symmetry

MDPI

Article

An Intelligent Based Symmetrical Classification of Online Shop Selling Counterfeit Products

Shyh-Wei Chen [1], Po-Hsiang Chen [1], Ching-Tsorng Tsai [1] and Chia-Hui Liu [2,*]

[1] Department of Computer Science, Tunghai University, Taichung 407224, Taiwan
[2] Department of Applied Mathematics, Chinese Culture University, Taipei 11114, Taiwan
* Correspondence: ljh34@ulive.pccu.edu.tw

Abstract: In recent years, the social network has become popular and people have started trading transactions on the Internet. Many counterfeit websites have begun to appear which create websites with counterfeit products or use the digital advertiser's services to promote their websites on social media. Malicious sellers disguise high-quality products to attract consumers since buyers cannot receive transparent information. If there is asymmetry information, a secondary market will be formed. To solve the above problems, this research explored the machine-learning-based method to classify counterfeit and legitimate websites with symmetry information. The data set is 1612 websites used in this paper and a total of 15 feature values and takes 804 counterfeit websites and 808 legitimate websites. The Random Forest and Deep Neural Network algorithms were used to classify fake websites. This study also used statistical tests, such as Chi-square and ANOVA detection, to compare the importance of features in feature selection. The experiment results show that the RF accuracy is 99.2% and the DNN accuracy is 93.2%. The RF Precision and Recall are 100% and 98.5%, respectively. The DNN Precision and Recall are less than RF. Then, the RF F1-score is 99.2% which is higher than DNN.

Keywords: machine learning; counterfeit website; random forest; deep neural networks; asymmetry information

Citation: Chen, S.-W.; Chen, P.-H.; Tsai, C.-T.; Liu, C.-H. An Intelligent Based Symmetrical Classification of Online Shop Selling Counterfeit Products. *Symmetry* **2022**, *14*, 2132. https://doi.org/10.3390/sym14102132

Academic Editors: Charles Tijus, Jih-Fu Tu and Teen-Hang Meen

Received: 30 August 2022
Accepted: 4 October 2022
Published: 13 October 2022

1. Introduction

One of the most worrisome issues on the Internet is fraud and digital advertising has suffered such a worrisome threat for a long time [1,2]. Counterfeit products are generally considered high-imitation products, which are almost the same as real products in appearance. They are fakes or unauthorized replicas of the real product. It is commonly used in the sale of clothing, accessories, shoes, and other products. In the process of e-commerce transactions, consumers will generally judge whether the product is sold on an official website and whether the product description or information is detailed, etc. However, these malicious people will not only imitate fake websites, but also provide very detailed product information and exquisite photos, so that consumers will buy without knowing it, which will lead to financial losses, and even affect the balance of the entire online e-commerce [3]. In addition, with the popularity of social networking sites, such as Facebook, Instagram, and Line, many people will shop through social networking sites in order to pursue new products or even popular products. The counterfeit network has also begun to defraud consumers through social digital media. Even digital advertisers have suffered. Malicious sellers disguise high-quality products to attract consumers because buyers do not understand the products sold by sellers [4]. Therefore, if there is information asymmetry, a secondary market will be formed. It is through information asymmetry that counterfeit Internet fraudsters use digital advertisers to publish many fake websites on many social platforms to attract more uninformed consumers. It is often difficult for digital advertisers to identify these fake sites, which in turn affects the advertiser's credibility [5]. Fraudsters create websites with counterfeit goods and use fake audiences to

attract advertisers or use the digital advertiser's services to advertise their websites on the broad social network media. The digital advertising ecosystem is shown in Figure 1 [6]. These ads are published by advertising agent companies to assist websites to increase exposure rate which could bring more online customers. Though they are now being abused by fraudulent websites, causing the reputation of the advertising agency and the services they provide to suffer [7].

