World Journal of Engineering Research and Technology

WJERT

www.wjert.org

SJIF Impact Factor: 5.218

AN APPROACH TO TACKLE PHISHING AND SMISHING ATTACKS

***[1]Harshita Shewale and [2]Prof. G. A. Patil**

[1]D. Y. Patil College of Engineering & Technology, Kasaba Bawada, Kolhapur, Maharashtra, India.

[2]Department of Computer Science and Engineering, D. Y. Patil College of Engineering & Technology, Kasaba Bawada, Kolhapur, Maharashtra, India.

Article Received on 09/04/2019 **Article Revised on 30/04/2019** **Article Accepted on 20/05/2019**

***Corresponding Author**
Harshita Shewale
D. Y. Patil College of
Engineering & Technology,
Kasaba Bawada, Kolhapur,
Maharashtra, India.

ABSTRACT

Increasing threats of phishing attacks on mobile computing platforms is the severe problem of recent years. Hardware limitations of mobile devices and habits of mobile users are the main reasons for phishing attacks. This paper includes a comprehensive study on mobile phishing attacks that covers identity extraction using OCR and detection of phishing attacks done through SMS; Smishing. The implementation of Mobifish on smart phones running the android 5.0 operating system and above versions is experimented.

INDEXTERMS: Phishing attack, identity extraction, smishing.

INTRODUCTION

Phishing attacks are the one by which an attacker steal private information of user by impersonating a legitimate entity. They steal user name, password, credit card details, account number etc. Phishing sites stay online only for some time and so it is hard to find the attacks. Attackers frequently change their techniques so that new techniques are able to overcome existing anti- phishing tools. Now a days mobile platforms have become new target of phishing attacks. It is only because of hardware limitations such as small screen size as compared to PCs, application switching, inconvenience of user input, lack of identity indicators, user's wrong habits of mobile handling etc.

There are two types of phishing detection schemes: one is heuristic based scheme and second one is blacklist based scheme. The blacklist based scheme detects only those sites that are enlisted as blacklisted sites. As phishing sites stay only for sometimes, known as zero day phishing attacks, they are not recognized by black list based scheme. The heuristic based approach deals with features extracted from URL and HTML source code. For mobile apps, it is hard to find or to check that where the user's credentials go; means to which server the actual information go. If it goes to legitimate site's server then it is authenticated or if not then it is confirmed that it is a fake site. Hence it is very essential to develop effective scheme for detection of phishing attacks.

Related Work

Longfei Wu, Xiaojiang Du and Jie Wu[1] proposed Mobifish which is a lightweight anti-phishing scheme. It is a scheme for mobile devices which is capable of defending against phishing attacks on mobile web pages, apps, and persistent accounts. MobiFish aims to solve the essential problem of identity masquerade, without reliance on HTML source code, search engine, or machine learning techniques.

Kinda and Krugel[2] proposed Antiphish technique which tracks the sensitive information of a user and generate warnings whenever user attempts to give away this information to a website that is considered to be un-trusted.

M Dunlop et al[3] defined Gold Phish that utilizes the OCR technique for phishing detection in PC browser. OCR is used to extract text from images found in web pages (e.g., the company logo), and then, it is compared to the top-ranked domains from Google's search service. This lightweight scheme works with mobile browsers, and it does not depend on external search engines.

M Chou, R. Ledesma, Y. Teraguchi & J. C. Mitchell[4] proposed Spoof Guard technique that uses URLs, images, links and domain names to check the similarity between given page and the pages sorted previously.

El-Alfy and AlHasan[6] have proposed a model for filtering text messages for both email and SMS. They have analyzed different methods in order to finalize a feature set such that complexity can be reduced. They have used two classification algorithms i.e. support vector machine(SVM) and Naïve Bayes and 11 features like URLs, likely spam words, emotion symbols, special characters, gappy words, recipient address, subject field and spam domain.

Naïve Bayes text classification,[7] The Bayesian classification is used as probabilistic learning method.

