Electronics and Electrical Engineering

AIMS Electronics and Electrical Engineering, 5(1): 93–116.
DOI: 10.3934/electreng.2021006
Received: 25 January 2021
Accepted: 08 March 2021
Published: 11 March 2021

http://www.aimspress.com/journal/ElectrEng

Research article

Analysis of phishing emails

Ladislav Burita*, Petr Matoulek, Kamil Halouzka and Pavel Kozak

Department of Informatics and Cyber Operations, University of Defence, 65 Kounicova Street, 66210 Brno, Czech Republic

* **Correspondence:** Email: ladislav.burita@unob.cz; Tel: +420973442172; Fax: +420973442327.

Abstract: This research aims to describe and analyze phishing emails. The problem of phishing, types of message content of phishing emails, and the basic techniques of phishing email attacks are explained by way of introduction. The study also includes a review of the relevant literature on Web of Science and analyzes articles that deal with the threat of phishing attacks and defense against them. Data collected within a time interval of two months from two email accounts of one of the authors of the study was used for the analysis of 200 email messages. Data has been resented in tabular form, to allow further statistical processing using functions such as sum, average and frequency analysis. The core part of the study involved the classification and segmentation of emails according to the main goals of the sent message. The text analytical software Tovek, was used for the analysis, Contribution of the manuscript is in the understanding of phishing emails and extending the knowledge base in education and training in phishing email defense. The discussion compares the results of this research with those of the studies mentioned in the "Introduction" and "Literature review" sections. Furthermore, the emerging problems and limitations of the use of text analytical software are described, and finally the issue is devoted to problems with obtaining personal data from recipients' emails. The "Conclusion" section summarizes the contributions of this research.

Keywords: phishing email; analysis; statistics; segmentation (business, charity, fund, transfer, others), text analytical SW Tovek

1. Introduction

Phishing emails are a type of targeted email attack where social engineers lure the recipient into performing specific actions such as clicking on a malicious link, opening a malicious attachment, or

visiting a web page and entering their personal information [1]. Phishing attacks seek to trick recipients into believing that an email is legitimate, in order to solicit sensitive information (e.g., usernames, passwords, and credit card numbers) or install malware. As a result, phishing is a fundamental component of many cyber-attacks and is often used as a first step in advanced persistent threats [2].

Phishers use many different techniques to initiate phishing attacks, the main methods used are email, SMS, social media, instant messaging, search engines and malicious websites. Phishers always modify their methods to use any communication method available to reach their victims. The spear phishing is a highly targeted of phishing attack. Rather than sending more phishing emails to anyone, the phisher sends spoofed emails to consumers that appear to originate from somebody they know [3].

Phish Threat security experts focused on which phishing templates is the best, or more precisely the worst results. Whether corporate email users succumb more to sugar or whips. Threats or free offers. Specific instructions or useful suggestions. Wording with "you must" or "you might like." The answers cover a wide range of topics around phishing, but they have one thing in common: none of them fall into the category of threats. Most of the test fraudulent emails that users responded to dealt with common, not dramatic, issues that were obviously interesting or important topics. Nothing on the top ten list was really urgent or frightening, and all the messages sounded likely and uncomplicated enough to be worth resolving quickly. The top ten list, resp. the worst [4], is given below:

1. Code of ethics—report from the HR department representing the new code of ethics.
2. Delayed delivery of the annual tax statement—simulated warning to employees.
3. Scheduled server maintenance.
4. You have been assigned a task—partially targeted phishing.
5. Test the new email system.
6. Updating the rules for going on holidays.
7. Car lights on—the building manager obviously wants to be very helpful.
8. Undelivered shipment by courier service.
9. Secure document—it is said to be a "secure document" from the HR team.
10. Report from social networks—for example: simulated LinkedIn notification telling recipient: "You have unread messages from Josef."

Phishing detection is a subset of digital literacy that K–12 educators could include as part of the curricula on critical thinking and information literacy. K–12, from kindergarten to 12th grade, is an American expression that indicates the range of years of supported primary and secondary education found in the United States, which is similar to publicly supported school grades prior to college in several other countries [5].

The aim of this paper is to analyze the text of phishing emails and results of the research confront with research results described in the Introduction and in Literature review. Acquired knowledge of the study should help in recognition of phishing emails, defense against phishing emails attacks, and could contribute to education and training in cyber security.

This paper is organized as follows. The “Introduction” provides some phishing definitions and lists the top ten phishing attacks. The research methodological elements and steps are presented under the “Materials and methods” section. It includes the literature review, data collection, statistical surveys, classification and segmentation of phishing emails, and application of text

analytical SW. The section "Results of phishing email analysis" is a core part of this paper. It presents the statistical survey results and detailed analysis of phishing emails under the following segments: Business, Charity, Fund, Transfer, and Others. The last part of this section is "Analysis of phishing email duplicity," which deals with repeatedly sent emails. The "Discussion" section follows, where the results of the current research are compared with those of the studies mentioned in the "Literature review" subsection. We discuss the question: What is the reason for obtaining personal information? We also discuss the problems associated with using text analytical SW in analysis of personal data. Then we conclude the paper.

2. Materials and method

This section describes methodological elements and steps of the research. These steps include the literature review, data collection, statistical surveys, classification and segmentation of phishing emails, and application of text analytical SW.

2.1. Literature review in the theme "Phishing amail attackts and defense against them"

The literature review made it possible to find out what topics other researchers have dealt with in phishing. The literature source was publications indexed on Web of Science and publications recommended by a reviewer of the manuscript. The sub-chapter is divided into five parts:

1. Research on the susceptibility of respondents to phishing attacks.
2. Results of efforts to improve phishing email detection.
3. Analysis of email content and development of phishing recognition capabilities.
4. URL-based phishing attack detection.
5. Summary of the literature review.

2.1.1. Research on the susceptibility of respondents to phishing attacks

The study [6] used a role-play scenario-based methodology to investigate why some email phishing attacks are successful, and why some people are more susceptible to them. To examine which email phishing attacks are most successful, was utilized a social influence framework, based on six principles of influence; namely: authority, consistency, liking, reciprocity, scarcity. and social proof. Participants were exposed to both genuine and phishing emails which contained these influence principles. Results indicate that participants were quite poor at correctly judging the safety of a link, regardless of whether the email was phishing or genuine. These findings have practical applications for phishing education, training and awareness programs.

The research [7] determined the effect of Internet user age and email content such as weapons of influence and life domains on spear-phishing susceptibility. Six life domains phishing emails can refer to financial, health, ideological, legal, security, and social. In total, 100 young and 58 older users received, without their knowledge, daily-simulated phishing emails over 21 days. A browser plugin recorded their clicking on links in the emails as an indicator of their susceptibility. Forty-three percent of users fell for the simulated phishing emails, with older women showing the highest susceptibility. It was found that older women were particularly susceptible to phishing. In addition, young users' susceptibility decreased across the course of the intervention, while older

users' susceptibility did not decrease. The effectiveness of phishing emails varied depending on the weapons of influence and life domains. In particular, susceptibility was highest for scarcity and legal emails and lowest for social proof and financial emails. Further, the relative effectiveness of specific weapons of influence differed between young and older users. While young compared to older users showed greater susceptibility to scarcity and authority emails, older compared to young users showed greater susceptibility to reciprocation and liking emails. Older users showed lower susceptibility awareness than young users.

