CONTENT BASED PHISHING DETECTION

Author: Prajakta Saraf
Co-Author / Corresponding Author: Jimoh Abdulganiyu

CHAPTER ONE

GENERAL INTRODUCTION

1.1 INTRODUCTION

Phishing is used by malicious actors (attackers) which masquerade as legitimate institutions and contact targeted victim or victims via emails, messages, phone calls (also known as vishing) asking for their personal, confidential credentials like banking and credit card details, passwords, personally identifiable information that can later be misused by the attacker to steal identities, rob bank accounts, demand ransom for the stolen information, sell the stolen information on the black market, political purposes etc. Phishing is a cyber crime committed by the attacker with various mottos such as financial benefit to the phisher, stealing confidential, sensitive data with respect to politics (political advantage), competitive benefit.This is the most simple and common attack that has resulted in a hefty monetary loss of the victims till date. Inorder to prevent such attacks, systems implementing Machine Learning (ML) techniques have been devised which cater to provide considerably good results but might prove to be inefficient in certain circumstances as the attackers use dynamic techniques. The dataset provided to the ML algorithms may or may not contain all possible malicious data as their data points and might prove to have a certain percentage of induced error.

(Malwarebytes Labs, n.d.), unlike other threats, phishing does not require any sophisticated technical expertise. “Phishing is the simplest kind of cyber attack and, at the same time, most dangerous and effective. That is because it attacks the most vulnerable and powerful computer on the planet: the human mind”. Phishing uses social engineering and does not aim at exploiting technical vulnerabilities. Social engineering is the use of deception (mind games) to manipulate individuals and lure them into providing sensitive, confidential information. No operating system is completely safe from phishing be it Windows, Macs, Android, iPhone. Attackers often resort to phishing instead of cracking through several layers of security. The weakest link in a security system is not a glitch in a computer's code but more often it’s a human mind that does not check where an email(message) came from. These kinds of emails(messages) generally enclose scary messages with links or attachments. To overcome this fear, the victims are demanded to go to a website and fill in the required information. If the users take the bait and click the link, they are redirected to an imitation of a legitimate website. Here they are asked to login and provide personal information that might directly go to the phisher and be misused.

Whittaker et al. (2010), despite increasing public awareness, phishing continues to be a major threat to internet users. According to Gartner’s estimation, phishers stole $1.7 billion in 2008 while the Anti-Phishing Working Group identified roughly twenty thousand unique, new phishing sites each month between July and December of 2008. Inorder to combat phishing, Google published a blacklist of phishing URLs and phishing patterns. For the sake of its effectiveness, it must be comprehensive, error-free and timely. A blacklist lacking comprehensiveness fails to protect a portion of its users. A blacklist with errors gives unnecessary warnings and thus trains its users to ignore the warnings which is undesirable. One that is not timely fails to warn its users about a phishing page thus resulting in heavy losses. Considering that phishing pages only remain active for an average of approximately three days, timely warnings play a vital role. Majority of the phishing pages last for a day or even less than a day. A delay of only a few hours can significantly degrade the quality of the blacklist.

Garera et al. (2007), according to these authors, phishing is a form of identity theft that combines social engineering and sophisticated attack vectors to harvest financial information from unsuspecting consumers. Often this is executed by luring victims into clicking an URL (which generally is in form of hyperlinks masqueraded as an URL of a legitimate institution) which on clicking redirects the victims to a rogue page and asks for sensitive information that is stolen by the phisher. It has been found that it is often possible to tell whether a URL belongs to a phishing attack or not, without requiring any knowledge of the corresponding page data. This can be done by describing several features that can be used to distinguish a phishing URL from a benign one. These features are used to model a logistic regression (ML technique) filter that is efficient and has a high accuracy. This kind of filter has been used to perform thorough measurements on several URLs and thus give a quantification of the prevalence of phishing on the Internet today.

This can put a check on URLs only and does not provide a solution for offline phishing attacks. Also this technique malfunctions at times of no occurrences of features (currently present in the URL) in the dataset. These kinds of shortcomings have been witnessed by most of the existing preventive measures.

1.2 BACKGROUND OF THE STUDY

Phishing has become the most common cyber security attacks and are exponentially growing across the globe. Phishing is a method of trying to gather personal information using deceptive e-mails and websites. Phishing is a cyber attack that uses disguised email as a weapon. The goal is to trick the email recipient into believing that the message is something they want or need like a request from their bank, or a note from someone in their company which instructs them to click a link or download an attachment. The main feature of such phishing notes or messages is their form. The attackers masquerade as a trusted entity of some kind, often a real or plausibly real person, or a company the victim might do business with. It is one of the oldest types of cyber attacks, dating back to the 1990's, and it is still one of the most widespread and pernicious, with phishing messages and techniques becoming increasingly sophisticated (Fruhlinger, J., 2020).

Goodchild, J., (2019), new research from Brigham and Women's Hospital in Boston finds hospital employees are extremely vulnerable to phishing attacks. The study highlights just how effective phishing remains as a tactic. Also, it enlightens the need for defense against and awareness of email scams is more critical than ever. The research was a multi-center exercise that looked at results of phishing simulations at six anonymous healthcare facilities in the US. Research coordinators ran phishing simulations for close to seven years and analyzed click rates for more than 2.9 million simulated emails. Results revealed that 14.2 percent of phishing emails were clicked which accounts for a rate of one in seven. This new research on phishing in healthcare puts a spotlight on the vulnerability of this kind of data.

In view of these increasing, sophisticated, widespread phishing attacks several countermeasures have been devised. Also several studies out there have inculcated awareness among people regarding cyber crimes, especially phishing. But, because of these attacks being sophisticated it becomes difficult to recognize phishing emails, websites, messages, etc. Various methods, concepts belonging to machine learning find applications in devising countermeasures against phishing, like usage of skewed data sets in order to detect offensive or sensitive content in online communities.

