Similarity measure fuzzy soft set for phishing detection

Rahmat Hidayat [a,b,1,*], Iwan Tri Riyadi Yanto [a,c,2], Azizul Azhar Ramli [a,3], Mohd Farhan Md. Fudzee [a,4]

[a] Faculty of Computer and Information Technology, Universiti Tun Hussein Onn Malaysia, Malaysia
[b] Department of Information Technology, Politeknik Negeri Padang, Padang, Indonesia
[c] Departement of Information System Universitas Ahmad Dahlan Yogyakarta, Indonesia
[1] rahmat@pnp.ac.id; [2] yanto.itr@is.uad.ac.id; [3] azizulr@uthm.edu.my; [4] farhan@uthm.edu.my
* corresponding author

ARTICLE INFO

Article history
Received December 29, 2020
Revised March 31, 2021
Accepted March 31, 2021
Available online March 31, 2021

Keywords
Similarity measure
Fuzzy soft set
Phishing detection
Classification

ABSTRACT

Phishing is a serious web security problem, and the internet fraud technique involves mirroring genuine websites to trick online users into stealing their sensitive information and taking out their personal information, such as bank account information, usernames, credit card, and passwords. Early detection can prevent phishing behavior makes quick protection of personal information. Classification methods can be used to predict this phishing behavior. This paper presents an intelligent classification model for detecting Phishing by redefining a fuzzy soft set (FSS) theory for better computational performance. There are four types of similarity measures: (1) Comparison table, (2) Matching function, (3) Similarity measure, and (4) Distance measure. The experiment showed that the Similarity measure has better performance than the others in accuracy and recall, reached 95.45 % and 99.77 %, respectively. It concludes that FSS similarity measured is more precise than others, and FSS could be a promising approach to avoid phishing activities. This novel method can be implemented in social media software to warn the users as an early warning system. This model can be used for personal or commercial purposes on social media applications to protect sensitive data.

1. Introduction

Phishing is an attack vector that deploys technical subterfuge and social engineering to surreptitiously obtain otherwise personal and sensitive information such as credit card pins and user identities [1]. Unsuspecting users are lured by criminal elements, masquerading as legitimate entities via electronic communication media to divulge vital, personal, often, financial information, which may, in turn, be used illegally by the criminals without the knowledge or consent of the real owner. Phishing is an instance of identity theft [2]. The phishing cycle often starts with an email that replicates the identity of a trusted associate or organization. It is usually with a bogus but juicy claim to a reward for the unsuspecting recipient, or in other instances, a dubious revalidation exercise by elements posing as financial institutions, demanding that users supply their authentication details. To take the bait, the user is made to fill out personal data such as bank account PIN, social security number, or some other useful authentication details, which may be used by the criminals to perpetrate illegal transactions later.

Phishing attacks pose severe risks to both individuals and corporate entities and have dire consequences on global security and the economy [3]. It is even more so dangerous, as it appears that phishers continue to perfect means to outmaneuver also the knowledgeable and security-conscious [4].

https://doi.org/10.26555/ijain.v7i1.605

Technology giants such as Google and Facebook have lost about $100 million to phishing emails from hackers who were impersonated as hardware vendors in 2017. The phishing attack's economic effect is enormous; a report gathered for five years by the FBI internet crime complaint center showed that financial loss occasioned by phishing attacks exceeds $12 billion globally [5].

Phishing attackers are increasingly becoming more resilient over the years due to alarming attack volume and its innovativeness that were being implemented. Security specialists and phishers are in a vicious circle because it becomes very complicated to catch phishers. Phishers are continually changing their tactics to beat anti-phishing techniques [6]. The total number of phishing sites detected by APWG in the second quarter of 2019 was 182,465, and it was marginally up from 180,768 in the first quarter that Significantly increased from 138,328 in the fourth quarter of 2018 and 151,014 in the third quarter of 2018 [7]. The email has also been identified as the top phishing target; consequently, a phishing email attack aimed at individuals and corporate bodies is on the rise [8]. To safeguard the sensitive information of users, an adequate means of spotting phishing emails must be developed.

Anti-phishing development has been conducted by a previous study [9] to prevent users from phishing scams. Today, numerous email filters continue to use certain static approaches; they are insufficiently resilient to comply with emerging phishing trends and could only comply with established phishing activities. It caused email users vulnerable to different phishing attacks. Since the impostor is not static in his activities, this is a loophole; As often as possible, they change the operating mode not to be detected [10]. This has inspired several researchers to investigate additional successful strategies for combating both proven and emerging fraud. Additionally, the techniques have been implemented that contributed to the Data Mining algorithm invention [11]–[14]. One of the Data Mining approaches is Classification that could be useful for predicting phishing websites [15]–[17]. Phishing is a prevalent classification issue in data mining to create a classifier based on huge website features. Phishing attacks, phishing classification, detection, and future challenges have been described in [18]–[20].

There are two important concepts in classification problems in applying soft set theory, specifically, the idea of decision-making based on fuzzy soft set (FSS) and the theory of comparing the similarity of two fuzzy soft sets [21]. Maji *et al.* [22] studied the soft set decision-making issue as a basis for classification implementation. Furthermore, Handaga [23] has suggested an extended classification approach, called the Fuzzy Soft Set Classifier (FSSC) based FSS, which uses the two soft sets' similarity. As compared to soft set classification based on decision-making problems, FSSC has low computational complexity and a high degree of accuracy.

Based on these findings, this study's main objective was to investigate the FSS to classify phishing websites. We hope to get early detection of phishing activity from the results of this study. A classification model is constructed using a feature set. For instance, in this case, web page information is required, such as URLs and network features. These features and classification or machine learning techniques collection in this category could be extracted [16]. The best feature sets are identified with high demands when mined. Thus, the prediction accuracy of classifiers can be improved [15].

