Original Article

Analyzing the behavior of internet customers based on social engineering

Sara Hajighorbani [1], Changiz Valmohammadi [2*], Kiamars Fathi Hafshejani [3]

[1] Department of Information Technology Management, South Tehran Branch, Islamic Azad University, Tehran, Iran.
[2] Department of Industrial Management, South Tehran Branch, Islamic Azad University, Tehran, Iran.
[3] Department of Management, South Tehran Branch, Islamic Azad University, Tehran, Iran.
Corresponding author and reprints: Changiz Valmohammadi [2], Associate Professor, Department of Industrial Management, South Tehran Branch, Islamic Azad University, Tehran, Iran.
Email: ch_valmohammadi@azad.ac.ir
Received: 02 Dec 2021 **Accepted:** 14 Jan 2022 **Published:** 18 Feb. 2022

Abstract

Background: The customers' opinions about the features and experience of using the products are considered as a valuable and reliable source for comparison and decision-making. Thus, the present study was an attempt to analyze the behavior of Internet customers based on social engineering.

Methods: This study is applied research in the area of social networks. The statistical population of this study included Amazon social network users. The data includes XML and txt files brought to the programming environment. To analyze the behavior of Internet customers, a method based on the ensemble learning technique was implemented in MATLAB software. The common criteria that were used in data mining applications such as accuracy, sensitivity, and F-score.

Results: The proposed model compared to other ensemble methods (support vector machines, Naive Bayes, ensemble neural networks, and decision tree ensemble) is in the priority in all three criteria for recognizing real and non-real users and has a better function. This method had high accuracy, precision, sensitivity, and F-criteria compared to other methods and it has a good status in evaluation criteria. The performance of the proposed model was much better than single algorithms and is the priority in terms of data mining evaluation criteria, but the training time for this model was much longer than other methods.

Conclusion: The use of the proposed model in any organization that provides a product or service online, is quite promising and better results can be achieved with more studies.

Keywords: Behavior; Engineering; Internet; Social networking.

Cite this article as: Hajighorbani S, Valmohammadi C, Fathi Hafshejani K. Analyzing the behavior of internet customers based on social engineering. *Soc Determinants Health*. 2022;8(1):1-13. DOI: http://dx.doi.org/10.22037/sdh.v8i1.36938

Introduction

A network needs to provide services appropriate to its customers and can identify real customers and classify them. Some users try to disrupt the network by using social engineering tools (1).

Social engineering, which obtains the necessary information through abusing the weaknesses of individuals and acquiring their personality habits, is defined as a clever abuse of the natural human tendency to trust, which persuades a person to disclose information or do certain works with the help of a set of techniques. (2).

In the study Yi & Liu, the hybrid recommender system was examined. User-based recommendation, item-based recommendation, and content-based recommendation are some of the

recommender methods discussed in this article. (3).

In the study Khalid et al., machine-learning methods were examined and the upgraded support vector machine and the boost gradient support vector machine were used. (4).

In a study conducted by Gefen et al., machine-learning rules, especially neural networks, were used to predict the trend of gold, silver and crude oil stock market to examine users (5).

The model proposed in Zhou et al., in the first stage, after breaking the text into words and finding their roots, the sentence was converted into words. In the second stage, the words were labelled with appropriate speech components, and in the third stage, the scores were calculated, and in the last stage, the neural network was trained and predicted future prices (6).

Evermann et al. used deep learning and behavior analysis to predict the next behavior, which improves management. In this method, unique features were used to form the feature vector and network input (7).

Huang et al. showed that the deep artificial neural network was successful in recent years in the areas of pattern recognition and machine learning. In this study, he evaluated a variety of deep neural network methods, including supervised, unsupervised and boost networks, and compared the results with each other (8).

At the international level and in Iran, despite an increase in the use of social networks and Internet sites, they do not pay attention to the level of security and reliability, and even when customer behavior is analyzed, they do not pay attention to unusual users (11). Now the question is asked that how to determine the level of trust in customers and increase the reliability of the customer cycle by analyzing their behavior and how to detect phishing by analyzing customers and their behaviors.