The purpose of the paper [8] was to explore user susceptibility to phishing by unpacking the mechanisms that may influence individual victimization. The focus was on the characteristics of the e-mail message, users' knowledge and experience with phishing, and the manner in which these interact and influence how users cognitively process phishing e-mails. A field experiment was conducted where 194 subjects were exposed to a real phishing attack. The experimenters manipulated the contents of the message and measures of user traits and user processing were obtained after the attack. Phishing susceptibility was predicted by a particular combination of both low attention to the e-mail elements and high elaboration of the phishing message. Finally, individual factors such as knowledge and experience with e-mail increased resilience to the phishing attack.

2.1.2. Results of efforts to improve phishing email detection

The contribution of the paper [9] is the ability of the proposed Phishing Email Detection System (PEDS) to adapt itself to reflect changes in the environment. The novelty claim stems from the fact that was introduced a new approach that used Reinforcement Learning (RL). An algorithm called Feature Evaluation and Reduction (FEaR) was developed to explore the new behavior as well as to rank a selected list of features. In the field of online phishing email detection, the number of important features is always changing. The algorithm is dynamically changing the number of important features and extract them from next email. A Neural Network (NN) is used as the core of the classification model, and an algorithm called Dynamic Evolving Neural Network using RL allows the NN to evolve dynamically. Through investigating the previous studies, a very limited number of studies have been built to handle zero-day (first using) phishing attacks. Any model that supposed to detect zero-day phishing attack need to have the ability to dynamically adapt the detection model to reflect changes in phishing emails. In addition, it should have the ability to explore new behaviors in newly received email in the online mode.

Targeted spear phishing attacks have been implicated in many major security breaches. Email filtering systems are the first line of defense against such attacks. These filters are typically configured with uniform thresholds for deciding whether or not to allow a message to be delivered to a user. However, users have very significant differences in both their susceptibility to phishing attacks as well as their access to critical information and credentials that can cause damage. Work presented in the paper [10] has considered setting personalized thresholds for individual users based on a Stackelberg game model. First, in the model user values can be substitutable, modeling cases where multiple users provide access to the same information or credential. Second, it was considered attackers who make sequential attack plans based on the outcome of previous attacks. For multiple-credential scenarios, it was formulated a bi-level optimization problem for finding the defense strategy and then reduce it to a single level optimization problem. Experimental results lead to significant higher defender utilities than two existing benchmarks in different parameter settings.

The study [11] examines overconfidence in phishing email detection. Authors believe that overconfidence can lead to one's adopting risky behavior in uncertain situations. Study focuses in the experiment with 600 subjects and tests what leads to overconfidence in phishing detection. Each subject of the experiment judged a set of randomly selected phishing emails and authentic business emails and were examined two metrics of overconfidence (i.e., over precision and overestimation). Results show that cognitive effort decreased overconfidence, while variability in attention allocation, dispositional optimism, and familiarity with the business entities in the emails all increased overconfidence in phishing email detection.

2.1.3. Analysis of email content and development of phishing recognition capabilities

Phishing emails use a range of influence techniques to persuade individuals to respond, such as promising a monetary reward or invoking a sense of urgency. The study [12] explored a number of factors that may affect the persuasiveness and trustworthiness of emails by examining participant judgements. Participants of the experiment were recruited from Central Washington University to complete an online study in marketing communications. One hundred and twenty-four participants were female and fifty-four were male. The majority of participants being in the 18–24 years age group category (168 participants), the remaining 10 participants were over the age of 24. The experiment was conducted online using the platform www.qualtrics.com. Recent advances in phishing susceptibility research have expanded current understanding of how people make decisions regarding suspicious emails. However, the precise role of various message-specific factors, including how and why they influence people's judgements and decisions, remains unclear. It was investigated how three of these factors, which have not been extensively examined in previous research, influence judgements of email trust and persuasiveness, specifically the use of loss and reward-based influence techniques, authentic design cues, and referencing a salient current event. The use of loss-based influence techniques and the presence of authentic design cues was found to increase perceived trust and persuasiveness, with a number of psychological mechanisms identified that may account for these findings. It is hoped that these findings will provide a basis from which to systematically explore the potential role of these various underlying mechanisms. Only by understanding how people evaluate email communications will it be possible to understand why phishing emails work and how best to mitigate them.

Phishing emails success rely on social engineering techniques; they exploit human psychology and convince a victim to give away personal information and money. The authors believe that phishing emails can contain different types of principles of persuasion and techniques that can increase their effectiveness. The paper [13] builds a unique list of principles of persuasion. Persuasion is typically defined as the "human communication that is designed to influence others by modifying their beliefs, values, or attitudes". First, persuasion involves the intent to achieve a goal on the part of the message sender. Second, communication is the means to achieve that goal. Five principles of persuasion in social engineering:

1. Authority: Society trains people not to challenge authority but to respond without questioning.
2. Social Proof: People tend to mimic what the majority of people do or seem to be doing, so let their guard and suspicion down and prefer to share the same responsibilities and risks.
3. Liking, Similarity and Deception: People prefer to follow or relate to other people whom they know, like, are attracted to, or who seem familiar or similar to themselves.

4. Distraction: When people focus on what they can gain, lose or need, on strong emotional states or on whether an item will soon be unavailable or is restricted, this can heighten people's emotional state and make them forget other important considerations when making decisions.
5. Commitment and Reciprocation: Reciprocating a favor or responding to some action can be an automatic response that is linked to a sense of commitment with a previous situation.

Subject lines were chosen over the entire email texts due to not only small size and objectivity, but especially because it is a good practice and highly recommended that this type of attack is preferably detected earlier in the interaction with the user. Indeed, subject lines are the first contact point with the email recipient and a main piece of information that triggers the user into deciding whether an email should be opened or not. The study shows that subject line in phishing email can be cleverly crafted to include different persuasive content and turn simple sentences into triggers to influence users to open an email. The authors believe that it is essential to improve techniques designed to prevent and detect phishing emails within the initial users' interactions and stress the relevance of investing in studies on this topic [13].

The aim of the research [14] is authorship analysis of phishing emails. The authorship analysis of phishing emails consists of three steps: 1) data pre-processing; 2) dimensionality reduction; and 3) cluster analysis. In data pre-processing phase is to convert email documents into a numeric space using combination of term frequencies of words (TF)s and the WordNet semantic similarity measure. The data obtained in this step is very sparse and high dimensional, therefore one needs to apply dimensionality reduction before applying clustering techniques. Main contributions are in a new data pre-processing procedure is designed based on combination of TFs and the matrix of path-similarity distance measure; in an algorithm for finding groups of similar emails in phishing emails datasets. This algorithm is based on the combination of clustering and feature selection algorithms. Research demonstrates that the use of accurate clustering algorithms in combination with the feature selection algorithms can help to identify meaningful groups in phishing emails datasets and to design an effective defense mechanism to prevent phishing attacks.

Every electronic message poses some threat of being a phishing attack. If recipients underestimate that threat, they expose themselves, and those connected to them, to identity theft, ransom, malware, or worse. If recipients overestimate that threat, then they incur needless costs, perhaps reducing their willingness and ability to respond over time. Metacognition is described generally as "cognition about cognition" it refers in the paper [15] to individuals' understanding of their ability to detect phishing emails. That is a special case of the metacognitive ability to navigate online systems, in which limited bandwidth - messages may be misleading, not just because of poor design. In field experiments, was examined the appropriateness of individuals' confidence in their judgments of whether email messages were legitimate or phishing, using calibration and resolution as metacognition metrics.