The thwarting phishing attack studies are currently challenging, while researchers focus on phishing attack prevention and identification. Therefore, In this paper, we proposed a novel approach to phishing website detection. In this study, we choose the complete Classification of anti-phishing solutions as the research methodology. The experiments conducted to explore fuzzy soft set (FSS) at several similarities focus on determining the phishing dataset's classification performance. This paper also describes the basic theory and definitions of fuzzy set (FS), soft set (SS), fuzzy soft set (FSS) [24], Similarity measure, and Classification. In addition, FSS and new related results are presented, and open-ended questions are provided for further investigation.

2. Method

2.1. Fuzzy Soft Set

This part is intended only to introduce the main definitions and preliminaries which was used in the sequel in the following set theory's extensions, respectively: fuzzy set (FS), soft set (SS), soft matrix (SM), and fuzzy soft set (FSS).

Definition 2.1 Fuzzy set (FS) [25]: Given U as the universal set of point or object spaces. The set characterized by function $f_X: U \rightarrow [0,1]$ as a fuzzy set (class) X upward U. Furthermore, f_X defines a membership function, the fuzzy set X as an indicator function, and the value of $f_X\ (u)$ as the membership grade of $u \in U$ in X. A fuzzy set X over U (a universal set) could be written as in (1).

$$X = \{(f_X\ (u)/u): u \in U, f_X\ (u) \in [0,1]\} \tag{1}$$

Definition 2.2 Soft set (SS) [26][27]: Given U and E are a universal set and a set of parameters, respectively. Suppose that $A \subseteq E$, the formula $P\ (U) = 2^U$ is used to express the power set of U, then a pair (F, A) is to express the soft set of U, and is defined F_A as the set of ordered pairs (2).

$$F_A = \{\big(\big(e, F_A\ (e)\big): e \in E, F_A\ (e) \in P\ (U)\} \tag{2}$$

where F is the mapping that formulates by $F: A \rightarrow P\ (U)$. The support of F_A is A where $F_A\ (e) \neq \phi$, $\forall e \in A$ and $F_A\ (e) = \phi\ l\ \forall e \notin A$. It could be defined as the relatives parameters of the set U is the soft set (F, A) of U.

Example 2.1 The problem of making a decision to buy a car is given based on the "attractiveness of the car," which is then expressed as a soft set (F, E). Assume that the universal set U contains five cars (c), denoted as $\mathrm{U} = \{\mathrm{c}_1, \mathrm{c}_2, \mathrm{c}_3, \mathrm{c}_4, \mathrm{c}_5\}$, and $\mathrm{E} = \{\mathrm{e}_1, \mathrm{e}_2, \mathrm{e}_3\}$ with $\mathrm{e}_\mathrm{i}(\mathrm{i} = 1, 2, 3)$ were the notation used to express the parameters in the meaning of the words: "beautiful", "expensive", and "luxurious", respectively. Furthermore, the soft set (F, E) over U could be written in the relation: $(\mathrm{F}, \mathrm{E}) = \{(\mathrm{e}_1, \{\mathrm{c}_1, \mathrm{c}_2\}), (\mathrm{e}_2, \{\mathrm{c}_1, \mathrm{c}_3, \mathrm{c}_4\}), (\mathrm{e}_3, \{\mathrm{c}_1, \mathrm{c}_2, \mathrm{c}_5\})\}$. The description form of this soft set is presented in Table 1.

Table 1. The soft set (F,E) representation form

U	e_1	e_2	e_3
c_1	1	1	1
c_2	1	0	1
c_3	0	1	0
c_4	0	1	0
c_5	0	0	1

Definition 2.3 Soft matrix (SM) [28]: Given a soft set over U (a universal set), namely $(\mathrm{F}_\mathrm{A}, \mathrm{E})$. Then a subset of $\mathrm{U} \times \mathrm{E}$ is defined as R_A (3) that described a relation from $(\mathrm{F}_\mathrm{A}, \mathrm{E})$.

$$\mathrm{R}_\mathrm{A} = \{(\mathrm{u}, \mathrm{e}): \mathrm{e} \in \mathrm{A}, \mathrm{u} \in \mathrm{F}_\mathrm{A}(\mathrm{e})\} \tag{3}$$

Then, the of R_A characteristic function as $\eta_{\mathrm{R}_\mathrm{A}}: \mathrm{U} \times \mathrm{E} \rightarrow \{0,1\}$, where

$$\eta_{R_A} = \begin{cases} 1, & if\ \ (u, e) \in R_A \\ 0, & if\ \ (u, e) \notin R_A. \end{cases} \tag{4}$$

Table 2 shows the R_A produced by $U = \{u_1, u_2, \cdots, u_n\}$ and $A \subseteq E = \{e_1, e_2, \cdots, e_m\}$.

Table 2. Tabular representation of R_A.

