Detecting phishing e-mails using Text Mining and features analysis

Luisa Franchina[1], Serena Ferracci[1] and Federico Palmaro[2]

[1]*Hermes Bay S.r.l.*
[2]*Prisma S.r.l.*

Abstract

Phishing e-mails are used by malicious actors with the aim of obtaining sensitive information from a victim, deceiving or blackmailing them. An inattentive or uninformed user may often fail to recognise if an e-mail is sent by an authentic sender or is a scam. We therefore sought to develop a method that can effectively and efficiently detect phishing e-mails and report them to the user. We analyse all the information available on receipt of the e-mail both statically and performing text mining on the content and subject of the e-mail. In addition to indicating weather e-mails are suspicious, the degree of accuracy with which the above statement is made is also reported, and the aspects of the e-mail that are characteristic of a phishing e-mail are highlighted. Excellent results were achieved with our methodology, reaching 99.2% accuracy.

Keywords

Keywords: Text Mining, Static Analysis, Cyber-Security, Phishing, Software Security, Classification.

1. Introduction

Phishing is a type of Internet scam in which attackers try to trick a victim into providing personal information, financial data or access codes by posing themselves as a trustworthy entity in a digital communication. According to Internet records, the first time the term 'phishing' was used and registered was on 2 January, 1996 [1]. This type of attack has spread very quickly, growing 3x in a decade (2008-2018, latest data available May 2019 [2]). Also with the current need for smart working in more and more sectors, many operations that were previously carried out in person are now carried out by e-mail, increasing the themes and templates for phishing attacks and the likelihood of the user taking the bait.

In this respect, it is necessary to create a detection system that is both effective and efficient. Since this system must protect the user, but also, considering the average number of incoming e-mails in a company's mailbox, it must still ensure usability. Other works have dealt with the same problem by considering different aspects of the e-mail or using different techniques to determine whether the analysed e-mail is phishing or not. Unfortunately, some of these focus on just one aspect of the e-mail, such as the phishing site that the e-mail links to you to or the body of the e-mail that in the case of phishing presents particular characteristics.

ITASEC21: Italian Conference on CyberSecurity (ITASEC), April 07–9, 2021

l.franchina@hermesbay.com (L. Franchina); s.ferracci@hermesbay.com (S. Ferracci); f.palmaro@prismaprogetti.it (F. Palmaro)
https://github.com/serenaferracci (S. Ferracci)

The problem with these solutions is that by focusing on a single aspect, these may not recognise some phishing e-mails. In particular, there are phishing e-mails that only contain an attachment of the size of the window opened by the user, so any click made on the page will redirect to the phishing site. In this case, the e-mail does not present any body to be analysed and therefore will not be identified as phishing by the detection system.

Our aim is to make the best possible use of all the information we have on phishing e-mails. This includes both the metadata describing the composition of the e-mail and its content, i.e. the body and the subject line. Starting with an e-mail in its raw format, the metadata and the content are analysed separately using different techniques, in the first case a static analysis is performed to get information about the composition of the e-mail, while in the second case a text mining analysis is performed that allows us to characterise the nature of the text and how close it is to a phishing e-mail. During the analysis, a score is assigned which represents the level of dangerousness of the e-mail. The higher the score, the greater the probability that the e-mail is a phishing e-mail.

The rest of the paper is structured as follows: related work is presented in Section 2, the proposed methodology is introduced in Section 3, followed by the presentation of results and discussion in Section 4, while concluding remarks and future work are illustrated in Section 5.

2. Related Work

The solution proposed so far can be divided in three main areas: blacklist, heuristics and data mining [3]. Blacklists are the simplest and lightest, they are frequently updated lists of previously detected phishing URLs, Internet Protocol (IP) addresses or keywords. Whitelists, on the other hand, are the opposite, and could be used to reduce false positive rates. Blacklists do not provide protection against zero-day attacks (i.e., attacks that were not seen previously), as a site must be previously detected to be included in the list. Heuristics are characteristics that are found to exist in phishing attacks in reality, however the characteristics are not guaranteed to always exist in such attacks. If a set of general heuristic tests are identified, it can be possible to detect zero-day phishing attacks. The last one is data mining, solution that exploit machine learning techniques to classify or cluster (using algorithms such as k-Nearest Neighbors (k-NN), k-means, Support Vector Machines (SVM)) the incoming e-mail. To be able to use this type of detection we need both a list of characteristics to be taken into account during the analysis of the e-mail and a suitable and labelled dataset for the training and testing phase.

In conclusion, detection by heuristics is a good balance between efficiency, since very promising results are obtained both for attacks already seen and for zero-day attacks, and effectiveness, because it does not need a training phase or time-consuming computation.

In the literature there are many heuristics based methodologies. Most of them, however, only consider selective parts of the e-mail and exclude some important aspects. The most popular approach for this kind of analysis focuses on the URLs or destination websites.

For phishing detection, Pandey and Ravi in [4] extract 17 features from the source code and URL of the websites. Also, the URL is used for checking the results of the search engine, blacklist and SSL certificate of a website. The combination of these two approaches proved potent in detecting phishing websites accurately.

Basnet [5] introduces a technique that provides insights into the effectiveness of using different machine learning algorithms for the purpose of classification of phishing e-mails. They focus mainly on URLs and metadata, extracting 10 characteristics from the metadata and also obtaining information from external sources (e.g., WHOIS), which could affect performance, but they only partially analyse the text of the e-mail by searching for a few keywords chosen statically in advance. The authors managed to achieve the best results using Biased SVM and Artificial Neural Networks which gave 97.99% accuracy.

Afroz and Greenstadt [6] proposes a phishing detection approach called Phishzoo that uses profiles of trusted websites' appearances built with fuzzy hashing techniques to detect phishing. This method makes profiles of sites that consist of fuzzy hashes of several common content elements (e.g., URL, images, most used texts, HTML codes, script files, etc.), which are related to their structure and appearance. The profiles are stored in a local database and are matched against all sites at the time of loading. It uses 636 phishing sites from www.phishtank.com and 20 profiles of legitimate sites and detects attacks with 97% accuracy.