Participants in experiments had reasonable calibration but poor resolution, reflecting a weak correlation between their confidence and knowledge. Of the 40 emails that participants reviewed, 19 were phishing emails (adapted from public archives), 19 were legitimate emails (adapted from real ones), and 2 were attention checks. Although a 50% base rate of phishing emails is not realistic (less than 1% of actual emails are phishing), that rate was used to reduce the burden on participants and the time required to collect sufficient data for analysis. The order of the emails was randomized for each participant. Each phishing email contained one or more of the features often associated with phishing: impersonal greeting, suspicious URLs, unusual content based on the stated sender and

subject, requests for urgent action, and grammatical errors or misspellings. Understanding the relationship between metacognition and phishing detection is critical for improving training and education. Analyses to identify how metacognition differs for phishing and legitimate emails, the relationship between metacognition and individual factors for phishing and legitimate emails, and the relationship between metacognition and real-world vulnerability [15].

2.1.4. URL-based phishing attack detection

Interesting research, published in [16] and further developed in [17] was oriented to phishing emails detection, based on URL links analysis. This research was oriented to one of the critical phish emails risk moment, based on detection malicious URL addresses with almost 100% accuracy using convolutional neural networks (CNN).

URL is an address that allows locating a website on the Internet. The user encounters it mainly when using distinguish them. Dictionaries containing predefined words are not an optimal solution in the case of URLs because each minimal change in the letter in the address can refer to an impersonation attempt. The average address length was 186 characters, and the longest address was 1149 characters. To optimize the network at this stage were used 256 characters to encode the URLs. The used architecture was very good at analyzing the natural language processing (NLP); however, the analysis of URL addresses in terms of the occurrence of phishing attacks seemed to be novel. An idea to sensitize the network, so that it can detect the address distortions to optimize the architecture presented by us so that the network can be implemented in mobile devices with limited memory and computing power. An embedding layer was used in the network to change the representation of the input data from a one-hot vector to a real-valued vector from the input element. This technique has perfect effects in NLP where by changing the representation of words, was obtained better results for a given classifier.

Authors used a modern version of Recurrent Neural Network (RNN) called Long Term Short Term Memory (LSTM) and it was proposed the method of identifying phishing websites based solely on the URL address text by a deep neural network with convolutional layers and encoded URLs as one-hot character-level vectors and presented them as inputs to a CNN. There were checked many variants of CNNs against testing error to achieve the best data generalization. Moreover, it was found out that the use of the embedding layer improved the results. The results presented show that the CNN network dealt with the classification better than LSTM. In experiments, can be observe the minimal advantage of CNN over LSTM in terms of accuracy.

2.1.5. Summary of the literature review

The literature review confirmed the interest in cyber security research – phishing emails and defense against them. Three papers that are of interest in social influence and explore user susceptibility in phishing attacks fell under the first area of interest, “Research on the susceptibility of respondents to phishing attacks”. There were three papers oriented to the second interest area, “Results of efforts to improve phishing email detection”. For these papers, methods of artificial intelligence and development systems of email filtering were used in the field experiments.

Studies that fell under the third interest area, “Analysis of email content and development of phishing recognition capabilities,” are closest to the objectives of our article. They were made up of

four papers and involve research into the reasons for accepting/rejecting phishing email based on influence techniques and principles of persuasion and metacognition. The object of analysis is the subject line of phishing email and its authorship. Themes for result discussion include phishing templates, life domains of phishing emails, and features often associated with phishing.

2.2. Data collection and statistical survey

The phishing emails were collected from the email accounts of one of the authors in the months of October and November 2020. The text of each potential phishing email was included into one file; its range reached at the end almost 70 pages. Only the plain text of the fraudulent messages was chosen for data collection. Emails whose text could only be accessed by clicking on a URL or downloading an attachment were excluded for security reasons. The total number of data was 200 emails, each introduced in a file with its own serial number. The text file, composed of the emails, was further divided into 200 separate files, for better processing by analytical software.

A simple statistical survey (summation, average, frequency analysis) was applied to determine the following parameters of the phishing emails.

- The number of emails sent to the business (university) or personal account.
- Identity of the sender: male, female, or not specified; company (corporate) or personal account; and country of the sender.
- Whether immediate response was required; whether the answer should contain personal data.
- Whether money or other wealth (gold, diamonds) were promised.
- Length of email in words, its language, nationality of sender.

A frequency analysis showed how many phishing emails were sent per day and per week and the trend of the emails.

2.3. Classification, segmentation and application of text analytical SW

The phishing emails were classified and segmented manually based on their characteristics. Then the text analytical SW, Tovek, was used and results were organized in a table. The set of characteristics:

1. The main goal (message) of the email; form of address (greeting).
2. A person or company sender; requested information or action from sender.
3. Promised result after sending the answer or fulfilling the requested action.

The text analytical software (SW), Tovek Tools (TT), is used to easily find information in text from various sources (files, emails, databases, etc.) and in different formats. TT consists of five modules, namely Index Manager, Tovek Agent, Query Editor, Info Rating, and Harvester. The first step of data processing is data indexing in the Index Manager module so that searches can be performed quickly in the Tovek Agent module using the prepared options or automatic search for entities (name, URL address, date, phone number, geographic information).

Simple queries contain only a few search keywords and operators. Complex queries should be created using the module Query Editor, and results can be in form of search archives. The Tovek language formulates queries in a precise format with various search parameters. The TT module for context analysis is Info Rating, which makes it possible to find the context in documents with respect

to context queries. The results of document search can be exported to the Harvester module for content analysis. It is possible to discover word connections and prepare word frequency analysis and graphs of word contexts [18].

3. Results of phishing emails analysis

3.1. Statistical analysis of phishing emails

Table 1 depicts the results of the statistical survey of the selected parameters of the phishing emails. It contains the total amount of emails, the sender of the email (male, female, or not specified) (see Figure 1), and the number of emails sent to the business (university) or personal account. It also includes the number of emails sent by organization (corporate) or personal accounts, whether the email required immediate response and whether the answer should contain personal data, the amount of money promised, and the language of the emails.

Table 1. Statistical survey.

Parameter of emails (unit)	Amount	Parameter of emails (unit)	Amount
Total number	200	Name of sender not available (%)	8
Male sender (%)	54	Female sender (%)	38
Sent to business account (%)	89	Sent to personal account (%)	11
Corporate mail (%)	26	Personal mail (%)	74
Immediate response required (%)	86	Answer with personal data (%)	37
Total promised money (mil. USD)	1736	Money promised per mail (mil. USD)	8,7
Total promised gold (kg)	537	Average length of email (words)	190
Language English (%)	90	Language Czech or Slovak (%)	10

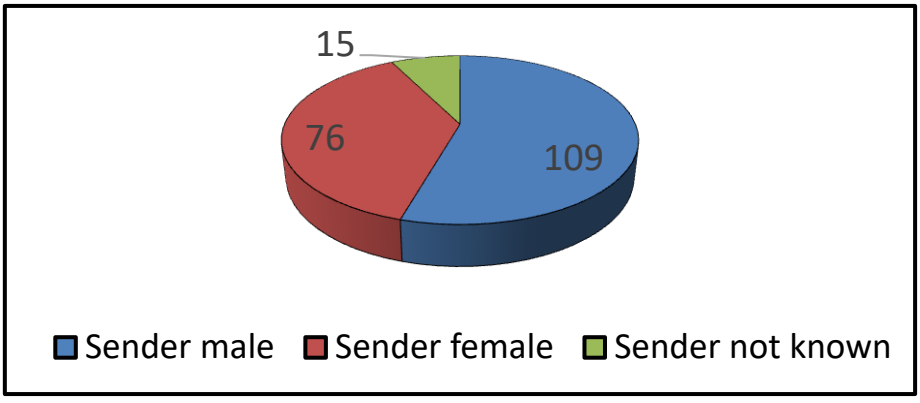


Figure 1. Number of emails by sender.