R_A	e_1	e_2	$\cdots$	e_m
u_1	$\eta_{R_A}(u_1, e_1)$	$\eta_{R_A}(u_1, e_2)$	$\cdots$	$\eta_{R_A}(u_1, e_m)$
u_2	$\eta_{R_A}(u_2, e_1)$	$\eta_{R_A}(u_2, e_2)$	$\cdots$	$\eta_{R_A}(u_2, e_m)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
u_n	$\eta_{R_A}(u_n, e_1)$	$\eta_{R_A}(u_n, e_2)$	$\cdots$	$\eta_{R_A}(u_n, e_m)$

Let $a_{ij} = \eta_{R_A}(u_i, e_j)$, with $i = \{1,2,\ldots,n\}$, and $j = \{1,2,\ldots,m\}$. We can define the soft matrix of order $n \times m$ of the soft set (F_A, E) over U in the form as in (2).

$$[a_{ij}]_{n\times m} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix} \tag{5}$$

Example 2.2 Based on Example (2.1), the soft matrix of the soft set is written as

$$[a_{ij}]_{5\times 3} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Definition 2.4 Fuzzy soft set (FSS) [29][30]: Given U and E as a universal set and a set of parameters. Assume $\mathrm{A} \subseteq \mathrm{E}$, A pair (F, A) states the a fuzzy soft set (FSS) over U, with mapping F is formulated as $\mathrm{F}: \mathrm{A} \rightarrow \mathcal{F}(\mathrm{U})$, and $\mathcal{F}(\mathrm{U})$ defines the power set of fuzzy sets of U. The fuzzy subset of U is formulated as map $\mathrm{f}: \mathrm{U} \rightarrow [0,1]$.

Example 2.3 Suppose the interval [0,1] instead of 0.1. as a membership function that assigns a real number to each element. It can characterize Example 2.1. We can write $(\mathrm{F}, \mathrm{E}) = \{\mathrm{F}(\mathrm{e}_1) = \{(\mathrm{c}_1, 0.2), (\mathrm{c}_2, 0.7)\}, \mathrm{F}(\mathrm{e}_2) = \{(\mathrm{c}_1, 0.6), (\mathrm{c}_3, 0.8), (\mathrm{c}_4, 0.4)\}, \mathrm{F}(\mathrm{e}_3) = \{(\mathrm{c}_1, 0.3), (\mathrm{c}_2, 0.5), (\mathrm{c}_5, 0.9)\}\}$ as FSS the FSS (F, E) over U. The FSS representation is presented in Table 3.

Table 3. The FSS (F, E) representation.

U	e_1	e_2	e_3
c_1	0.2	0.6	0.3
c_2	0.7	0	0.5
c_3	0	0.8	0
c_4	0	0.4	0
c_5	0	0	0.9

2.2. Similarity measures

2.2.1. Matching function

In this section, the fuzzy soft set (FSS) is redefined for larger computational facilities. It is also U, and E are finite. Furthermore, we define an FSS as follows:

Definition 2.5 Given universal set U and a set of parameter E. Suppose that the collection of all fuzzy subsets of U is written with the notation U^I. An FSS over U is stated as a pair (F, E), with F is a mapping formulated by $F : E \rightarrow U^I$.

Basically, definitions 1 and 4 are the same if we take the exact subset A from E and assign the e-approximation $F(e) = \underline{0}\ \forall e \in E \setminus A$, then the FSS (F, A) and (F, E) has the same meaning. We can formulate an FSS over U as a matrix. An example is given to illustrate this process. Look again at example 2.2. In the fuzzy membership matrix, the $(i, j)^{th}$ entry is filled the value of membership $F(e_i)(e_j)$ if $e_i \in A$, and it is equal to 0 if $ei \notin A$. Therefore, a fuzzy membership matrix can be written as:

$$\hat{A} = \begin{bmatrix} 0.5 & 1.0 & 0 & 0 \\ 0.9 & 0.8 & 0 & 1.0 \\ 0 & 0.7 & 0 & 0 \\ 0 & 0 & 0.6 & 0 \\ 0 & 0 & 0 & 0.3 \end{bmatrix}$$

Based on the mentioned interpretation above, Matrix A represents FSS (F, A) and could be written as $(F, A) = \hat{A}$. It is clear that (F,A) has complement $(F,A)^C$ that could be expressed by another matrix $\hat{B}$ as follows:

$$\hat{B} = \begin{bmatrix} 0.5 & 0 & 1.0 & 1.0 \\ 0.1 & 0.2 & 1.0 & 0 \\ 1.0 & 0.3 & 1.0 & 1.0 \\ 1.0 & 1.0 & 0.4 & 1.0 \\ 1.0 & 1.0 & 0 & 0.7 \end{bmatrix}.$$

Further, the fuzzy membership matrix column is denoted by the vector $\vec{F}(e_1)$ or directly could be written as $F(e_1)$, e.g., the vector $\vec{F}(e_1)= F(e_1)= (0.5,0.9,0,0,0)$ retrieved from the Matrix $\hat{A}$. Furthermore, we define the similarity measure based on the match function.

Definition 2.6 Given two fuzzy softs over U, (F, E) and (G,E). The similarity between them, denoted by $S(F,G)$ or $S_{F.G}$ is formulated by (3).

$$\Sigma(\Phi,\Gamma)= S_{F.G} = S_{F.G} = \frac{\sum_{i=1}^{n} \{\vec{F}(e_i)\cdot\vec{G}(e_i)\}}{\sum_{i=1}^{n} \{(\vec{F}(e_i))^2 \mathrm{v} \overrightarrow{(G}(e_i))^2\}} \tag{6}$$

An example is given below to illustrate the Definition 2.6.

