing of unhealthy webpage information is the core of the unhealthy webpage discovery system based on a convolutional neural network. The original unstructured webpage information has been converted into a text eigenvector through the webpage information preprocessing module, and unhealthy topics are generated by clustering the eigenvectors of the text. The text classification function in the webpage information preprocessing module is based on the simplified implementation of TensorFlow on the data set. It combined with the convolutional neural network CNN to classify the text captured from the Internet. This CNN structure can prevent over-fitting errors in the model. We trained the classifier on four categories of unhealthy webpage information: violence and terrorism, pornographic text, politically sensitive, and normal.

## Convolutional Neural Network Construction and Parameter Design

The construction of a convolutional neural network requires a specific framework. Now, the commonly used mainstream deep learning frameworks are PyTorch and TensorFlow. PyTorch is a deep learning framework that has emerged in recent years. Its chief advantages are that it is easy to understand, and the readability of the code is increased, but the framework is immature and has only been used in recent years. TensorFlow is a deep learning framework developed earlier. Although the network construction process is cumbersome, it has stable performance and powerful functions. It has a place in deep learning through continuous version updates.

The classification of text data information is based on the simplified implementation of TensorFlow on the data set. It combined with the convolutional neural network CNN algorithm to perform multiple training tests on the captured text to obtain the correct classification type. The neural network layer is implemented using the underlying interface functions in constant mode, and the model is built directly in constant mode, such as a full connection layer, activation function layer, pooling layer, and convolution layer.

The construction of the neural network model of the research object in this paper uses a three-layer neural network, which includes a convolutional layer and two hidden layers. A pooling layer follows convolution. The first hidden layer is a fully connected layer containing 128 neurons using

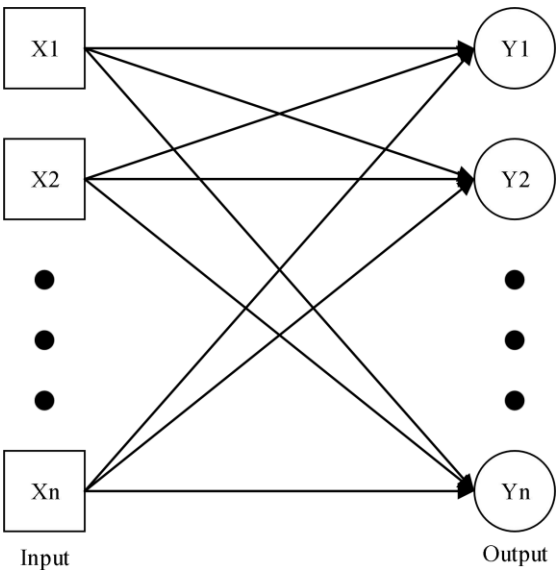
the ReLU activation function. The second hidden layer contains four neurons using a fully connected layer with a softmax activation function.

- **Convolutional layer:** The convolutional layer is a compact unit that makes up a convolutional neural network. The convolutional unit constitutes the smallest unit of the convolutional layer. The convolutional unit is obtained through the backpropagation algorithm. A convolution operation is generally used to extract distinct features of the transmitted data information. The smaller the number of convolution layers, the more blurred the extracted features are. We iteratively generated larger layers from edge features obtained from lower layers.
- **Fully connected layer:** We fully connected Each node of the fully connected layer to each node of the previous layer to extract the previous features. Because of the need to connect with each node in the upper layer, the parameters of the fully connected layer are maximum, and of course, the memory consumption is also large. Let $x$ denote the input of the fully connected layer, and $y$ denotes the output of the fully connected layer; then, the relationship between $x$ and $y$ is $y=w*x+b$, where $w$ and $b$ respectively represent the weight of the fully connected layer and the bias of it. The values of $w$ and $b$ control the optimization of the model. We show the fully connected model diagram in Figure 3.
- **Softmax function:** Also known as the normalized exponential function. Neural networks with classification problems generally add a softmax function to the output layer so that the output signal range is between *(0,1)* and the sum of the output signals is *1*, as shown in formula (1):

$$y_i = \frac{e^{a_i}}{\sum_{k=1}^{C} e^{a_k}} \tag{1}$$

- **ReLU activation function:** Additionally, known as the modified linear unit, it is like the display of the positive semiaxis part of the x-axis by the one-dimensional linear proportional function, where x comes from the input of the previous layer. Because it can avoid the problems of gradient