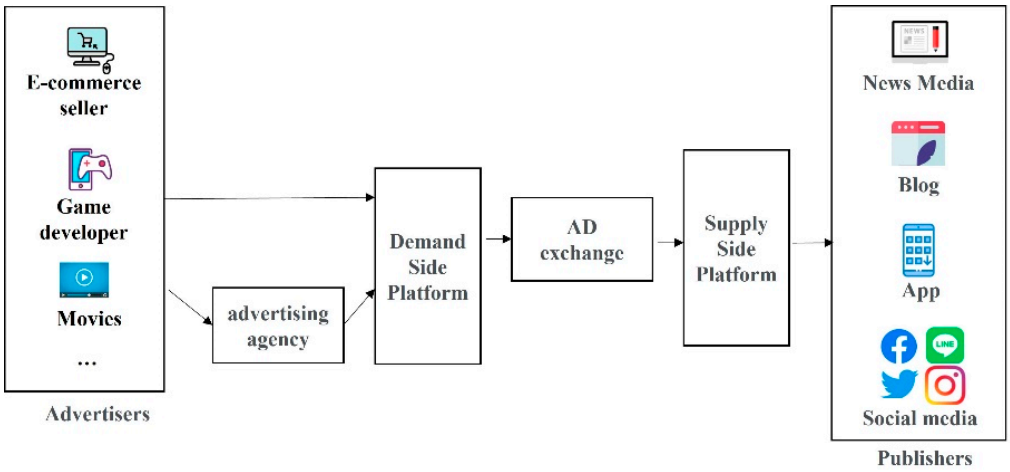

Figure 1. Digital Advertising Ecosystem.

This phenomenon is caused by the difficulty of preventing fraudulent online vendors. Most consumers cannot confirm whether an online shop is fraudulent or not. In addition, they also cannot state who exactly operates these websites and distinguish whether those popular brands is counterfeit or not. Consumers may receive counterfeit goods after purchasing from a fraudulent online shop. Sometimes, even the digital advertisement agent still cannot recognize those fraudulent online shops. Due to the fraudulent online shop growing more and more, it is necessary to classify whether the online shop is counterfeit or not. Traditionally, to strengthen information security measures, most companies take measures such as blacklisting, encrypting customer data, fraud monitoring, etc., to prevent the above problems. However, it is still unable to effectively identify fraudulent websites in real time. Therefore, this paper proposes a machine-learning method to classify whether the website is selling counterfeits or not. The contribution is successful and efficient in assisting advertising agents or general consumers to judge fake websites and reduce the risk of buying and selling.

This paper uses the collected feature data of legitimate websites and counterfeit websites. The collection time is from March to November 2021, and Random Forest [8] and Deep Neural Networks (DNN) [9] are used, respectively. There are 1612 websites used in this paper and a total of 15 feature values and take 804 fake websites and 808 legitimate websites as our feature extraction training set. In this paper, there are two steps of classifying whether the websites are selling counterfeit or not: (1) To find the features of fake websites, where the Chi-square test and ANOVA are used to filter useful features. (2) Using machine-learning models to distinguish counterfeit websites from legitimate websites, in which Random Forest [5] and Deep Neural Networks (DNN) are used as training models. As for the result, on a DNN model with three hidden layers and on a RF model using 100 decision trees, 93.2% and 99.2% accuracy are achieved, respectively. Therefore, we would use the RF model to classify whether the website is selling a fake product.

As for the rest of the paper, it is organized as follows: The background knowledge required and related works are discussed in Section 2. Next, we introduce the proposed method and the data collection and feature extraction in Section 3. Then Section 4 illustrates the experiment's results. Afterwards, the conclusion and the future work of this thesis are shown in Section 5.

2. Literature Review

2.1. Random Forest

Random Forest is a concept of an ensemble method which is the supervised learning algorithm of machine learning. It is suitable for classification and regression problems which solve and improve performance by combining multiple classifiers. Random Forest contains a lot of decision trees to improve the predictive accuracy of the given dataset by using the average. The decision tree is often used in classification or regression problems which have nonlinear data classifications. The analysis result is presented in a tree architecture. The decision tree starts from the root node of the tree and gradually expands according to the classification problem. The internal node of the tree represents the test problem. Each node represents an attribute, the branch represents the result obtained by the problem, and the leaf node represents the category of the classification. Random Forest is composed of decision trees. Figure 2 demonstrates the concept of the Random Forest algorithm.

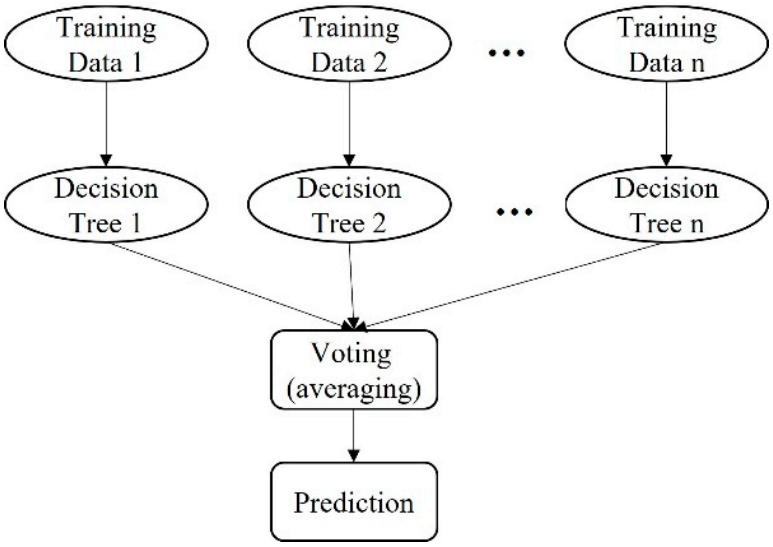

Figure 2. The concept of the RF.