Example 2.4 Given two fuzzy soft sets over U, (F,E) and (G,E), where $U = \{x_1, x_2, x_3, x_4, x_5,\}$ and $E = \{e_1, e_2, e_3, e_4,\}$. The fuzzy membership matrix is written as :

$$\hat{A} = \begin{bmatrix} 0.2 & 0.5 & 0 & 0 \\ 0.7 & 0.3 & 0 & 0.7 \\ 0 & 1.0 & 0 & 0 \\ 0 & 0 & 0.95 & 0 \\ 0 & 0 & 0 & 0.2 \end{bmatrix} \text{and } \hat{B} = \begin{bmatrix} 0.13 & 0.4 & 0 & 0 \\ 0.6 & 0.1 & 0.1 & 0.4 \\ 0.1 & 0.3 & 0.3 & 0 \\ 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 0.1 \end{bmatrix}$$

Then $S(F,G)= S_{F.G} = S_{F.G} = \dfrac{\sum_{i=1}^{4} \{\vec{F}(e_i)\cdot\vec{G}(e_i)\}}{\sum_{i=1}^{4} \{(\vec{F}(e_i))^2 \mathrm{v} \overrightarrow{(G}(e_i))^2\}} \cong 0.617.$

Proposition 2.1 Given two fuzzy soft sets over U, (F,E) and (G,E). The following holds: (i) $S_{F.G} = S_{G.F}$, (ii) (F,E)= (G,E) $\Rightarrow S_{F.G} = 1$, (iii) (F,E) $\cap$ (G,E) = $\phi \Leftrightarrow S_{F.G} = 0$, and (iv) if (F,E) $\subseteq$ (H,E) $\subseteq$ (G, E), then $S_{F.G} \leq S_{H.G}$ Proof. Trivial.

2.2.2. Similarity measure

Given $U = \{x_1, x_2, \dots\dots.., x_n,\}$ and $E = \{e_1, e_2, \dots\dots\dots, e_m\}$. Suppose that two FSS's over (U,E). are $\hat{F} = (F,E)$ *and* $\hat{G} = (G,E)$. The e_i-th approximations of $\hat{F}$ is denoted by $F(e_i)$ and is formulated as $\hat{F} = \{F(e_i) \in P(U)e_i;\ \in E\}$, while $\hat{G} = \{G(e_i) \in P(U); e_i\ \in E\}$ with $G(e_i)$ is the e_i -th approximations of $\hat{G}$. The notation of $P(U)$ is stated all fuzzy subsets of U collection.

Given the similarity within the soft $\hat{F}$ *and* $\hat{G}$ as $M\ (\hat{F},\ \hat{G})$. Calculate the *e* – approximations to determine the similarity between $\hat{F}$ and $\hat{G}$. To do that, We defines $M_i\ (\hat{F},\ \hat{G})$ to state the similarity between the two e_1 approximations $F(e_1)$ *and* $G(e_1)$.

Definition 2.7 Let us define $\mathrm{M_i}\ (\hat{\mathrm{F}}, \hat{\mathrm{G}})$ as in (4).

$$\mathrm{M_i}\ (\hat{\mathrm{F}}, \hat{\mathrm{G}}) = \frac{\sum_{j=1}^{n} (\mathrm{F_{ij}} \wedge \mathrm{G_{ij}})}{\sum_{j=1}^{n} (\mathrm{F_{ij}} \vee \mathrm{G_{ij}})} \tag{7}$$

where $\mathrm{F_{ij}} = \mathrm{F(e_1)\ (x_1)} \in \mathrm{I}$ and $G_{ij} = G(e_1)\ (x_1) \in I$. *Then* $M_{F.G} = M\ (\hat{F},\ \hat{G}) = max_i\ M_i\ (\hat{F},\ \hat{G})$. The definition could be illustrated by Example 2.5.

Example 2.5 Examine the two FSS with $U = \{x_1, x_2, x_3, x_4\}$ and $E = \{e_1, e_2, e_3, e_4\}$:

$$\hat{F} = \begin{pmatrix} 0.2 & 0.5 & 0.9 & 1.0 \\ 0.1 & 0.2 & 0.6 & 0.5 \\ 0.5 & 0.4 & 0.3 & 0.2 \\ 0.1 & 1.0 & 0.3 & 0.4 \end{pmatrix} \text{ and } \hat{G} = \begin{pmatrix} 0.4 & 0.3 & 0.2 & 0.9 \\ 0.6 & 0.5 & 0.2 & 0.1 \\ 0.4 & 0.3 & 0.2 & 0.1 \\ 1.0 & 0.9 & 0.8 & 0.7 \end{pmatrix}$$

Then, $M_1 = \frac{0.8}{2.5} = 0.32.\ M_2 = 0.71.\ M_3 = 0.35,\ M_4 = 0.63$

Hence $M_{F.G} = \max\ \{M_1, M_2, M_3, M_4\} = M_2 = 0.71$

Proposition 2.2 Given two FSS over $(U,\ E)$ as $\hat{F} = (F, E)$*and* $\hat{G} = (G,E)$. Then, the conditions apply

(i) $M_{F.G} = M_{G.F}$,

(ii) $\hat{F} = \hat{G} \Rightarrow M_{G.F} = 1$,

(iii) $\hat{F} \cap \hat{G} = \phi \Leftrightarrow M_{G.F} = 0$, and

(iv) $\hat{F} \subset \hat{H} \subset \hat{G} \Rightarrow M_{F.G} \leq M_{H.G}$.

Proof. Proven by definition becomes easier.

Note 2.1 Also here $M_{F.G} = I$ did not imply $\hat{F} = \hat{G}$.

2.2.3. Similarity measure based distance

Given two fuzzy sets denoted as A and B. If the distance between the two sets is d, the similarity between them can be formulated as $S = \frac{1}{1+d}$. Again, an FSS is a group of its fuzzy sets' *e – approximations*. Furthermore, the distance between two fuzzy sets can be defined as

$$d_{(A,B)} = max_i\ |a_1 - b_2| \tag{8}$$

where $A = (a_1, a_2 \dots, a_n)$, and $B = (b_1, b_2 \dots, b_n)$. Then, the similarity between A dan B will be formulated as $T(A,B) = \frac{1}{1+d\ \infty\ (A,B)}$.

