Evaluating the Feasibility of Using Classifiers in Detecting Social Engineering Fraud

Clifford Kengocha
Kabarak University, 13 P.O. Box Private Bag, Kabarak, 20157, Kenya
Tel: +254 721681468, Email: cogeto@kabarak.ac.ke

Abstract

Social engineering fraud is among the most notorious forms of fraud through which people continue to lose money and sensitive information. Its increasing prevalence is negatively affecting strides made in mobile and digital banking. Despite efforts in creating public awareness, its mitigation has not been effective as the tricks used by swindlers keep evolving. Virtually all existing solutions to the problem are based on human interventions such as manually reporting and blacklisting phone numbers. This approach is slow and inefficient due to the huge number of incidents reported relative to the limited existing human resource capacity. This paper presents an evaluation of the feasibility of using classifiers to detect voice-based social engineering fraud. Findings suggest the possibility of using natural language processing and machine learning to automate the detection of voice-based social engineering fraud. Outcomes of the study can be used to develop automated real-time SEF detection systems.

Keywords: Machine learning, social engineering fraud, natural language processing, classifier

1. Introduction

Social engineering fraud (SEF) refers to any form of fraud that involves tricking victims to divulge sensitive information or authorize payments (Meinert, 2016). Before the mass penetration of information and communication technologies, such fraud occurred through face-to-face interactions. However, today, this form of fraud can be conducted over the phone, email or using messaging services. Phone-based SEF is particularly common as it takes relatively shorter times to get responses from victims and thus, complete the attack successfully (Nturibi, 2018). While it may be easy for some people to detect attempts to defraud them, it remains difficult for the elderly, illiterate or otherwise less-knowledgeable individuals to identify an impending fraud.

Various interventions including blacklisting suspect phone numbers and tightening law enforcement have been forged in attempts to control the problem. A particularly noteworthy effort in the prevention of related frauds is the introduction of voice biometrics by a Kenyan mobile money service provider to facilitate customer identification (Chetalam, 2018). This service allows customers to use their voices as a factor of authentication, thereby, preventing impersonation attacks. Nonetheless, the tricks and tactics used by swindlers keep changing such that countermeasures have not been effective. Therefore, there is a need to enhance existing measures or develop new approaches to address the problem.

One of the promising solutions to this problem is the use of machine learning to continuously collect data on phone-based fraud instances and use it to identify new events. Classifiers can be used for this purpose. Classifiers are a type of supervised machine learning algorithms that use input data sets to categorise new observations. Unlike regression algorithms which approximate the closeness of a given sample to an expected output, classifiers provide either “match” or “no match” outputs only. Thus, they have a chance of incorrectly classifying an observation. Nevertheless, this type of machine learning has been extensively applied in image classification problems with significant success (Litjens et al., 2017). This paper evaluates the applicability of the technology in voice classification to detect SEF.

2. The Problem

The high prevalence of social engineering fraud has resulted in substantial financial losses to end consumers and reputation damage to service providers. If this situation persists, the uptake of mobile money transfer technology may decline resulting in economic stagnation. Additionally, it may impact the adoption of other technologies by affected end users.

3. Objectives

The main aim of this research was to evaluate the feasibility of using classifiers to detect voice-based social engineering fraud. This aim was achieved by addressing the following objectives:

f) To identify and define indicators of looming fraud (IOLF) in voice calls.
g) To establish an approach for classifying SEF using indicators of looming fraud.
h) To establish the potential accuracy of classifiers in categorizing a call as fraudulent.
i) To recommend applications and areas improvement of the approach.

4. Literature review

Given that social engineering, in general, is an old problem, there exists vast literature on the subject. Some researchers have even developed appliances that help to minimize risks associated with social engineering. This literature review provides an analysis of relevant literature aimed at identifying the gaps in addressing social engineering fraud.