Methods

The present study was applied research in the area of social networks. The statistical population of the present study included Amazon social network users. The data included XML and text files brought to the programming environment. The number of samples examined in this study included more than ten thousand records of data related to Amazon users. The model proposed in this study can help improve the quality of social networking services and increase their security. The proposed model is presented to predict customer behavior based on social engineering. In this model, real and unreal and phishing users are detected by recognizing social engineering tools. In this model, an ensemble of classifiers is used. Classifier ensemble includes a set of classifiers that ensemble or combines the opinions and decisions of each member of the group to classify new samples. This combining of single classifiers in classification problems was done by voting among the models and in regression problems by weighted or unweighted averaging among them. In the present study, k-fold, f-score and ensemble classifier algorithms including cost-sensitive decision trees and neural networks were used. In fact, the study problem was the data mining problem of classification type that aimed at identifying unreal and phishing users. The data set was randomly changed to the training and testing section using the k-fold technique. Then, on the randomly selected sets in each stage, the best features that led to the highest efficiency and precision of the algorithm were selected to the extent that the features that had the most repetition in the best criteria were identified using f-score criteria. Then different sets of decision trees were formed, in each branch of which the neural network was considered as the decision-maker, and a number of them were randomly used in decision-making each time to select the best set of trees. Accordingly, the precision of estimation on the entire data set was maximized. All these

stages and evaluations were performed by MATLAB software. The proposed model was a comprehensive model that resulted in a comprehensive analysis of Internet customers.

In this study, a random forest algorithm was used to analyze the behavior of Internet customers. It also examines behaviors from different dimensions and recognizes the nature of the customer from a social engineering perspective. Since information related to customers was unstructured data and could not be used directly in the development of learning models, in the proposed method, the existing data were pre-processed and necessary features were extracted from them. In the proposed algorithm for training each branch, sampling is done from the available data set and after training all branches, voting among branches is done to predict any new behavior. Our studied problem is a classified data-mining problem that aims at identifying real or unreal users and detecting phishing. We have two main stages to classify customer behavior. In the first stage, the information in the data set is processed. In this stage, to create a new data set that can be used to train the learning model, we also have a few separate stages ahead, which are described below. The output of this stage is the data in a new feature space so that it can be completely used to train learning models. After creating a new data set, the training stage of the proposed learning model in this research, which is the random forest, is performed. Finally, after completing the training of the model learning process, it is used for a new classification. The detailed steps of each stage are described in the following sections.

Customer behavior classification

After completing the pre-processing phase, the customers' behavior is analyzed and they are classified. At the end of the training, to classify a new sample, voting is performed among the trees and the class with the highest number of votes is selected for the new sample. Thus, using the data generated in the first stage, we train and evaluate the random forest model in the following way. This model aimed to analyze the behavior of Internet customers to determine whether they are real or unreal and to detect phishing. The proposed random forest algorithm is described in detail below.

The general process of generating random forest in the proposed method is described below.

1. We sample the training data subsets based on the number of decision trees specified.
2. To estimate the error rate, we separate a part of the data and discard it. We identify the error using these data.
3. We develop every tree in the forest until its training is completed.
4. We apply the new sample to all the trees in the forest and the tagged one will be the vote of the majority of the trees.

Characteristics of the proposed method:

The proposed method has the following functions and features.

-The proposed method checks the accuracy of the information and increases it and provides good generalizability

• It applies to any data set and is not limited to specific data.

• It applies to any number of features and variables.

•The developed forest can be trained in parallel.

Random forest

One of the efficient tools used in problems related to the classification or estimation of the target variable is the decision tree. Each decision tree uses a top-down recursive partitioning strategy. A decision tree divides the input space into a set of separate areas and assigns a response value to each area. In general, the single decision tree is prone to overfitting and has little generalizability. Generalizability means that the model can adapt well to new data and has minimum detection or prediction error. We say that one model has an overfitting problem when there is another model that is less adaptable to training

samples but in general more adaptable to all samples, both training and non-training (new). As stated before, with increasing the size of the tree, the precision of the tree on the training samples increases, but the precision of detection on other data decreases. When there is an error or noise in the training data, the decision tree grows larger and becomes more complex due to error to adapt to all the training data and precision on training data increases but decreases during testing. The problem of overfitting is a major problem in tree learning and many learning methods. In most cases, the problem of overfitting reduces precision by 10 to 25 per cent. Another disadvantage of the single decision tree is the instability of its results in the presence of noise in the input data. When a decision tree is formed, a small change in the training samples can cause fundamental changes in its structure. To overcome these problems, the random forest algorithm, which is a learning method based on a class of decision trees, has been proposed. Random forest algorithm is known as an ensemble learning method and is used for both classification and regression applications. In this method, several decision trees are trained and voting among the trained trees is used to predict or classify a new data sample. When a new data sample is viewed, it moves from the root to a leaf, depending on its features. In the classification issue, the new data tag is a tag that many trees have predicted. In the regression problem, the mean (weighted or unweighted) of the values predicted by all trees is introduced as the predicted value for the new sample. The basic idea of ensemble learning methods, such as random forests, is that a group of "weak learners" can build a "strong learner."