To establish multiple decision trees with differences for ensemble learning, the data set must first be differentiated to generate multiple differentiated decisions tree since there is only one data set. The concept of the Ensemble Method is to combine multiple classifiers to solve a complex problem. The greater number of decision trees results in higher accuracy and prevents overfitting. One of the ensemble algorithms is Bootstrap aggregating (Bagging) which was proposed by Breiman [8] in 2001 and based on an algorithm proposed by Bootstrap. Random Forest is an extension of Bagging which uses different training data rather than one sample. It employs the bagging method to generate the required prediction. The concept of Bagging is shown in Figure 3.

It assumes that the number of training data is *N* and sampling for M datasets. Each dataset contains N′ data. It uses the sampling method to build a lot of datasets and to train several functions. The following procedure is dependent on the different problems. If the problem is a kind of regression problem, it takes the average of the functions as the result. Otherwise, if it is a kind of classifier problem, it uses voting to decide the result. During the training process, there are a series of questions in the decision tree, such as whether the weather is good or not, whether the lunch is delicious, etc. Starting from the root node, the data is divided into diverse parts along with the feature of the data. The principle of division is that the segmentation should be able to obtain the maximum Information Gain (IG, Equation (1)). That is, the amount of information obtained equals the original amount

of information minus the amount of information after division. Information Gain is defined as how much a feature of information can bring to the classification system.

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j) \quad (1)$$

f: the feature of the node which is used to partition
D_p: the data of the parent node
D_j: the data of the jth child node
N_p: the number of the parent node
N_j: the number of the jth child node
I: the impurity measure

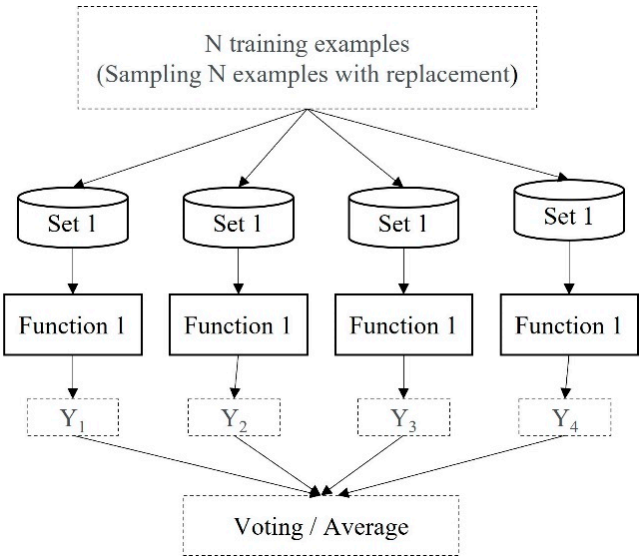

Figure 3. The concept of Bagging.

2.2. Deep Neural Networks (DNN)

The Deep Neural Network is a branch of Deep Learning which can be taken as a relatively deep neural network. The various types of layers are established, such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN). A neural-like network is an algorithmic model that mimics the animal nervous system. In a deep neural network, there are numerous layers: input layer, hidden layer, and output layer. Each layer will have many neurons. The neurons will add up the input of the previous ones. The transition into the activation function is categorized as the output of the neuron. Each neuron is connected to the neuron of the next layer, and each neuron is assigned a weight value so that the output value of the neuron of the previous layer is passed to the neuron of the next layer after a calculation with the weight. The activation function is usually chosen to be a nonlinear function. The common one is the sigmoid function or the hyperbolic tangent function, and the more popular Relu function in recent years [10]. The architecture of a DNN refers to the settings of the number of layers, the number of neurons in each layer, the connection method of neurons between layers, and the activation function. These parameter settings need to be set manually when building the model, and different parameter settings may greatly affect the performance of the neural network. The learning and training process of the neural network is to optimize each weight value. The neural network optimizes each weight value, mainly in the way of forward-propagation and back-propagation [11]. The former propagation calculates the connection weights corresponding to neurons for weighted sum operation. The equations for the weighted

sum and activation of the final layer are shown as Equations (2) and (3). Equations for the weighted sum and activation of the final layer.

