SN Computer Science (2020) 1:361
https://doi.org/10.1007/s42979-020-00377-8

ORIGINAL RESEARCH

Detecting Phishing SMS Based on Multiple Correlation Algorithms

Gunikhan Sonowal[1]

Received: 20 May 2020 / Accepted: 16 October 2020 / Published online: 2 November 2020

Abstract
The SMS phishing is another method where the phisher operates the SMS as a medium to communicate with the victims and this method is identified as smishing (SMS + phishing). Researchers promoted several anti-phishing methods where the correlation algorithm is applied to explore the relevancy of the features since there are numerous features in the features corpus. The correlation algorithm assesses the rank of the features that is the highest rank leads to the more relevant to the appropriate assignment. Therefore, this paper analyses four rank correlation algorithms particularly Pearson rank correlation, Spearman's rank correlation, Kendall rank correlation, and Point biserial rank correlation with a machine-learning algorithm to determine the best features set for detecting Smishing messages. The result of the investigation reveals that the AdaBoost classifier offered better accuracy. Further analysis shows that the classifier with the ranking algorithm that is Kendall rank correlation appeared superior accuracy than the other correlation algorithms. The inferred of this experiment confirms that the ranking algorithm was able to reduce the dimension of features with 61.53% and presented an accuracy of 98.40%.

Keywords Phishing · Smishing · Correlation Algorithm · Machine Learning Algorithm

Introduction

Phishing is an entirely crucial attack these days where attackers snitch the credentials from the users using social engineering with technologies [44, 46]. Social engineering is the practice of influence and persuasion to deceive victims for acquiring information or performing some operation [16, 38]. The United Nations reports 350% rise in phishing websites during the COVID-19 pandemic [48]. Currently, phishing is expanding rapidly and according to the report concerning the Anti-Phishing Working Group(APWG) [4], the number of unique phishing websites detected in January-June 2020 is shown in Fig. 1.

Nowadays, attackers employ numerous communication mediums to communicate with the victims such as email, text message(SMS), telephone, and others [5]. However, SMS is one of the feasible mechanisms to effectively communicate with others through mobile phones without the internet. It is generally accepted that every person possesses mobile phones and the number of mobile phone users was estimated at 5.15 billion in 2020 [14].

According to the CallHub, the response rate of 98% SMS messages is 45% in comparison to the email is 28-33% which indicates that the email is 6% lower response rate than SMS [7, 10]. Attackers exploit this service and sending the phishing SMS to the users which are similar to the legitimate SMS to steal the credentials [23, 37, 52]. According to [33], smishing is a variant of phishing in which attackers send instant messages instead of emails, which appear to have been sent by a genuine organization and demand that the clients tap on a link or disclose the credentials through the text message reply. It is well known that the SMS messages are less expensive, and with the modest SMS package, the phishers are capable of sending a substantial number of messages to the users [15]. As per the report of the security firm, Cloudmark, approximately 30 million smishing messages are dispatched to the mobile users across North America, Europe [22].

Smishing is effective research where several researchers are imparting methods to detect smishing messages. While most of the methods are particularly providing guidelines consisting of awareness of the unknown link in text messages and others. However, very few publications can be found in the literature that addresses the issue of detection of