The number of emails delivered to the work account was significantly higher than the number of emails delivered to the personal account. This disparity is not surprising, because work accounts are certainly more interesting to phishers than personal accounts. It was not possible to use simple automatic search to statistically determine data. The text of phishing emails is often quite complicated, so it was necessary to read them individually and enter the detected values in the tables.

The number of emails from men was more than half of the total emails. This result is awaiting, because phishing emails often encourage various businesses and more often men work in this area. Phishing emails also often require the recipient to send an immediate response; otherwise, there is a risk of delay.

The country or nationality of the email sender is given in Table 2 and Figure 2. The determination was performed using the text analytical SW and the results were corrected manually. Often, names of different countries crop up for one sender and it was difficult to decide the actual country. In some cases, the country was determined using the international prefix of the telephone number used or the registered office address of the company whose representative the sender claimed to be. In almost 40% of emails, the country or nationality of the sender could not be ascertained.

The frequency analysis in Figure 3 shows how many phishing emails were sent per week. There was a slightly increasing trend. The last week had only 5 days. The maximum number of emails sent on any given day was ten and the minimum was zero.

Table 2. Country or nationality of email senders.

Country	%	Country	%
None	39.5	France	1.5
USA	13.5	New Zealand	1.5
UK	9.5	Thai	1.5
Burkina Faso	5.0	Togo	1.5
Czech or Slovak	4.5	India	1.0
Cote D'Ivore	4.0	Israel	1.0
Benin Republic	3.5	Libya	1.0
Turkey	2.5	Other	6.5
South Africa	2.0		

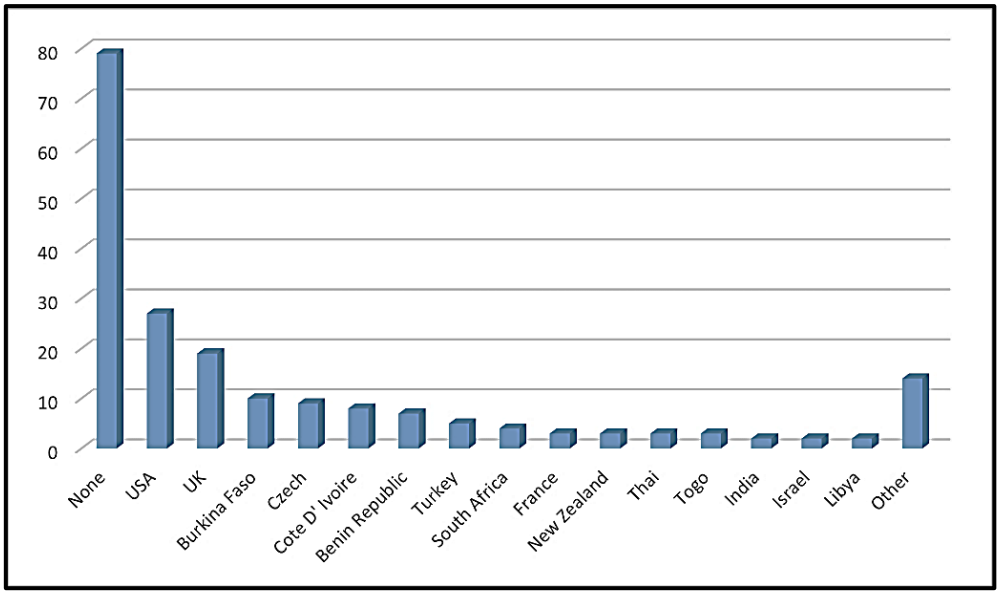


Figure 2. Number of emails by country of nationality of sender.

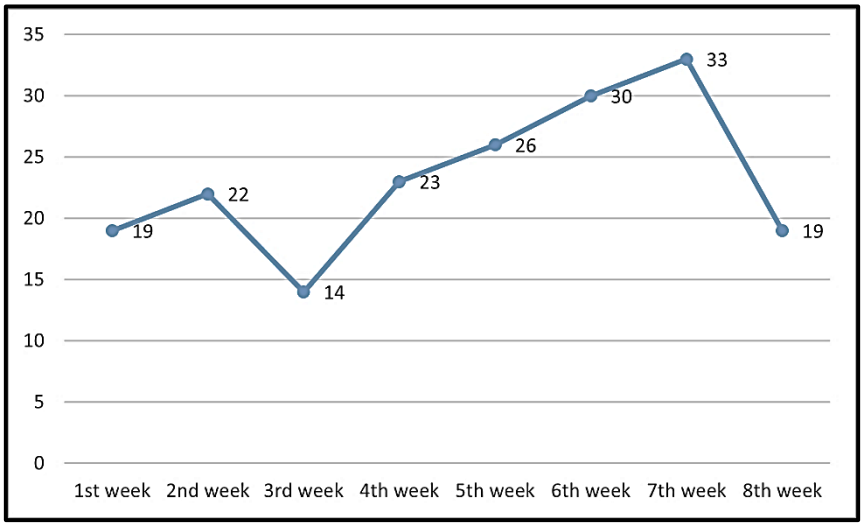


Figure 3. Weekly frequency of incoming emails.

3.2. Classification by the main goal (message) of the phishing email

Depending on the main goal (message), the phishing emails were classified into five segments: 1) Business, 2) Charity, 3) Fund, 4) Transfer, and 5) Others. In the following subsections we explain the segments in detail (in parentheses beside the title of each subsection is given the number of emails and its percentage of the total number).

After extraction of the relevant phishing emails in any segment, some analysis was performed to find out what personal information (name, address, and phone number) is requested by the emails. The three parts of the query (1) were first joined by the "or" operator and then by the "and" operator:

(receiver name or your name or full name) or **(contact address or home address or house address or your address)** or **(your tel or phone number or mobile number or mobile telephone or direct cell or direct telephone)** (1)

3.2.1. Business segment (64, 32%)

The Business segment included a wide range of phishing emails that offer cooperation on a project, investment in the recipient's country, execution of a contract, or realization of a business opportunity. Some emails also required a partner to trade in medical and supplier products, or offered opportunities in the development of IT services. The last set of emails included in this segment offered the recipient work in various areas with a promise of a high salary and many lucrative benefits. Querying of the email files was performed using a hierarchical query processed in the Query Editor (see Figure 4).

The result of the query was a set of files (emails) that could be sorted by file name (Figure 5) or by score (Figure 6), which expresses the degree of fulfillment of the query conditions. In addition, it could contain selected entities, in Figure 5 with email addresses and in Figure 6 with detected states of the selected emails.

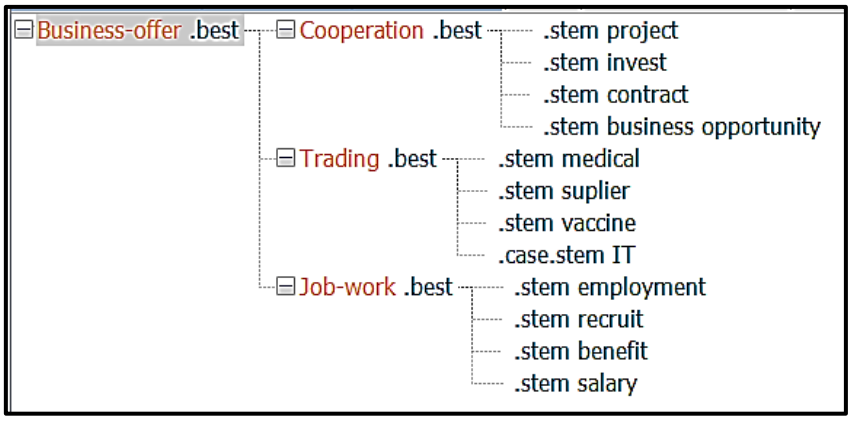


Figure 4. Hierarchical query in the Business segment.