This model is one of the most accurate learning models available. In most cases, single decision trees have higher variance or bias. Bias in decision trees refers to a set of hypotheses that along with inferential training data confirm the tag attributed to the samples by the learner model. The two hypotheses in the decision tree can be expressed as follows: 1. Shorter trees are preferred to taller trees. 2. Trees that give features with the highest information yield closer to the root are preferred to other trees. Each tree in this model is built completely independently, so a random forest can be trained more quickly with parallel processing. In this model, two parameters are determined by the user. The first parameter is the number of decision trees built in the forest. This parameter varies depending on the number of training data and the number of features in each problem. The second parameter is known as m. When selecting the failure feature at each node of the tree, the m features should be randomly selected and among them features, the best feature for data segmentation should be selected with the help of the mentioned criteria. This parameter is considered in the whole process of random forest training and among all fixed trees. Typical values for this parameter are: Sqrt (nVariable), Log (nVariable) +1. nVariable refers to the number of features in the data set or problem variables. n addition, for training each tree, N training samples is randomly selected among the initial training data set. The N parameter is usually considered as the size of the entire primary data set available. Thus, there may be commonalities among the training subsets. The considerable point is that this model can manage the overfitting phenomenon well and have more generalizability in its problem space, meaning that it is not limited to training data, is not over-fitted, and is general. In other words, due to the random selection of training subsets, and especially in the selection of m features, the phenomenon of overfitting is avoided.

The general process of building a random forest is described below.

Sample training data subset among the initial training data as the number as the desired decision tree. In each training set, separate and discard some data to estimate the error rate. The discarded data is called

out-of-bag or OOB. The error calculated using this data is an internal unbiased estimate for the generalization error. In other words, another strength of this model is revealed here. In each training subset, develop a decision tree without pruning until the end of the training process. In each node, instead of selecting the best failure feature among all features, randomly select them feature and among these m features, select the best of them as the failure feature. Apply the new data to all trees and the final tag of the data is the vote of a majority of the trees. In the regression problem, the final predicted value is the average of the values predicted by all trees.

The above algorithm describes the second phase of the proposed model.

Prediction error estimation

In random forests, a separate test set is not needed to estimate unbiased error or the precision of the final diagnosis of the model. This error is estimated at runtime. At the time of building each tree, almost one-third of the training data belonging to that tree is discarded as OOB data and is not included in the process of tree building. Classification error estimation using OOB data is performed as follows.

For each of the OOB data belonging to the i^{th} tree, predict its class with an i^{th} tree.

After completing training all forest trees, give a tag, predicted by the majority of trees, for any data placed in the OOB class, and compare it to its real tag.

Consider the ratio of the number of OOB data that have been misclassified to the total number of data in the OOB class throughout the training process as a classification error. For the regression problem, the square of the OOB data error divided by the total number of OOB data is considered as the prediction error.

Extractable information from the model

Using random forests, in addition to determining the precision or percentage of detection, more information can be obtained about the problem variables, which are described below.

To determine the importance of a variable such as m, after building each decision tree, consider data from the OOB set that that tree predicted correctly. Then, change the values of the variable m in these data and re-apply these new data to the tree for detection.

Deduct the number of data predicted correctly in the new data set from the number of initial OOB data that the tree correctly predicted them and you did not change them. Repeat this process for all trees. The mean of calculated values shows the raw score of the variable m. Finally, normalize the calculated values in the range of zero and one. In fact, due to the existence of OOB data and their use at runtime, it is possible to identify simultaneously the important variables to which the dependent variable of the problem is sensitive. This feature does not exist in any of the learning algorithms.

If the number of problem variables is very large and we face a complex problem space, a random forest can be created first using all the features. After determining the most important features, the forest can be built with only the important features. If the criterion for identifying the best failure feature in each node is the Gini index when a node failure occurs on the feature m, the Gini Impurity Criterion created for the sub-nodes is lower than the parent node. By adding this reduction in impurities to each m variable across the forest, it provides a level of importance for that variable that is consistent with the previous method. It is one of the most useful tools of this model. One matrix is N * N, and N is the total number of initial training data sets. Once a decision tree was created, apply both the training data and its OOB data to the tree. If sample i is similar to sample j in the similar final node, add one element (i, j) to the adjacency matrix. Finally, normalize the matrix elements by dividing by the total number of trees. An adjacency matrix can be used to define data structure or unsupervised learning.

Results

To examine the effect of different parameters of the proposed model on the precision of detection, we have designed and implemented experiments. Using these experiments, acceptable values are determined for the free parameters of the model. To show the strength and efficiency of the proposed model, it is compared with other existing methods and the results are presented.

Tested data set

The data used in this study are collected from BPI 2012 and brought to the MATLAB programming environment in CSV format. The existing data are converted and presented in the form of XML files, which after initial processing and loading into a format usable for processing, they are entered as input to the first phase of the proposed model.