Proposed Approach

Mobifish is a lightweight anti- phishing scheme in which we use OCR- Optical Character Recognition technique to extract text from the screenshot taken. Then the identity of the given page is checked with claimed identity and the actual identity. If both of them do not match then system gives warning to the user about phishing attack. The second one is the phishing attacks (smishing) which done through SMS as it is the most popular communication service. People use SMS service to communicate rather than emails as sending SMS doesn't need internet connection and it is simple and efficient.

The proposed work is to design a system which consists of identity extraction and SMS phishing. The system architecture is as shown in Fig.1 that includes other phishing detection techniques also, like web page phishing, app phishing, account registry phishing and voice call phishing.

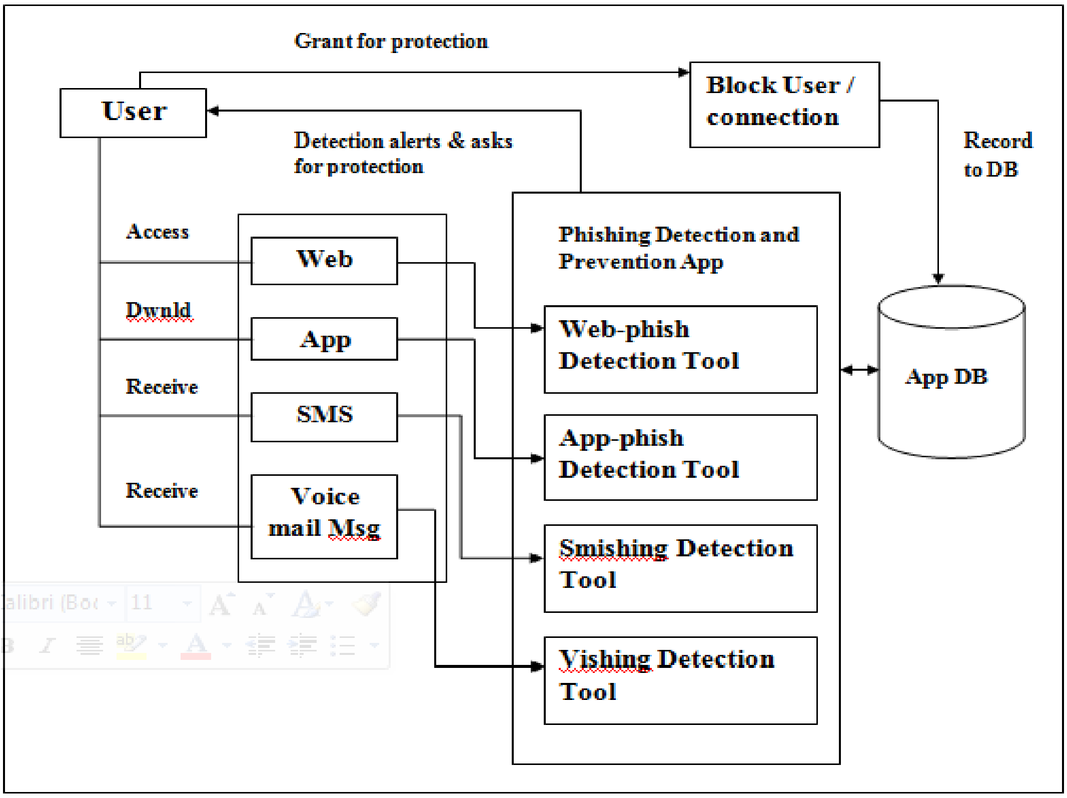


Fig 1: Architecture of phishing techniques.

A. Identity extraction

It works on the concept of Optical Character recognition. The image of typed, handwritten or printed text is converted electronically into a machine encoded text. It is done from a scanned document or a photo of a document. The screenshot may contain entire login page or the majority content of the page. Another observation is that the brand names and the company logo (identity) are located at the apparent places in the login page, which can be easily captured and extracted from the screenshot. A complete login page or a part of the page or an URL is selected as an input to the OCR. Most well known enterprises use their brand names as the SLD (Second Level Domain name) of their official website. It is valuable for phishing attack detection as long as high quality OCR solution is used.