Skóre	Jméno souboru	Entita email
25	007.docx	admin@hiflyingllc.com
25	008.docx	info@info.com
87	009.docx	exxonmobiloilemployment@hotmail.com, exxonmobiloilgas-usa@engineer.com
58	011.docx	alexdaniel0982@gmail.com
25	012.docx	abdul.hamza03@gmail.com, markmarketing922@yahoo.com
44	013.docx	lawrrence1020@gmail.com
25	014.docx	mrpeteralexandra3033@outlook.com, mrpeteralexandra3@gmail.com
25	016.docx	c.butor@aol.com, mrs.butor.car12@gmail.com
25	017.docx	elizabethmarvin19@yahoo.com, elizabethmarvin31@yahoo.com
44	019.docx	mraisha.elgaddfi0.1@gmail.com
25	022.docx	mrmikeload123@aol.com, ubanig252@gmail.com, united_nation_121@outlook.c...
25	023.docx	admin@ppid.subulussalamkota.go.id, cardinalglobalfinanuk@gmail.com
25	025.docx	nelsonativor3@gmail.com
25	028.docx	rosettadouglass3@gmail.com
44	030.docx	foffice498@yahoo.com, romansoika080@gmail.com

Figure 5. Result of the query in the Business segment, sorted by file name.

Skóre	Jméno souboru	Entita stát
87	054.docx	Ghana
87	009.docx	United States
68	032.docx	Australia, United States
58	094.docx	United States
58	053.docx	
58	051.docx	Mali, United States
58	011.docx	Cote d'Ivoire
44	188.docx	Benin
44	185.docx	South Africa
44	183.docx	Georgia
44	167.docx	Benin
44	145.docx	
44	123.docx	
44	100.docx	
44	072.docx	United States

Figure 6. Result of the query in the Business segment, sorted by score.

The analysis of personal information in the Business segment using query (1) with the "or" operator gave 34 emails (53%) and analysis with the "and" operator gave only six emails (9%).

Example of email (file 053):

Jacob info@artemisiaoliveoil.com: Contract Arrangement (From J.W); From: Wrench, Jacob
Attn: Sir, I am, Jacob Wrench (Mr), a Level 2 Director with the Contract Awards and Monitoring Committee of Ministry of Urban and Rural Development. I got your information from business network online, my duty as empowered by the British Government is to monitors provision the basic amenities, social recreational activities in urban and rural areas, this programs include assistance to deprived local communities and to co-ordinate project and development at the national levels, further more from this projects we have been able to secure some reasonable amount of money to the tune of Three Million One Hundred and Sixty Thousand British Pounds (GBP3,160,000.00) as commission from various contractors resulting from over invoicing, hence all the necessary approvals has been completed. This approved fund is now securely deposited with a financial institution here in U.K for onward transfer to its destination. This fund is deposited with the reason that it is payment for a foreign expatriate/company {contractor} as we are Government officials we are not allowed to operate or open a foreign bank account hence you need to stand as the beneficiary and claim this funds on our behalf from the bank. I am making this contact with you with faith of making as a reliable person/company with high integrity/dignity one with conscience that will claim this funds on our behalf as the beneficiary {the foreign contractor},and we have agreed to offer you 30% of the total sum as commission for your assistance/effort and 5% will be used to settle every expense each of us might incur on the process, we will use our 65% to invest under your recommendations and guide and go into joint venture business with you. I would greatly appreciate your assistance, I look forward to hear from you be informed that confidentiality is the nature of this business because I would not be happy to lose my job. For confidentiality I request you contact me via my email below. When I hear from you I will inform you further on how to proceed if you are willing to part of this project.
Best regards Jacob Wrench, Email: j.wrench@mail.co.uk

3.2.2. Charity segment (41, 21%)

Here is the typical scenario of most phishing emails classified under the Charity segment: The sender is an old woman, a widow, without children. The deceased husband left a large fortune (millions of USD), which she wants to donate to charity. She describes with brutal openness that she is seriously ill (cancer) and has a few weeks left to live. The letter is written in the Christian spirit, asking the recipient of the donation to set up a charity fund and get a reward for this mission.

The following was the query (2) for selecting emails in the Charity segment:

charity, cancer, illness, disease, hospital, widow, God, Christ, Lord (2)

Remark: The operator "," works in the Tovek Agent first as the "or" operator in a query and then as the "and" operator in the search result to determine the score.

The analysis of personal information in the Charity segment using query (1) with the operator "or" gave 15 emails (37%) and the analysis with the "and" operator gave six emails (14%).

Example of email (file 058), displayed in Tovek Viewer with marked keywords and entities:
Mrs. Melinda Cruz<melindacruz40798@yahoo.com>
From Mrs. Melinda Cruz Charitable Donation.
Dearest in Christ, Greetings in the name of our lord Jesus Christ. I am Mrs. Melinda Cruz, an aging widow (69 years old) suffering from long time illness breast Cancer and Cancer of the lungs, I am currently admitted in a private hospital here in Abidjan Cote D' Ivoire. I have some funds I inherited from my late loving husband, the sum of US$ $ 2 million dollars which he deposited with a bank in Abidjan. I am looking for a honest and God fearing person that can use these funds for God's work also creating a charitable organization for the less privileged, helping the elderly, the poor, also the war and HIV/AIDS victims, and 15% out of the total funds will be for your compensation for doing this work. I found your email address from the internet and I decide to contact you after my prayers hoping that you will not betray the trust. If you could be able to do this work, please email me back for more detail on how it could be done. I wait for your positive reply.
Your Sister in Christ, Mrs. Melinda Cruz.

3.2.3. Fund segment (35, 18%)

This segment included phishing emails that promised the recipient money obtained from a fund (compensation, scam or fraud), a financial gift, or assets from inheritance.

The following was the query (3) for selecting emails in the Fund segment:

fund and (compensation or scam or fraud), gift, inheritance (3)

List of mentioned funds (or organization) in the phishing emails:

- Bill Gates Foundation.
- Compensation fund for scam victims, Bank of Holland.
- Compensation fund, The International Police & the Financial Crimes Enforcement Network.
- Compensation funds with ORA Bank.
- Fund in the Bank of America, New York.
- The UN office compensation fund for scam victims, DNB Bank, London, UK.
- Compensation fund, United Bank for Africa, Lomé, Togo.
- The Charles Koch Charitable Foundation.
- The International Monetary Fund, First National Bank, South Africa.
- The International Monetary Fund, the World Bank Group.
- The United Nations Compensation Unit, West Africa.
- The United Nations Development Fund (UNDF).

The analysis of personal information in the Fund segment using query (1) with the operator “or” gave 17 emails (49%) and analysis with the “and” operator yielded 5 emails (15%). The graphs of Figures 7 and 8 are the result of the email content analysis prepared with the TT module Harvester and show phrases (word connections) in the emails. The number at the link between the words depicts the "strength" of the connection (number of occurrences).