$$Z^L = W^{L-1} * a^{L-1} + b^{L-1} \tag{2}$$

$$a^{(L)} = \sigma\left(Z^L\right) \tag{3}$$

L: the layer, W: weight, b: bias, a: activation function

2.3. The Related Works

In recent years, in addition to e-commerce websites, there are also many detective analyses of fake websites. Koepke [12] used several algorithms to build their models, such as Naïve Bayes, Logistic Regression, and Simple Bayesian classifier. They used the IP address and DNS records as the feature data, which are mainly used to detect counterfeit websites. In 2009, Abbasi [13] proposed a detection software which is AZProtect. They used Support Vector Machine as the main algorithm to extract features according to web page content to identify fraudulent websites. The comparison of these studies is shown in Table 1. Koepke [12] has a good idea to use multiple models for analysis, but it is difficult to improve the accuracy to the greatest extent. The reason is that the dataset is not large enough and there is no importance of detection features for model training. Abbasi [13] used SVM as the main model, and they collected a lot of data. However, they should pay attention to the balance of the data. They collected 100 legitimate websites and 350 fake websites, resulting in various indicators of legitimate websites. The unbalanced dataset may affect the judgment of the entire model.

Table 1. Confusion Matrix. The comparison of different models.

	Model	Data	Accuracy (%)	Precision (%)
Koepke [7]	Naïve Bayes	300	91.2	88
	Logistic Reg.		93.8	93.2
	Bayesian		91.6	88.6
Abbasi [8]	SVM	450	89.11	67.84

3. The Proposed Method and the Data Models

3.1. The Proposed Method

This paper adopts the Random Forest and DNN methods to classify whether the website is fraudulent or legal. Those methods are supervised learning algorithms which make predictions based on samples. For example, it can use a historical headcount to estimate the future headcount. In supervised learning, the input variable contains the labels of the training data and your desired output. Using an algorithm to analyze training data is a process of analyzing the input data, and finally calculating a function to obtain a prediction or classification result. Therefore, we configure a corresponding standard answer with the feature value obtained in the websites. The problem is to solve whether it is a fraudulent website, which is a dichotomy problem. Thus, it defines legitimate websites as 0 and fake sites as 1. There are four components in the proposed scheme, which are web application, database, message broker, and process pools. To identify a good model in machine learning, it is necessary to use several indicators to determine and filter the obtained results. The indicators have the following methods: Accuracy, Precision, Recall, and F1-Measure. To calculate several indicators, the formulas are listed in the Confusion Matrix (Table 2). This paper will use the above four metrics for the evaluation of our models. Table 3 shows the formulas of the four indicators.

Table 2. Confusion Matrix.

	Actual Values Positive (1)	**Actual Values Negative (0)**
Predicted Value Positive (1)	True Positives TP	True Negatives TN
Predicted Value Negative (0)	False Negatives FN	True Negatives TN

Table 3. The related measurement formulas from the confusion matrix.

The Four Formulas
Accuracy $= \frac{TP+TN}{TP+FP+TN+FN}$
Precision $= \frac{TP}{TP+FP}$
Recall $= \frac{TP}{TP+FN}$
F1 Measure $= \frac{2*Presision*Recall}{Precision+Recall}$

3.2. The Data Collection and Feature Extraction

In this study, the fraudulent websites are collected from three different resources: (1) the Google search engine: this is also used in the previous research [4,14], (2) Facebook: this is a reliable resource which lists the known fraudulent websites, (3) Whoscall application in Taiwan: this an online system that allows victim consumers provide fraudulent websites which the user needs to provide the related information. As for the legitimate websites, we first collect numerous products from Yahoo and MOMO which are online-shopping platforms in Taiwan. Then, we search and collect their online stores URL from the Google search engine according to the collected brands. After that, using the web crawl technique to the required HTML files and we can analyze the content and capture the HTML. Using various techniques such as searching, regular string processing, segmentation, and replacement we can filter and capture the required data. Based on website analysis, this study adopts 17 features in which there are 14 features proposed in [14] and three self-defined features. Those features are shown in Table 4. The process is to search for the required eigenvalues automatically, for example, whether it is registered in a particular country or whether it is a website within one year. These features must be applied to JAVASCRIPT to crawl the WHOIS Server. WHOIS Server [15] is the most advanced domain name server in the Internet Domain Name System (DNS) and is responsible for returning the authoritative domain name server address of the top-level domain. After pointing to the WHOIS server of the domain name registry, we can see whether the website is registered in a particular country. The 17 features are detailed as follows:

Since it must extract 17 features in this study, those features should be compared during training models. The Chi-square test [16] and ANOVA (analysis of variance) [17] are used in this paper as feature comparison algorithms. The Chi-square test is a widely used statistical hypothesis test method for enumeration data, which is suitable for analyzing the association between two groups of categorical variables. Pearson's Chi-squared test is used to determine whether there is a significant difference in between categorical variables and independent variables or not. In this way, suitable features are selected, and the calculation formula of the Chi-square value is as follows Equation (4):

$$\sum X_{i,j}^2 = \frac{(O-E)^2}{E} \tag{4}$$

O: the observed number, E: the expected number

Table 4. The 17 features of the testing data.