Now, suppose that $(F, E) = \{F(e_i), i\ =\ 1,2 \cdots, n\}$ and pair $(G, E) = \{G(e_i), j\ =\ 1,2, \cdots, n\}$ are two FSS, where $F(e_i)$ is the e_i -th *approximations* of (F,E) and $G(e_i)$ is the e_i -th *approximations*

$F(e_i)$ *and* $G(e_i)$. So $T_i(F,G) = \frac{1}{1+d_\infty^1}$, where d_∞^1 as the distance between $F(e_i)$ and $G(e_i)$ approximations. The similarity measure $T\,(F,G)$ between (F,E) and (G,E) is formulated as $T(F,G) = min_1\, T_1(F,G)$.

Example 2.6 Consider the two fuzzy soft sets $F$and G, with $U = \{x_1, x_2, x_3\}$ and $E = \{e_1, e_2, e_3\}$:

$$F = \begin{pmatrix} 0.2 & 0.9 & 1.0 \\ 1.0 & 0.1 & 0.5 \\ 0.4 & 0.2 & 0.4 \end{pmatrix} \text{ and } G = \begin{pmatrix} 0.6 & 0.9 & 0.1 \\ 0.7 & 1.0 & 0.5 \\ 0.1 & 1.0 & 0.4 \end{pmatrix}$$

Then, d_∞^1 = 0.4, d_∞^2 = 0.9 and d_∞^3 = 0.9. Hence $T_1 = \frac{1}{1+0.4}$ = 0.71, $T_2 = \frac{1}{1+0.9}$ = 0.53 and $T_3 = \frac{1}{1+0.9}$ = 0.53.
$\therefore T_{F,G} = \min_i\, T_i = 0.53$

Proposition 2.3 Given two FSS over (U,E) as (F,E) and (G,E). Then the relation can be applied as:

(i) $T_{F,G} = T_{G,F}$,

(ii) (F,E) = (G,E) ⇔ $T_{F,G} = 1$,

(iii) $\hat{F} \subset \hat{H} \subset \hat{G} \Rightarrow T_{F,G} \leq T_{H,G}$, for any soft set (H,E) over (U,E).

Note 2.2 The following properties do not apply here:

(i) $\hat{F} \cap \hat{G} = \Phi \Leftrightarrow T_{F,G} = 0$

3. Results and Discussion

3.1. Fuzzy Soft set classification

The steps of the classification algorithm consist of the learning (training) and classification step. Before the two steps are done, firstly, fuzzification and formation of the fuzzy soft set are applied. These two steps yield all data's feature vectors as well as the training and testing dataset. The data set is split into two parts which are used and testing training and testing. Each experiment splits the data randomly into nine different percentages of training and testing data as the data training and testing sample size variations, respectively, as shown in Table 4. The training aims to produce a fuzzy soft as each class fixed model. The data will be learned based on the data class group [31]. The Learning step is to obtain each class center. Data $U = \{u_1, u_2, \dots, U_N\}$, there is C class of data with $n_r; r = 1,2,\dots,K$ data of each class where $\sum_{r=1}^{K} n_r = N$, and $A \subseteq E, A\{e_i, i = 1,2,\dots,m\}$ with E is a set of parameters,. Suppose the set of r-th class FSS as F_{C_r}. Then the class center vector is denoted as P_{C_r} can be defined as in (9).

$$P_{C_r} = \frac{1}{n_r} \sum_{j=1}^{n_r} \mu_{C_r(e_i)}(u_1), i = 1,2,\dots,m; r = 1,2,\dots,k. \quad (9)$$

Classification is a technique for assigning unknown data to a target class. The new data generated by the training phase will be used to evaluate the classes in the new data, specifically by comparing two sets of acquired class center vector fuzzy soft sets and the new data. This comparative study uses the formula for similarity measure (10).

$$S\left(F_{P_{C_r}}, \mathrm{F}_G\right) = 1 - d_i\left(F_{P_{C_r}}, \mathrm{F}_G\right) \quad (10)$$

where d_i is the similarity and distance measure that have been discussed, i.e., Similarity measure, Distance measure, Matching function, and Comparison table.

After obtaining each class similarity value, it will determine which class label is most suitable for the new data F_G by calculating the maximum value of the similarity result for all classes. The class label could be written as in (11).

$$label_{clas} = \arg\left[\max_{r=1}^{k} S\left(F_{P_{C_r}}, F_G\right)\right] \quad (11)$$

Table 4. The composition training and testing dataset

Case	Training (%)	Testing (%)
1	60	10
2	60	20
3	60	30
4	60	40
5	70	10
6	70	20
7	70	30
8	80	10
9	80	20

3.2. Computational experiment

The algorithm, for experimentation, is built in MATLAB R2016a (9.0.0.34136) version that runs on an Intel Core i5 1.80GHz processor and of 8GB RAM under macOS High Sierra 10.13.1 operating system. A fuzzy soft set (FSS) algorithm was used to measure the algorithm's precision, recall, and response times when running the experimental datasets. The result is summarized and shown in Fig. 1 to Fig. 3. Fig. 1 shows that the accuracy results. It can be seen that the FSS based Similarity measure has the best performance than the other measurement. Meanwhile, the lowest one is based on a comparison table.