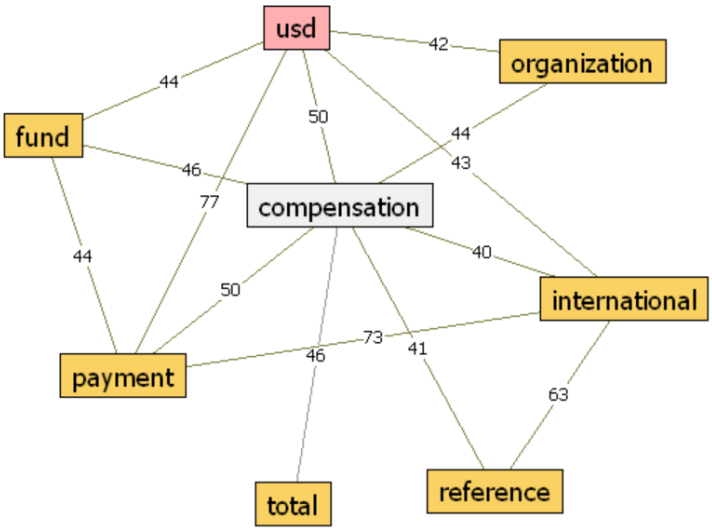


Figure 7. Content analysis of the word compensation.

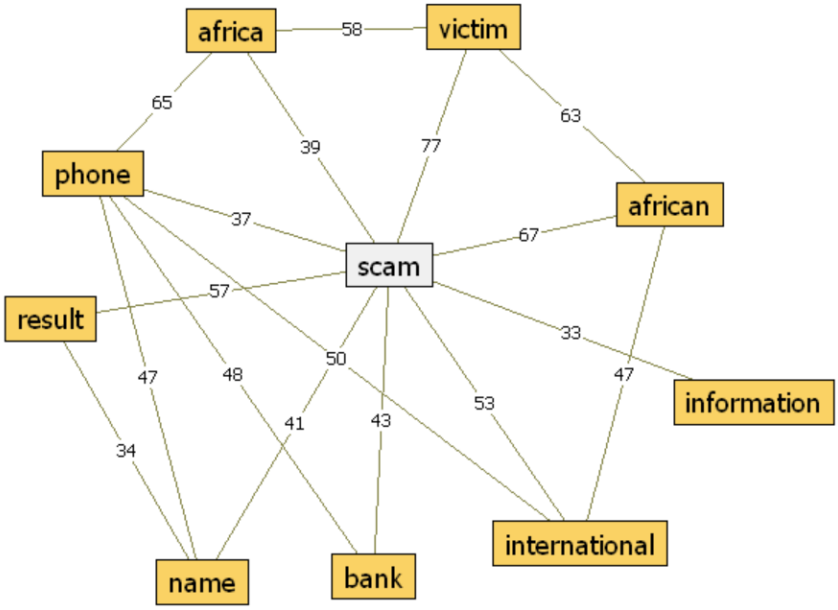


Figure 8. Content analysis of the word scam.

Example of email (file 178):

Mrs. Angela Blake<federalministryfinance460@gmail.com> Good Day Sir / Mrs.

How are you doing, this is to let you know your name was shortlisted among the foreigners to be compensated for Scam Victim, The total sum of US$5.000,000.00 which was approved by INTERNATIONAL MONETARY FUND (IMF) INVESTIGATION AND DEBT SETTLEMENT COMMITTEE to be release to you at your preferable mode of payment by the Investigation and Debt Settlement Committee, The Payment is as a result of compensation Inheritance Fund for scam victim

which is due for payment. Approved by the Investigation and Debt Settlement Committee / International Monetary Fund, All is required from you is your details for Authorization and Endorsement of your Name on the said fund US$5 Million for payment.

DETAILS REQUIRED FOR OPENING OF CLAIM FILE IN YOUR NAME FOR PAYMENT:

Your Full Name:
Your Home Address:
Your Direct Phone Number:
Your Nationality:
Your Occupation:
Your Age / Marital Status:
Your Nearest Airport:
Copy of your ID:

Please acknowledge the receipt of this message for more details, as you furnish this office your details stated below required for proper Verification and Documentation and opening of the claim file of your name on the legal documents as the rightful beneficiary for payment. Waiting to hear from you and God Bless you for your understanding.

Yours Faithfully, MRS. ANGELA BLAKE, Federal Ministry of Finance Consultant.

3.2.4. Transfer segment (29, 15%)

This segment included phishing emails in which the sender requested cooperation for money or other asset transfer. The commission for transfer ranged from 30 to 60 % of the transferred amount. The initiator was usually a bank clerk who claimed to have discovered a free money account and that the intended transfer would be completely risk-free.

The query (4) for selecting emails in the Transfer segment was:

(transfer or shipment) and (money or gold or diamonds) (4)

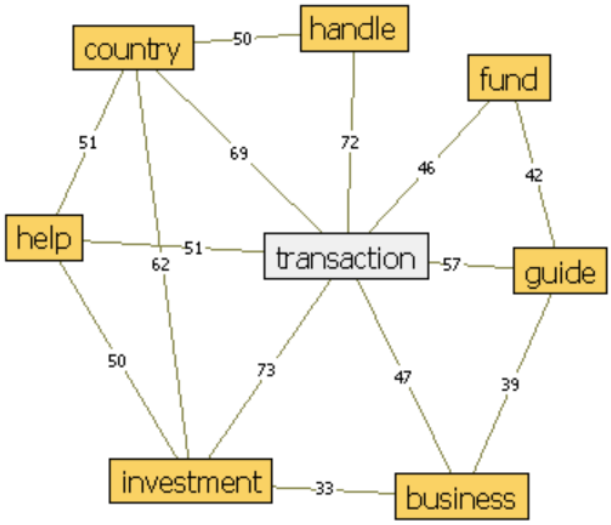
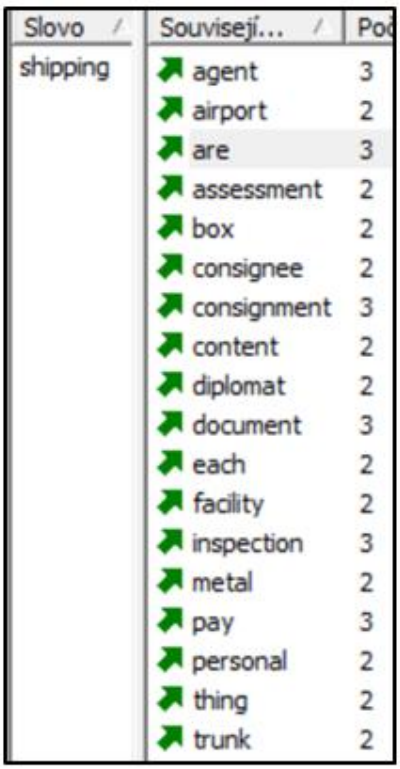


Figure 9. Content analysis of the word transaction.



Slovo	Souvisejí...	Poč
shipping	agent	3
	airport	2
	are	3
	assessment	2
	box	2
	consignee	2
	consignment	3
	content	2
	diplomat	2
	document	3
	each	2
	facility	2
	inspection	3
	metal	2
	pay	3
	personal	2
	thing	2
	trunk	2

Figure 10. Connections of the word shipping.

The graph of Figure 9 and list of words on Figure 10 are the result of content analysis of the emails grouped under the Transfer segment. They were prepared with the Harvester module and show phrases (word connections) in the emails.

The analysis of personal information in the Transfer segment using query (1) with the operator "or" yielded 14 emails (48%) and analysis with the "and" operator yielded 5 emails (17%).