Finally, Aburrous et al. [7] presents a classification model for detecting e-banking phishing websites based on fuzzy logic combined with data mining algorithms to characterize e-banking phishing website factors. The paper investigates phishing techniques by classifying the phishing types and defining six e-banking phishing website attack's criteria with a layered structure.

Other methods are based on text analysis, without considering other characteristics, which are easily bypassed if the e-mail consists of a single clickable attachment and no text.

Peng, Harris, & Sawa [8] develop a system, SEAHound, to analyse e-mails' text and detect inappropriate statements which are indicative of phishing attacks. They focus on the natural language text contained in the attack, performing semantic analysis of the text to detect malicious intent. Natural language processing (NLP) is used to parse each sentence and identify the semantic roles of important words in the sentence in relation to the predicate, determining if the sentence is a question or command. Potential topics are extracted by finding (verb-direct object) pairs which are evaluated by whether they are contained in a topic blacklist of malicious pairs. The blacklist is generated using supervised machine learning based on the pairs found in a training set of phishing and legitimate e-mails.

The approach proposed in [9] defines an automated method to identify malicious e-mails employing a recurrent neural network (RNN) content classifier, which identifies malicious features without human input. Their approach takes into account exclusively the text and the textual structure of the e-mail. Their results have 96.74% accuracy.

In [10], the authors use text mining techniques to extract word-based features from a dataset consisting of HamCorpus from the SpamAssassin [11] project, for legitimate e-mails, and Phishing-Corpus [12], for phishing e-mails. The method proposed does not imply the separation of the e-mail into header and footer. Tokenization is performed to separate words from the e-mail by using white space (space, tab, newline) as the delimiter, stop word removal to remove unimportant words, and stemming to remove infixation ending. Then each e-mail is converted into an equivalent vector by constructing a term-document-frequency (TDF) matrix where each row corresponds to a document (e-mail) and each column corresponds to a term (word) in the document.

The solution presented in this work differs from previous studies for the following reasons: (i) we selected the characteristics of an e-mail to be studied both from the metadata and from the text, body and subject, already in our possession without finding additional information whose search could affect performance (ii) we did not select a priori the number of keywords to be taken into account for the text analysis, but a text mining technique was used.

3. Methodology

The methodology reported in the following attempts to overcome the weaknesses, which have been exposed in the survey of Almomani et al. [13], affecting the solutions proposed so far. The main weaknesses reported concern the dateset, the time-consumption related to the number of features taken into consideration and the results obtained with the algorithm.

With regard to the dataset, there are solutions that conduct their tests on small datasets, for example the work of Chandrasekaran et al. [14] uses a dataset of only 200 items, or on datasets that are not considered standard. We collected more than 4000 e-mails (including both legitimate and phishing e-mails) from datasets widely used in this field of research and partitioned it in such a way as to comply with a realistic scenario. The detailed description of the dataset is given in the section 3.2.

Furthermore, it was chosen to use only features and information that can be obtained offline from the e-mail and only those that made a significant contribution to the analysis were taken into account. These kinds of features are also called basic features: they are those that can be extracted directly from an e-mail not yet processed. Basic features can be categorized as follows:

- Structural features: Structural features can be extracted from an HTML tree.
- Link features: Link features represent different features of URL links embedded in an e-mail, such as the number of links with IP, number of deceptive links (URL visible to the user), number of links behind an image, number of dots in a link and, so on.
- Element features: Element features represent the type of Web technology used in an e-mail such as HTML, scripting, particular JavaScript, and other forms.
- Word list features: A list of words may possibly characterize a phishing e-mail.

In this way, we limit both the time needed to analyse the individual e-mails and the memory required to save the data, without running into the problem of time consumption and high memory requirements that the solution proposed by Bergholz et al. [15] suffers from.

Finally, the competitive results compared to the literature obtained according to the methodology explained below are set out and analysed in section 4.

3.1. Text Mining

Text mining is often used in conjunction with text analytics. According to this approach, text data (keywords, concepts, verbs, nouns, adjectives, etc.) are extracted through the text mining process and are then used in the text analytics phase to extract useful information from the data attained. Given a dataset, the main function of a text mining algorithm is to extract the concepts that best describe the starting dataset. In order to achieve this, the following steps must be taken:

- *Collecting data:* this first phase involves collecting and selecting documents that may be useful for the subsequent analysis.
- *Pre-processing the text:* this phase involves adapting the raw text into analysable text. In particular, pre-processing and cleaning operations are carried out to detect and remove anomalies, e.g misspelled words, slang, numbers, etc. This way, the true essence of the available text is captured and its size reduced. In this phase, the following further steps are applied [16]:
  - Tokenisation, which breaks down the sequence of characters into words / phrases called tokens [17];
  - Filtering, which removes unnecessary parts of text (i.e., stopwords [18] and hapaxes [19]);
  - Lemmatisation, which considers the morphological analysis of words, i.e., grouping the various inflected forms of a word so that they can be analysed as a single entity. Lemmatisation is preferred to stemming, because the former has proved to outperform the latter [20].
- *Applying text mining techniques:* this is the most interesting phase where textual data (keywords, concepts, verbs, nouns, adjectives, etc.) are extracted using text mining techniques.

Text mining techniques are aimed at finding the thematic information hidden in a text, to facilitate the process of archiving and building a logical knowledge map. These techniques are based on certain algorithms that select the relevant parts of a document and eliminate the irrelevant ones. Text categorisation, information extraction, clustering, text summarisation are among the most common. In particular, the one of interest to us is information extraction. This is a technique that extracts meaningful information from a large amount of text. Usually this information is taken from unstructured and/or semi-structured machine-readable documents and transformed into structured information. In most cases, this activity concerns the processing of natural language texts.