The Classification	The Features
URL	length_of_fqdn
	replica_in_fqdn
Content element	num_of_currencies_seen
	num_of_duplicate_pricess_seen
	percent_savings
	contain_emails
	large_iframes
	contain_phone_numbers
	has_mobile_app
	has_social_media
	has_payment_option
Content structure	node_counts
	dom_height
	text_length
WHOIS	in_top_one_million
	country_registered
	under_a_year

Analysis of variance (ANOVA) is a common statistical model in data analysis. It is used to explore the relationship between the dependent variable of continuous data type and the independent variable of categorical data type. In statistics, analysis of variance (ANOVA) is a general term for a series of statistical models and their associated processes, in which the variance of a variable can be decomposed into parts attributed to different sources of the variable. The total variation of a set of data is divided into several parts according to the sources of possible variation. By measuring these sources of variation, it is possible to know whether there is a difference between each variation. When the factors of the independent variable contain equal to or more than three categories, the statistical mode to test whether the mean of each category is equal. In the Chi-square test, if $X2$ is large, it means that the observed value deviates too much from the theory number. It indicates that there is a significant difference between the comparative data. The obtained $X2$ and degrees of freedom (df) to look up the table to calculate the p-value. According to R.A. Fisher [18], the number of p is less than 0.05 which means the feature has significant importance. After testing, two tables can be obtained as shown in Tables 5 and 6. In this paper, the threshold of the p-value is set to 0.05. In these result tables, only two features are greater than 0.05. Therefore, the two features (relica_in_fqdn and contain_phone_numbers) were chosen to be deleted. We only use 15 features for training models in this study which are shown in Tables 5 and 6.

Table 5. Correlations with categorical variables: Chi-square test.

The Features	p-Value
replica_in_fqdn (It was chosen to be deleted)	0.09233
contain_emails	0.00134
large_iframes	0.00541
under_a_year	0.00292
country_registered	0.01590
in_top_one_million	0.00813
contain_phone_numbers (It was chosen to be deleted)	0.07877
has_mobile_app	0.02050
has_social_media	0.00193
has_payment_option	0.00891

Table 6. Correlations with Continuous Variables: ANOVA test.

The Features	*p*-Value
length_of_fqdn	0.00315
num_of_currencies_seen	0.0192
num_of_duplicate_pricess_seen	0.0078
percent_savings	0.0291
node_counts	0.0061
dom_height	0.0102
text_tag_ration	0.0071

3.3. Data Preprocessing

The data can be divided into categorical variables and continuous variables. The category variables are general such as: male, female or breakfast, lunch, dinner and other named items. The continuous variables are numerical items that can be compared in size. Data preprocessing is the processing of raw data before training the model. The importance of data preprocessing is that it affects the quality of the data, and thus, determines the limits of model performance. In this paper, two steps are used in preprocessing data, one is filling missing values and the other is data normalization. The purpose of filling in missing values is to make the dataset as complete as possible. To handle missing values, this paper will pad the missing values of each feature with 0 for features with missing values including in these features: under_a_year, percent_savings, country_registered, has_payment_option, and contain_phone_numbers. There are different ranges between features, which can lead to some large-scale feature models that sometimes have the opposite effect when making predictions. Therefore, data normalization is scaling the features to be in the same range so that each feature has the same effect. This study uses the Min-Max Scaler to normalize the features to the range of 0 to 1.as shown in Equation (5).

$$X' = \frac{X - X_{min}}{X_{man} - X_{min}} \tag{5}$$

4. Experiment Result

The hardware of the experimental system used the Windows10 operating system, equipped with GPU TITAN V and a 64 GB running memory, installed by Anaconda 3. In the experiment, we first divided into 0–200 trees, 100 trees as shown in Figure 4. The blue line is the training accuracy (Train ACC) and the red line is the test accuracy (Test ACC). It found that the convergence state was reached after 50 trees. However, the training accuracy will continue to fluctuate. After several experiments, we found that taking 100 trees is very convergent. Therefore, from this experiment result, this study takes 100 trees in the Random Forest training model.