Phishing attacks result in data breaches, information security breach, ransomware attack, virus downloads, SQL injections, Trojan horses, vishing, worms, and much more. They adversely affect victims with respect to loss of money, property, personal information, etc. Several such phishing incidents have been recorded up till date. The motives behind such attacks can range widely from political benefit, monetary benefit, social benefit, to self satisfactory benefits. Prevention of phishing attacks has become critical for internet users, rather everybody. Effective and timely prevention of phishing attacks can ensure safe and trustworthy information storage for all and can thus reduce loss of money, property, theft of identity, disruption of reputation of legitimate institutions, thus ensuring safe, protected, well encrypted online transactions be it with respect to money, information, or any sorts of communication that might take place via online media.

1.3 STATEMENT OF THE PROBLEM

According to Malwarebytes Labs (2019), phishers are coming for the organization’s employees and customers. Phishing attacks are on a rise. It’s high time for all organizations to fasten their seat belts and work out on their anti-phishing strategies. A system where phishers are concerned, it does not matter whether the technique being used is revolutionary or old hat. Somebody, somewhere is going to fall for it. Here’s where social engineering comes into picture which is the most powerful tool of phishers and is the most complex loophole for the cyber security workforce. Many revolutionary techniques have been devised but they are not able to cope up with the social engineering aspect of vulnerability posed upon the organization’s employees and customers. It has become crucial for organizations and their employees to make sure if their business is secure and that their customers are performing safe email practices. If their customers are logging into fake portals, eventually they are going to tie up their support channels asking for help, refunds, reorders and more. If their employees are stung, they open doors to data theft, network infiltration, ransom demands, spying and thus resulting in a massive dent in the organization’s reputation. According to the author, all of these are poor directions to head in.

As mentioned earlier, social engineering is the most powerful weapon of all phishers. Social engineering is the term used for a broad range of malicious activities accomplished through human interactions. It uses psychological manipulation to trick users into making security mistakes or giving away sensitive information. Social engineering attacks are carried out using various techniques such as baiting, scareware, pretexting, phishing, spear phishing.

Putting a check on social engineering is next to impossible. All the countermeasure techniques devised in order to prevent phishing have this shortcoming in common. There hasn’t been any technological advancement as such in detecting and putting a check on social engineering which continues to be a measure drawback of all techniques that have been devised till date.

Another major concern of the cyber security workforce is offline phishing. Offline phishing generally refers to vishing (phishing via phone calls), messaging (scary text messages that may contain links or attachments). A system resistant to offline phishing has still not evolved. This again remains to be a major challenge before the cyber security workforce.

This study thus throws light on the shortcomings of the existing techniques as well as strives to provide solutions to various existing shortcomings.

1.4 MOTIVATION FOR THE STUDY

There is a rapid increase in cyber crimes since ages. The appearance of each and every cyber crime has changed rapidly with time. With these dynamic changes taking place from time to time, it has become difficult for the cyber security workforce to cope up with the dynamic nature of the attacks. A system that dynamically adjusts to these changes is a challenge for the cyber security professionals. The cyber attacks have caused a lot of loss of money and property due to lack of timely alerts. Time also is one of the key features towards prevention of phishing (cyber crimes).

Various methods have already been proposed to control phishing. Some of these Machine Learning techniques while others use victim based countermeasures. Victim based countermeasures include creating awareness among customers, employees, creating blacklists or using already existing ones to block the blacklisted sites and much more. A system that could dynamically adapt to changes and generate timely alerts is what I would focus on. Such a system can actually contribute largely towards preventing phishing attacks effectively without causing further consequences of phishing.

1.5 AIM AND OBJECTIVES OF THE STUDY

1.5.1 AIM

The aim of this research work is to detect phishing using techniques of Machine Learning (ML).

1.5.2 OBJECTIVES

Inorder to achieve the above mentioned aim, some objectives that will be followed have been listed as follows:

i. Machine Learning techniques will be used for detecting phishing.
ii. Content based analysis of the available data.

1.6 METHODOLOGY OF THE STUDY

This study emphasizes on providing countermeasures against offline as well as online phishing by adopting a combination of various machine learning techniques that together can help us control all types of phishing effectively. In order to put a check on online and offline fishing, following steps have been put forth:

I. **Input**- The input given to the designed system will be of 4 different categories namely,
   a. Email/SMS contents
   b. Audio (Call recording)
   c. Other online communications

II. **System**- As mentioned in the above point, input given to the system will be of 3 different types and based on these types, the system will apply separate techniques or algorithms for each different type of input. These are machine learning algorithms that have proved to be most effective for this purpose. In this type of system, machine learning algorithms with a few modifications will be used for better results with respect to time as well as accuracy.

III. **Output**- The machine learning algorithms used by the system uniquely for each type of input type has an output dependent on the algorithm that has been used for predicting the result. But the final output will be in the form of alert messages in case of any suspicious source of input or else it will just inform the user that the source of input is safe enough for further use.

PICTORIAL REPRESENTATION

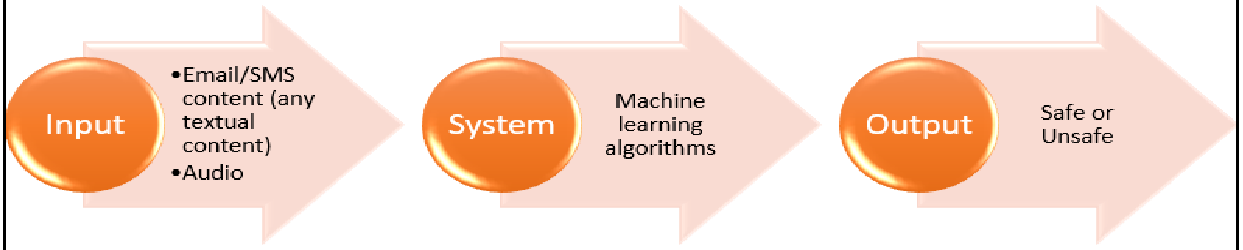

Input
•Email/SMS content (any textual content)
•Audio
System
Machine learning algorithms
Output
Safe or Unsafe