3.2. Data Collection

The methodology adopted consists of two types of analysis: metadata analysis and content analysis through text mining. Both analyses need a dataset that is partitioned into a preparatory sub-dataset and a test sub-dataset. In particular, the chosen dataset consists of a total of 6055 e-mails of which 2246 are phishing e-mails [12] and 3809 are legitimate e-mails [21]. The two datasets, before being pre-processed to extract the content of the e-mails are analysed to obtain other information. So unlike other searches that apply pre-processing at this stage or use a dataset that has already been processed, we decided to analyse all e-mails in their HTML format with their respective metadata, so we have the HTML structure, with the various fields (e.g., Content-Type, Subject), that allows us to find the necessary information in a short time and to lose the least amount of information available. As mentioned, the dataset obtained was divided at the beginning into two parts. The preparatory dataset to be used to obtain a description of the composition of the phishing e-mail consists solely of phishing e-mails, 1834 e-mails to be exact. On the other hand, the test dataset consists of 412 phishing e-mails and 3809 legitimate e-mails.

Many works create and employ a datasets that are exactly half phishing and half legitimate (e.g., [22] , [23]). However, in reality, a user does not receive a phishing e-mail for every legitimate e-mail in his or her inbox. Hence, the fifty-fifty division does not reflect reality and consequently the results of the classification in percentage of false positives and false negatives may be biased.

3.3. Technique

The analysis for the classification of e-mails is mainly a static analysis, which is divided into metadata analysis and content analysis, body and subject of the e-mail, through text mining. The methodology followed to complete the analysis is illustrated in Figure 1 and described below.

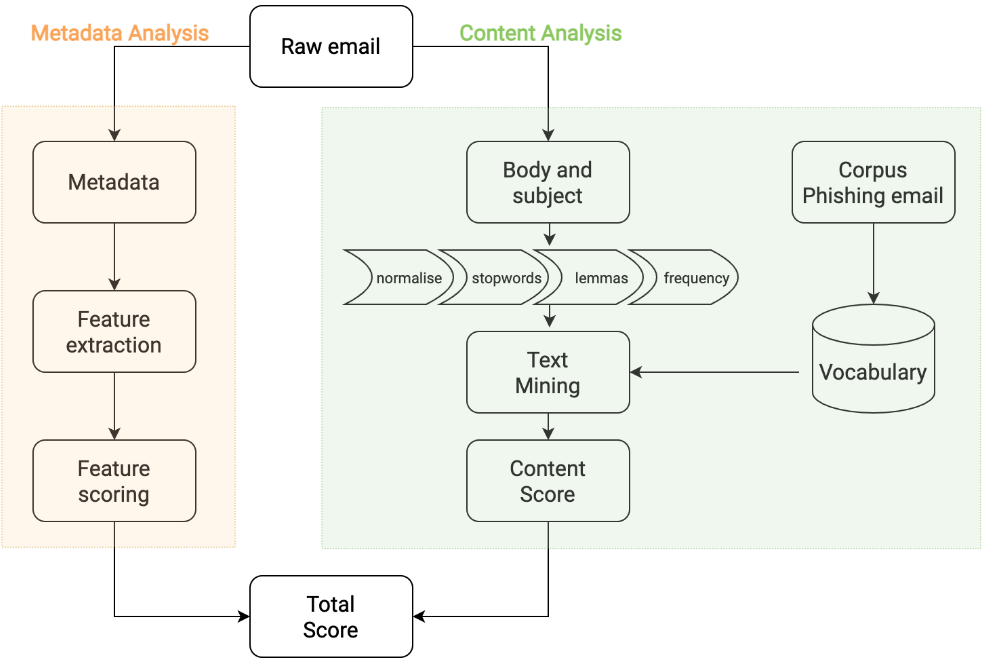
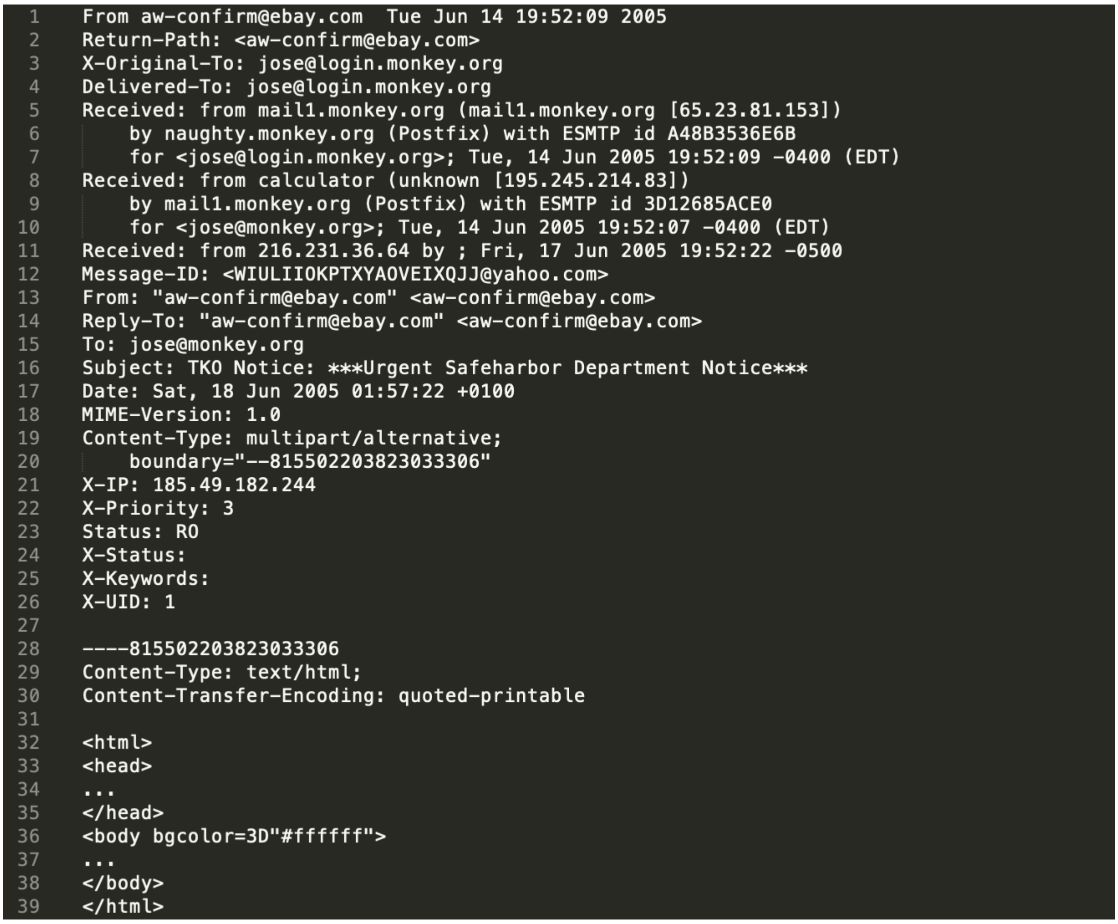