Once we pass the screenshot taken to identity extraction step, the OCR actually works in step by step manner.

1. A scanner generates an image of the document and the text is intelligently extracted from that image. At this step the document image is only a meaningless cloud of intense points, ‘pixels’ on lighter background.
2. Colors and grayscale of an image is converted into black and white images by the process known as intelligent binarization routine.
3. The next step is page analysis. Interested area of screenshot to be recognized is marked on the page. The OCR software extracts the text information from the black and white pixels of the selected zones. The shapes are recognized and accordingly characters are assigned.

This is done in several steps.

1. Line segmentation consists of slicing a page of text into its different lines. This step also analyzes interline spacing, line skew, drop letters, and separates touching lines.
2. The word segmentation isolates one word from another.
3. The Character segmentation- it does not apply when word image decoding is used. It separates various letters of a word.
4. This step organizes the dots of scanned image into characters. If the characters have the same width, character segmentation is easy.
5. The actual character recognition extracts characteristics out of each isolated shape and assigns a symbol.

Mainly, there are two different ways of character recognition. First one is by recognizing characters entirely known as pattern recognition. Second one is by detecting the individual lines and strokes by which characters are made, known as feature detection.

Pattern Recognition

Monospace font is a concept where every letter has exactly the same width and it is carefully designed so that each letter could be easily recognized and distinguished from all other letters. By standardizing on one simple font, OCR becomes a relatively easy problem to solve.

Feature Detection

The name for feature detection is Intelligent Character Recognition (ICR) and is more sophisticated way of spotting characters. It detects the individual component features like angled lines, crossed lines, from which the character is made, instead of recognizing the complete pattern of a character. Most modern OCR programs work by feature detection rather than pattern recognition.

Character Recognition

ICR converts an image of every character to the appropriate character code. Sometimes this results into several character codes for uncertain images. For example recognition of the image of t he ‘‘I’’ character can produce the codes for ‘‘I’’, ‘‘l’’, ‘‘1’’, ‘‘|’’, the final character code will be selected later. Here dictionary support can improve the recognition quality. Dictionary can help to make the decision.

The OCR is implemented from the library available at com. google. android. gms. samples. vision.

The class Detector is the base class for implementing specific detector instances such as Barcode Detector, Face Detector, Multi Detector and Text Recognizer. Here we use class Text Recognizer which extends class Detector.

B. Smishing

It is a type of phishing attack done through SMS. These SMS are in the form of URL padding or the front loading of web addresses of malicious site with legitimate domain name. The message style is panicky.

The approach adopted is to identify and classify the spam SMS messages when it is received on the mobile phone, regardless of newly created spam message (zero-hour attack). In this, firstly dataset is collected and the features for our experiment are finalized. After finalizing features, the features are extracted from the messages (ham and spam) to create a feature vector. These feature vectors are used for training and testing purposes. Our proposed system takes the decision based on nine features.

The detection workflow is shown in Fig 2.