Examining different parameters in the model

In general, in any algorithm, there are several free and effective parameters in the performance of the algorithm. The proposed algorithm is not an exception in this regard and has several parameters that optimally adjust them to increase the precision and power of the algorithm in classification. In this section, we have shown the effect of these parameters in the form of experiments. OOB error estimation is also used for evaluation, which is a method for evaluating this algorithm, and graphs have been drawn according to this criterion. Estimation of this error is one of the unique features of this algorithm. However, there is no need to use other evaluation methods such as X-validation, and this error is a good estimate of the generalizability of the algorithm. Also, to prove this claim and to compare and evaluate this method with other methods, some criteria for evaluating classification algorithms such as accuracy, error rate and response time considered in most data mining applications, are used.

Parameter m

One of the parameters that can be adjusted in the model is known as m. It should be noted that this parameter means the number of features extracted at each node of the tree and randomly extracted from the entire set of problem features, and by using this selected subset, the best feature to failure is selected. During the model development, this parameter is considered for all fixed trees and the typical values selected for it are Sqrt (nVariable) or Log (nVariable). To observe the effect of this parameter and select the best possible value for it, we performed the following experiment:

Experiment 1: Parameter m

In this experiment, we implemented the model with the following conditions:

1. Number of trees: 150
2. Failure feature determination criteria: Gini index
3. The parameter m: It is variable and takes three separate values. 1. The total number of extracted features. 2. The square root of the total number of features in each data set. 3. The logarithm of the total number of features in each data set.

In this experiment, the desired features are extracted and quantified. Then, one training model is created by using them.

Table 1 presents the results of the model evaluation in these three different modes.

Table 1. Evaluation of the model in terms of different values of the parameter m

Parameter m	Feature space	OOB error	Accuracy
nVariable	1-gram	9.6%	0.89
	2-gram	13.3%	0.87
	3-gram	17.16%	0.81
	MRC Fea.	20.63%	0.77
Sqrt(nVariable)	1-gram	12.1%	0.85
	2-gram	15%	0.82
	3-gram	20.6%	0.73
	MRC Fea.	26%	0.67
Log(nVariable)	1-gram	12%	0.85
	2-gram	14.3%	0.83
	3-gram	19.1%	0.74
	MRC Fea.	24%	0.69

Also, Figure 1 compares the results of applying different values of the m parameter based on the OOB error.
Based on the above results, reducing the number of features to sqrt (nVariable) and log (nVariable) values does not result in promising results, since the OOB error has increased and other evaluation criteria have decreased.

The criterion for determining the failure feature

One of the most important parameters in the proposed model and all methods that operate based on the decision tree is the selection of an appropriate criterion for determining the point of failure at the time of decision tree construction. Since the core of the random forest is the decision trees in it, it is important to select the best criterion for determining the point or feature for data segmentation. There are several criteria in this regard, which are described in detail.
To observe the effect of this parameter and select the best criterion for selecting the point of failure, the second experiment is performed as shown in Figure 1:
Experiment 2: Criteria for determining the failure characteristic
In this experiment, we have implemented the model with the following conditions:
1. Number of trees: 150
2. m: nVariable. Given the previous experiment and the fact that reducing the value of m may hurt the performance of the algorithm, we consider its value as the number of variables or the same features in the data. In this experiment, as in Experiment 1, the desired features are extracted and quantified. Then, a model is developed by using them. The size of the feature space for different values of n is the same as in Experiment 1.
3. Criteria for selecting the point of failure: It is variable and each time one of the criteria of information gain, gain ratio and Gini index is selected and tested.
Table 2 shows the results of the model evaluation in these three different modes.

Table 2. Evaluation of the model based on different parameters for determining the point of failure

Criteria for segmentation of data	Feature space	OOB error	Accuracy
Information Gain	1-gram	11%	0.87
	2-gram	14%	0.84
	3-gram	21%	0.79
	MRC Fea.	22%	0.77
Gain Ratio	1-gram	11%	0.87
	2-gram	14%	0.84
	3-gram	19%	0.78
	MRC Fea.	21%	0.75
Gini Index	1-gram	9.6%	0.89
	2-gram	13.3%	0.87
	3-gram	17.16%	0.81
	MRC Fea.	20.63%	0.77

Based on the results, it can be seen that the efficiency of the Gini index in selecting the failure feature is better than the other two criteria. It has not been proven so far which criterion performs better than the other criteria, and the performance of these criteria depends on the data used and the studied problem. Also, Khalid et al. has been shown that most of the time the information gain and Gini index have the same performance and the Gini index works better only in 2% of cases. However, in this study, it was observed that the use of the Gini index led to better results (4).
By experimenting with 1, we tried to find the best value for the m parameter in this algorithm. Based on the results in Table 1, if we value this parameter to the number of