Figure 1: Methodology for detecting phishing e-mails

Metadata Analysis

Starting from a raw format e-mail, an e-mail in HTML format that is not processed yet, an example is proposed in Figure 2. As can be seen from the figure, the e-mail is divided into two sections: the fields containing all the information relating to the e-mail and the actual content of the e-mail still in HTML format. The metadata is composed of `name: value` and does not undergo any processing, through the name the most relevant fields are selected according to the objective of our analysis and the value is saved for the evaluation of the e-mail. The content analysis on the other hand consists of two parts, the HTML is analysed first to obtain information about the text, e.g. punctuation, presence of attachments or URLs. Once this phase is finished, the HTML is converted into text by means of a parse so that it is ready to be analysed by means of the text mining techniques explained below.



```
From aw-confirm@ebay.com  Tue Jun 14 19:52:09 2005
Return-Path: <aw-confirm@ebay.com>
X-Original-To: jose@login.monkey.org
Delivered-To: jose@login.monkey.org
Received: from mail1.monkey.org (mail1.monkey.org [65.23.81.153])
    by naughty.monkey.org (Postfix) with ESMTP id A48B3536E6B
    for <jose@login.monkey.org>; Tue, 14 Jun 2005 19:52:09 -0400 (EDT)
Received: from calculator (unknown [195.245.214.83])
    by mail1.monkey.org (Postfix) with ESMTP id 3D12685ACE0
    for <jose@monkey.org>; Tue, 14 Jun 2005 19:52:07 -0400 (EDT)
Received: from 216.231.36.64 by ; Fri, 17 Jun 2005 19:52:22 -0500
Message-ID: <WIULIIOKPTXYAOVEIXQJJ@yahoo.com>
From: "aw-confirm@ebay.com" <aw-confirm@ebay.com>
Reply-To: "aw-confirm@ebay.com" <aw-confirm@ebay.com>
To: jose@monkey.org
Subject: TKO Notice: ***Urgent Safeharbor Department Notice***
Date: Sat, 18 Jun 2005 01:57:22 +0100
MIME-Version: 1.0
Content-Type: multipart/alternative;
    boundary="--815502203823033306"
X-IP: 185.49.182.244
X-Priority: 3
Status: RO
X-Status:
X-Keywords:
X-UID: 1

----815502203823033306
Content-Type: text/html;
Content-Transfer-Encoding: quoted-printable

<html>
<head>
...
</head>
<body bgcolor=3D"#ffffff">
...
</body>
</html>
```

Figure 2: Example of raw format e-mail

At the end of this evaluation of the metadata fields, which we called features extraction, the complete list we obtained is the following:

- the e-mail address domain. Most organisations, except some small operations, will have their own e-mail domain and company accounts. For example, legitimate e-mails from Google will read `@google.com`. The use of a public domain is therefore suspect.
- the presence of unsafe URLs. The presence of URLs using insecure protocols, e.g. the presence of http://, can be an indicator of a phishing e-mail for the external website does not provide a secure, encrypted connection.
- the use of particular punctuation marks ('!' and '$'). Phishing e-mails are most often characterised by poor punctuation. They often use ! to emphasise the urgency of the e-mail, e.g. to obtain the credentials of the victim's bank they ask for immediate access to

a fake link, or the symbols of various currencies often appear if the purpose of the e-mail was to infect the PC and demand a ransom.

- the amount of misspelled words. In addition to poor punctuation, phishing e-mails are often characterised by poor grammar.
- the presence of attachments. The attacker may use several vectors to damage the victim. One of the most commonly used is an external link to a fake website, for which we perform URL analysis. The other vector is attachments, which may contain viruses that aim to directly or indirectly damage the computer, exploit the computer to perform future attacks or even steal information from the victim.

A score is assigned to each item on the list by assessing the harmfulness to the user of the presence of the item in question, the higher the resulting score the more dangerous the e-mail.