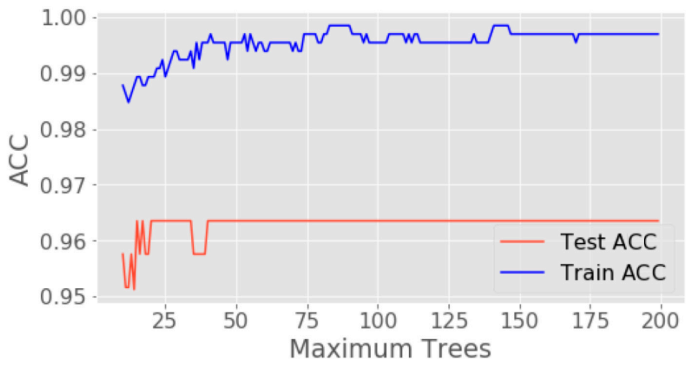

Figure 4. The Test and Train accuracy of 0–200 trees.

Using another training model, the Deep Neural Networks experiment, this paper considers the requirements of DNN for data complexity and number of transactions, and the parameter selection is constructed by building four hidden layers. The number of neurons in the four hidden layers is 256 for hidden layer 1, 128 for hidden layer 2, 64 for hidden layer 3, and 32 for hidden layer 4. Between the first three layers, there is a drop ratio of 25%, the activation function uses Relu, the last layer uses SoftMax, and the optimizer uses Adam. This paper adopts a classification method in machine learning, which is supervised learning. The collected information must have a standard answer, which is generally called a label. Therefore, it configures a corresponding standard answer with the feature value obtained on the website. The problem of this study is to solve whether it is a counterfeit website, which is a dichotomy problem. Therefore, legitimate websites are defined as legit and counterfeit websites are defined as fake. There are 1612 websites used in this paper and a total of 15 feature values and we took 804 fake websites and 808 legitimate websites as our feature extraction training set. This paper used two models: Random Forest and Deep Neural Networks and used 80% of data for training and 20% for testing. The results are shown in Tables 7 and 8 which are those 15 features training and test results.

Table 7. The experiment result of the Random Forest.

	Accuracy	**Precision**	**Recall**	**F1-Score**
Training	99.2%	100%	98.5%	99.2%
Testing	96.9%	97.6%	96.4%	97%

Table 8. The experiment result of the Deep Neural Network.

	Accuracy	**Precision**	**Recall**	**F1-Score**
Training	93.2%	95.8%	91.2%	93.4%
Testing	91.9%	88.9%	93.5%	91.1%

The accuracy of RF in training data is 99.2%, precision is 100%, recall is 98.5%, and the F1 Score is 99.2%. The accuracy of DNN in training data is 93.2%, precision is 95.8%, recall is 91.2%, and the F1 Score is 93.4%. The training process of DNN is shown in Figures 5 and 6 which includes Accuracy and Loss. From the result of the experiment, it can tell that using RF to solve whether it is a counterfeit website or not is better than using DNN.

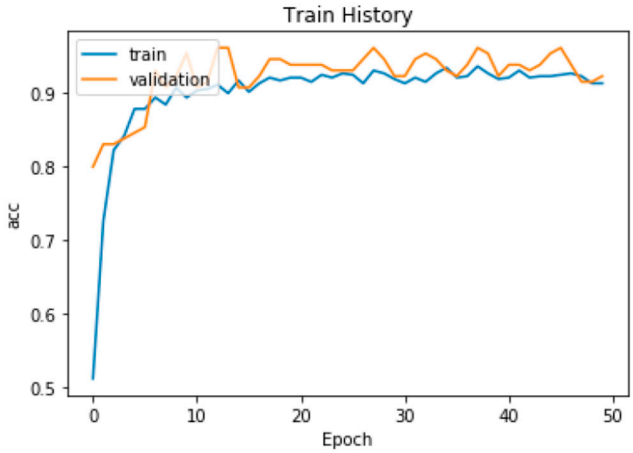

Figure 5. The accuracy of the DNN.

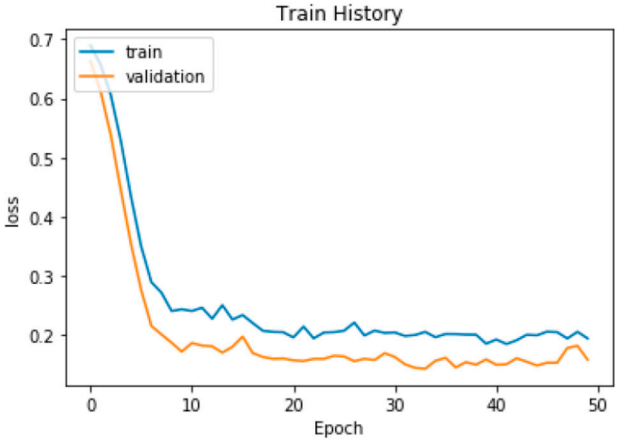

Figure 6. The loss of the DNN.