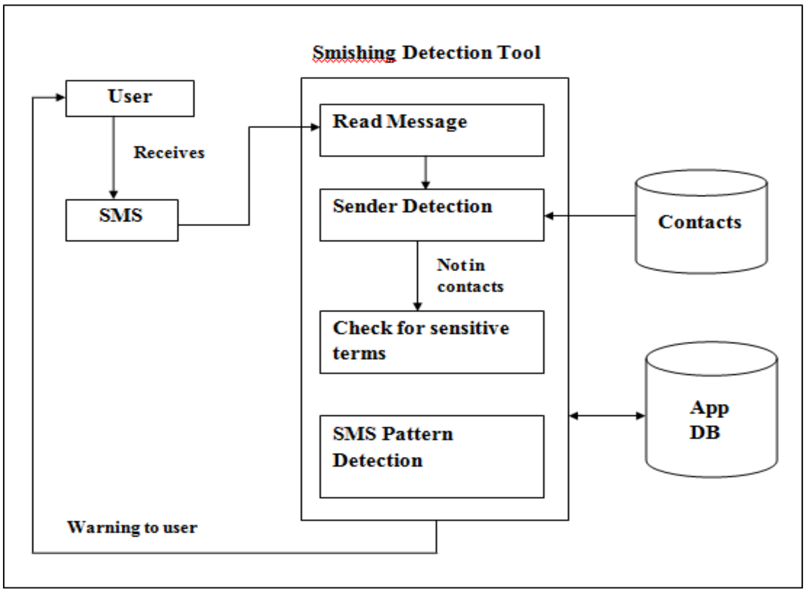


Fig. 2: Smishing detection workflow.

The Classification algorithm used is Naïve Bayes Classification algorithm as a smishing detection technique. This technique represents a supervised learning method as well as a statistical method for classification.

Steps Involved

Step I: Preprocessing

In this step collection of dataset and finalizing of the features is done.

Feature selection and extraction

Feature selection is a very important task for the SMS Spam filtering. Selected features should be correlated to the message type such that accuracy for detection of spam message can be increased. There is a length limit for SMS message and it contains only text. We study the characteristics of spam messages in depth and find some features, which are useful in the efficient detection of spam SMS. The features that we have extracted and evaluated for our proposed approach are summarized as follows:-

1. Presence of Mathematical Symbols: +, -, <, >, /, ^.
2. Presence of URLs: http or www.

3. Presence of Dots: to separate the sentence or words.
4. Presence of special symbols: $,!, #, *, &.
5. Lowercased words: used to seek user's attention.
6. Uppercased words: WON, ATTENTION, FREE, etc.
7. Presence of mobile number: ask the users to dial on the given number.
8. Specific Keywords: presence of suspicious keywords.
9. Message length: total length of the message including space, symbols, special characters etc.

Step II: Classifier training

In training phase, a binary classifier is generated by applying the feature to the message. There are different algorithms used for classification, accurate out of which we have used Naïve Bayes classification algorithm.

The Table 1 shown below give details of features extracted and their values to be assigned.

Table 1: Features extracted and their value.

Feature (S)	**Value= 0**	**Value= 1**
1.Mathematical Symbol (S1)	Mathematical Symbol Absent	Mathematical Symbol present
2. URL (S2)	URL Absent	URL Present
3. Presence of Dots (S3)	No dots in message	Dots are present
4. Special Symbol (S4)	No any special Symbol	Special Symbol Present
5. Lowercased words (S5)	Lowercased words are absent	Lowercased words are present
6. Uppercased words (S6)	Uppercased words are absent	Uppercased words are present
7. Mobile Number (S7)	No any mobile number within message	Mobile number present in Message
8. Spam Keywords (S8)	Spam Keywords are absent	Spam Keywords are present
9. Message Length (S9)	Counts total length of the message	

Step III: Classifier testing

In the testing phase, the classifier determines whether a new message is a phishing or not according to the binary values we get in classifier training for each feature extracted.

The Table 2 shows how all the features are considered when a message is checked in Naïve Bayes classification for detecting whether it is a ham message or spam message. Example messages:

"You have bin selected for a $1000 Walmart GiftCard, Enter code "FREE" at http://www.walmart.com.f.biz/wm/ to claim your prize: 161 left!" This is a message which has most of the features same as the spam message and “Have you finished work yet?” is another simple text message which does not has spam features. for understanding and evaluating many learning algorithms. It calculates explicit probabilities for hypothesis and it is robust to noise in input data.