✉ Gunikhan Sonowal
gunikhan.sonowal@gmail.com

[1] Department of Computer Science, Pondicherry University, Pondicherry, India

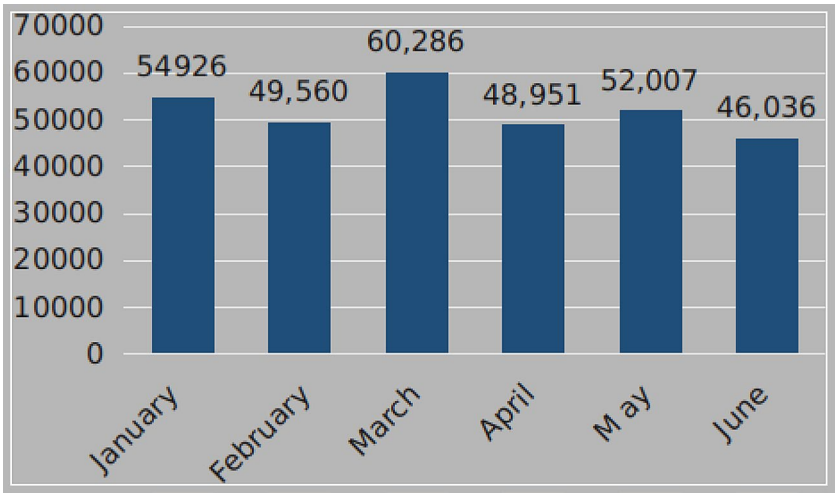

Fig. 1 Number of unique phishing Web sites detected

smishing messages. As a result, this paper proposed a model that detects smishing messages with a machine learning algorithm. The motivation of this paper is to determine the best feature set using more than one correlation algorithms. [47] proposed a model smidca which used the Pearson correlation algorithm with the machine learning algorithm to detect the phishing SMS. The weakness of the SmiDCA model was presented with low accuracy. Therefore, this paper applied several correlation algorithms: Pearson rank correlation, Kendall rank correlation, Spearman rank correlation, and the Point-Biserial correlation, and finds the best correlation algorithm which affords the superior accuracy.

This paper is organized as below: section "Literature Review" overviews the background works of smishing. Section "Methodology" explains the methodology of the paper. In section "Correlation Algorithms", the correlation algorithm will be discussed. Section "Experimental Analysis" experiments the proposed methodology and result is depicted. Section "Discussion" discusses the outcome of the result and the paper is concluded in section "Conclusion".

Literature Review

As referred to above that the attackers practiced several communication mediums to communicate with the victims and SMS is as well as an imperious medium where the phishers attempt mobile phone users. For this purpose, this paper primarily concentrates on the phishing SMS detector methods and recently several researchers are presenting advanced techniques to detect phishing SMS. This section overviews some of the anti-phishing SMS techniques in the remaining parts.

[19] proposed a mobile phone specific anti-phishing solution that distinguished the essential measure to improve the tactics to combat phishing assault on mobile. In this solution, the authors identified several versions of phishing attacks regarding mobile devices such as Bluetooth phishing, Smishing, Vishing, and mobile web application phishing, and so forth. This category of a model normally employs the heuristic-based approach where distinct features are extracted to detect phishing attacks.

It is known that the characters of the text messages are limited based on the communication protocol. Therefore, most of the attacker sends a short URL to victims, and it is difficult to verify which file or webpage the short URL interfaces to users. Therefore, [35] proposed a method that composes the destination information of the short URL. Furthermore, the method analyzed the webpage and measured the risk of the webpage and blocked the short URL by comparing with predefined threshold. The shortcoming of this model is the determination of the threshold.

Another method entitles S-Detector (Smishing detector) was proposed by [27] which differentiates the genuine messages from Smishing messages. This method initially investigated the presence of a URL in the messages. If the URL is detected in the message then the model verifies the URL whether the URL is a short URL or not. If the URL is short then converts into the long URL and verifies the APK file. If the file is present then terminates the investigating otherwise employed the morphological analyzer where the noun words are selected as features. The model implements the Naive Bayesian classifier to blocks the smishing messages and notifies the users regarding the smishing messages. The outcome of the investigation describes that this model provides protection, accessibility, and reliability in preventing shrewder and more malignant security threats. The weakness of this model is the usage of only keywords which is unsuitable for phone number email-id and other attacks.

Some more investigation was examined by [6] on the concept of the time duration of the spam messages per day and noticed that the attackers communicate the highest

peaks of spam messages from 10 am to 4 pm. Further analysis shows that the familiar words of the spam messages that are *candidate, congressmen, election, candidate number, and information* of the smishing messages. An operation was conducted by the authors using the contents of the spam sent by each spammer and based on the keywords present in the URL, they were able to classify the smishing messages from the legitimate messages. The model used only militated features which is insufficient for detecting all categories of the smishing messages.

In another study to recognize smishing messages [36], the authors assembled seven features and analyzed these features with a random forest classifier and the outcome of the experiment reveals that the classifier achieved the accuracy of 92%. Lee et al. [30] incorporated *Cloud based virtual environment* to identify the suspicious URL by verifying whether the URL possesses a position with the downloading APK record or an application without reference. The method further enhanced the probability of smishing identification by practicing the method.

Mishra and Soni [34] proposed a recent model that contains multiple filters where SMS Content Analyzer inspects the instant message substance. Naive Bayes Classification Algorithm arranges the malicious substance and keywords present in the instant message. URL Filter assesses the URL to recognize malicious features. Source Code Analyzer looks at the source code of the site to distinguish the unsafe code installed in it. Apk Download Detector distinguishes whether any malicious record is downloaded while conjuring the URL. The results of the analyses show an accuracy of 96.29%. Although the model employed multiple stages for detecting the smishing messages, it requires other classification algorithms for increasing accuracy.

One fashionable model smidca [47] which collected 39 features to distinguish the smishing messages. This model operated a random forest classifier with the features selection algorithm which achieved 96.16% accuracy and the model was capable to lessen the feature dimension over 40.71% with the help of feature selection algorithm. The weakness of this model is low accuracy.