It can be determined that Random Forest has significant accuracy in this dataset. In addition, it also has better performance in various indicators. Moreover, we ranked the importance of features in the training process of Random Forest. According to each feature's importance to each decision tree in Random Forest, it took the average. Then it can be ranked by its contribution to each feature, as shown in Figure 7. From this figure, there are the top three features proposed in this study, node_counts, dom_height, and text_length in the training process of Random Forest. From this, those features of this dataset play a role in the training process.

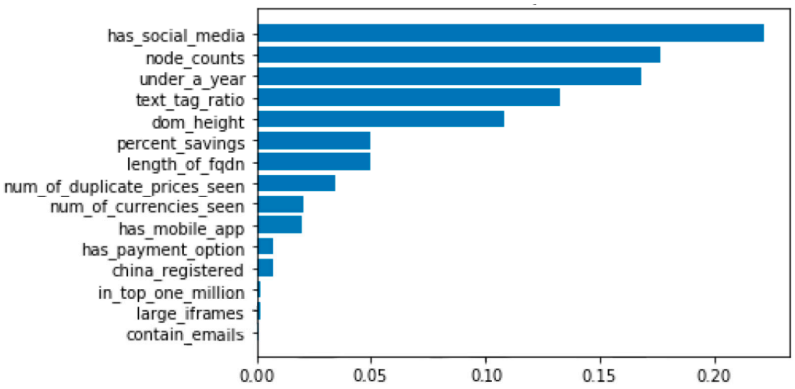

Figure 7. Feature Importance.

5. Conclusions

5.1. Analysis

Regarding the experiment results, the performance of the RF model is better than the DNN model in accuracy, precision, recall, and F1-score, respectively. The reason is that, given the time constraint and the limited data set, the random forest model is less computationally expensive. In addition, the RF model does not need a GPU to complete training. A RF can have a different interpretation of a decision tree but with better performance. Every Neural Networks' hyperparameter is critical in DNN. The cause of underfitting is using too few neurons in the hidden layers. This happens when there aren't enough neurons in the hidden layers to accurately detect the signals. In addition, using too many neurons in the hidden layers can result in several problems. Too many neurons in the hidden layers could result in overfitting. When a neural network has too big of an information-processing capacity, the training set cannot be enough to train all neurons in the hidden layers, and overfitting occurs. Despite sufficient training data, a second problem may arise. A training model with an inordinate number of neurons can take a long time. Eventually, the training time will become so prolonged that the neural network will not be able to be trained

adequately. There must obviously be a balance between the numbers of the neurons in the hidden layers. In our experiment, those various indicators of the DNN model are still good enough that they are also over ninety percent. It is an acceptable performance. If we want to improve the performance of the DNN model, we can increase the hidden layers or the data set scale. However, it will cost a lot of computation resources and take much more time to achieve the result. Considering the goal of this paper is to be efficient and accurate to classify whether a website is selling counterfeits or not. Taking the RF model in the proposed method is better than the DNN model. Considering the training time and accuracy, Random Forests are easier to use than neural networks in this study.

5.2. Conclusions

For the asymmetry information, a secondary market will be formed. Internet fraudsters used counterfeit websites on many social platforms to attract more uninformed consumers. To solve those problems, this study classified counterfeit websites by using machine learning. The data set of this study is 1612 websites used in this paper and a total of 15 features value and take 804 counterfeit websites and 808 legitimate websites. It is trained by 15 characteristics. After that, we use two machine learning algorithms for model training: Random Forest and Deep Neural Networks. The contributions of this paper are as follows:

- This paper used statistical tests, Chi-square and ANOVA detection, to compare the importance of features in feature selection. Then, we selected 15 features of training models.
- In the paper, we used the Random Forest method within 100 decision trees and the Deep Neural Network with three layers to classify whether the websites selling fake products. Regarding the two training model results, the performance of RF is better than DNN.
- Moreover, the importance of features was ranked in the training process of Random Forest in this study. The result shows that there are top three features proposed in this study (node_counts, dom_height, and text_length). From this, the features of this dataset play a role in the training process.

As for the result, using the Random Forest algorithm to classify counterfeit websites can reduce the advertisers who published the fraudsters' ads and further reduce their reputation. It can also prevent the consumers who are misled by asymmetric information and buy counterfeit products.