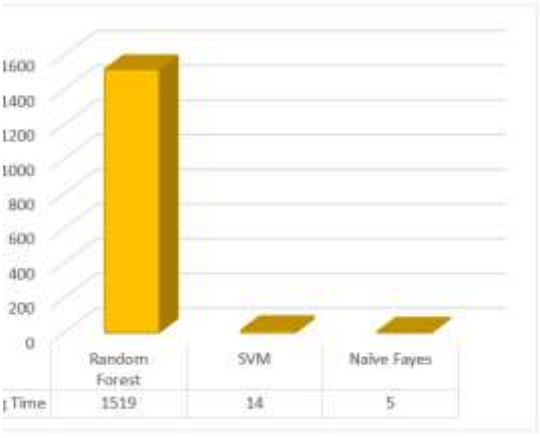

Figure 1. Comparison of the proposed model with single learning algorithms in terms of response time

features in the data set, we will obtain the least OOB error and the best conditions in other evaluation criteria.

Finally, the goal of Experiment 2 was to determine the best criterion for selecting failure features in tree nodes. Based on the results shown in Table 3, we found that using the Gini index criterion provided the least error compared to other criteria.

After determining the best values for the parameters of the proposed model, we evaluate this method with other methods. For this purpose, we evaluate the proposed method with other single and ensemble algorithms, as well as with the best methods proposed so far.

Based on the results of Table 3, it can be seen that the performance of the proposed model is much better than single algorithms. In Table 4, we compare learning methods based on the time required to build the model.

Then, we compared the proposed method with ensemble models. The results are presented in Table 5. Figure 2 Also presents the results based on the 14 times required to build the model and response time.

Table 3. Evaluation of the proposed method in identifying real and unreal customers with single learning methods

Method	Accuracy	Error rate	Time required to build the model (seconds)
Proposed model	0.89	9%-20%	1519
Support Vector Machine	0.66	Unable to determine error	14
Naive Bayes	0.72	Unable to determine error	5
neural network	0.74	Unable to determine error	347
decision tree	0.78	15%-38%	884

Table 4. Evaluation of the proposed method in detecting phishing with single learning methods

Method	Accuracy	Error rate	Time required to build model (seconds)
Proposed model	0.92	9%-20%	1720
Support Vector Machine	0.67	Unable to determine error	20
Naive Bayes	0.70	Unable to determine error	5
neural network	0.74	Unable to determine error	330
decision tree	0.76	15%-38%	850

Table 5. Evaluation of the proposed method in detecting real and unreal users with other ensemble models

Method	Accuracy	Error rate	Time required to build the model (seconds)
Proposed model	0.89	9%-20%	1519
Support Vector Machine	0.73	Unable to determine error	1645
Naive Bayes	0.78	Unable to determine error	336
neural network ensemble	0.81	Unable to determine error	1042
decision tree ensemble	0.82	15%-38%	1285

Table 6. Evaluation of the proposed method in detecting phishing with other ensemble models

Method	Accuracy	Error rate	Time required to build model (seconds)
Proposed model	0.94	9%-20%	1720
Support Vector Machine	0.70	Unable to determine error	1650
Naive Bayes	0.75	Unable to determine error	844
neural network ensemble	0.80	Unable to determine error	1168
decision tree ensemble	0.83	15%-38%	1436

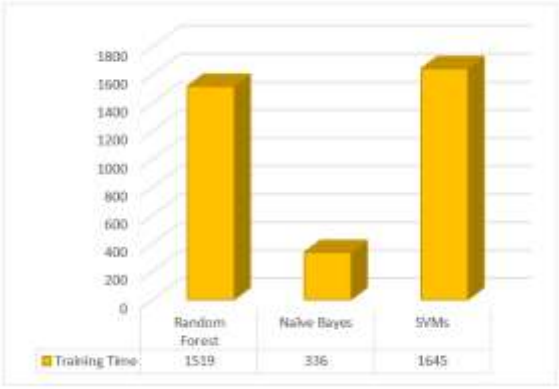

Figure 2. Comparison of the proposed model with other ensemble learning algorithms in terms of response time

Based on the results presented in Table 6 and Figure 2, it can be seen that the proposed algorithm in this research is more efficient than other ensemble methods and has a favorable status in the evaluation criteria. Based on the results presented in Table 6, it can be seen that the proposed model in this research is in the priority among in three criteria and has a better performance.