**Figure 3. Fully Connected Layer Model**



explosion and gradient disappearance and simplify the calculation process, it is widely used instead of other activation functions, as shown in formula (2):

$$y = \max(0, x) \tag{2}$$

### Response Module

The unhealthy information response module mainly includes two parts: save the detected unhealthy webpage classification in the form of a report locally or send the detected unhealthy webpage information to the user's mailbox. The naming method of the detection content saved to the local is in the form of "unhealthy webpage detection + time", and the content is the web page address where the unhealthy webpage information is detected and the type of unhealthy webpage information. The content of the sent mail is the report saved to the local, and this part is sent using the SMTP protocol.

## KEY TECHNOLOGIES FOR REALIZING UNHEALTHY WEBPAGE DISCOVERY SYSTEMS BASED ON CONVOLUTIONAL NEURAL NETWORKS

### Construction of Convolutional Neural Network Model

The analysis of unhealthy webpage information mainly uses convolutional neural network technology, constructing neural networks, and training neural network models.

First, we build the neural network model to construct a multilayer RNN network, which includes two hidden layers and an output layer. The first hidden layer contains 128 neurons using the ReLU activation function in the fully connected layer, and the second hidden layer contains four neurons using the softmax activation function in the fully connected layer. Then, the neural network model is trained for classifier training and model verification.

The convolutional neural network is used to create the classifier training set. First, the classification is created, and then the information captured from the website webpage is stored for later integration processing because the text format captured on the webpage and the text needs to be processed before being used by the neural network. Then, the word vector mapping is set to the maximum pooling and collapse dimensions to prevent the over-fitting phenomenon. Next, the RNN neural network is created using the ReLu and softmax activation functions to prevent gradient explosion and gradient disappearance. Later, the use of cross-entropy prediction may be wrong. This time, we must calculate the predicted classification and correct classification errors. Finally, the error calculation accuracy of the previous step is optimized.

We built the neural network model using a three-layer neural network, which includes a convolution layer and two hidden layers. The pooling layer followed convolution. The first hidden layer is a fully connected layer containing 128 neurons using the ReLU activation function. The second hidden layer contains four neurons using the full connection layer with the softmax activation function.

### Realization of Unhealthy Webpage Discovery System Recognition Based on a Convolutional Neural Network

The main task of neural network model training is unhealthy webpage recognition discovery, including generating a classification list, generating vocabulary, word vector mapping, constructing a multilayer convolutional neural network, classifier prediction, a loss function, using an optimizer for optimization, model accuracy calculation, and other steps. The classifier uses four categories, i.e., violence, pornography, political sensitivity, and normality. Each category has up to 6500 data points. The data set is divided into three types: verification set, training set, and test set.

The steps in the neural network model classifier training process include:

**Step 1:** Generate a classification list: political sensitivity, violence, illegal, normal, pornographic text.
**Step 2:** Generate vocabulary. If not, the build _ vocab () function is used to construct a vocabulary list, where PAD is used to complete the length of vocabulary.
**Step 3:** Word vector mapping, embedding shape is (5000, 64), (*5000* is the size of the vocabulary, 64 is the dimension of the word vector), set the maximum pooling layer and "collapse dimension" reduction _ indices=[1] to prevent over-fitting.
**Step 4:** Construct a multilayer convolutional neural network, which includes two hidden layers and an output layer. The first hidden layer adopted 128 neurons activated by ReLu in the fully connected layer, and the second hidden layer contained four neurons activated by softmax in the fully connected layer.
**Step 5:** The classifier predicts and determines which of the four categories the text information belongs to.
**Step 6:** Loss function, cross-entropy. The main purpose is to reduce the error that occurs when calculating the predicted classification and the correct classification.
**Step 7:** Use the optimizer for optimization, mainly to update the weights.
**Step 8:** Calculate the accuracy of the model.

Convolutional neural network training configuration, including elements such as the word vector dimension, sequence length, and the number of convolution kernels, are shown in Table 1.