Social Engineering Detection

Social engineering is an art of deception. In the context of this study, it involves tricking an individual in attempts to lead them to provide information or authorize a payment. Its use in leading individuals to authorize payments, such as transferring money, has been noted with concern in the recent past (Castle et al., 2016). In US Patent 9,123,027, Srivastana, Walker and Olson present methods and systems for detecting social engineering in emails. The invention by the three scholars involves the extraction of both semantic and non-semantic data items from an email message. Semantic data refers to the meanings of the words used while non-semantic data refers to other properties of the email such as the sender’s address. Using this approach, the researchers developed a functional appliance that can detect some level of social engineering in emails (Srivastana, Walker & Olson, 2015). This study indicates the significance of semantic data in computer-aided detection of malicious communication.

The use of natural language processing techniques in detecting social engineering has also been researched. Sawa, Bhakta, Harris and Hadnagy (2016) identify questions and commands in speech as two fundamental components of natural language. By processing natural language to extract questions and commands, the scholars theorize that it is possible to identify questions directed at obtaining sensitive information or commands intended at leading a user to perform unintended actions. Their approach involves comparing extracted questions and commands against a blacklist of common tricks. The authors reported a high rate of detection accuracy with few false positives. This research is consistent with Castle et al.'s patent inasmuch as both employ the use of semantic data extracted from a communication occurring in a natural language. Therefore, it is evident that using natural language processing to extract semantic data is a feasible method for obtaining indicators of looming fraud.

In another study, Bhakta and Harris further explore the concept of using semantic data items in the identification of IOLFs. Since a robust detection approach should be applicable to a variety of attack vectors, the authors propose the use of a topic blacklist to identify a potential fraud communication. According to their model, an attack is detected if a topic in a communication matches a topic on the blacklist (Bhakta & Harris, 2015). One of the major challenges with this approach is the difficulty in describing blacklist topics. Options for defining an unambiguous blacklist may be limited to keyword identification and the detection of word sequences. Clearly, this method would result in a non-exhaustive list of entries in the blacklist which will likely result in low detection rates. However, the method can be improved by allowing ambiguous definition of topics then adding another layer of verification to address the ambiguity. For instance, if a topic is identified as ambiguous, a non-semantic data item, such as the caller's phone number, can be used in the next step to rate its suspiciousness. The feasibility of using both semantic and non-semantic properties is evidenced in Castle et al.'s patent discussed above. Therefore, one can conclude that the concept of using predefined topic blacklists is a valid parameter for describing IOLFs.

Data quality is another factor that affects the success of machine learning systems. Poor data quality negatively impacts machine learning. As such, a high quality set of training data is paramount to the success of this approach. Quality data can be defined as data that exhibits consistency. Excessive heterogeneity in a data set can make it difficult for a machine learning system to accomplish its objectives. Similarly, excessive heterogeneity can result in relatively high error rates. A classifier, which is a typical machine learning system, is trained by running two sets of data through the algorithm until it can acceptably identify the set to which a given data sample belongs. In the context of this study, such data can be effectively obtained from users who can recognize fraudulent calls, either by recalling a similar incident or otherwise. Heartfield and Loukas (2018) aver that such social engineering training data for machine learning can be reliably obtained from end users. In a concept they dubbed as human-as-a-security-sensor, the researchers implemented Cogni-Sense, a prototype application for Microsoft Windows that "enables and encourages users to report semantic social engineering against them" (2018). The researchers conduct an experiment to test the effectiveness of Cogni-Sense deploying human sensors. They defined at least one user report as the criterion for detection. From the experiment, the scholars reported to have found less than 10% missed detections when human sensors are used compared to 81% where only technical sensors are used. There could be some variations in the rate of effectiveness of the sensors, but nevertheless, their study demonstrates the validity of collecting machine learning data from users.

Adedoyin et al. (2017) also evaluate the use of case-based reasoning, an application of machine learning, to predict fraud in mobile money transfer. Case-based reasoning is also a type of classification. Unlike in previously evaluated literature, Adedoyin et al.'s research does not refer to social engineering. Nevertheless, the researchers identify classification as one of the valid methods of predicting fraud. In case-based reasoning, past cases and their solutions are stored. When a similar or approximate case is encountered in future, the algorithm will use the past decision to predict the value of the expected output. This study further demonstrates that classification has the potential use in detecting fraud including social engineering fraud.