Example of email (file 167):

Mrs.Katrina Phillip<office.payments54@yahoo.com> WELCOME TO RIA MONEY TRANSFER: FROM THE DESK OF RIA MONEY TRANSFER REMITTANCE DEPT; Our Ref:RMTT0XX2/987
WEB SITE.www.riamoneytransfer.com, Send Money Worldwide. Ria Money Transfer is a convenient way to send Money Fast. It's Easy, Reliable and you Have a choice of Great Options.
ATTN. Dear Beneficiary, Please we need your current Home Address and Telephone Number to Enable us Start Sending you $7.5 Million Dollars to you Through Ria Money Transfer as we were instructed by The Minister of Finance Benin Republic this Morning To Be Sending you $5,000.00 each day Till the Whole Amount $7.5 Million Dollars, Call Rev. COOLEY WILLIAM now and ask him to give you MTCN and every Other Information you Need to pick up your $5,000.00 Today. Here is what he may require from you.
Your Receiver name----------
Your Country----------
Your City----------
Your Tel----------
Your Test Question----------
Answer----------

Your Id----------
occupation_____
Forward the Information here; (ria.moneyt@aol.com) Or Call Rev. COOLEY WILLIAM On Telephone +229 -64751435
Thanks Mrs.Katrina Phillip, Director, Chair of the Compensation and Benefits Committee

3.2.5. Other segment (31, 16%)

This segment included emails with marginal significance due to their volume in the whole set of emails. The commonest categories among this segment were:

- Repeated contact (7, 4 %).
- Loan offer (7, 4 %).
- Erotic offer (6, 3 %).
- Undelivered package (2, 1 %).

Example of email (file 176), displayed in Tovek Viewer with marked keywords and entities:
dhloffice983@gmail.com
We have registered your ATM MASTER CARD of US $(92.55M)with DHL
Courier Company with the registration code of (Shipment Code awb33xzs) please kindly Contact them with your full delivery information such as follows to enable you receiver your ATM MASTER CARD in six working days from the DHL Courier Company department as well.
Your Full Name....
Your Full Address.....
Your Telephone Number.....
Your Passport copy.....
Your Age.....
Your Occupation.....
Name of Director: MR.RICHARD MARK
Call me :+229-51037495
Contact E-mail: rm8401926@gmail.com
Best Regard
MR.RICHARD MARK

3.3. Analysis of the phishing emails duplicity, repeatedly sending

The most frequent duplicate corporate phishing emails (10 times) was sent by Exxon Mobil, UK; in one case the email was from Murray Bell (Male), who posed as the HR Manager (file 9), and in all other cases from Peter Alexandra (Male), who posed as the Operations and Corporate Affairs Officer. Each email number, email address and the date the email was sent are provided in Table 3. The content of file 9 was a job offer, whereas in the other cases the content requested the recipient's full cooperation and partnership to re-profile funds amounting to US $ 12.2M. All emails the have been included in the Business segment. Exxon Mobil does really exist, available at https://corporate.exxonmobil.com/.

Table 3. The email number, email address and date email was sent.

No	Email address	Date	No	Email address	Date
9	exxonmobiloilemployment@hotmail.com	8.10.	104	vs@sphpl.com	5.11.
14	MrPeterAlexandra3033@outlook.com	7.10.	109	vs@sphpl.com	7.11.
47	ybjeon@cmvn.biz	18.10.	149	mramadan@psmchs.edu.sa	15.11.
93	benergo@benergo.ru	2.11.	155	paul@it.ca	17.11.
99	scanner@prwnetwork.com	4.11.	194	m.opatija@aci-club.hr	29.11.

The next sending of duplicate corporate phishing emails (two times) was made through:

- Bank of America New York (does exist), files 15 and 77, Transfer segment.
- Bill Gates Foundation (does exist), files 66 and 79, Fund segment.
- United Nations Compensation Unit, Tunde automotive Marina, Lagos, Nigeria (does not exist, it is scam), files 22 and 121, Fund segment.

The duplicity in phishing personal emails are listed in Table 4. The characteristics of the emails include name of sender (Name), their sex (Sex: Male, Female), the relevant segment to which the email belongs (Segment), file resp. email numbers (File number), information about the content of the duplicate emails (Content), and number of duplicates (Dup).

No regularity or intent could be traced in sent duplicate phishing emails.

Table 4. Characteristics of the duplicate emails.

Name	Sex	Segment	File numbers	Content	Dup
David Kane	M	Fund	92,122	identic	2
Dianka	F	Others	4,24,114	little differences	3
Grace William	F	Charity	110,125	identic	2
Jean Phillip	M	Transfer	15,77	identic	2
Kim Chan Ouk	F	Business	39,154	identic	2
Lida Oum	F	Business	182,193	identic	2
Mehmet Osman Pisir	M	Others	82,159	identic	2
Melinda Cruz	F	Charity	33,58	little differences	2
Rita Van Zyl	F	Others	105,108,144	identic	3

4. Discussion

In this discussion, we will first compare the results of our research with those of the studies mentioned in the “Literature review” section. Next, we explore the problems with using the text analytical SW for the analysis of personal data.

The top ten phish threats [4] provide a breakdown of phishing threats that correspond to our results in only one item, "Undelivered shipment by courier service," and this is only a marginal topic in the Others segment. The explanation for this difference is probably that the phishing attacks according to [4] probably belong to spear phishing, which is not covered in this study.

The research [7] defines the six life domains of phishing emails: financial, health, ideological, legal, security, and social. These domains are difficult to compare to the phishing email segments of

our research. The Business segment does not correspond to any domain; the Charity segment could belong to the health and social domains; the Fund segment could belong to the financial, security and social domains; and the Transfer segment could belong to the financial and legal domains. The breakdown according to [7] is too general and does not correspond to the current situation of phishing email attacks.

On the contrary, our research completely confirms features often associated with phishing [15]: "impersonal greeting, suspicious URLs, unusual content based on the stated sender and subject, requests for urgent action, and grammatical errors or misspellings". In our study, impersonal greeting was used in 100% of the phishing emails analyzed. The most frequent was "NO greetings" (31%); followed by the greeting "Hello" (25%) combined with such a word as "Beloved," "Dear," "Friend," "Partner," or "Sir / Madam"; then the greeting "Dear" or "Dearest" (18%) combined with "Beloved," "Customer," "Friend," "in Christ," or "Sir / Madam." The greeting "Good Day" (8%) was often combined with "Sir / Madam". Other greetings (Attention, Hey, Hi, May Allah Bless You, etc.) were used less frequently.

Suspicious URLs were used in any emails, but were not analyzed for security reasons. There were requests for urgent action in 86% of the emails. We were able to detect grammatical errors and/or misspellings in a large group of the emails, despite that we are not native English speakers.

The offer of a large amount of money can be considered as unusual content. In all the segments analyzed, the amount offered was 1736 million USD, which was almost 9 million USD per email, a sum of money most people cannot dream of earning even in a lifetime. In the Business segment, the ease of creating a business partnership (by just replying to an email) can be considered unusual content. In fact, business partnerships are built gradually and only gradually can mutual trust grow.

In the Fund segment, it can be considered as unusual content that funds could be raised to compensate fake victims of fraud. In the Transfer segment, receiving an invitation to pick up boxes full of money from the airport can be considered as unusual content if you, so can a bank clerk informing you that "forgotten" money could be transferred from the bank to your account without any risk.

In the Charity segment, a typical scenario of the phishing email (written in the Christian spirit) repeating the phrase or word "an old woman," "widow," "without children," "seriously ill (cancer)" and/or "has a few weeks left to live" can be considered as unusual content. Also unusual are emails with content to the following effect: the sender's deceased husband left a large fortune (millions USD), which the sender wants to donate to charity; the recipient of the donation should set up a charity fund and get a reward. We think that people who are so seriously ill have other worries than creating charitable funds.