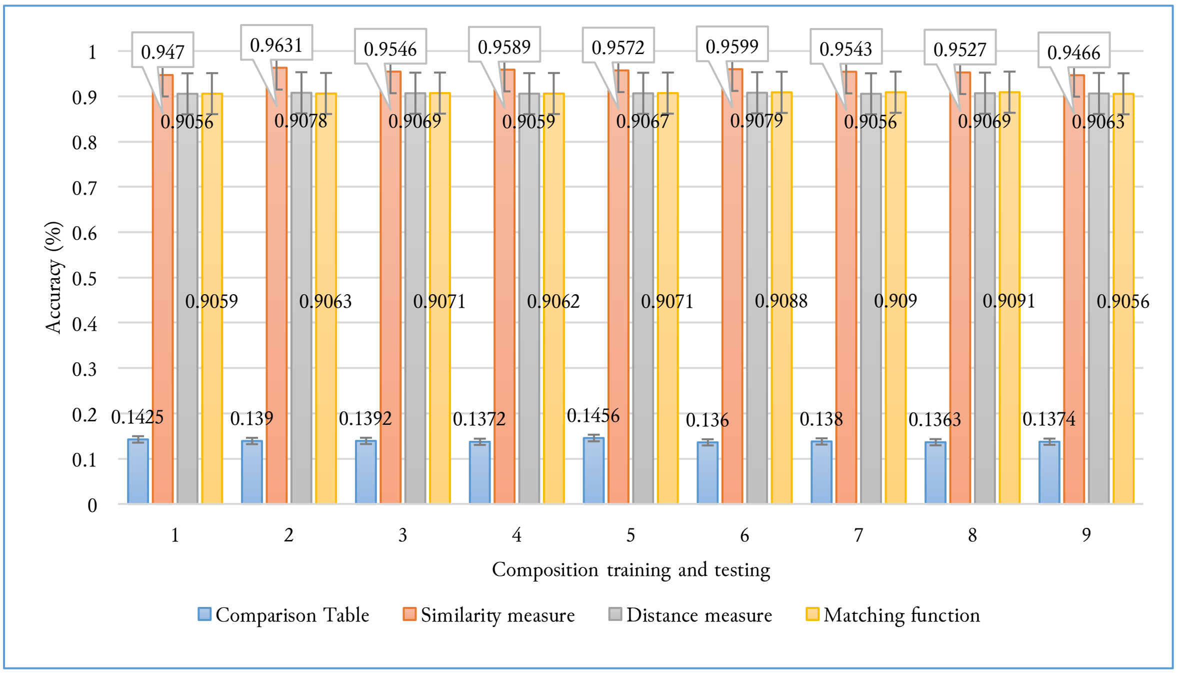

Fig. 1. Results of Accuracy.

Fig. 2 shows that the highest recall is FSS based on the Similarity measure. It proved that the Similarity measure could select the most widely relevant item to predict the phishing case with the highest accuracy. Even though, refer to the response time shown in Fig. 3, the Similarity measure placed on the second faster than the others. However, the time response for each similarity measure in this experiment is almost similar, i.e., up to 0.45 on average.

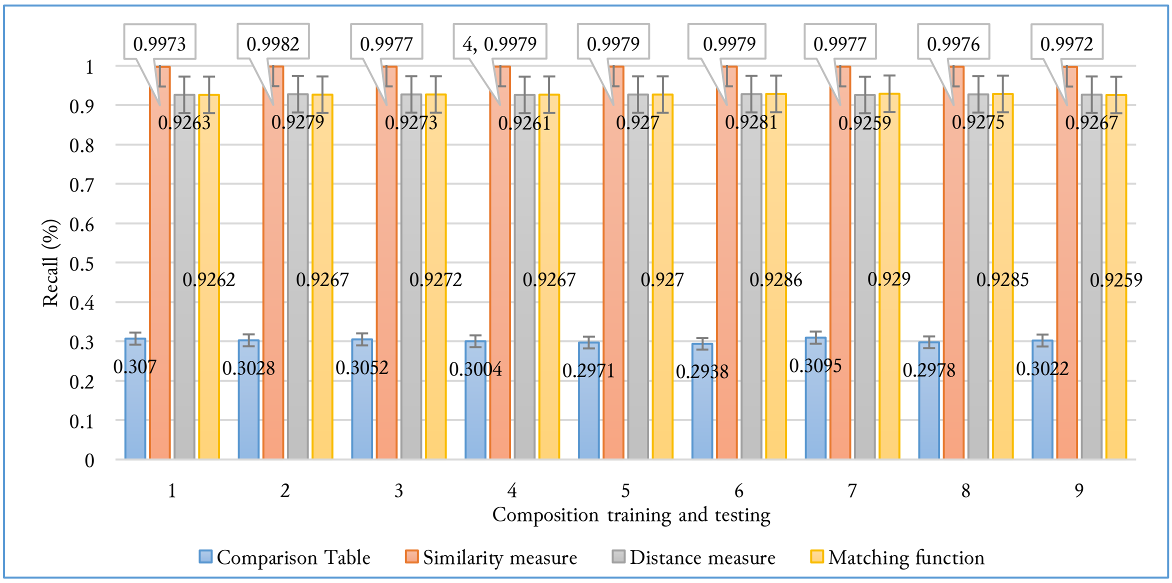

Fig. 2. Results of Recall.