Discussion

The present study provided a model for analyzing the behavior of Internet customers in social networks based on social engineering. In this model, two main phases were implemented to classify customers and their behavior. In data mining problems and in general in text mining tasks, since the user data are unstructured texts, a pre-processing phase is performed to extract the feature from these unstructured data. In the present study, in the first phase of the proposed model, redundant punctuation marks, numbers and URLs were deleted from the data and tagged word by word. In the proposed model, two approaches called "linguistic approach" and "semantic approach" as well as "social engineering" and "machine learning "were used. The present study was designed and implemented to evaluate the proposed experimental model. The first class of these experiments was to determine the appropriate values for the free parameters in the model.

These parameters included the number of trees in the forest, the criterion for determining the point of failure and the m parameter (the number of features that become a candidate to determine the point of failure). We observed that if 150 trees are built in a forest according to the Gini index, and all features become a candidate to select the best point of failure in each node of the tree, a model with the best performance would be obtained. Then, the built model was evaluated with single and ensemble algorithms. According to the results, we found that single models would not be very efficient in a complex feature space. Selection of the basic learning models is also an important point in the ensemble learning methods, as we observed in the fourth chapter that the proposed model has a better performance than other ensemble methods due to the advantages of the decision tree model over other models. Finally, we evaluated the proposed model of this research with the performed studies. In this evaluation, we also found that in addition to the ensemble nature of the algorithm, which is effective in its performance, machine-learning methods generally work better than other methods. Finally, it can be concluded that the use of the proposed model in any organization that provides a product or service online is quite promising and better results can be achieved with more studies. We evaluated the proposed model with single and ensemble algorithms as well as with the best methods proposed so far. These methods included single support vector machine, ensemble support vector machine, naive Bayes, Bayesian ensemble, and random forest.

The results revealed that the proposed model in the accuracy criterion improved by 23%, 17%, 16%, and 11%, respectively, compared to the single support vector machine method, naïve Bayes method, ensemble support vector machine, and Bayesian ensemble method. The proposed

model in the sensitivity criterion improved by 21%, 17%, 10%, and 6%, respectively, compared to the single support vector machine method, naïve Bayes method, ensemble support vector machine, and Bayesian ensemble method. Also, the proposed model in the precision criterion improved by about 19%, 18%, 7%, and 8%, respectively, compared to the single support vector machine method, naïve Bayes method, ensemble support vector machine, and Bayesian ensemble method. The proposed model in F- score criterion improved by about 20%, 17%, 8%, 7%, respectively, compared to the single support vector machine method, naïve Bayes method, ensemble support vector machine, and Bayesian ensemble method. The performance of the proposed model is much better than single algorithms. It is also observed that the proposed model was more efficient than ensemble methods. A new classification method applies the principal component analysis (PCA) to rotate the main feature vector to obtain different training data sets for learning-based classifiers (9).

A great number of studies focus on the design of multiple classification systems, which are based on similar classification models, and a subset of different data or a subset of features has been trained to these classifications. Such multiple classification systems are often referred to as classifier groups. One of the most known effective methods in classification models is the bagging method and random forest (9).

Kouloumpis et al. examined Twitter users through hashtags. In this experiment, two databases, including HASH and EMOT, which include Twitter messages, were used for machine learning. Also, a database called ISIEVE classified manually was used for evaluation. Initially, the data were pre-processed, which the pre-processing stage includes three stages of 1- Tokenization 2- Normalization, in which the abbreviation of the words is found and replaced with its correct meaning 3- Part-of-speech (POS) tagging. (10).

The advantage of the method used in Kouloumpis et al. is the extraction of effective features and the use of ensemble learning, which increases the precision of the prediction, but the extraction of effective features is time-consuming (10).

the question is asked that how to determine the level of trust in customers and thus increase the reliability of the customer cycle by analyzing their behavior and how to detect phishing by analyzing customers and their behaviors (11).

A review of research literature suggests that no study has been conducted so far with the method proposed in the present study and the studies have been based on other methods. Among these studies, we can refer to the study conducted by Pavlou & Fygenson, to predict customer behavior using artificial neural network technique. In this article, a data mining technique was presented to identify the desired products of customers based on their purchases (12).

In a study conducted by Khalid et al., machine learning methods were examined and they proposed a method based on a boost gradient support vector machine. After pre-processing and breaking the sentences into the words that make them up, the weight of each word is calculated and given to the data classification function to make a decision. In this method, the core function of the support vector machine is gradient function. This method has a very good performance in detecting positive samples and minimizing decision error (4).

In a study conducted by Tsiakis & Sthephanides, to examine the users, machine-learning rules, especially neural networks, were used to predict the trend of gold, silver and crude oil stock market. For example, in the model proposed in the Tsiakis & Sthephanides, in the first step, after breaking the text into words and finding their roots, the sentence was converted into words. In the second step, the words were tagged with appropriate parts of speech, and in the third step, the scores were calculated, and in the last step,

the neural network was trained and predicted future prices (13).