The aforementioned smishing detection models primarily focus on the URL and the attacker employs the short URL to hide the malware file. The shortcomings of these models are the usage of the only URL and limited keywords. The accuracy of the above model is providing low accuracy. Therefore, this paper collected 52 features for detecting phishing SMS and four different ranking algorithms are used to rank the features. With the assistance of the AdaBoost classifier, the model evaluated the accuracy of 98.40% using Kendall rank correlation even though 61.53% features are pruned from the feature corpus.

Methodology

This paper consists of three components: Feature collection, Feature ranking, and searches the best feature set using a machine learning algorithm. The feature collection component collects features from the existing and novel features from the SMS and builds a feature vector. The feature vectors are sent to the ranking algorithm to search the relevancy of the features. In the ranking algorithm, the proposed model employs four correlation algorithms where the highest rank features indicate more relevant under the specific task. After arranging the features based on the ranking, the proposed model employs a sequential forward feature selection algorithm for searing the best feature set.

Feature's Collection

This paper collects 52 features from the SMS where 39 features from the paper [47] and 13 are novel features. The features which are used in this model as explained below:

- Bag of words (F1–F20): The phisher commonly employs some words to deceive the victims such as F1: please, F2: SMS, F3: Account, F4: Customer, F5: Card, F6: Email, F7: Apple, F8: Details, F9: Update, F10:iPhone, F11: Online, F12: Bank, F13: Link, F14: Message, F15: Call, F16: Store, F17: Today, F18: Nationwide, F19: Refund, and F20: Due.
- SMS size (F21): The size of SMS is a significant feature because the size is used in genuine SMS based on the organization standard. Hence, the phishing SMS and legitimate SMS are different in size.
- SMS consists of Email-id (F22): Email-id is another way of transferring information through the internet. The phisher assigns Email-id inside the SMS for demanding credentials.
- SMS consists of URL (F23): The phisher creates a phishing site by impersonating the genuine sites and the URL of the phishing sites is presented in the SMS for encouraging to visit the phishing sites.
- SMS consists of Phone Number (F24): The phisher registers phone number to communicate with the victims, and the phone number is sent with SMS for requesting credentials.
- SMS consist of Special character (F25): Many phishing SMS employs Special character to support the legitimacy of the SMS such as currency symbol and others.
- Misspelled words (F26): The words used in the genuine SMS are analyzed by experts and hence, the genuine organization avoids the misspelled words in the SMS.

However, most of the phishing SMS contains misspelled words.

- Number of Parts of Speech (F27–F33): Parts of speech are an important feature in smishing messages. The number of parts of speech that appear in phishing SMS is different from the legitimate SMS. This paper practices the English part of speech, such as F27: Noun, F28: Pronoun, F29: Adjective, F30: Verb, F31: Adverb, F32: Proposition, F33: Conjunction.
- Readability algorithms(F34–F39): The readability algorithm measures the level of understanding of the English text [43]. The text style of phishing messages and legitimate messages is different which leads to the discrimination features for detecting the phishing messages. Six algorithms are employed in the feature corpus which is explained below:
  - Automated readability index(F34): Smith and Senter [42] proposed the Automatic readability index for measuring the readability score. The equation of the automatic readability index is shown in equation (1)

$$\mathrm{ARI} = 4.71\left(\frac{L}{W}\right) + 0.5\left(\frac{W}{S}\right) - 21.43 \tag{1}$$

    where L be the number of letters and numbers, W is the number of spaces, and S is the number of sentences.
  - Flesch Reading-Ease Score and Flesch-Kincaid Grade Level (F35–F36): Flesch [18] proposed the Flesch-Kincaid Readability Test which for evaluating the difficulty of a text in English. Two tests are primarily conducted: Flesch Reading-Ease Score and Flesch-Kincaid Grade Level.

    F35: Flesch reading-ease score (FRES) test is shown in equation (2)

$$\mathrm{FRES} = 206.835 - 1.015\left(\frac{\mathrm{TW}}{\mathrm{TS}}\right) - 84.6\left(\frac{\mathrm{Tsy}}{\mathrm{TW}}\right) \tag{2}$$

    where TW be the total words, TS be the total sentence, Tsy be total syllables and Tsy be the total syllables

    F36: Flesch-Kincaid Grade Level (FKGL) is shown in Eq. (3)

$$\mathrm{FKGL} = 0.39\left(\frac{\mathrm{TW}}{\mathrm{TS}}\right) + 11.8\left(\frac{\mathrm{Tsy}}{\mathrm{TW}}\right) - 15.59 \tag{3}$$

  - Gunning Fog Index (F37): Gunning [21], an American businessman developed this readability test.