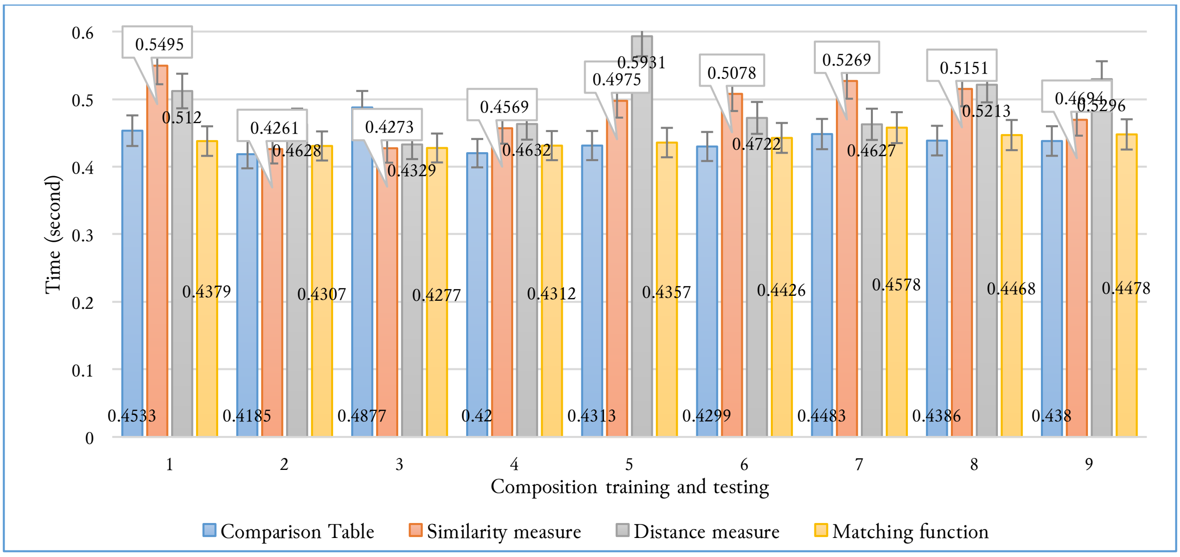

Fig. 3. Results of Response Time.

The overall average of all techniques in terms of accuracy, recall, and timely response is summarized in Table 5. It shows that the Similarity measure based on FSS has good performance raising to 0.9549 and 0.9977 in accuracy dan recall. This result concludes that the Similarity measure has the best precise of the other measurements, although its response time was not better than Matching Function.

Table 5. The summarized measurement results

Measurement	Accuracy	Recall	Response times
Comparison Table	0.139	0.3018	0.4406
Similarity Measure	**0.9549**	**0.9977**	0.4863
Distance Measure	0.9066	0.927	0.4944
Matching Function	0.9072	0.9273	**0.4398**

4. Conclusion

In this article, we have carried out an analysis of the proposed technique. Phishing data collection on web pages and important application areas in web mining are part of Data Classification. Some similarity measures based on a fuzzy soft set have been applied to the phishing dataset. The experimental results based on the accuracy and recall show that the best classifier is the Fuzzy soft set (FSS) based Similarity measure. It means that FSS has a promising approach in phishing detection in this study, although its response time was not better than the Matching Function. Future work could also include a hybrid classification model combining multiple web mining techniques such as attribute selection and grouping.

Acknowledgment

The authors are grateful to all those who have helped in completing this paper. The first author is obliged to supervisors for their supervision during Ph.D study at Post-Graduate, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia. This research is funded by Universiti Tun Hussein Onn Malaysia under Fundamental Research Grant Scheme (FRGS) with code FRGS/1/2019/ICT01/UTHM/02/2.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding statement. Universiti Tun Hussein Onn Malaysia funds this research under Fundamental Research Grant Scheme (FRGS) grant with code FRGS/1/2019/ICT01/UTHM/02/2.

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper.

References

[1] E. O. Asani and A. A. Adegun, "Maximum Phish Bait: Towards Feature Based Detection of Phising Using Maximum Entropy Classification Technique," in *iSTEAMS Research Nexus Conferenc*, 2014. Available at: Google Scholar.

[2] G. Xiang, "Toward a phish free world: A feature-type-aware cascaded learning framework for phish detection." Carnegie Mellon University, 2013. Available at: Google Scholar.

[3] R. Islam and J. Abawajy, "A multi-tier phishing detection and filtering approach," *J. Netw. Comput. Appl.*, vol. 36, no. 1, pp. 324–335, Jan. 2013, doi: 10.1016/j.jnca.2012.05.009.

[4] R. Lucky, "Clickphobia [Reflections]," *IEEE Spectr.*, vol. 48, no. 1, pp. 25–25, Jan. 2011, doi: 10.1109/MSPEC.2011.5676377.

[5] FBI, "Business E-mail Compromise The 12 Billion Dollar Scam," *FBI Field Office*, 2018. [Online]. Available: https://www.ic3.gov/Media/Y2018/PSA180712.

[6] I. R. A. Hamid and J. H. Abawajy, "Profiling Phishing Email Based on Clustering Approach," in *2013 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, 2013, pp. 628–635, doi: 10.1109/TrustCom.2013.76.

[7] "Most phishing attacks come from one group," *Comput. Fraud Secur.*, vol. 2010, no. 5, p. 2, May 2010, doi: 10.1016/S1361-3723(10)70045-2.

[8] PhishLabs, "2018 Phishing Trends and Intelligence Report: Hacking the Human.", 2018, [Online]. Available: https://info.phishlabs.com/hubfs/2018 PTI Report/PhishLabs Trend Report_2018-digital.pdf.

[9] S. Chanti and T. Chithralekha, "Classification of Anti-phishing Solutions," *SN Comput. Sci.*, vol. 1, no. 1, p. 11, Jan. 2020, doi: 10.1007/s42979-019-0011-2.

[10] A. A. Akinyelu and A. O. Adewumi, "Classification of Phishing Email Using Random Forest Machine Learning Technique," *J. Appl. Math.*, vol. 2014, pp. 1–6, 2014, doi: 10.1155/2014/425731.

[11] N. Rtayli and N. Enneya, "Enhanced credit card fraud detection based on SVM-recursive feature elimination and hyper-parameters optimization," *J. Inf. Secur. Appl.*, vol. 55, p. 102596, Dec. 2020, doi: 10.1016/j.jisa.2020.102596.