The problem of this method is its inefficiency in noisy data. In research conducted by Hemmatian & Sohrabi, to understand the existing challenges and solutions well for exploring Internet users, the different available methods of machine learning were examined, and the advantages and disadvantages of each were stated: 1- Machine learning methods have low supervision rate and 2 - Very much dependent on tagged training data. 3- It requires human effort and linguistic knowledge, 4- It is expensive. However, the machine learning methods with supervision have a very high precision of classification, are resistant to noise and are effective in identifying the subject of the text. Among semantic classification methods, lexicon-based methods have relatively little precision, are unable to identify the idea of words that are about a particular subject and do not exist in the lexicon, and have low precision in examining different domains. However, one of the advantages of this method is easy to access words in the lexicon and its implementation is very fast (14).

Everman et al., in another article, did another work based on deep learning. They tried to predict the next behavior by using deep learning and behavior analysis to improve the management. In this method, unique features were used to form the feature vector and network input. The results of this algorithm were tested on two different data sets. Although the proposed method had acceptable functionality, there was a problem of losing some data and information in it (15).

In the study conducted by Poria et al., a 7-layer deep convolutional neural network was used to examine all aspects of a text. Speech patterns were also ensembled with the neural network to obtain better results. Their results showed that this method was more effective than popular methods, but the ensemble of speech patterns with deep neural networks to improve the pattern of diagnosis increased complexity and time (16).

Wafi et al. also proposed a new method in their article. In their paper, 7 important features have been extracted and results of two modes of Multilayer Perceptron and Simplified Fuzzy ARTMAP Neural Networks were compared. The results showed that in each method, 90% of the classification was correct. In a specific method, which was the ensemble of Perceptron and Lundberg, the accuracy of the results increased up to 94% (17).

Then, the proposed model was compared with the ensemble of random forest model and a linguistic model and the ensemble of random forest model and a semantic model in terms of different performance criteria.

The present study also showed that the proposed model improved the accuracy criterion by 19% compared to the ensemble of random forest method and linguistic model and 4% compared to the ensemble of random forest method and semantic model. The proposed model also improved in the sensitivity criterion by about 10% compared to the ensemble of the random forest method and the linguistic model and by about 9% compared to the ensemble of the random forest method and the semantic model. The proposed model also improved the precision criterion by about 20% compared to the ensemble of random forest method and linguistic model and by about 2% compared to the ensemble of random forest method and semantic model. The proposed model also improved the f-score by about 13% compared to the ensemble of random forest method and linguistic model and by about 6% compared to the ensemble of random forest method and semantic model. No research was found to compare the results in this regard. Based on the results, we found that in a complex feature space, single models would not be very efficient. The selection of basic learning models is also an important point in ensemble learning methods. The proposed model of the present study has a better performance than other ensemble methods

due to the advantages of the decision tree model over other models. Finally, we evaluated the proposed model of this study with the studies performed. In this evaluation, we also found that in addition to the ensemble nature of the algorithm, which is effective in its performance, machine-learning methods generally work better than other methods.

Conclusion

The present study provided a model for analyzing the behavior of Internet customers in social networks based on social engineering. In this study, we first designed and performed experiments to find the best values for the parameters set in this algorithm. Then, we evaluated the proposed algorithm and compared it with other available methods as well as the ensemble algorithm. By performing the final experiment and evaluating the algorithm, we found that the method proposed in this study worked with more power and was quite satisfactory and yielded good results. The information used in the presented study included data on Internet customers, which the model training features were extracted by pre-processing them. Then, for customer classification based on social engineering, real and unreal customers and phishing were detected and customers were classified. Then, the proposed model was evaluated. All the mentioned steps and calculations were performed in MATLAB software. The results of these comparisons showed that the model proposed in the present study had a more appropriate function in all criteria.

Future Recommendations

Since the OOB error decreases with an increasing number of trees and the model detection power increases, it is possible to increase the number of trees and then prune them. Training data can be selected definitively during sampling to avoid increasing and decreasing oscillating error. Creating an unrelated dependency between features results in misleading and incorrect training of the learning pattern, and the efficiency of the classification algorithm decreases, and OOB error, compared to when each independent word is considered a feature, increases. Thus, each word can be considered as a feature. Also, the dimensions of the feature space can be reduced by appropriate feature extraction methods, provided that it does not adversely affect the performance of the learning algorithm.

Acknowledgements

All the professors and experts who have contributed to the compilation of this article are sincerely appreciated.