Figure 1. Pictorial representation of the proposed generalized model

1.7 PROPOSED MODEL

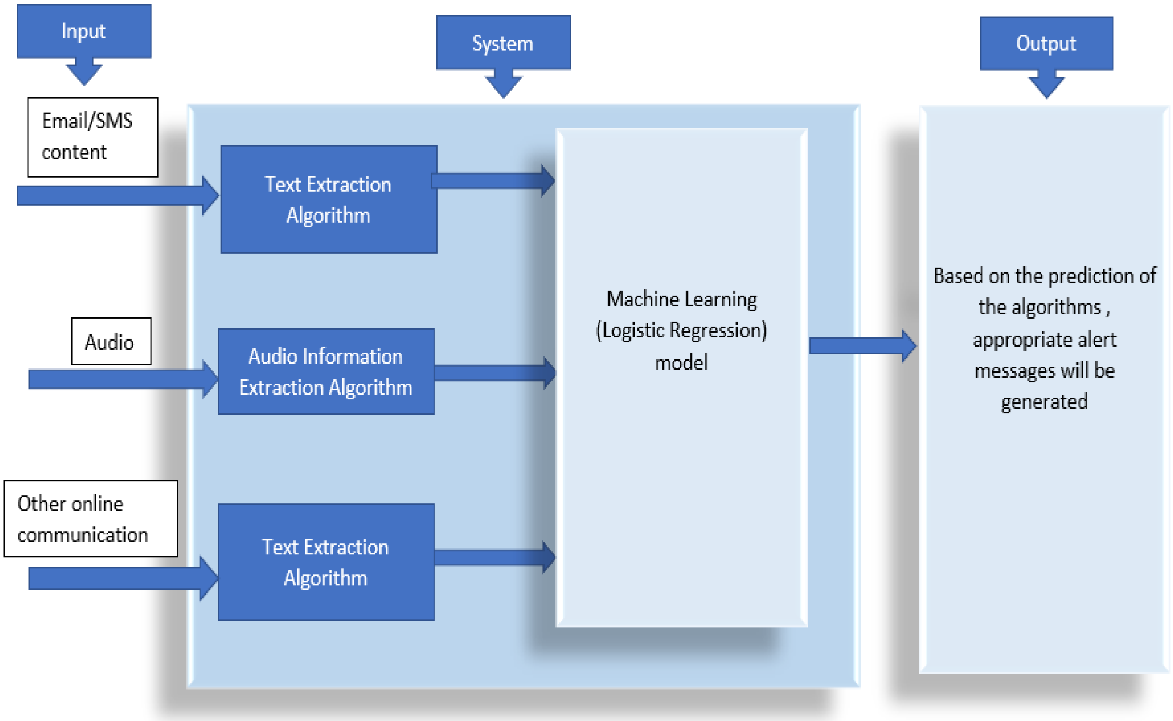

Input
Email/SMS content
Audio
Other online communication
System
Text Extraction Algorithm
Audio Information Extraction Algorithm
Text Extraction Algorithm
Machine Learning (Logistic Regression) model
Output
Based on the prediction of the algorithms , appropriate alert messages will be generated

Figure 2. Conceptual model of the study

1.8 SCOPE OF THE STUDY

This study largely focuses on providing solutions for almost all types of phishing i.e. online as well as offline. In online phishing, phishing that is carried out most commonly is in the form of emails containing scary messages with attachments or hyperlinks that demand urgent action and also nowadays any other kind of online communication can also be a bait for the users. This is increasingly exponentially. In other communications, social media, websites, etc have been included. In the offline type of phishing, gathering information via SMS (sending scary messages with hyperlinks demanding urgent action) and phone calls (victims receive phone calls from the phishers masqueraded as legitimate institutions demanding for personal information of the victim) are included.

Note: Compelling the victim to provide personal information via any means falls under social engineering.

PICTORIAL REPRESENTATION OF SCOPE OF THE STUDY

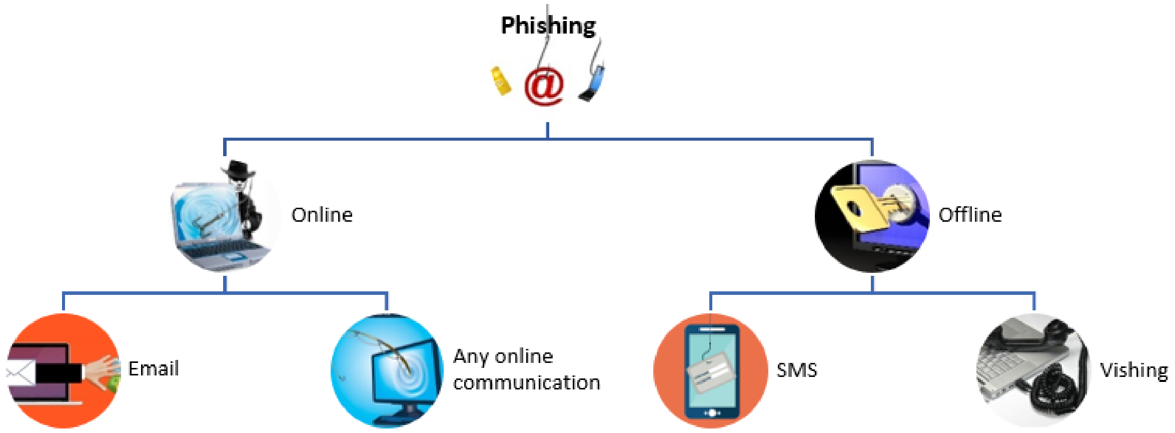


Figure 3. Scope of the Paper

1.9 SIGNIFICANCE OF THE STUDY

The already existing studies do not give a solution for all the problems (online as well as offline phishing) collectively. This study would provide an overall solution to almost all types of phishing attacks. Specifically:

I. Majority of the studies focus on classifying URLs as phishing or not phishing. That too in altogether different ways. This study gives the most optimal method for online as well as offline phishing.

II. As time passes on, the cyber attacks have intensified and have started targeting all types of online transactions or communications. We must start devising techniques to put a check on phishing carried in such scenarios as well.

III. High percentage of phishing attacks take place via offline means also with the help of SMS or simply phone calls. These types of attacks have also been reported by a large number of victims and this study caters to providing solutions for this type of phishing attacks as well.

1.10 DEFINITIONS OF TERMS

Phishing: Phishing is a fraudulent practice where individuals are indulged into providing their personal, confidential information such as passwords, important documents, bank account numbers etc. in response to an email or SMS or phone call that appears to be sent from a legitimate institution.