Content Analysis

Content Analysis is necessary for the determination of whether the content of an e-mail is phishing or legitimate. In particular, it is determined through text mining how similar the style and words of the analysed e-mail are to those of a phishing e-mail. The process involves the creation of two vocabularies, one for the corpus and the other for the subject of the e-mail. The vocabularies are built separately using a dataset of 1834 phishing e-mails. For both we execute the following phases:

- *Normalisation:* this is used to allow correct word recognition and thus resolve cases of ambiguity. For this purpose, excess spaces are removed and instead added after punctuation in case they are missing. The text is then converted either to upper case or lower case, in order to reach a certain homogeneity (in our case we have chosen to convert everything to lower case) and all characters that are not useful for analysis are deleted, e.g., emoticons, numbers, non-printable characters, etc.
- *Removal of stop-words and hapaxes:* in addition to the character filter applied in the first step, a word filter is also introduced. In particular, all stop-words are eliminated. For instance, common words, given their high frequency in a language, are usually considered insignificant, and hapaxes, words that appear only once in the corpus and are therefore of little use for the analysis we wish to carry out. In addition to these, URLs are also removed because when considered as words and not hypertext they have no useful meaning for this analysis.
- *Lemmatisation:* the process of reducing a inflected form of a word to its canonical form, called lemma. The lemmatisation process means that verb forms are converted to the present infinitive, nouns and adjectives to the masculine singular, articulated prepositions to their form without articles, and so on. This conversion attempts to limit the number of words that will later have to be taken into account for the creation of the vocabulary and at the same time allows for a cleaner vocabulary.
- *Frequency analysis:* a second word filter is applied. Words are divided into frequency bands. Three bands can be considered: high, medium and low frequencies. The high frequency band is where each word has a different number of occurrences from every

> other word. It generally consists of about 30 or 50 forms (depending on the size of the corpus) and, among these, at most 4 or 5 are main words, while the others are grammatical words. The boundary between high and medium frequencies lies just above the first parity. The mid-frequency band is characterised by having within it words with different conditions of both parity and number of occurrences. Starting then from the bottom of the word list, the boundary between medium and low frequencies is identified by the first gap in the consecutive number of increasing occurrences [24]. We consider only high and medium frequencies.

At the end of this process, we would attain, as a result, two lists of words, also called keywords, which best describe the two sections of phishing e-mails. These vocabularies are sorted in descending order according to the frequency and it is divided into three range, words in the first range are considered more suspicious than the others and so on.

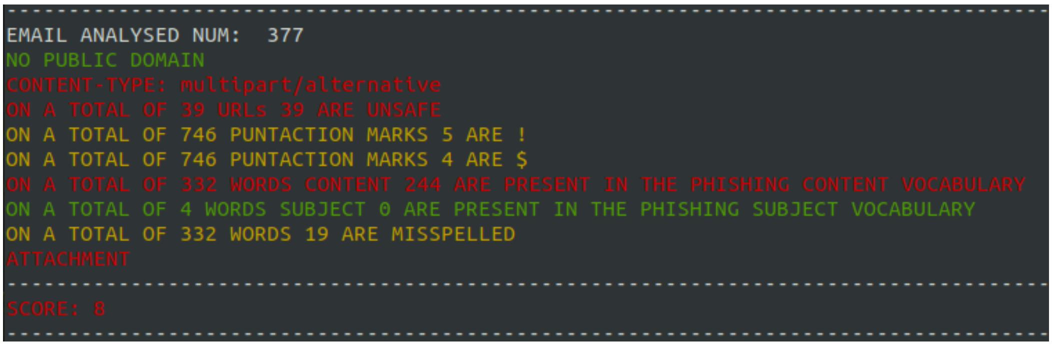
For the analysis of the testing e-mail, the body and the subject are extracted separately. The two sections are processed using the same steps used for the vocabularies except for the frequency analysis. This is necessary to bring the analysed text to a level of homogeneity that allows comparison with the respective vocabulary. At this point, for the two sections of the e-mails, we check how many words are present in the vocabularies and assigning a score in relation to the vocabulary's range. The resulting score is then normalize. In particular, the value is remapped to a fixed range of values. The starting range is from 0 to the number of words contained within the e-mail and the target range is from 0 to a maximum value set a priori. In this way, for each analysed e-mail we have a maximum score that can be assigned to each characteristic so as to favour comparisons between different e-mails and to have homogeneity between the scores obtained. The resulting value is then added to metadata score in order to obtain the final score that can vary between a minimum of 0 (the e-mail does not present any characteristics of phishing e-mails) to a maximum of 10 (the e-mail with considerable certainty is phishing).

Once all the considered e-mail characteristics have been analysed and scored, a report is presented to the user. In this way the user not only has an evaluation of the incoming e-mail, but also knows which characteristics are suspicious and can analyse them in detail if he or she so wishes.

4. Results & Discussion

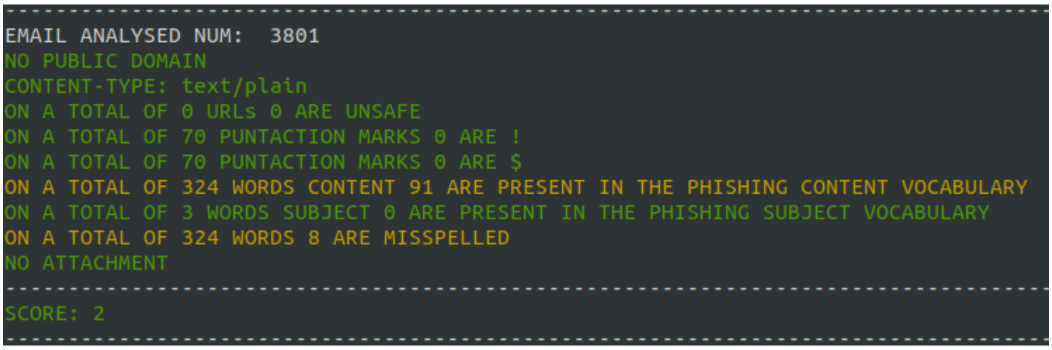
The resulting report summarises the analysis carried out presenting in detail the characteristics taken into account and highlighting their importance for the final classification. Unlike related works that only present the user with the binary result of the classification (phishing or legitimate), we have decided to provide the user with a detailed overview of the analysis performed and the characteristics that led to the classification of the e-mail as phishing or legitimate.