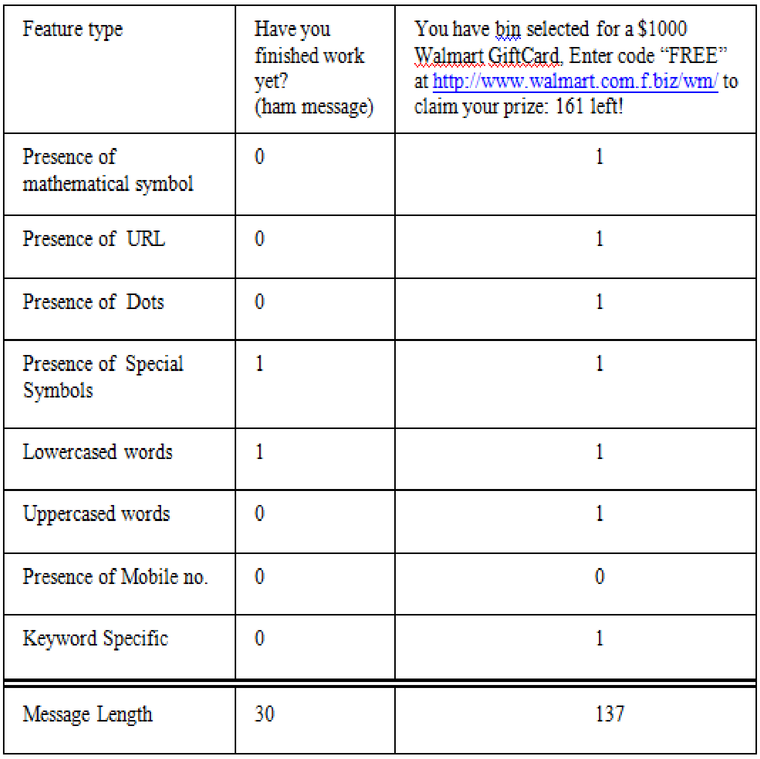
For this smishing module, suppose,
X= (x1, x2, x3,…xn) set of features extracted
C= (C1, C2) types of total messages i.e. ham messages and spam messages.
Naïve Bayesian classification predicts X belongs to class Ci iff
P (Ci/X)> P(Cj/X).
P (Ci/X) = P(X/Ci) P(Ci) / P(X).

$$P(X\,/.Ci) = \prod_{k=1}^{n} P(x_k\,/\,Ci)$$

P(X/Ci) = P(x1/Ci) * P(x2/Ci) *…* P(xn/ Ci)

Maximize P(X/Ci) P(Ci) as P(X) is constant. Naïve Bayes Assumption of “Class Conditional Independence”.

Table 2: SMS message feature value for ham and spam messages.