Phisher: A phisher is an attacker or cyber criminal that indulges individuals into providing their personal details by exploiting any existing vulnerability of the victim’s system or simply by social engineering with motives like monetary theft, identity theft, political purposes, self satisfaction and many more.

Social engineering: This is a technique used by phishers that makes use of deception (mind games) to manipulate individuals and lure them into providing sensitive, confidential information.

Cyber crime: It is a crime that involves computer, internet as a tool for committing thefts of various types like monetary theft, proprietary information theft, identity theft, etc. It is also known as computer-oriented crime.

Vishing: It is a type of phishing involving phone calls, voice messages or recent techniques like Voice over IP (VoIP) as a tool for carrying out information theft.

Machine learning: Machine learning or ML is a recent technique wherein machines learn from their experiences (training datasets) like humans and then give outputs in the form of predictions from their learnings.

Algorithm: An algorithm is a set of instructions or set of steps to be followed in order to solve a particular problem especially used by a computer.

Classification: It is a type of supervised learning that classifies an input variable into a class that it best fits into with the help of previous knowledge, experience (training dataset) that has been given to the algorithm.

Supervised learning: It is a type of learning employed by machines working on machine learning techniques that aim at mapping an input to a label based on previously fed combinations of features and labels.

Logistic regression: It is a type of machine learning algorithm, more specifically a classification algorithm that gives the probability of an input variable belonging to a particular class or label with the help of statistical methods.

Skewed dataset: A skewed dataset refers to a data where the majority of the data points belong to a single class. Thus the other class has a lesser number of datapoints as far as a 2-class model is considered.

Information extraction: Information extraction is a process of extracting some information from a larger information source. Herein information can be in the form of text or audio.

CHAPTER TWO

REVIEW OF LITERATURE

2.1 RELATED WORKS

Abuzuraiq et al. (2020), in their paper- Intelligent Methods for Accurately Detecting Phishing Websites, have reviewed various machine learning techniques that have been used in previous studies. In this paper, different phishing detection approaches were considered and a conclusion was drawn based on the accuracies and performances of each employed approach. These approaches were classified into three main groups namely, Content-Based approach, Heuristic-Based approach and Fuzzy rule-based approach. The approach of specifying weights to the words that draw out from URLs and HTML contents such as Brand name had a drawback of dependency on third party server i.e. Yahoo Search and gave an accuracy of 98.20%. Another approach of utilizing a logo image to determine the identity of the web page by matching real and fake web pages gave an accuracy of 93.40% and a drawback of dependency on a third party server which is google image search. Usage of URLs heuristics and website rank, took a long time in extracting the features and website ranks thus, resulting in 97.16% of accuracy. A combination of two algorithms namely, KNN (K- Nearest Neighbors) and SVM (Support Vector Machine) resulted in 90.04% accuracy that was quite low as compared to similar other studies. Usage of a fuzzy logic system took considerably higher amounts of time as it was implemented in five phases that resulted in 98.17% accuracy. Combination of fuzzy systems and neural networks proved to be 99.6% accurate though setting rules and membership functions were a challenging task.

Basnet et al. (2014), in their paper have proposed a method for detecting phishing URLs wherein the input to the system is URL (phishing or non-phishing), here phishing URLs are considered as positive class while non-phishing are considered as negative class, which is then redirected to a feature collector that checks for keywords, search engines, reputation, lexical and extracts certain URL features that are given to a classifier which produces the final output in terms of whether the URL is a phishing or a non-phishing one. This method employs various techniques for giving out predictions. It employs detection of phishing URLs by using the information present on the URL alone without looking at the actual web page contents regardless of the context or medium of the URL. It examines the publicly available information about a URL in evaluating whether the URL is phishing or not. Various studies employing different machine learning classifiers have been compared in this paper and the best method suitable for classifying URLs into phishing, non-phishing has been determined. The paper has tried to meet the requirements of the methodology that can be used for near real-time applications in detecting phishing URLs. A dynamic model that would fit to all types of modifications that might take place in the structure and contents of a phishing URL, has been modelled in this research work.

Marchal et al. (2016), focuses on a very crucial topic- limitations of phishers. From the observations enlisted in this paper, phishers try to make a phishing web page that resembles its target but they do not have unlimited freedom structuring the phishing web page. Also,a webpage can be characterized by a small set of key terms, how these key terms are used in different parts of a web page is different in the case of legitimate and phishing websites. Based on these observations, a phishing detection system with several notable properties has been designed in this piece of work. This system requires little training data, scales well to much larger test data, is language-independent, fast, resilient to adaptive attacks and implemented entirely on client-side. Also, this paper focuses on identifying the target websites that a phishing web page is attempting to mimic. In order to increase their chances of success, phishers try to make their phish mimic its target closely and obscure any signal that might tip off the victim. However, in crafting the structure of the phishing web page, phishers are restricted in two significant ways. First, external hyperlinks in the phishing web page, especially those pointing to the target, are to domains outside the control of phishers. Second, while phishers can freely change most parts of the phishing page, the later page of its domain name is constrained as they are limited to domains that phishers control. A web page can be represented by a collection of key terms that occur in multiple parts of the page such as its body text, title, domain name, other parts of the URL etc that can further be used in differentiating legitimate and phishing web pages. These were some of the most significant observations highlighted in this paper.

Zhao et al. (2020), while there exists a rich body of work on active learning, this paper focuses on problems with two distinguishing characteristics: severe class imbalance (skew) and small amounts of training data. Both of these problems occur with surprising frequency in many web applications. For instance, detecting offensive or sensitive content in online communities (pornography, violence and hate-speech) is receiving enormous attention from industry as well as research communities. These kinds of problems have both - a vast, majority offensive content, so the number of positive examples for such content is orders of magnitude smaller than the negative examples. In order to address both of these issues, a hybrid active learning algorithm (HAL) that balances exploiting the knowledge available through the currently labelled training examples with exploring the large amount of unlabelled data available. Classifiers trained on the examples selected for labelling by HAL easily out-performs the baselines on target metrics like recall at a high precision threshold and area under the precision-recall curve given, the same budget for labeling examples. HAL offers a simple, intuitive and computationally tractable way to structure active learning that can significantly amplify the impact or reduce the cost of human labelling for a wide range of web applications.