    The equation of the Gunning Fog Index is shown in Eq. (4)

$$\mathrm{GFI} = 0.4\left[\left(\frac{\mathrm{Words}}{\mathrm{Sentences}}\right) + 100\left(\frac{\mathrm{Complex\ Words}}{\mathrm{Words}}\right)\right] \tag{4}$$

  - SMOG Index (F38): Mc Laughlin [32] developed this SMOG index for testing the health messages. The equation of Smog to test readability score is shown in (5).

$$\mathrm{SMOG} = 1.0430\sqrt{\mathrm{TP} \times \frac{30}{\mathrm{TS}}} + 3.1291 \tag{5}$$

    where TP be the total number of Polysyllables and TS be the Total sentence.
  - Coleman Liau Index (F39): Coleman and Liau [13] developed the Coleman-Liau Index to calculate the readability score . The Coleman-Liau index (CLI) is shown in Eq. (6)

$$\mathrm{CLI} = 0.0588L - 0.296S - 15.8 \tag{6}$$

    L denotes the average number of letters per hundred words and S denotes the average number of sentences per hundred words.
- Character count (F40): The number of characters used in genuine SMS based on the communication protocol, unlike the phishing SMS.
- Number of the alphabet (F41): Most of the genuine SMS contains only alphabets unlike the phishing SMS contains alphanumeric, special characters, phone number, and others. Therefore, the number of phishing's alphabet is different from the legitimate alphabet.
- Number of uppercase letters (F42): The sentence is used in a genuine organization with a practiced proper uppercase letter, but the phishing SMS contains an abnormal uppercase letter.
- Number of Digits (F43): Digit is adopted by phishers for multiple purposes such as currency, winning prize, and others.
- Number of Spaces (F44): Space is related in SMS for separating between two words or numbers and the phishers employed unnecessary spaces, unlike the genuine SMS.
- Number of Punctuation Marks (F45–F52): The genuine organization constructs the sentence with proper punctuation marks in the SMS but the phishing SMS contains the unclear punctuation marks. This feature used eight punctuation marks such as F45:″,″, F46:″.″, F47:″?″, F48:″!″, F49:″ :″, F50:″;″, F51:″′″, F52:″″.

Correlation Algorithms

The term "correlation" is approved to evaluate the relationship between quantities[1]. Several correlation algorithms are employed to rank the features for evaluating the relevancy of the features to reduce the dimension of the feature corpus. This paper examines four kinds of correlations: Pearson rank correlation, Kendall rank correlation, Spearman rank correlation, and the Point-Biserial correlation [11, 12, 28]. Assume, $X = \{x_1, x_2, \ldots x_n\}$ is the feature vector and $Y = \{y_1, y_2, \ldots y_n\}$ is the decision vector. The correlation algorithm computes the rank of the feature by comparing the feature vector with the decision vector.

Pearson Correlation Coefficient

Pearson correlation coefficient is broadly admitted in the feature selection algorithm to determine the best feature set[47]. The PCC is defined by ρ and equation is shown in (7).

$$\rho(P, Q) = \frac{\text{cov}(P, Q)}{\sigma P, \sigma Q} \tag{7}$$

where $\text{cov}(P, Q)$ denotes the covariance of P, Q, and σP and σQ are the standard deviation of P and Q.

$$\text{cov}(P, Q) = \frac{1}{n-1} \sum_{i=1}^{n} (p_i - \bar{p})(q_i - \bar{q}) \tag{8}$$

where $\bar{p}$ and $\bar{q}$ are mean of P and Q.

Spearman Rank Correlation

Several researchers practiced Spearman rank correlation for feature selection algorithms [41, 50]. The equation of Spearman's coefficient is similar to the Pearson and the simplified version is shown in Eq. 9.

$$S(P, Q) = 1 - \frac{6 \sum_i^n (P_i - Q_i)^2}{n(n-1)} \tag{9}$$

where S Spearman rank correlation, n number of observations

Kendall Rank Correlation

Unlike Spearman's coefficient, Kendall's τ does not consider the difference between ranks—only directional agreement [20, 29]. The equation of the Kendall rank correlation is shown in the Eq. (10).

$$\tau = \frac{n_c - n_d}{0.5 * n(n-1)} \tag{10}$$

where n_c number of concordant, n_d number of discordant

Point Biserial Rank Correlation

The point-biserial correlation is related to the Pearson correlation equation except that one of the factors is dichotomous [9, 31]. The equation of the point biserial rank correlation is shown in the Eq. (11).

$$r_{rb}(X, Y) = \left(\frac{\bar{Y}_1 - \bar{Y}_0}{S_y} \right) \sqrt{\frac{n\bar{X}(1-\bar{X})}{n-1}} \tag{11}$$

where S_y is standard deviation, $\bar{Y}$, $\bar{X}$ are mean values.