5.3. Future Studies

In future work, the research attempt is to improve the performance of the model. The models used in the current study are for classifying counterfeit websites specifically. Nevertheless, the internet is also full of fraudulent websites of all kinds. As a result. an updated workflow should be implemented as time progresses. Furthermore, the dataset used for the training model consists of almost exclusively Asia-based online shopping websites. Future studies could investigate whether the selected model's performance is improved when other countries' websites are included. From the dataset used for training models, it consists of mostly online shopping websites from Asia. Future studies can collect more different websites from other countries and investigate whether the characteristics of them are the same as this study or not.

Author Contributions: Conceptualization, S.-W.C., P.-H.C., C.-T.T. and C.-H.L.; methodology, S.-W.C., P.-H.C., C.-T.T. and C.-H.L.; software, S.-W.C. and P.-H.C.; validation, C.-T.T. and C.-H.L.; writing—original draft preparation, S.-W.C. and P.-H.C.; writing—review and editing, C.-T.T. and C.-H.L.; visualization, S.-W.C. and P.-H.C.; supervision, C.-T.T. and C.-H.L.; project administration, C.-T.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Edelman, B. Pitfalls and fraud in online advertising metrics: What makes advertisers vulnerable to cheaters, and how they can protect themselves. *J. Advert. Res.* **2014**, *54*, 127–132. [CrossRef]
2. Flosi, S.; Fulgoni, G.; Vollman, A. If an advertisement runs online and no one sees it, is it still an ad? Empirical Generalizations in Digital Advertising. *J. Advert. Res.* **2013**, *53*, 192–199. [CrossRef]
3. Riek, M.; Bohme, R.; Moore, T. Measuring the influence of perceived cybercrime risk on online service avoidance. *IEEE Trans. Dependable Secure Comput.* **2016**, *13*, 261–273. [CrossRef]
4. George, A.A. The market for "Lemons": Quality uncertainty and the market mechanism. *Q. J. Econ.* **1970**, *84*, 488–500.
5. Asha, S.M.; Deepa, S.P.; Chandra, M.M.; Venugopal, K.R. Detection of fraudulent and malicious websites by analysing user reviews for online shopping websites. *Int. J. Knowl. Web Intell.* **2016**, *5*, 171–189.
6. Chen, G.; Cox, J.H.; Uluagac, A.S.; Copeland, J.A. In-depth survey of digital advertising technologies. *IEEE Commun. Surv. Tut.* **2016**, *18*, 2124–2148. [CrossRef]
7. Ram, D.G.; Afrouz, H.; Raymond, A.P. Analysis of third-party request structures to detect fraudulent websites. *Decis. Support Syst.* **2022**, *154*, 113698.
8. Breiman, L. Random Forests. *J. Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
9. Dan, C.; Ueli, M.; Juergen, S. Multi-column deep neural networks for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3642–3649.
10. Vinod, N.; Geoffrey, E.H. Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th International Conference on Machine Learning (ICML'10), Madison, WI, USA, 21–24 June 2010; pp. 807–814.
11. Lecun, Y. A Theoretical Framework for Back-Propagation. In *Artificial Neural Networks*; Mehra, P., Wah, B., Eds.; IEEE Computer Society Press: Los Alamitos, CA, USA, 1992.
12. Jason, K.; Siddharth, K.; Ahmed, A. Exploratory experiments to identify fake websites by using features from the network stack. In Proceedings of the IEEE International Conference on Intelligence and Security Informatics, Washington, DC, USA, 11–14 June 2012.
13. Ahmed, A.; Hsinchun, C. A comparison of tools for detecting fake websites. *Computer* **2009**, *42*, 78–86.
14. Claudio, C.; Giovanni, R. Learning to detect and measure fake ecommerce websites in search-engine results. In Proceedings of the International Conference on Web Intelligence (WI '17), New York, NY, USA, 23–26 August 2017; pp. 403–410.
15. Harrenstien, K.; Stahl, M.; Feinler, E. *RFC 812: NICNAME/WHOIS*; IETF: Fremont, CA, USA, 1982.
16. McHugh, M.L. The Chi-square test of independence. *Biochem. Med.* **2013**, *23*, 143–149. [CrossRef] [PubMed]
17. Steven, F.S. Analysis of variance: The fundamental concepts. *J. Man. Manip. Ther.* **2009**, *17*, 27E–38E.
18. Fisher, R.A. Statistical Methods for Research Workers. In *Breakthroughs in Statistics*; Kotz, S., Johnson, N.L., Eds.; Springer: New York, NY, USA, 1992.