(Analytics India Magazine, 2018). This research work emphasizes on putting together common types of classification algorithms namely Logistic Regression, Naive Bayes, Stochastic Gradient Descent, K-Nearest Neighbours, Decision Tree, Random forest, Support Vector Machine. Classification can be performed on structured or unstructured data. It is a technique of categorizing data into a given number of classes, for instance, predicting whether the URL given to the system as an input is phishing or non-phishing. The main aim of a classification problem is to identify the category or class to which a new data will fall under. This research work compares all the above mentioned algorithms and talks about their respective accuracies and f-scores. It finds that Logistic Regression accounts for 84.6% of accuracy and has an f-score of 0.6337. While, Naive Bayes, Stochastic Gradient Descent, K-Nearest Neighbours, Decision Tree, Random Forest, and Support Vector Machine have accuracies of 80.11%, 82.20%, 83.56%, 84.23%, 84.33% and 84.09% respectively. Also, they account for f--scores of 0.6005, 0.5780, 0.5924, 0.6308, 0.6275 and 0.6145 respectively.

(MonkeyLearn, n.d.). Text classification is the process of assigning tags or categories to text according to its content. It’s one of the fundamental tasks in Natural Language Processing (NLP) with broad applications such as sentiment analysis, topic labelling, spam detection and intent detection. Unstructured data in the form of text is everywhere: emails, chats, web pages, social media, support tickets, survey responses, and more. Text classification is the task of assigning a set of predefined categories to free-text. Text classifiers can be used to organize, structure and categorize pretty much anything. Text classification can be done in two different ways: manual and automatic classification. In the former, a human annotator interprets the content of text and categorizes it accordingly. This method usually can provide quality results but it's time-consuming and expensive. The latter applies machine learning, natural language processing, and other techniques to automatically classify text in a faster and more cost-effective way. There are many approaches to automatic text classification, which can be grouped into three different types of systems:

1) Rule-based systems
2) Machine Learning based systems
3) Hybrid systems

All the above mentioned works have certain drawbacks as well. Offline phishing and phishing carried out by various different means like SMS, any online communication, etc. Most of the studies focus on classifying URLs as phishing or not phishing. Though there are restrictions on the phishing URLs, techniques employed by the phishers are of varied types and, in order to put a check on these dynamic types of attacks there is a need of a system that dynamically adjusts to these changes or a system that captures the features that cannot be omitted by any phishing URL, sites, or rather any such phishing attempt. Also, there is a certain percentage of errors that every technique induces. These discrepancies have to be eliminated and a compatible, accurate system for all types of phishing attempts has to be designed for a safer, secured future. Vishing control mechanisms are not yet efficient. There is no such mechanism that gives alerts while the call goes on. This remains to be a major challenge before technicians.

Summarizing the whole thing, systems that can dynamically adapt to the changing and advancing features and techniques employed phishers (social engineering being the deadliest of them all), systems that provide solutions to almost all types of phishing or phishing attempts, systems that can give timely alerts so as to avoid further damage (in all scenarios like, offline phishing-SMS, vishing, social media, or rather any online or offline communication) is the need of the hour.

CHAPTER THREE

METHODOLOGY AND MODEL

3.1 METHODOLOGY

Methodology is the term that refers to the different mechanisms adopted for a specific field of study. It is also involved in analysis of the adopted mechanisms. It comprises providing theoretical basis to a study that further includes implementation and detailed analysis of the employed techniques or mechanisms. Methodology generally includes the detailed analysis of the adopted techniques or mechanisms with respect to the theoretical analysis. The methods included in the methodology of a study give an overview of the final results or outputs have been obtained. There are various steps or various methods that are employed by any study. These employed techniques are reading through the lines of any problem and coming up with a research problem, collecting related data and related literature for the designed research problem, formulating your own hypothesis based on existing works in that specific field of study, collecting data through the power of experimentation, drawing conclusions based on these experiments, conducting surveys, etc.