## SYSTEM TESTING AND ANALYSIS

### Test Environment

The test environment is the basis of system testing. It is necessary to ensure that the environment is configured correctly to test a system. Therefore, the correct test environment is a critical step in testing. We show the environment configuration required for this system test in Table 2.

It can be seen that the CNN model proposed in this paper has low hardware requirements, and can run smoothly under low computer configuration. This is due to the neuron structure of the convolutional neural network, which reduces computational costs and the resources consumed. Furthermore, with this configuration, the webpage is crawled and detected, and many webpages can be detected and filtered quickly.

### System Performance Test

The performance of a system determines whether the system needs to be optimized again. Therefore, the number of sensitive words is a key factor in determining the accuracy and omission rate. The

**Table 1. Convolutional Neural Network Configuration**

| | |
|---|---|
| Word vector dimension | 64 |
| Sequence length | 600 |
| Number of categories | 4 |
| The number of convolution kernels | 256 |
| convolution kernel size | 5 |
| Glossary size | 5000 |
| Fully Connected Layer Neurons | 128 |

**Table 2. Test Environment**

| Configuration item | Specific configuration |
|---|---|
| Operating system | Windows10 |
| Memory capacity | 4 GB |
| Hard disk capacity | 1 TB |
| CPU | Core i7 980X |
| Network bandwidth | 200 M |

following describes the effect of the number of sensitive words on the accuracy and omission rate of the classifier training and validation sets.

The test shows that when the number of sensitive words is *100*, the accuracy of the training set is *38%*, and the omission rate is *39%*. When the number of sensitive words exceeds *900*, the accuracy of the training set is close to *95%*, and the omission rate is close to *18%*. Therefore, in using this system, at least *900* sensitive words should be set.

The test shows that when the number of sensitive words is *100*, the accuracy of the validation set is *21%*, and the omission rate is *43%*. When the number of sensitive words exceeds *900,* the accuracy of the training set is close to *96%*, and the omission rate is close to *18%*. Therefore, in the use of this system, at least *900* sensitive words should be set.

## Comparison With Other Models

This paper compares the effectiveness of the CNN model with the model of unhealthy webpage discovery created by different models. We show the specific results in Table 3.

As seen from Table 3, the CNN model in this paper is superior to the KNN model and the Bayesian decision theory model. The accuracy of the CNN model was *4%* higher than the CUB-SVM model; *9%* higher than the MINSVM model; *15%* higher than the SMOTE-SVM model; *2.8%* higher than the Bayesian Decision Theory; and *14%* higher than the improved KNN model.

**Figure 4. The Relationship Between the Number of Sensitive Words and the Accuracy and Omission Rate of the Training Set**

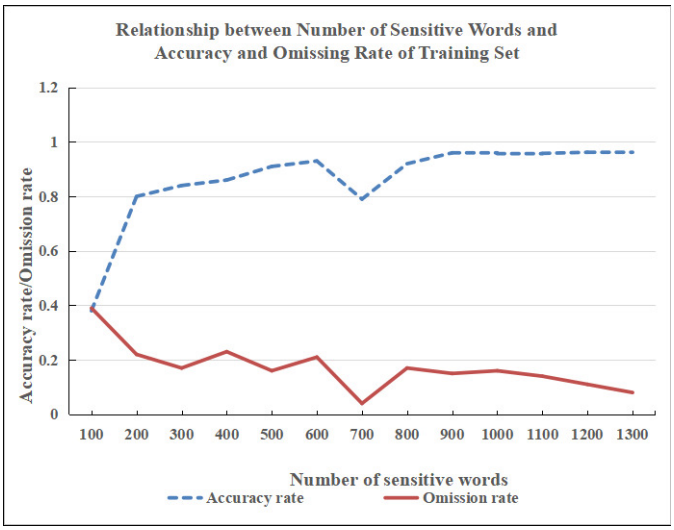


**Figure 5. The Relationship Between the Number of Sensitive Words and the Accuracy and Omission Rate of the Validation Set**