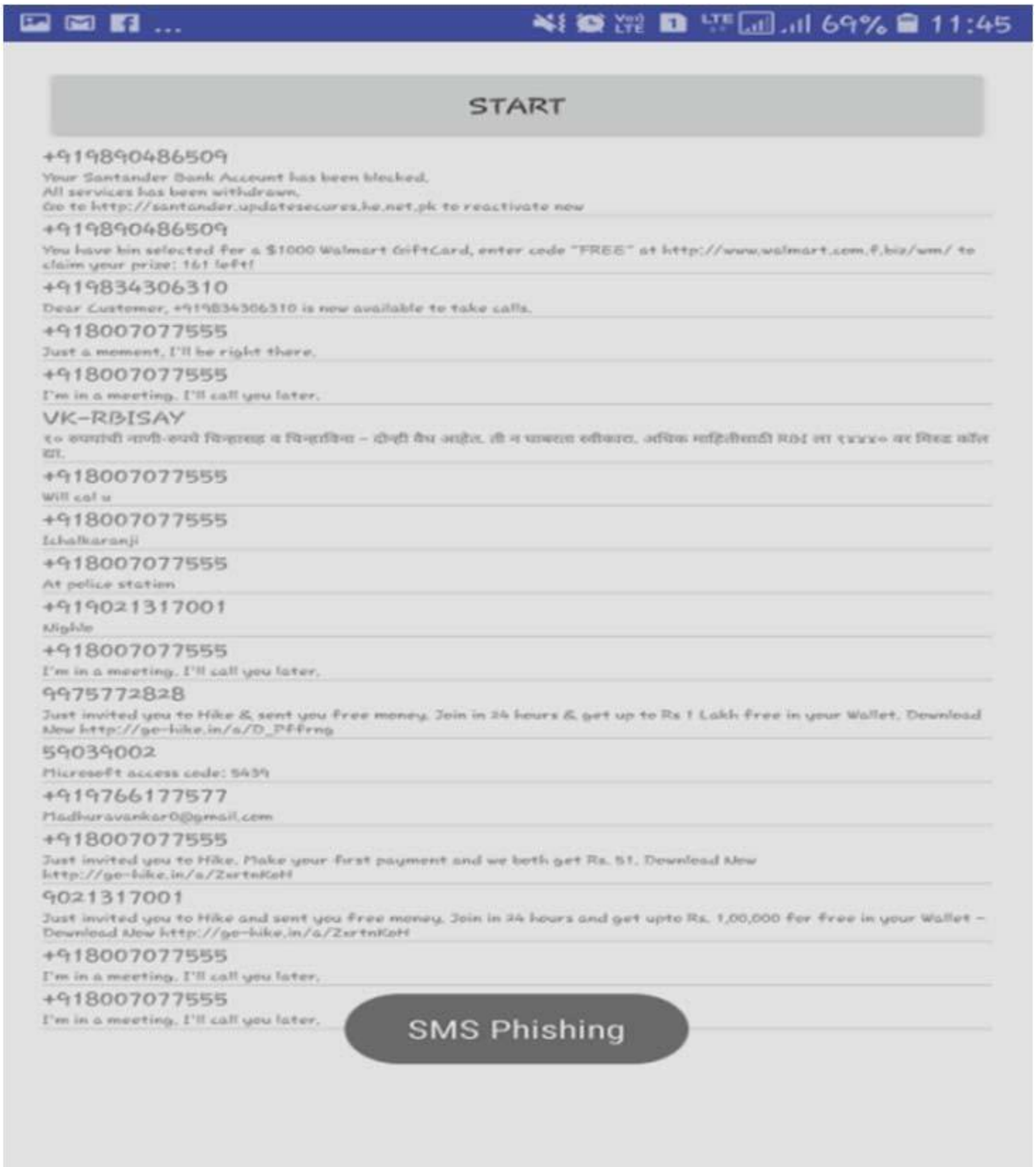
Feature type	Have you finished work yet? (ham message)	You have bin selected for a $1000 Walmart GiftCard. Enter code “FREE” at http://www.walmart.com.f.biz/wm/ to claim your prize: 161 left!
Presence of mathematical symbol	0	1
Presence of URL	0	1
Presence of Dots	0	1
Presence of Special Symbols	1	1
Lowercased words	1	1
Uppercased words	0	1
Presence of Mobile no.	0	0
Keyword Specific	0	1
Message Length	30	137

RESULTS

We implemented OCR technique and Smishing on the latest version of Android operating system. For smishing detection we give the entire list of messages from user's inbox. Fig 3 shows the list of input messages.

Naïve Bayesian Classification

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. It provides practical learning algorithms and prior knowledge and observed data can be combined. It also provides a useful perspective.



+919890486509
Your Santander Bank Account has been blocked.
All services has been withdrawn.
Go to http://santander.updatesecures.he.net.pk to reactivate now

+919890486509
You have bin selected for a $1000 Walmart GiftCard, enter code "FREE" at http://www.walmart.com.f.biz/wm/ to claim your prize: 161 left!

+919834306310
Dear Customer, +919834306310 is now available to take calls.

+918007077555
Just a moment, I'll be right there.

+918007077555
I'm in a meeting. I'll call you later.

VK-RBISAY
१० रुपयांची नाणी-रुपये [illegible] RBI ला १४४४० वर मिस्ड कॉल द्या.

+918007077555
Will cal u

+918007077555
Ichalkaranji

+918007077555
At police station

+919021317001
Nighle

+918007077555
I'm in a meeting. I'll call you later.