Voice-Based Authentication

Voice based authentication is increasingly being adopted as an alternative to password authentication. In the context of social engineering fraud, the major advantage of voice-based over passwords is the difficulty in executing impersonation attacks. As such, it has been proposed as a more secure approach to authenticating users prior to authorizing financial transactions in mobile money transfer platforms (Chetalam, 2018). However, it must be noted that social engineering fraud attack vectors are not limited to impersonation. In fact, a significant proportion of these attacks aim at influencing an authorized user to perform an unintentional action. Therefore, without undermining the effectiveness of voice-based authentication, using voice biometrics in place of passwords may not fully address the problem. Regardless of the inadequacy of voice biometrics in addressing fraud, it demonstrates that speaker recognition is a practical approach to identifying users. This conclusion revives the idea of using non-semantic properties in detecting IOLFs. In other words, it suggests that potential fraud can be detected simply by identifying the voice of the caller. In a given scenario, assuming the concept of human sensors is implemented, a swindler can be detected effortlessly if they have been reported before as a potential con-person.

This literature review has analysed current research on social engineering fraud and the different approaches researchers are using to attempt to solve the problem. The use of both semantic and non-semantic data items has stood out as a potential method of identifying indicators of looming fraud. Additionally, the use of predefined blacklists has been seen as an effective, though limited, approach to defining IOLFs. To improve the quality of training data and hence, the effectiveness of machine learning, literature has shown that using human sensors to report suspect cases can result in high detection efficiency. All these studies strongly suggest that a combination of these technologies can be used to effectively address voice-based social engineering fraud.

5. Methodology

Methods –Literature Review

This study used a literature review approach to address its objectives. To recap, the study aimed at evaluating the effectiveness of using classifiers to detect voice-based social engineering fraud. The study required an analysis of literature from diverse subjects in social engineering, machine learning and natural language processing. Consequently, the literature divided into the following subjects:

ii) Social engineering detection –Literature on social engineering detection was essential in identifying existing solutions and the challenges faced in addressing the problem. Observations from this field of study were also useful in defining IOLFs.

iii) Automated fraud detection –literature from this field of study was included to focus on addressing the use of classifiers in identifying IOLFs.
iv) Machine learning and natural language processing: Literature from this field of study was expected to address the use of natural language processing techniques in extracting semantic data from communications. It also tackled the feasibility of extracting useful non-semantic data from speech, such as data that can be used to distinguish a speaker.

Conceptual Framework

Conclusions from the literature review strongly suggested that a combination of (i) the use of both semantic and non-semantic data; (ii) the use of predefined topic blacklists; and, (iii) the use of human sensors to report suspected or actual cases of fraud, could be effective in detecting voice-based social engineering fraud. This conclusion is represented in the following illustration.

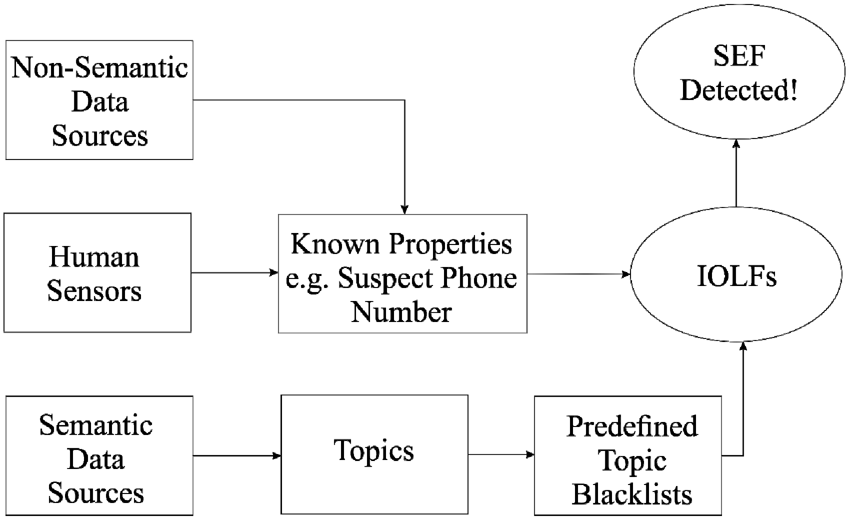
Non-Semantic Data Sources
SEF Detected!
Human Sensors
Known Properties e.g. Suspect Phone Number
IOLFs
Semantic Data Sources
Topics
Predefined Topic Blacklists