Author's contribution

Sara Hajighorbani and Changiz Valmohammadi developed the study concept and design. Kiamars Fathi Hafshejani acquired the data. Sara Hajighorbani and Changiz Valmohammadi analyzed and interpreted the data, and wrote the first draft of the manuscript. All authors contributed to the intellectual content, manuscript editing and read and approved the final manuscript.

Informed consent

Questionaires were filled with the participans' satisfaction and written consent was obtained from the participants in this study.

Funding/financial support

There is no funding.

Conflict of interest

The authors declare that they have no conflict of interests.

References

1. Adib PA. Monitoring, optimizing the accuracy of trust among online social network users using data mining technique in Apache Spark environment, the third national conference on distributed computing and big data processing. Shahid Madani University of Azerbaijan;2017.
2. Panahi A. Introduction to Social Engineering and its Tools. The First Regional Conference on the Application of Electrical and Computer Sciences in the Telecommunication Industry;2012.
3. Yi S, Liu X. Machine learning based customer sentiment analysis for recommending shoppers, shops based on customers' review. Complex

Intell. Syst. 2020;6(1):621-634. https://doi.org/10.1007/s40747-020-00155-2

4. Khalid M, Ashraf I, Mehmood A, Ullah S, Ahmad M, Choi GS. GBSVM: Sentiment Classification from Unstructured Reviews Using Ensemble Classifier. Applied Sciences. 2020;10(8):2788-2797. https://doi.org/10.3390/app10082788
5. Gefen D, Karahanna E, Straub DW. Trust and TAM in Online Shopping: An Integrated Model. MIS quarterly. 2003;27(1):51-90.
6. Zhou L, Dai L, Zhang D. Online shopping acceptance model - A critical survey of consumer factors in online shopping. Journal of Electronic Commerce Research. 2007;8(1):41-62. https://www.scirp.org/(S(i43dyn45teexjx455qlt3d2q))/reference/ReferencesPapers.aspx?ReferenceID=482812
7. Evermann J, Rehse JR, Fettke P. Predicting process behavior using deep learning", Decision Decision Support Systems, International Conference on Business Process Management. 2017;327-338.
8. Huang L, Ding B, Wang A, Xu Y, Zhou Y, Li X. User Behavior Analysis and Video Popularity Prediction on a Large-Scale VoD System. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM). 2018;14(1):1-24. https://doi.org/10.1145/3226035
9. Lasota T, Luczak T, Trawinski B. Investigation of Rotation Forest Method Applied to Property Price Prediction. International Conference on Artificial Intelligence and Soft Computing. 2012;403-411. DOI: https://doi.org/10.1007/978-3-642-29347-4_47
10. Kouloumpis E, Wilson T, Moore J. Twitter Sentiment Analysis: The Good the Bad and the OMG!. Proceedings of the International AAAI Conference on Web and Social Media. 2021;5(1):538-541. https://ojs.aaai.org/index.php/ICWSM/article/view/14185.
11. Proksch S, Lowe W, Wäckerle J, Soroka S. Multilingual Sentiment Analysis: A New Approach to Measuring Conflict in Legislative Speeches. Legislative Studies Quarterly. 2018;44(1):97-131. https://doi.org/10.1111/lsq.12218
12. Pavlou PA, Fygenson M. Understanding and Predicting Electronic Commerce Adoption: An Extension of the Theory of Planned Behavior. MIS Quarterly. 2006;30(1):115-143. https://doi.org/10.2307/25148720
13. Tsiakis T, Sthephanides G. The Concept of Security and Trust in Electronic Payments. Computers and Security. 2005;24(1):10-15. https://doi.org/10.1016/j.cose.2004.11.001
14. Hemmatian F, Sohrabi MK. A survey on classification techniques for opinion mining and sentiment analysis. Artif Intell Rev. 2019;52(1):1495-1545. https://doi.org/10.1007/s10462-017-9599-6
15. Everman J, Rehse JR, Fettke P. A Deep Learning Approach for Predicting Process Behavior at Runtime. International Conference on Business Process Management. 2016; 327-338. DOI: https://doi.org/10.1007/978-3-319-58457-7_24
16. Poria S, Cambria E, Gelbukh A. Aspect extraction for opinion mining with a deep convolutional neural network. Knowl. Based Syst. 2016;108(1):42-49. DOI: https://doi.org/10.1016/j.knosys.2016.06.009
17. Wafi NM, Sabri N, Yaakob SN, Nasir ASA, Nazren ARA, Hisham MB. Classification of Characters Using Multilayer Perceptron and Simplified Fuzzy ARTMAP Neural Networks. Advanced Science Letters. 2017;23(6):5151-5155. DOI: https://doi.org/10.1166/asl.2017.7330