9975772828
Just invited you to Hike & sent you free money. Join in 24 hours & get up to Rs 1 Lakh free in your Wallet. Download Now http://go-hike.in/a/D_Pffrng

59039002
Microsoft access code: 5439

+919766177577
Madhuravankar0@gmail.com

+918007077555
Just invited you to Hike. Make your first payment and we both get Rs. 51. Download Now http://go-hike.in/a/ZxrtnKoH

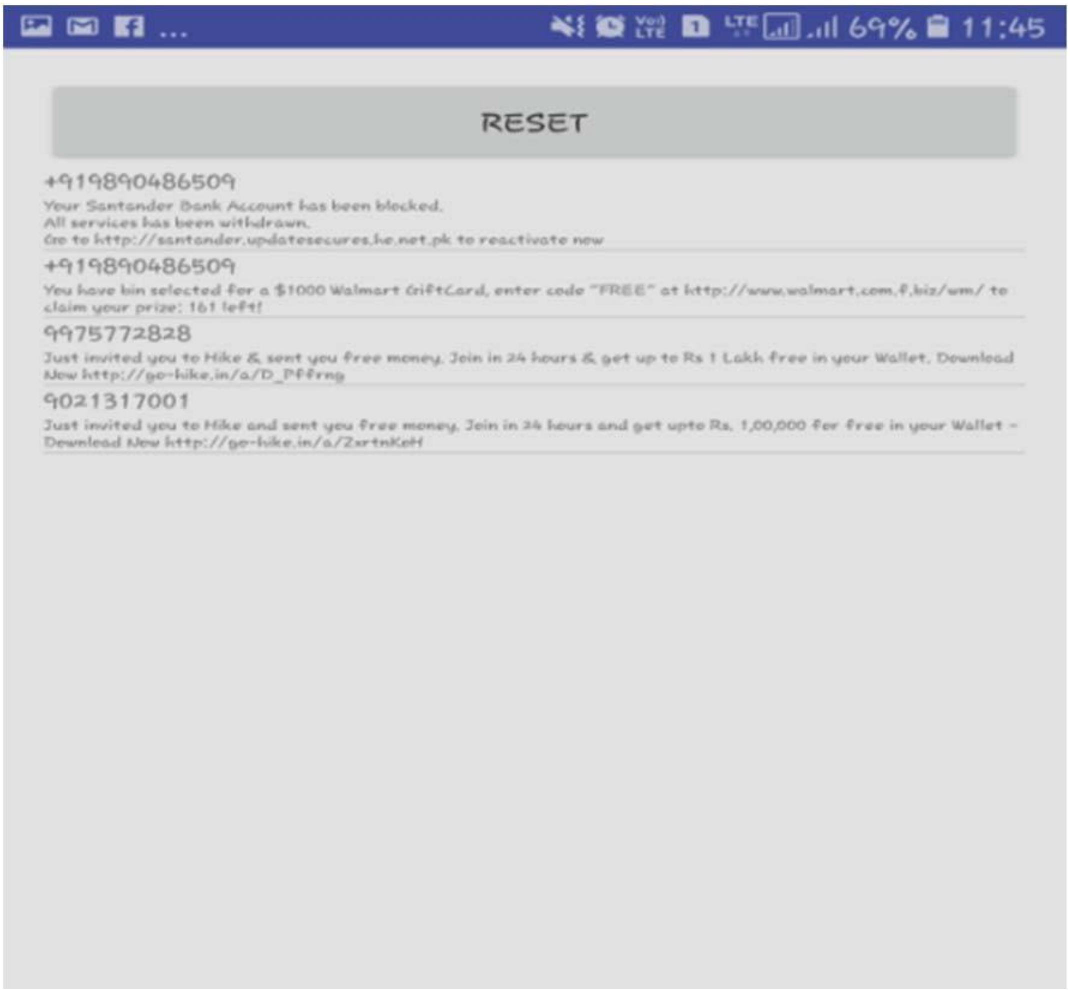
9021317001
Just invited you to Hike and sent you free money. Join in 24 hours and get upto Rs. 1,00,000 for free in your Wallet – Download Now http://go-hike.in/a/ZxrtnKoH

+918007077555
I'm in a meeting. I'll call you later.