Machine Learning Algorithm

The Machine learning algorithm is widely studied in the SMS classification. Numerous classification algorithms are applied with a specific end goal to recognize phishing SMS. This paper performs four well-known classifiers such as AdaBoost, random forest, Decision Tree, and Support Vector Machine.

Decision Tree The decision tree applies both categorical and continuous input and it separates the data into two or more homogeneous sets based on the most important splitter in input features [49]. The feature (attribute) in the decision tree is described by each node, the decision is accepted by each link (branch) and the result (discrete or continuous value) is evaluated by each leaf. The weakness of the decision tree is the determination of the feature for the root node in each level which is known as feature selection. Therefore, two major feature selection algorithms are adopted to determine the root: Information Gain (IG) and Gini Index (GI).

Random forest Algorithm The random forest algorithm is an ensemble classification algorithm; that is, a gathering of classifiers [1, 2, 8, 24, 45]. Rather than using only one classifier to foresee the target, in an ensemble, various classifiers to anticipate the target. In the random forest, these ensemble classifiers are the arbitrarily generated decision trees and each decision tree is a single classifier and the target prediction depends on the majority voting technique. Therefore, the target class receiving the majority number of votes regards as the final predicted target class.

Support Vector Machine (SVM) The support vector machine is operated based on the concept of locating a hyperplane that maximizes the margin between the two classes [17, 25, 51]. The vectors that represent the hyperplane are the support vectors. A hyperplane is a decision plane that divides the set of different classes and the margin is a gap between two lines which is computed using the

[1] https://www.statisticssolutions.com/correlation-pearson-kendall-spearman/

Table 1 English text messages

Total SMS	Phishing SMS	Ham SMS
5578	747	4831

perpendicular distance from the line to support vectors. If the margin is larger in between the classes, then it is considered an acceptable margin, a smaller margin is an unacceptable margin.

AdaBoost Boosting is a general ensemble technique that produces a reliable classifier from several weak classifiers. Hence, the AdaBoost is a boosting algorithm developed for binary classification and best utilized to promote the execution of decision trees [26, 39, 40]. A weak classifier is set up on the training data using the weighted samples; therefore, each decision stump settles on the decision on one input variable and outputs a + 1.0 or − 1.0 value for the first or second class value. The misclassification rate is computed from the trained model.

Feature Search Algorithm

Although, feature selection algorithms are implemented for multiple objectives such as enhanced accuracy, decreases complexity, faster training for machine learning algorithms, and others. However, this paper primarily operates the machine learning algorithm to improve accuracy. Once the features are ranked using the ranking algorithm, the features are allotted to the search algorithm to search the best feature set. This paper employs the sequential forward feature selection algorithm to explore the best feature set. The sequential forward feature selection combines features one by one to the features set according to the highest rank orders. However, the limitation of this algorithm is the termination point, otherwise, the algorithm would continue until the end of the features. Therefore, this paper adopted the policy defining by Sonowal and Kuppusamy [47] that if the classifiers provide constant accuracy or less accuracy continuously three times then the algorithm would terminate.

Experimental Analysis

Data Collection

This paper gathered data on phishing and Ham SMS from Tiago A. Almeida [3] is shown in Table 1. This data was employed for several machine learning methods in order to verify the performance of the proposed methods.

Table 2 Selecting the classifier

classifier	Precision	Recall	F1-scores	Accuracy
Random Forest	98.39	91.42	94.72	98.66
DecisionTreeClassifier	93.4	92.49	92.91	98.12
AdaBoostClassifier	97.75	92.23	94.86	98.67
Support Vector Machine	97.11	92.76	94.79	98.64

Table 3 Comparative analysis of different correlation algorithms

Rank algorithm	Number of features	Precision	Recall	F1-scores	Accuracy(%)
Pearson	21	95.49	90.21	92.72	98.12
Spearman's	18	96.6	91.02	93.67	98.37
Kendall rank	20	96.49	91.42	93.97	98.40
Point biserial	21	95.49	90.21	92.72	98.12

Experimental Result

The experiment is carried out with three steps: the first step is to determine the best classifier, the second step is to find the relevant feature using a correlation algorithm and the third step is to evaluate the best feature set using the best classifier with relevant features.