Fig. 1: Illustration showing the interaction of human sensors, semantic and non-semantic data sources and predefined topic blacklists in the automated detection of social engineering fraud

6. Results

This study was set to establish the extent to which the detection of social engineering fraud can be automated. In this quest, literature was systematically reviewed to address the objectives outlined initially. From the review, the following results were obtained:

- Indicators of looming fraud (IOLFs) can be obtained from both semantic and non-semantic data items. These may include reported phone numbers, names, aliases or specific tricks.
- Classification of SEF using IOLFs can be achieved using combination of (i) both semantic and non-semantic data; (ii) predefined topic blacklists; and, (iii) human sensors to report suspected or actual cases of fraud.
- The accuracy of any such classifier as above would largely depend on the data sources provided for training. More homogeneous data sets would result in higher accuracy of the classifier. It was also observed that the use of classifiers in other problem areas such as image classification have proven successful. Additionally, the

study found that the use of human sensors, as opposed to technical systems, can drastically improve the quality of data used for training the classifier.

7. Recommendations and Areas for Further Study

The outcomes of the study suggest feasibility and a high likelihood of success in the application of automated detection of social engineering fraud. The approach and objectives of the study were specific aimed at its application in detecting voice-based social engineering fraud. However, some of the discussions and results could be applied in frauds that use other attack vectors such as email. The findings can only be applied by both phone service providers and end-user application developers.

Although the research extensively covered the problem area, there are areas of study that could not be addressed due to limited scope. First, the definition of IOLFs is a continuous process which improves as more data is collected and tested. Therefore, the study recommends that further research be conducted to develop a more robust definition of IOLFs. Secondly, the study did not examine different classification approaches. It should be noted that different classification algorithms have varying applicability and accuracy. Therefore, it would be important to explore which algorithms are best suited for this application. This can only be achieved through tests with actual systems and actual data to compare the efficiency of different algorithms.

8. Conclusions

The study was successful insofar as it was able to determine that classifiers can be used in the automated detection of voice-based social engineering fraud. Some of the key concepts that would facilitate such tasks include the use of semantic and non-semantic data items, predefined topic blacklists and human sensors. Regardless of the success of the research, a few areas of further study were identified. These include strategies for widening the definition of IOLFs and further examination of suitable algorithms for this application.

References

Nturibi, B. M. (2018). A Mobile Money Social Engineering Framework for Detecting Voice & SMS Phishing Attacks-A Case Study Of M-Pesa (Doctoral dissertation, United States International University-Africa).

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, *42*, 60-88.

Castle, S., Pervaiz, F., Weld, G., Roesner, F., & Anderson, R. (2016). Let's Talk Money: Evaluating the Security Challenges of Mobile Money in the Developing World. In *Proceedings of the 7th Annual Symposium on Computing for Development* (p. 4). ACM.

Srivastava, M. K., Walker, W. A., & Olson, E. A. (2015). *U.S. Patent No. 9,123,027*. Washington, DC: U.S. Patent and Trademark Office.

Sawa, Y., Bhakta, R., Harris, I. G., & Hadnagy, C. (2016, February). Detection of social engineering attacks through natural language processing of conversations. In *2016 IEEE Tenth International Conference on Semantic Computing (ICSC)* (pp. 262-265). IEEE.

Bhakta, R., & Harris, I. G. (2015). Semantic analysis of dialogs to detect social engineering attacks. In *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)* (pp. 424-427). IEEE.

Chetalam, L. J. (2018). *Enhancing Security of Mpesa Transactions by Use of Voice Biometrics* (Doctoral dissertation, United States International University-Africa).

Adedoyin, A., Kapetanakis, S., Samakovitis, G., & Petridis, M. (2017, December). Predicting fraud in mobile money transfer using case-based reasoning. In *International Conference on Innovative Techniques and Applications of Artificial Intelligence* (pp. 325-337). Springer, Cham.

Meinert, M. C. (2016). SOCIAL ENGINEERING: The Art of Human Hacking. *American Bankers Association. ABA Banking Journal*, *108*(3), 49.