The aim is to increase the level of awareness and education about phishing attacks in general and phishing e-mails in particular. Contrary to the examples given in the survey [13], we do not propose interactive education. The user does not have to search for information from external sources and we do not try to test their skills, we simply expose them to the result in such a way that they understand the risk factors of the e-mail under analysis.



```
--------------------------------------------------------------------------------
EMAIL ANALYSED NUM:  377
NO PUBLIC DOMAIN
CONTENT-TYPE: multipart/alternative
ON A TOTAL OF 39 URLs 39 ARE UNSAFE
ON A TOTAL OF 746 PUNTACTION MARKS 5 ARE !
ON A TOTAL OF 746 PUNTACTION MARKS 4 ARE $
ON A TOTAL OF 332 WORDS CONTENT 244 ARE PRESENT IN THE PHISHING CONTENT VOCABULARY
ON A TOTAL OF 4 WORDS SUBJECT 0 ARE PRESENT IN THE PHISHING SUBJECT VOCABULARY
ON A TOTAL OF 332 WORDS 19 ARE MISSPELLED
ATTACHMENT
--------------------------------------------------------------------------------
SCORE: 8
--------------------------------------------------------------------------------
```

Figure 3: Example of result for phishing e-mails



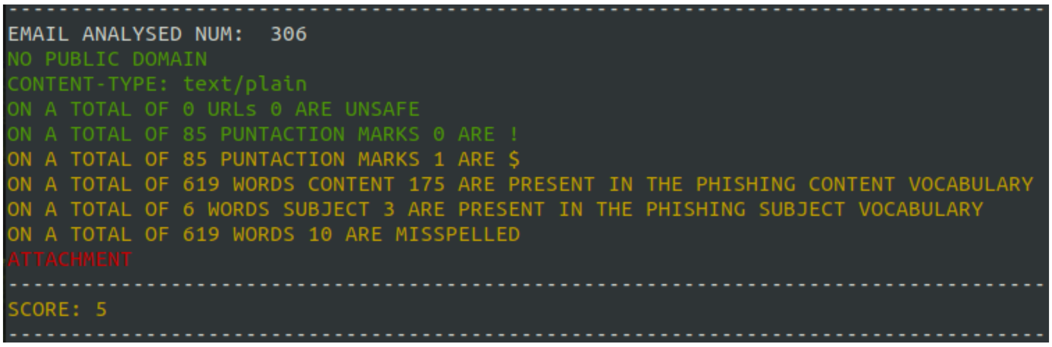
```
--------------------------------------------------------------------------------
EMAIL ANALYSED NUM:  3801
NO PUBLIC DOMAIN
CONTENT-TYPE: text/plain
ON A TOTAL OF 0 URLs 0 ARE UNSAFE
ON A TOTAL OF 70 PUNTACTION MARKS 0 ARE !
ON A TOTAL OF 70 PUNTACTION MARKS 0 ARE $
ON A TOTAL OF 324 WORDS CONTENT 91 ARE PRESENT IN THE PHISHING CONTENT VOCABULARY
ON A TOTAL OF 3 WORDS SUBJECT 0 ARE PRESENT IN THE PHISHING SUBJECT VOCABULARY
ON A TOTAL OF 324 WORDS 8 ARE MISSPELLED
NO ATTACHMENT
--------------------------------------------------------------------------------
SCORE: 2
--------------------------------------------------------------------------------
```

Figure 4: Example of result for legitimate e-mails

Figure 3 and Figure 4 show two examples of the report returned to the user, one for a phishing e-mail and the other for a legitimate e-mail respectively. Just by presenting the most relevant features for the classification of the e-mail, the user can get an idea of the nature of the e-mail. In addition, each selected feature is assigned a colour according to the score derived from the analysis. The range of scores that can be assigned to each characteristic is divided into three bands: safe, suspicious and dangerous, which correspond respectively to the three colours we can see in the report: green, yellow and red.

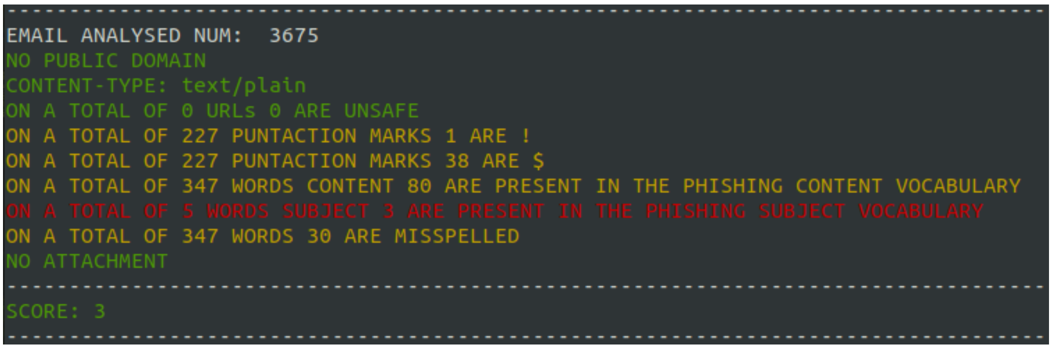
Heterogeneous feature scoring are used and the importance is evident from Figure 5 and Figure 6. Both e-mails have the same number of highlighted features: four suspicious features and a dangerous one. If all the selected features had the same score, for example 1 if the features is suspicious, 0 otherwise, then both e-mails would have been wrongly classified as phishing or legitimate, based on the chosen threshold, increasing the false positives or false negatives respectively.