[12] G. L. Gray and R. S. Debreceny, “A taxonomy to guide research on the application of data mining to fraud detection in financial statement audits,” *Int. J. Account. Inf. Syst.*, vol. 15, no. 4, pp. 357–380, Dec. 2014, doi: 10.1016/j.accinf.2014.05.006.

[13] S. Bagga, A. Goyal, N. Gupta, and A. Goyal, “Credit Card Fraud Detection using Pipeling and Ensemble Learning,” *Procedia Comput. Sci.*, vol. 173, pp. 104–112, 2020, doi: 10.1016/j.procs.2020.06.014.

[14] R. S. Moorthy and P. Pabitha, “Optimal Detection of Phising Attack using SCA based K-NN,” *Procedia Comput. Sci.*, vol. 171, pp. 1716–1725, 2020, doi: 10.1016/j.procs.2020.04.184.

[15] S. Nandhini and V. Vasanthi, “Extraction of Features and Classification on Phishing Websites using Web Mining Techniques,” *IJEDR*, vol. 5, no. 4, 2017. Available at: Google Scholar.

[16] G. Varshney, M. Misra, and P. K. Atrey, “A survey and classification of web phishing detection schemes,” *Secur. Commun. Networks*, vol. 9, no. 18, pp. 6266–6284, Dec. 2016, doi: 10.1002/sec.1674.

[17] A. Yasin and A. Abuhasan, “An Intelligent Classification Model for Phishing Email Detection,” *Int. J. Netw. Secur. Its Appl.*, vol. 8, no. 4, pp. 55–72, Jul. 2016, doi: 10.5121/ijnsa.2016.8405.

[18] Z. Dou, I. Khalil, A. Khreishah, A. Al-Fuqaha, and M. Guizani, “Systematization of Knowledge (SoK): A Systematic Review of Software-Based Web Phishing Detection,” *IEEE Commun. Surv. Tutorials*, vol. 19, no. 4, pp. 2797–2819, 2017, doi: 10.1109/COMST.2017.2752087.

[19] B. B. Gupta, A. Tewari, A. K. Jain, and D. P. Agrawal, “Fighting against phishing attacks: state of the art and future challenges,” *Neural Comput. Appl.*, vol. 28, no. 12, pp. 3629–3654, Dec. 2017, doi: 10.1007/s00521-016-2275-y.

[20] S. Purkait, “Phishing counter measures and their effectiveness – literature review,” *Inf. Manag. Comput. Secur.*, vol. 20, no. 5, pp. 382–420, Nov. 2012, doi: 10.1108/09685221211286548.

[21] R. Hidayat, I. Tri Riyadi Yanto, A. Azhar Ramli, M. Farhan Md. Fudzee, and A. Saleh Ahmar, “Generalized Normalized Euclidean Distance Based Fuzzy Soft Set Similarity for Data Classification,” *Comput. Syst. Sci. Eng.*, vol. 38, no. 1, pp. 119–130, 2021, doi: 10.32604/csse.2021.015628.

[22] P. K. Maji, A. R. Roy, and R. Biswas, “An application of soft sets in a decision making problem,” *Comput. Math. with Appl.*, vol. 44, no. 8–9, pp. 1077–1083, Oct. 2002, doi: 10.1016/S0898-1221(02)00216-X.

[23] B. Handaga, T. Herawan, and M. M. Deris, “FSSC: An algorithm for classifying numerical data using fuzzy soft set theory,” *Int. J. Fuzzy Syst. Appl.*, vol. 2, no. 4, pp. 29–46, Oct. 2012, doi: 10.4018/ijfsa.2012100102.

[24] I. T. Riyadi Yanto, E. Sutoyo, A. Rahman, R. Hidayat, A. A. Ramli, and M. F. M. Fudzee, “Classification of Student Academic Performance using Fuzzy Soft Set,” in *2020 International Conference on Smart Technology and Applications (ICoSTA)*, 2020, pp. 1–6, doi: 10.1109/ICoSTA48221.2020.1570606632.

[25] L. A. Zadeh, “Fuzzy sets,” *Inf. Control*, vol. 8, no. 3, pp. 338–353, Jun. 1965, doi: 10.1016/S0019-9958(65)90241-X.

[26] P. K. Maji, R. Biswas, and A. R. Roy, “Soft set theory,” *Comput. Math. with Appl.*, vol. 45, no. 4–5, pp. 555–562, Feb. 2003, doi: 10.1016/S0898-1221(03)00016-6.

[27] D. Molodtsov, “Soft set theory—First results,” *Comput. Math. with Appl.*, vol. 37, no. 4–5, pp. 19–31, Feb. 1999, doi: 10.1016/S0898-1221(99)00056-5.

[28] H. Aktaş and N. Çağman, “Soft sets and soft groups,” *Inf. Sci. (Ny).*, vol. 177, no. 13, pp. 2726–2735, Jul. 2007, doi: 10.1016/j.ins.2006.12.008.

[29] P. K. Maji, R. Biswas, and A. R. Roy, “Fuzzy soft sets,” *J. Fuzzy Math.*, vol. 9, no. 3, pp. 589–602, 2001, doi: 10.4236/am.2014.59127.

[30] Y. Celik, C. Ekiz, and S. Yamak, “Applications of fuzzy soft sets in ring theory,” *Ann. Fuzzy Math. Informatics*, vol. 5, no. 3, pp. 451–462, 2013. Available at: Google Scholar.

[31] J. Han, M. Kamber, and J. Pei, “Data Mining Trends and Research Frontiers,” 2012, pp. 585–631, doi: 10.1016/B978-0-12-381479-1.00013-7.