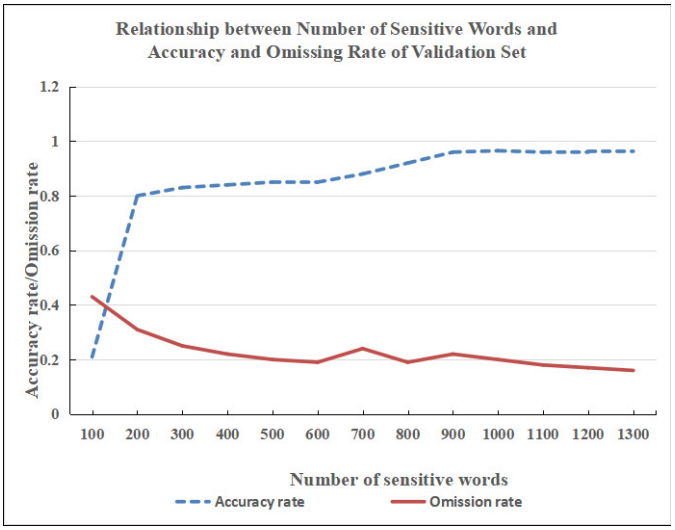



**Table 3. Comparison With Other Models**

| Paper | Model | Accuracy |
|---|---|---|
| (Ruxianguli et al., 2019) | CUB-SVM | 0.92 |
| (Ajeebi et al., 2013) | MINSVM | 0.87 |
| (Chawla et al., 2002) | SMOTE-SVM | 0.81 |
| (Sun et al., 2012) | Bayesian Decision Theory | 0.93 |
| (XU, 2013) | Improved KNN | 0.82 |
| Ours | CNN | 0.96 |

Such good accuracy is because the convolutional neural network adds a local connection mode compared to the traditional SVM and Bayesian Decision Theory, ensuring that the learned convolution kernel has the strongest response to the input spatial local pattern. The neural units of different layers are locally connected, so we only connected the neural units of each layer with some neural units of the previous layer. Each neural unit only responds to the region within the receptive field and does not care about the region outside the receptive field. This approach can effectively improve the accuracy of prediction and reduce resource utilization. Compared with the KNN algorithm, the algorithm proposed in this paper avoids the low prediction accuracy of rare categories with unbalanced samples through the initial transmission of neurons, thus improving the prediction accuracy of the complete model.

## CONCLUSION

This paper introduces a convolutional neural network-based detection system for unhealthy webpages. First, the web crawler is used to crawl the unhealthy webpage information according to the user's wishes, or the specified rules, and the crawled information is saved. Then, the convolutional neural network is used to create a classifier training set, integrate the captured information, and continuously train it to identify unhealthy webpages with greater accuracy. In addition, the simplified implementation

of TensorFlow on the data set combined with the convolutional neural network CNN to classify the text grabbed from the network makes the information classification more accurate. In the era of the rapid development of the Internet, the technology introduced in this article can meet all the requirements for discovering unhealthy webpages and will play a greater role in the future.

## REFERENCES

Ajeeb, N., Nayal, A., & Awad, M. (2013). Minority SVM for linearly separable imbalanced datasets. *Proceedings of the International Joint Conference on Neural Networks* (pp. 1-5). Institute of Electrical and Electronics Engineers. doi:10.1109/IJCNN.2013.6707030

Camara, J. R., Tomihama, R. T., Pop, A., Shedd, M. P., Dobrowski, B. S., Knox, C. J., Abou-Zamzam, A. M. Jr, & Kiang, S. C. (2022). Development of a convolutional neural network to detect abdominal aortic aneurysms. *Journal of Vascular Surgery Cases and Innovative Techniques*, *8*(2), 305–311. doi:10.1016/j.jvscit.2022.04.003 PMID:35692515

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. doi:10.1613/jair.953

Ertam, F. (2018). Deep learning based text classification with Web Scraping methods. *International Conference on Artificial Intelligence and Data Processing (IDAP)*, 1–4. doi:10.1109/IDAP.2018.8620790

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Graham, C. A., Shamkhalichenar, H., Browning, V. E., Byrd, V. J., Liu, Y., Gutierrez-Wing, M. T., Novelo, N., Choi, J.-W., & Tiersch, T. R. (2022). A practical evaluation of machine learning for classification of ultrasound images of ovarian development in channel catfish (*Ictalurus punctatus*). *Aquaculture (Amsterdam, Netherlands)*, *552*, 738039. doi:10.1016/j.aquaculture.2022.738039 PMID:35296028