+918007077555
I'm in a meeting. I'll call you later.

Fig. 3: List of Input message dataset.

When user clicks on the "start" button, the Mobifish gives the list of phishing messages as a result of smishing. Fig 4 shows all the messages which are phishing messages. User gets warning about not providing any credentials to the suspicious links and mobile numbers present in these phishing messages.



+919890486509
Your Santander Bank Account has been blocked.
All services has been withdrawn.
Go to http://santander.updatesecures.he.net.pk to reactivate now

+919890486509
You have bin selected for a $1000 Walmart GiftCard, enter code "FREE" at http://www.walmart.com.f.biz/wm/ to claim your prize: 161 left!

9975772828
Just invited you to Hike & sent you free money. Join in 24 hours & get up to Rs 1 Lakh free in your Wallet. Download Now http://go-hike.in/a/D_Pffrng

9021317001
Just invited you to Hike and sent you free money. Join in 24 hours and get upto Rs. 1,00,000 for free in your Wallet - Download Now http://go-hike.in/a/ZxrtnKeH

Fig. 4: List of phishing messages.

CONCLUSION

In this paper we have represented two aspects used in mobile phishing detection. Identity extraction gives only warning about phishing webpage or URL extracted from screenshots taken. Smishing detection is done through Naïve Bayes classification technique which gives the list of probable phishing SMS from the list of input message of user's message inbox.

In addition, mobile phishing attacks could also be in the form of phishing voice calls. The voice phishing (Vishing) uses the voice over internet protocol (VoIP) technique, in which the phone number is dynamically generated. Efforts to tackle such attacks is in progress.

REFERENCES

1. Longfei Wu, Xiaojiang Du and Jie Wu "Effective Defense Schemes for Phishing Attacks on Mobile Computing Platforms" IEEE transactions on vehicular technology, 2016; 65(8).
2. E. Kirda and C. Kruegel, "Protecting users against phishing attacks with AntiPhish," in Proc. 29th Annu. Int. COMPSAC, 2005; 517–524.
3. M. Dunlop, S. Groat, and D. Shelly, "GoldPhish: Using images for content-based phishing analysis," in Proc. 5th ICIMP, 2010; 123–128.
4. N. Chou, R. Ledesma, Y. Teraguchi, and J. C. Mitchell, ‘‘Client-side defense against web-based identity theft,’’ in *Proc. 11th Annu. NDSS*, Feb., 2004.
5. S. Garera, N. Provos, M. Chew, and A. D. Rubin, "A framework for detection and measurement of phishing attacks," in Proc. ACM WORM, 2007; 1–8.
6. El-Alfy, E.S.M., AlHasan, A.A.: Spam filtering framework for multimodal mobile communication based on dendritic cell algorithm. Future Gen. Comput. Syst., 2016; 64: 98– 107. doi:10.1016/j.future.2016.02.018.
7. Naive Bayes text classification (http://nlp.stanford.edu/IR- book/html/htmledition/naive-bayes-text- classification-1.html).
8. http://www.how-ocr- works.com/OCR/OCR.html.
9. https://developers.google.com/android/refer ence/com/google/gms/vision/detector.
10. https://developers.google.com/android/ref erence/com/google/gms/vision/text/TextRecognizer.
11. http://www.explainthatstuff.com/how-ocr- works.html.
12. Naïve Bayes Classification Algorithm.
13. "Towards filtering of spam messages using machine learning based techniques." Neelam Choudhary (&) and Ankit Kumar Jain Computer Engineering Department, National Institute of Technology, Kurukshetra, Haryana, India.
14. "Comparative Analysis of Mobile Phishing Detection and Prevention Approaches". Choudhary, N., Jain, A.K.
15. SMS Blocker Award. https://play.google.com/store/apps/details?id= com.sms Blocker &hl=en.
16. Text Blocker. https://play.google.com/store/apps/details?id=com.thesimpleandroidguy.app. Messageclient & hl=en.