Once the features are obtained, the model initially selected all the features in the first step and experimented with the four well-known classifiers that are Random forest, decision tree, AdaBoost, and support vector machine to explore the best classifier as explained in section "Machine Learning Algorithm". Table 2 shows that the AdaBoost performed slightly better accuracy than other classifiers. Therefore, AdaBoost is selected for further experiments.

In the second step, the model applies the correlation algorithm to rank the features which imply that more rank produces more relevant to the particular assignment. As defined above that this paper adapted four types of correlation algorithms as explained in section "Correlation Algorithms". The different correlation algorithms recognize the different features as relevant.

The model employs the sequential forward feature selection algorithm to determine the best features set in the third step. The sequential forward feature selection algorithm takes the feature one by one based on the ranked of the features and evaluates the accuracy with the best classifier as explained in section "Feature Search Algorithm". The model evaluated the accuracy separately of all the correlation algorithms.

Table 3 shows that the diverse correlation algorithm's accuracy based on their ranking of the features. If the

Table 4 Comparative analysis with other methods

Methods	Number of features	Accuracy(%)
Smidca: anti-smishing [47]	20	96.16
Distributed System [36]	13	92.00
The proposed model	20	98.40

accuracy of the experiment is examined then it was observed that all the algorithms produced equivalent accuracy. However, a closer inspection revealed that the Kendall rank correlation offered slightly better accuracy(98.40%).

Furthermore, the number of features is additionally an imperative aspect of the methodology. The table demonstrates that Spearman's rank correlation used only 18 out of 52 features which indicates that this correlation reduced the features corpus (65.38%), while, the accuracy is slightly lesser than Kendall rank correlation. In the event, it has been noticed the rate reduction of features for Kendall then it has found that the (61.53%) are features have been pruned.

The features are selected by Kendall rank < F24, F39, F15, F34, F38, F41, F23, F40, F44, F45, F30, F1, F32, F4, F25, F27, F37, F21, F33, F43>. The present findings have important implications for solving this problem because of the recently added features such as <F38, F40, F44, F45, F43> which assist to improve the accuracy of the model.

To verify the result of the experiment, the outcome of the proposed method is compared with the other existing methods. The result of the comparative analysis with other methods is shown in Table 4. The result shows that the proposed method provided better accuracy in contrast with other methods. Further comparison, it can be seen that the proposed model required the same number of features as SmiDCA but provided better accuracy. From the result, it can be concluded that the proposed method has the potentials to detect phishing SMS adequately.

Discussion

The aim of this paper is to detect smishing messages using a correlation algorithm with a machine learning classifier. Initially, this paper collected 52 features from the different directions of the SMS and experimented through four well-known classifiers. The result shows that the AdaBoost classifier provided better accuracy with 98.67%. Although the accuracy was satisfactory to detect smishing messages, the feature dimension was too high that was 52 features. In this way, this paper adopted a feature selection algorithm to reduce the dimension of the features.

This paper used four ranking algorithms to rank the features and employed the sequential forward feature selection algorithm to search the best features set. The experiment shows that the Kendall ranking algorithm offered superior accuracy (98.40%) with AdaBoost classifiers. Furthermore, this algorithm has lessened the number of features with 61.53% that indicated the proposed model could able to prune more than half of the features.

Finally, the result of the investigation was contrasted with other anti-smishing methods and the comparative analysis demonstrate that the proposed model furnished better accuracy. From the examination, it tends to infer that the proposed model tends to detect the smishing messages.

Conclusion

These days, the smishing messages are hastily growing and it dominates cyber-attack in the cyberspace. Although, most of the researchers are imparting several advanced techniques to reduce the pace concerning these attacks, they are still failed to obtain complete detection. A large scale of features is practiced to increase the accuracy of detection. Whereas it is not true that the highest features would contribute better accuracy. Therefore, the feature selection algorithm appears in this scenario to reduce the feature dimension.

This paper employed four ranking algorithms to rank the features such as Pearson rank correlation, Spearman's rank correlation, Kendall rank correlation, and Point biserial rank correlation. Initially, this paper selected the machine learning classifiers and found that the AdaBoost classifier offered the better accuracy. Furthermore, with the ranking algorithm that is Kendall rank correlation offered superior accuracy than the other correlation algorithms. The inferred of this experiment shows that the ranking algorithm was able to reduce the feature dimensionality with 61.53% and provided an accuracy of 98.40%.

In the future, more features and an advanced feature selection algorithm would be applied. The target of the future model is to find the best feature set with less time.

Compliance with Ethical Standards

Conflicts of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants performed by any of the authors.