Jindal, I., Pressel, D., Lester, B., & Nokleby, M. (2019). *An effective label noise model for DNN text classification*. 10.48550/arXiv.1903.07507

Liu, Y., Zhu, C., Wu, Y., Xu, H., & Song, J. (2021). MMWD: An efficient mobile malicious webpage detection framework based on deep learning and edge cloud. *Concurrency and Computation*, *33*(18), e6191. doi:10.1002/cpe.6191

Namara, M., Sloan, H., Jaiswal, P., & Knijnenburg, B. P. (2018). The potential for user-tailored privacy on Facebook. *Symposium on Privacy-Aware Computing (PAC)*, 31–42. IEEE. doi:10.1109/PAC.2018.00010

Oram, E., Dash, P. B., Naik, B., Nayak, J., Vimal, S., & Nataraj, S. K. (2021). Light gradient boosting machine-based phishing webpage detection model using phisher website features of mimic URLs. *Pattern Recognition Letters*, *152*(C), 100–106. doi:10.1016/j.patrec.2021.09.018

Park, S., & Park, N.-W. (2020). Effects of class purity of training patch on classification performance of crop classification with convolutional neural network. *Applied Sciences (Basel, Switzerland)*, *10*(11), 3773. doi:10.3390/app10113773

Parvathi, P., & Jyothis, T. S. (2018). Identifying relevant text from text document using deep learning. *International Conference on Circuits and Systems in Digital Enterprise Technology (ICCSDET)*, 1–4. doi:10.1109/ICCSDET.2018.8821192

Patil, D. R., & Patil, J. B. (2019). Malicious web pages detection using feature selection techniques and machine learning. *International Journal of High Performance Computing and Networking*, *14*(4), 473–488. doi:10.1504/IJHPCN.2019.102355

Ruxianguli, A., Yasen, A., & Wenqiang, G. (2019). Reactionary text filtering method based on n-gram and class-unbalanced SVM for Uyghur webpages. *Jisuanji Yingyong Yanjiu*, *36*(11), 3410–3414. doi:10.19734/j.issn.1001-3695.2018.07.0410

Sajedi, H. (2019). An ensemble algorithm for discovery of malicious web pages. *International Journal of Information and Computer Security*, *11*(3), 203–213. doi:10.1504/IJICS.2019.099408

Social, W. A. (2022). Hootsuite: Digital 2022 Global Overview Report (2022). [Data set]. https://datareportal.com/reports/digital-2022-global-overview-report

Sun, Y., & Zhou, G. X. (2012). Research on webpage filtering based on rough set and Bayesian decision theory. *Journal of Chinese Information Processing*, *26*, 67–72.

Wang, R., Zhu, Y., Tan, J., & Zhou, B. (2017). Detection of malicious web pages based on hybrid analysis. *Journal of Information Security and Applications*, *35*, 68–74. doi:10.1016/j.jisa.2017.05.008

Xu, Y.-B., Li, Z., & Chen, J. (2013). Parallel recognition of illegal Web pages based on improved KNN classification algorithm. *Jisuanji Yingyong*, *33*(12), 3368–3371. doi:10.3724/SP.J.1087.2013.03368

*Zengyu Cai received his master's degree in computer application technology from Northeast Normal University, Changchun, China, in 2006. He is an associate professor at Zhengzhou University of Light Industry. His research interests include trusted computing, plan recognition and information security.*

*Chunchen Tan graduated from Zhengzhou University of Light Industry in 2021 and received a bachelor ' s degree. At present, he is studying for Master of Electronics and Information Science in Zhengzhou University of Light Industry. The main research directions are network security and artificial intelligence.*

*Jianwei Zhang received his Ph.D. degree in computer application technology from PLA Information Engineering University in 2010. He is a professor at Zhengzhou University of Light Industry. His research interests include broadband information network and network security.*

*Tengteng Xiao, a bachelor ' s degree from Zhengzhou University of Light Industry in 2020, is currently a network engineer of Hundsun Technologies Inc. The main research direction is network security in the new generation of networks.*

*Yuan Feng received her master's degree in computer application technology from Chongqing University of Posts and Telecommunications in 2006. She is an associate professor at Zhengzhou University of Light Industry. Her research interests include information security and artificial intelligence.*