3.2 ARCHITECTURAL MODEL

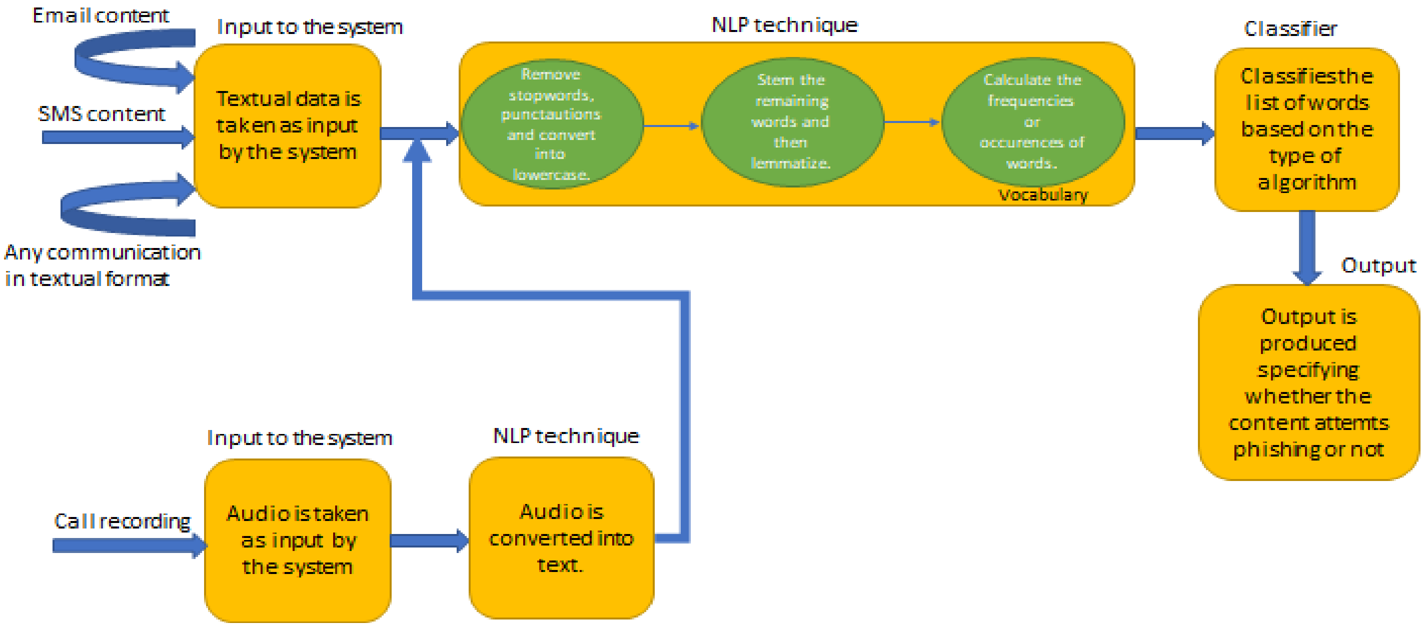

Email content
SMS content
Any communication in textual format
Input to the system
Textual data is taken as input by the system
NLP technique
Remove stopwords, punctautions and convert into lowercase.
Stem the remaining words and then lemmatize.
Calculate the frequencies or occurences of words.
Vocabulary
Classifier
Classifiesthe list of words based on the type of algorithm
Output
Output is produced specifying whether the content attemts phishing or not
Input to the system
Call recording
Audio is taken as input by the system
NLP technique
Audio is converted into text.

Figure 4. Architectural model

Architectural model, gives a glimpse of the system with respect to email, SMS, any online or offline communication, audio (call recording, any other voice communication). Recent studies have proven logistic regression to be a better option for classifying phishing URLs.

3.3 SYSTEM ARCHITECTURE

The proposed model involves the following steps-

1) Input- The input is expected from the user in the form of textual data or audio (call recording). Phishing is carried out via various ways namely email, SMS, calls (vishing) etc. In order to encounter offline phishing, provisions for testing SMS and call contents have been made in this system. There will be 2 separate input locations wherein input for textual data will be taken separately and also the acoustic data will be taken as input separately.

2) NLP technique (or precisely, the primary ML algorithm)- Here also, the textual and the acoustic inputs will be treated separately. The textual input will directly go through the NLP processes. First, the textual data will be treated as input to the NLP technique and then further processing upon that data would begin. The punctuation marks, stopwords (stopwords are words that are general purpose words and do not add much meaning to the text and these can be priorly defined by the programmer. Example, words like, a,an, the, upon etc will be treated as stopwords in this case) will be removed from the piece of text and then all the words will be converted into lowercase. This is preceded by a critical step known as stemming, wherein the remaining words (after stop words have been removed) will be reduced to their original forms. The suffixes and prefixes will be removed in this process called stemming. For example, carries, carried all will be reduced to their stem ‘carri’ by removing the suffixes es, ed respectively. From this example it is clear that there is no such word ‘carri’ in English. That's where another important technique called lemmatization comes into picture. Lemmatization helps in converting these stemmed words into words that exist in English language. This means that the stemmed word ‘carri’ will be converted to carry. Though lemmatization is a slower process as compared to stemming, it is evident from the above example that it is of prime importance. Then finally, words that occur more than once will be stored only once with their corresponding frequencies and this will create the vocabulary.

   In case of acoustic data, with the help of NLP, the speech will be converted into textual data and then as the textual data is processed, in the same way this data will also be.

3) Classifier (secondary ML algorithm)- It is difficult to conclude from the frequencies of words whether the input has an attempt of phishing being made. Hence, there is a need for another ML algorithm that would be able to give accurate predictions. Many studies have found out that out of all the classification techniques, regression works really well for classification. Therefore, here, regression will be used as the classifying technique.

As it is well known that regression gives values as output, here the probability of the input being an attempt of phishing will be the output of this algorithm.

4) Output- The output produced by the ML algorithm will be expressed in terms of percentage simply by multiplying the probabilities by 100. From this output the user himself can differentiate how harmful and how safe the email/ SMS/ call is.

3.4 NEED FOR THIS MODEL

The above mentioned model is of great use for all in the following ways:

1) A single model takes care of online as well as offline phishing.

2) Percentages are given as outputs so as to give a clear idea of the amount of data that is found to be suspicious and the one that is completely safe.

3) No additional storage or no other additional softwares are required for this model, just the application would suffice.

4) Easy to use for everyone.

3.5 SHORTCOMINGS OF THIS MODEL

The proposed model might fall short in the following situations-

1) Phishing cannot be controlled on the spot. Users will have to go to the application then check and then identify. Spontaneous response is not possible.

2) The outputs of ML algorithms are mere predictions that can have slight errors.

3) One must know how to record calls and use the app.