That is why we chose to use heterogeneous scores for the features, these scores were decided in accordance with reports from major companies (Google [25], Apple [26], Aruba [27]) in which the most common and dangerous features were listed and sorted. So we have that some characteristics are considered more dangerous than others, for example the presence of an attachment, which may link to a site created by the attacker to steal information from the



```
--------------------------------------------------------------------------------
EMAIL ANALYSED NUM:  306
NO PUBLIC DOMAIN
CONTENT-TYPE: text/plain
ON A TOTAL OF 0 URLs 0 ARE UNSAFE
ON A TOTAL OF 85 PUNTACTION MARKS 0 ARE !
ON A TOTAL OF 85 PUNTACTION MARKS 1 ARE $
ON A TOTAL OF 619 WORDS CONTENT 175 ARE PRESENT IN THE PHISHING CONTENT VOCABULARY
ON A TOTAL OF 6 WORDS SUBJECT 3 ARE PRESENT IN THE PHISHING SUBJECT VOCABULARY
ON A TOTAL OF 619 WORDS 10 ARE MISSPELLED
ATTACHMENT
--------------------------------------------------------------------------------
SCORE: 5
--------------------------------------------------------------------------------
```

Figure 5: Importance of heterogeneous scoring [phishing]



```
--------------------------------------------------------------------------------
EMAIL ANALYSED NUM:  3675
NO PUBLIC DOMAIN
CONTENT-TYPE: text/plain
ON A TOTAL OF 0 URLs 0 ARE UNSAFE
ON A TOTAL OF 227 PUNTACTION MARKS 1 ARE !
ON A TOTAL OF 227 PUNTACTION MARKS 38 ARE $
ON A TOTAL OF 347 WORDS CONTENT 80 ARE PRESENT IN THE PHISHING CONTENT VOCABULARY
ON A TOTAL OF 5 WORDS SUBJECT 3 ARE PRESENT IN THE PHISHING SUBJECT VOCABULARY
ON A TOTAL OF 347 WORDS 30 ARE MISSPELLED
NO ATTACHMENT
--------------------------------------------------------------------------------
SCORE: 3
--------------------------------------------------------------------------------
```

Figure 6: Importance of heterogeneous scoring [legitimate]

victim or be compromised and therefore damage the victim's computer [28], is evaluated more dangerous than the presence of a subject using words from the phishing vocabulary. Taking this into account we can correctly classify the former as phishing and the latter as legitimate.

We conducted tests on a dataset of 4220 e-mails, of which 9.7% were phishing e-mails. The results obtained are shown in the table and are very promising [13]. A threshold was specified for each row to differentiate the level of danger attributed to the incoming e-mail:

- score <= 4 → safe
- 4 < score <= 7 → suspect
- 7 < score < 10 → phishing
- score = 10 → very phishing

The threshold relates to the total score assigned to the single e-mail. At first glance, the threshold set at a value of 4 might seem low to classify an e-mail as suspected phishing, but evaluating the selected characteristics and the weight given to each of them, this is not the case. For example, an e-mail that exceeds a score of 4 might contain only unsafe URLs, an attachment and come from a public domain e-mail address. Furthermore, with this threshold we have a false positive value of 8.9% out of 3809 legitimate e-mails, which can be considered high if we do not consider that the user is provided with an explanatory report of the evaluation of the

Measure	Formula	Meaning	threshold > 4	threshold > 7	threshold = 10
Precision	$= \frac{\lvert TP\rvert}{\lvert TP\rvert+\lvert FP\rvert}$	The percentage of positive predictions that are correct	55.7%	95.7%	100%
Recall Sensitivity	$= \frac{\lvert TP\rvert}{\lvert TP\rvert+\lvert FN\rvert}$	The percentage of positive labeled instances that were predicted as positive	99.2%	87.4%	66.3%
Accuracy	$= \frac{\lvert TP\rvert+\lvert TN\rvert}{\lvert TP\rvert+\lvert TN\rvert+\lvert FP\rvert+\lvert FN\rvert}$	The percentage of correct predictions	91.8%	98.8%	96.7%
F-Measure	$= 2 \cdot \frac{precision \cdot recall}{precision+recall}$	A measure of a test's accuracy. It considers both the precision and the recall of the test to compute the score	71.3%	91.3%	75.7%
False Negative	$= \frac{\lvert FN\rvert}{\lvert FN\rvert+\lvert TP\rvert}$	The percentage of negative predictions that are wrong	0.8%	12.6%	33.7%
False Positive	$= \frac{\lvert FP\rvert}{\lvert FP\rvert+\lvert TN\rvert}$	The percentage of positive predictions that are wrong	8.9%	0.4%	0%

Table 1
Classification results using testing dataset

e-mail which can be used to decide the fate of the e-mail and as explained section 5 the tool can be used as a pre-filter for more elaborate methodologies, in order to reduce the overhead of the checks. On the other hand, in the state of the art we have false positive values ranging from 2.5% to 34% and many of the studies do not report false positive data, but only accuracy data.

The use of multiple thresholds was a deliberate choice because in some cases it is not easy to evaluate an e-mail and decide whether it is phishing or legitimate because there are many aspects to consider. Instead of using a single threshold, i.e., choosing a fixed score and if the score assigned to the e-mail is lower then it is legitimate otherwise it is phishing, we decided to give more nuance to the classification by assessing the degree of dangerousness of the e-mail. The intermediate band, with a score ranging from 4 to 7, is intended to signal to the user that the e-mail has certain characteristics that indicate that it could be dangerous, but that are not sufficient to classify it as phishing, so we invite them to pay attention to the content of the e-mail, indicating precisely which aspects to pay more attention to. In any case, the evaluation thresholds can be modified by the user to make the tool adaptable to their needs.

5. Conclusion

Phishing is a particular type of fraud perpetrated on the Internet by deceiving users. It mainly takes the form of misleading e-mails. The e-mail is sent, only apparently from financial institutions (banks or credit card companies) or from websites that require access after registration (web-mail, e-commerce, etc.). The message report registration or other problems to invite people to provide their confidential information and access data to the service. In our research, we created a methodology that uses as much purpose-relevant information as possible from an e-mail analysis, correctly classifying 99.2% of phishing e-mails. The described methodology could be further extended from the point of view of both research and presentation of the output. Regarding the research, as pointed out several times during the description, the methodology uses only basic features and is therefore very efficient. In order to make the classification more accurate, one could think of using the methodology as a pre-filter or layer to be placed before a tool that performs a deeper analysis. For instance, providing a detailed analysis of the destination website of the link present in the e-mail or the analysis of the attachment, which could be static or dynamic with the help of a sandbox. In this way, we are able to get a better analysis, reducing the analysis time because the number of e-mails that need in-depth analysis is reduced thanks to the filter. The presentation of the output is particularly important for user education. Our aim would be to visually present the entire e-mail to the user and not just the statistical data extracted from the analysis. Thus, the user will have at his disposal the received e-mail, including content and metadata, on which features and individual words, that are deemed suspicious as a result of the analysis, are highlighted. So, the user can become more aware of the structure of phishing e-mails and what are the most significant characteristics to look out for when receiving an e-mail.