References

1. Abu-Nimeh S, Nappa D, Wang X, Nair S (2007) A comparison of machine learning techniques for phishing detection. In:

Proceedings of the Anti-phishing Working Groups 2Nd Annual eCrime Researchers Summit, ACM, New York, NY, USA, eCrime '07, pp 60–69. 10.1145/1299015.1299021
2. Akinyelu AA, Adewumi AO. Classification of phishing email using random forest machine learning technique. J Appl Math. 2014;2014:2014.
3. Almeida TA. Ham and spam dataset. 2017. https://archive.ics.uci.edu/ml/datasets/sms+spam+collection. Accessed 2017.
4. Anti-Phishing Working Group (APWG). Phishing activity trends report. 2020. https://apwg.org/. Accessed 2020.
5. Arab M, Sohrabi MK. Proposing a new clustering method to detect phishing websites. Turk J Electr Eng Comput Sci. 2017;25(6):4757–67.
6. Baek M, Lee Y, Won Y. Property analysis of sms spam using text mining. In: Park JJJH, Chen SC, Raymond Choo KK, editors. Advanced multimedia and ubiquitous engineering. Singapore: Springer; 2017. p. 67–73.
7. Baglia M. Text marketing vs. email marketing: Which one packs a bigger punch. 2015. http://www.business2community.com/infographics/text-marketing-vs-email-marketing-one-packs-bigger-punch-infographic-01249186#7wxHqvEMDcCGqhlz.97. Accessed 2017.
8. Basnet RB, Sung AH. Classifying phishing emails using confidence-weighted linear classifiers. In: International conference on information security and artificial intelligence (ISAI), IEEE; 2010. pp. 108–112.
9. Calkins KG. Applied statistics: correlation coefficients. In: Andrews University Retrieved on June 5. 2005.
10. CallHub. 6 reasons why sms is more effective than email marketing—callhub. 2016. https://callhub.io/6-reasons-sms-effective-email-marketing/. Accessed 2018.
11. Chen PY, Popovich PM. Correlation: parametric and nonparametric measures; 2002. pp. 137-139, Sage.
12. Cheung MWL, Chan W. Testing dependent correlation coefficients via structural equation modeling. Org Res Methods. 2004;7(2):206–23.
13. Coleman M, Liau TL. A computer readability formula designed for machine scoring. J Appl Psychol. 1975;60(2):283.
14. DataReportal. Digital around the world. 2020. https://datareportal.com/global-digital-overview#:~:text=The%20number%20of%20mobile%20phone,latest%20data%20from%20GSMA%20Intelligence. Accessed 2020.
15. Delany SJ, Buckley M, Greene D. Sms spam filtering: methods and data. Expert Syst Appl. 2012;39(10):9899–908.
16. EC-Council. Ethical hacking and countermeasures: web applications and data servers, vol. 3. 1st ed. Boston: Course Technology Press. 2009.
17. Fette I, Sadeh N, Tomasic A. Learning to detect phishing emails. In: Proceedings of the 16th International Conference on World Wide Web, ACM, New York, NY, USA, WWW '07; 2007. pp. 649–656. https://doi.org/10.1145/1242572.1242660
18. Flesch R. A new readability yardstick. J Appl Psychol. 1948;32(3):221.
19. Foozy CFM, Ahmad R, Abdollah MF. Phishing detection taxonomy for mobile device. Int J Comput Sci Issues. 2013;10(1):338–44.
20. Geng X, Liu TY, Qin T, Li H. Feature selection for ranking. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, ACM; 2007. pp. 407–414.
21. Gunning R. The technique of clear writing. New York: McGraw-Hill; 1952.
22. Hale J. Protect yourself from smishing. 2012. https://www.cnet.com/news/protect-yourself-from-smishing-video/. Accessed 2018.
23. Hiremath R, Malle M, Patil P. Cellular network fraud & security, jamming attack and defenses. Procedia Comput Sci. 2016;78:233–40.
24. Ho TK. Random decision forests. In: Document analysis and recognition, 1995, Proceedings of the third international conference on IEEE, vol 1; 1995. pp. 278–282.
25. Huang H, Qian L, Wang Y. A svm-based technique to detect phishing urls. Inf Technol J. 2012;11(7):921–5.
26. Islam R, Abawajy J. A multi-tier phishing detection and filtering approach. J Netw Comput Appl. 2013;36(1):324–35.
27. Joo JW, Moon SY, Singh S, Park JH. S-detector: an enhanced security model for detecting smishing attack for mobile computing. Telecommun Syst. 2017;66:1–10.
28. Kendall M, Gibbons J. Rank correlation methods, trans. London: Edward Arnold; 1990.
29. Kendall MG. Rank correlation methods. London: Edward Arnold; 1955.
30. Lee A, Kim K, Lee H, Jun M. A study on realtime detecting smishing on cloud computing environments. In: Park JJJH, Chao HC, Arabnia H, Yen NY, editors. Advanced multimedia and ubiquitous engineering. Berlin, Heidelberg: Springer; 2016. p. 495–501.
31. Lev J, et al. The point biserial coefficient of correlation. Ann Math Stat. 1949;20(1):125–6.
32. Mc Laughlin GH. Smog grading-a new readability formula. J Read. 1969;12(8):639–46.
33. McAfee. Protect yourself from smishing. 2012. https://securingtomorrow.mcafee.com/consumer/family-safety/protect-yourself-from-smishing/. Accessed 2017.
34. Mishra S, Soni D. Smishing detector: a security model to detect smishing through sms content analysis and url behavior analysis. Future Gener Comput Syst 2020;108:803–15. https://doi.org/10.1016/j.future.2020.03.021, http://www.sciencedirect.com/science/article/pii/S0167739X19318758.
35. Mun HJ, Li Y. Secure short url generation method that recognizes risk of target url. Wirel Personal Commun. 2017;93(1):269–83.
36. Nair AES. Disributed system for smishing detection. 2013. Unpublished. http://hacnet.smu.edu/posters/201302_Ala.pdf, southern Methodist University, Dallas, Texas.
37. Panagiotis BP. Survey about attack and defence phishing techniques. 2018.
38. Peltier TR. Social engineering: concepts and solutions. Inf Secur J. 2006;15(5):13.
39. Ramanathan V, Wechsler H. Phishgillnet-phishing detection methodology using probabilistic latent semantic analysis, adaboost, and co-training. EURASIP J Inf Secur. 2012;1:1.
40. Ramanathan V, Wechsler H. Phishing detection and impersonated entity discovery using conditional random field and latent dirichlet allocation. Comput Secur. 2013;34:123–39.
41. Saeys Y, Abeel T, Van de Peer Y. Robust feature selection using ensemble feature selection techniques. In: Joint European conference on machine learning and knowledge discovery in databases, Springer; 2008. pp. 313–325.
42. Smith EA, Senter R. Automated readability index. Tech rep. 1967.
43. Sonowal G. Phishing email detection based on binary search feature selection. SN Comput Sci. 2020;1:4.
44. Sonowal G, Kuppusamy K. Phidma—a phishing detection model with multi-filter approach. Journal of King Saud University—Computer and Information Sciences. 2017. https://doi.org/10.1016/j.jksuci.2017.07.005, http://www.sciencedirect.com/science/article/pii/S1319157817301210.
45. Sonowal G, Kuppusamy K. Mmsphid: a phoneme based phishing verification model for persons with visual impairments. Inf Comput Secur. 2018a;26(5):613–36.
46. Sonowal G, Kuppusamy KS. Masphid: a model to assist screen reader users for detecting phishing sites using aural and visual