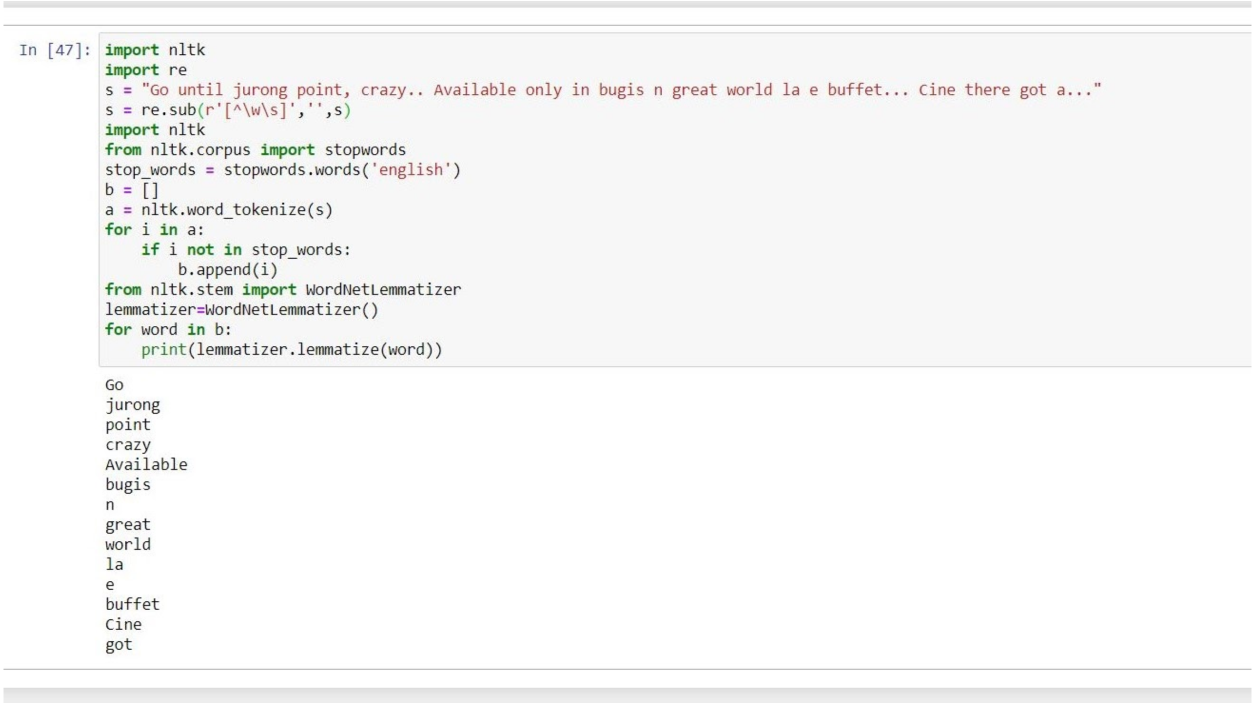
CHAPTER FOUR

IMPLEMENTATION, FUTURE WORK AND CONCLUSION

4.1 IMPLEMENTATION

4.1.1 PRE-PROCESSING

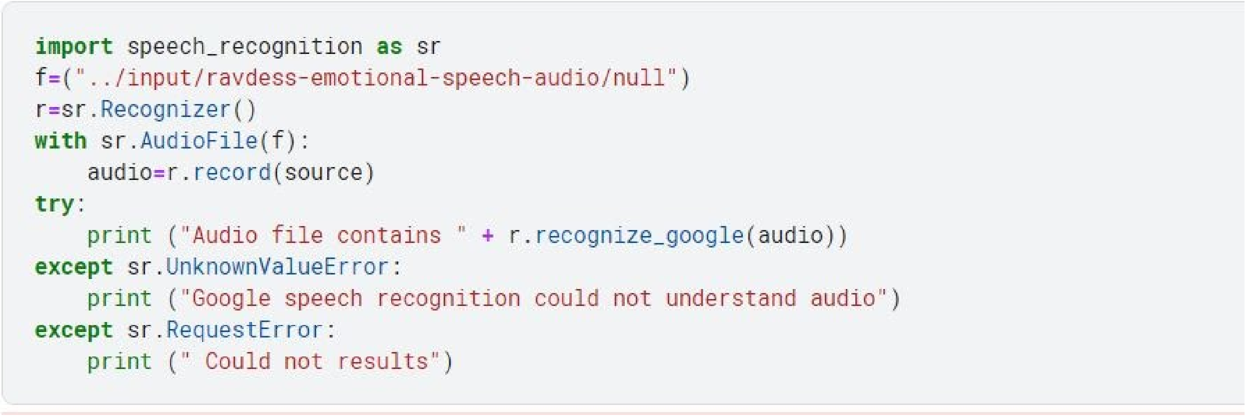
The textual input given by the user is first pre-processed by NLP techniques. The following code proves to be useful for this purpose:



```
In [47]: import nltk
         import re
         s = "Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got a..."
         s = re.sub(r'[^\w\s]','',s)
         import nltk
         from nltk.corpus import stopwords
         stop_words = stopwords.words('english')
         b = []
         a = nltk.word_tokenize(s)
         for i in a:
             if i not in stop_words:
                 b.append(i)
         from nltk.stem import WordNetLemmatizer
         lemmatizer=WordNetLemmatizer()
         for word in b:
             print(lemmatizer.lemmatize(word))
```

```
Go
jurong
point
crazy
Available
bugis
n
great
world
la
e
buffet
Cine
got
```

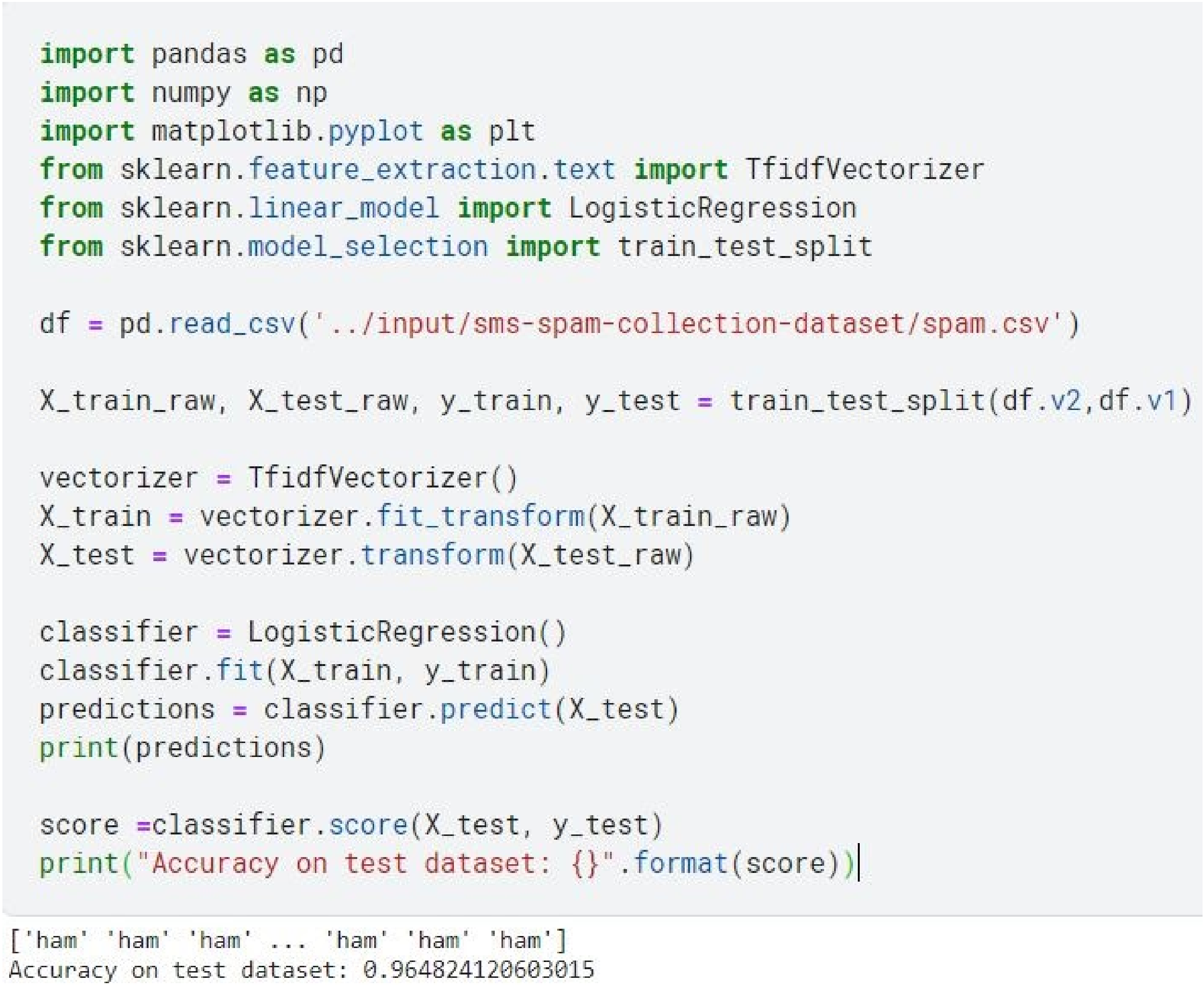
The audio file (call recording, etcetera) given as input by the user is first pre-processed by NLP techniques. The following code proves to be useful for this purpose:



```
import speech_recognition as sr
f=("../input/ravdess-emotional-speech-audio/null")
r=sr.Recognizer()
with sr.AudioFile(f):
    audio=r.record(source)
try:
    print ("Audio file contains " + r.recognize_google(audio))
except sr.UnknownValueError:
    print ("Google speech recognition could not understand audio")
except sr.RequestError:
    print (" Could not results")
```