References

[1] {www.phishing.org} history of phishing, https://www.phishing.org/history-of-phishing, 2020. [Online; accessed 25-January-2020].

[2] {Wikipedia} phishing, https://en.wikipedia.org/wiki/Phishing, 2020. [Online; accessed 25-January-2020].

[3] M. Khonji, Y. Iraqi, A. Jones, Phishing detection: a literature survey, IEEE Communications Surveys & Tutorials 15 (2013) 2091–2121.

[4] M. Pandey, V. Ravi, Text and data mining to detect phishing websites and spam emails, in: International Conference on Swarm, Evolutionary, and Memetic Computing, Springer, 2013, pp. 559–573.

[5] R. Basnet, S. Mukkamala, A. H. Sung, Detection of phishing attacks: A machine learning approach, in: Soft computing applications in industry, Springer, 2008, pp. 373–383.

[6] S. Afroz, R. Greenstadt, Phishzoo: An automated web phishing detection approach based on profiling and fuzzy matching, in: Proc. 5th IEEE Int. Conf. Semantic Comput.(ICSC), 2009, pp. 1–11.

[7] M. Aburrous, M. A. Hossain, K. Dahal, F. Thabtah, Intelligent phishing detection system for e-banking using fuzzy data mining, Expert systems with applications 37 (2010) 7913–7921.

[8] T. Peng, I. Harris, Y. Sawa, Detecting phishing attacks using natural language processing and machine learning, in: 2018 ieee 12th international conference on semantic computing (icsc), IEEE, 2018, pp. 300–301.

[9] L. Halgaš, I. Agrafiotis, J. R. Nurse, Catching the phish: Detecting phishing attacks using

recurrent neural networks (rnns), in: International Workshop on Information Security Applications, Springer, 2019, pp. 219–233.

[10] M. Zareapoor, K. Seeja, Text mining for phishing e-mail detection, in: Intelligent Computing, Communication and Devices, Springer, 2015, pp. 65–71.

[11] {spamassassin.apache.org} apache spamassassin, https://spamassassin.apache.org/, 2020. [Online; accessed 2-February-2020].

[12] {monkey.org} phishing corpus, https://monkey.org/~jose/phishing/, 2020. [Online; accessed 21-January-2020].

[13] A. Almomani, B. B. Gupta, S. Atawneh, A. Meulenberg, E. Almomani, A survey of phishing email filtering techniques, IEEE communications surveys & tutorials 15 (2013) 2070–2090.

[14] M. Chandrasekaran, K. Narayanan, S. Upadhyaya, Phishing email detection based on structural properties, in: NYS cyber security conference, volume 3, 2006.

[15] A. Bergholz, J. H. Chang, G. Paass, F. Reichartz, S. Strobel, Improved phishing detection using model-based features., in: CEAS, 2008.

[16] S. Kannan, V. Gurusamy, S. Vijayarani, J. Ilamathi, M. Nithya, Preprocessing techniques for text mining, International Journal of Computer Science & Communication Networks 5 (2014) 7–16.

[17] M. Panda, Developing an efficient text pre-processing method with sparse generative naive bayes for text mining, International Journal of Modern Education and Computer Science 10 (2018) 11.

[18] J. Kaur, P. K. Buttar, A systematic review on stopword removal algorithms, Int. J. Futur. Revolut. Comput. Sci. Commun. Eng 4 (2018).

[19] {Wikipedia} zipf's law, https://en.wikipedia.org/wiki/Zipf_law, 2020. [Online; accessed 25-January-2020].

[20] V. Balakrishnan, E. Lloyd-Yemoh, Stemming and lemmatization: A comparison of retrieval performances (2014).

[21] |Ham Corpus Enron| the enron spam and ham datasets, http://nlp.cs.aueb.gr/software_and_datasets/Enron-Spam/index.html, 2020. [Online; accessed 21-January-2020].

[22] M. Pandey, V. Ravi, Detecting phishing e-mails using text and data mining, in: 2012 IEEE International Conference on Computational Intelligence and Computing Research, IEEE, 2012, pp. 1–6.

[23] A. Almomani, T.-C. Wan, A. Altaher, A. Manasrah, E. ALmomani, M. Anbar, E. ALomari, S. Ramadass, Evolving fuzzy neural network for phishing emails detection, Journal of Computer Science 8 (2012) 1099.

[24] S. Bolasco, Analisi multidimensionale dei dati., Carocci, 2002.

[25] {Google Phishing} avoid and report phishing emails, https://support.google.com/mail/answer/, 2020. [Online; accessed 21-January-2020].

[26] {Apple Phishing} recognize and avoid phishing messages, phony support calls, and other scams, https://support.apple.com/, 2020. [Online; accessed 21-January-2020].

[27] {Aruba Phishing} phishing, la guida per capire cos'è e come riconoscerlo, https://www.aruba.it/magazine/email/phishing-la-guida-per-capire-cosa-e-come-riconoscerlo.aspx, 2020. [Online; accessed 21-January-2020].

[28] D. C. D'Elia, E. Coppa, F. Palmaro, L. Cavallaro, On the dissection of evasive malware, IEEE Transactions on Information Forensics and Security 15 (2020) 2750–2765.