similarity measures. In: Proceedings of the International Conference on Informatics and Analytics, ACM, New York, NY, USA, ICIA-16; 2016. pp. 87:1–87:6. https://doi.org/10.1145/2980258.2980443
47. Sonowal G, Kuppusamy KS. Smidca: an anti-smishing model with machine learning approach. Comput J. 2018b;61(8):1143–57. https://doi.org/10.1093/comjnl/bxy039.
48. The New Indian Express. Increasing cybercrime: un reports 350 per cent rise in phishing websites during pandemic. 2020. https://www.newindianexpress.com/business/2020/aug/08/increasing-cybercrime-un-reports-350-per-cent-rise-in-phishing-websites-during-pandemic-2180777.html. Accessed 2020.
49. Toolan F, Carthy J. Phishing detection using classifier ensembles. In: 2009 eCrime Researchers Summit, IEEE; 2009. pp. 1–9. https://doi.org/10.1109/ECRIME.2009.5342607.
50. Tsanas A, Little MA, McSharry PE. A simple filter benchmark for feature selection. J Mach Learn Res. 2010;2010:1–24.
51. Yearwood J, Mammadov M, Banerjee A. Profiling phishing emails based on hyperlink information. In: 2010 international conference on advances in social networks analysis and mining, IEEE; 2010. pp. 120–127. https://doi.org/10.1109/ASONAM.2010.56.
52. Yeboah-Boateng EO, Amanor PM. Phishing, smishing & vishing: an assessment of threats against mobile devices. J Emerg Trends Comput Inf Sci. 2014;5(4):297–307.