4.1.2 MACHINE LEARNING TECHNIQUE - LOGISTIC REGRESSION

The results of the pre-processing are then given to the Machine Learning (here, Logistic Regression) model which then gives a label (spam/ham) as the final output.The following code proves to be useful for this purpose:



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split

df = pd.read_csv('../input/sms-spam-collection-dataset/spam.csv')

X_train_raw, X_test_raw, y_train, y_test = train_test_split(df.v2,df.v1)

vectorizer = TfidfVectorizer()
X_train = vectorizer.fit_transform(X_train_raw)
X_test = vectorizer.transform(X_test_raw)

classifier = LogisticRegression()
classifier.fit(X_train, y_train)
predictions = classifier.predict(X_test)
print(predictions)

score =classifier.score(X_test, y_test)
print("Accuracy on test dataset: {}".format(score))
```

```
['ham' 'ham' 'ham' ... 'ham' 'ham' 'ham']
Accuracy on test dataset: 0.964824120603015
```

4.2 FUTURE WORK

In future, I plan to work on videos that promote or attempt phishing. This can be achieved by breaking up the video into frames and then analyzing those frames sequentially so as to make out the exact intended meaning out of it and thus, prevent innocent people from these kinds of phishing attempts as well. Considering the dynamic and evolutionary nature of phishing attacks or attempts, video analysis also becomes a crucial part in detecting phishing attacks and timely prevention of them so as to minimise loss of property, proprietary information, identity, or in that case any kind of loss caused post phishing.

4.3 CONCLUSION

A dataset containing spam and ham content was used to train the model followed by test train split. From the above code it is evident that the proposed model (Logistic Regression) has an accuracy of 96.48% which in turn indicates a good performance. Hence, this kind of a system or model can prove to be helpful in timely prevention phishing attempts.

REFERENCES

Whittaker, C., Ryner, B., Nazif, M. (2010). Large-Scale Automatic Classification of Phishing Pages, pg 1. https://storage.googleapis.com/pub-tools-public-publication-data/pdf/35580.pdf

Garera, S., Provos, N., Chew, M., Rubin, A.D.,(2007). A framework for detection and measurement of phishing attacks, pg 1. https://dl.acm.org/doi/10.1145/1314389.1314391

Malwarebytes Labs (n.d.). What is Phishing? Retrieved from
https://www.malwarebytes.com/phishing/.

Fruhlinger, J., (2020). What is phishing? How this cyber attack works and how to prevent it. Retrieved from
https://www.csoonline.com/article/2117843/what-is-phishing-how-this-cyber-attack-works-and-how-to-prevent-it.html

Goodchild, J., (2019). New research finds hospitals are easy targets for phishing attacks. Retrieved from
https://blog.malwarebytes.com/101/2019/03/new-research-finds-hospitals-are-easy-targets-for-phishing-attacks/

Malwarebytes Labs (2019). Businesses: It's time to implement an anti-phishing plan. Retrieved from https://blog.malwarebytes.com/101/2019/02/business-anti-phishing/

Abuzuraiq, A., Al-kasassbeh, M., Almseiden, M. (2020). Intelligent Methods for Accurately Detecting Phishing Websites, pg 1, 3.
https://www.researchgate.net/publication/340969538_Intelligent_Methods_for_Accurately_Detecting_Phishing_Websites

Basnet, R., Sung, A., Liu, Q. (2014). Learning to detect phishing URLs, pg 1, 3.
https://www.researchgate.net/publication/273302231_LEARNING_TO_DETECT_PHISHING_URLS

Marchal, S., Saari, K., Singh, N., Asokan, N. (2016). Know Your Phish: Novel Techniques for Detecting Phishing Sites and their Targets, pg 1.

https://www.researchgate.net/publication/306063325_Know_Your_Phish_Novel_Techniques_for_Detecting_Phishing_Sites_and_Their_Targets

Zhao, Q., Xie, J., Tata, S., Najork, M. (2020). Active learning for skewed data set, pg 1. https://research.google/pubs/pub49176/

Analytics India Magazine (2018). 7 Types of Classification Algorithms. Retrieved from https://analyticsindiamag.com/7-types-classification-algorithms/

MonkeyLearn (n.d.). Text Classification. Retrieved from https://monkeylearn.com/text-classification/