The analysis of the phishing email files using text analytic SW was accompanied by the problem of selecting suitable keyword representatives for queries, because many of the keywords were spread out in the emails regardless of their segmentation. We show this for the example of context analysis using the TT module Info Rating. The context query (Figure 11) detects the distribution of selected keywords in the emails. The result can be monitored in the context matrix (Figure 12), in which there is a number of documents (emails) in each node where the corresponding pair of context queries occur. In our example, the keywords are "Business" and "Fund", and they are found in 32 documents, out of which 14 belong to the Business segment, 13 to the Transfer segment, 4 to the Charity segment and 2 to the Fund segment.

Figure 11. The context query.

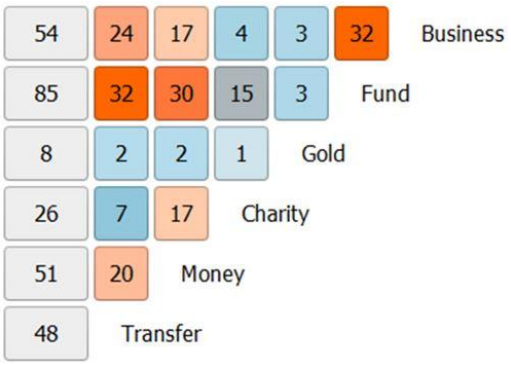


Figure 12. The context matrix.

Another example applies to the keyword "name", which should determine if this information is required in the email. But a large group of emails start with "My name is," and this keyword also appears elsewhere in the text of the emails. Therefore, with the knowledge of the content of emails, the word name must be supplemented by another adjective, e.g. full, receiver, your. This means that the result of each query by the analytical SW needs to be manually monitored and corrected.

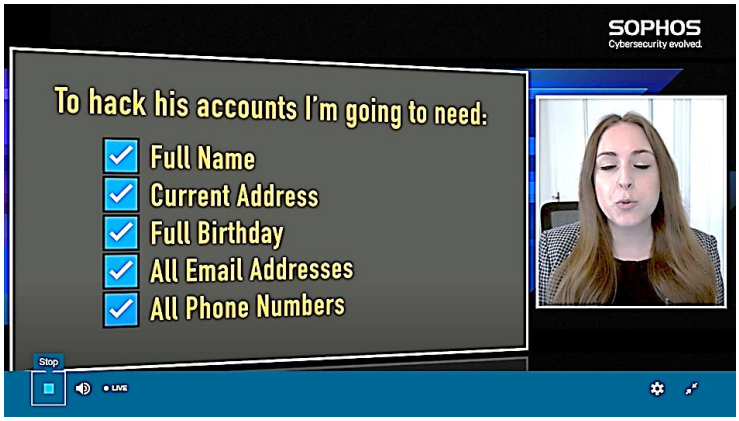


Figure 13. Hacking conditions of email account [19].

An interesting question is: What is the reason for obtaining personal information? Assume that this is for subsequent intrusion into the recipient's network account. Does the hacker have enough information to do this? Let's compare the information obtained from the phishing emails with the hacking conditions in Figure 13.

The analysis of personal information for the set of all phishing emails in query (1) with the "or" operator yielded the result 67 emails (34%) and the analysis with the "and" operator yielded 17 emails (9%). If we had added a condition for birthday, the search result would have been only one email (file 118). With respect to the hacking requirement of Figure 13, it can be concluded that hacking the email account of the receiver was not the primary target of the analyzed phishing attacks, but another form of cybercrime.

The request to obtain personal information contains in phishing emails, in addition to the data analyzed above (name, address, and phone number) a number of other less frequent requests for: age, country, fax number, marital status, nationality, occupation, sex. In addition, other means of identification were less often required, namely a scanned copy of the receiver's identity, passport, reference file, or letter of authorization. In emails classified under the Transfer segment, the following information was also required: receiver's bank name, account number, payment method, and credit card number.

5. Conclusions

The main goal of the study is to contribute to the understanding of phishing emails, while adding to the knowledge base on education and training in phishing email defense. An analysis of 200 emails was performed, including statistical survey, classification and segmentation by email content, and report on duplicate emails. The text analytical SW Tovek was used for the analysis, with manual corrections.

The novelty of the paper in comparing to the analyzed sources in the literature review consists mainly in the fact that a similar research topic, it was a detailed analysis of phishing emails, was not dealt with by any of them in the way carried out in our research. It was oriented to the detail understanding of phishing emails and extending the knowledge base for education and training in phishing email defense. The research of phishing emails included statistics, their classification, segmentation and analysis using the text analytical SW Tovek. In the discussion, the acquired knowledge was confronted with other sources, especially in the recognition characteristics of phishing emails.

The core part of the study involved classification of the phishing emails into the following five segments: Business, Charity, Fund, Transfer and Others. Each segment has been described in detail, highlighting its characteristics. The results of the analysis have been compared with results of some previous studies in the field.

Future research should be oriented to similar analysis, perhaps with smaller data samples, with the goal to discover changes in phishing emails after one year. It will also be of interest to find out if the same phishing emails, from the same workplace (same intranet network), are routed to different recipients' accounts.

Acknowledgements

The study presents results of the research in project [20] at the Department of Informatics and Cyber Operations, Faculty of Military Technologies, University of Defence, Czech Republic.

Conflict of interest

The authors declare that there is no conflict of interest.

References

1. Hong J (2012) The state of phishing attacks. *Communications of the ACM* 55: 74–81.
2. Singer PW, Friedman A (2014) *Cybersecurity: What Everyone Needs to Know.* 1st Eds, Oxford University Press.
3. Krombholz K, Hobel H, Huber M, et al. (2015) Advanced social engineering attacks. *J Inf Secur Appl* 22: 113–122.
4. Ducklin P (2020) Phishingové triky aneb 10 nejběžnějších podvodů roku 2020 (Phishing tricks or the 10 most common scams of 2020). *IT SYSTEMS*. Available from: https://www.systemonline.cz/it-security/phishingove-triky.htm
5. Becton L (2020) The Importance of Digital Literacy in K-12, Available from: https://www.educationcorner.com/importance-digital-literacy-k-12.html
6. Parsons K, Butavicius M, Delfabbro P, et al. (2019) Predicting susceptibility to social influence in phishing emails. *Int J Hum-Comput St* 128: 17–26.
7. Lin T, Capecci DE, Ellis DM, et al. (2019) Susceptibility to Spear-Phishing Emails: Effects of Internet User Demographics and Email Content. *ACM T Comput-Hum Int* 26: 1–28.
8. Adewumi OA, Akinyelu AA (2016) A hybrid firefly and support vector machine classifier for phishing email detection. *Kybernetes* 45: 977–994.
9. Sami S, Nauman A, Li Z (2018) Detection of online phishing email using dynamic evolving neural network based on reinforcement learning. *Decis Support Syst* 107: 88–102.
10. Zhao M, An B, Kiekintveld C (2016) Optimizing Personalized Email Filtering Thresholds to Mitigate Sequential Spear Phishing Attacks. In: *Proceedings of 30th Association-for-the-Advancement-of-Artificial-Intelligence (AAAI) Conference on Artificial Intelligence* 30: 658–664.
11. Wang J, Li Y, Rao HR (2016) Overconfidence in Phishing Email Detection. *J Assoc Inf Syst* 17: 759–783.
12. Williams EJ, Polage D (2019) How persuasive is phishing email? The role of authentic design, influence and current events in email judgements. *Behav Inform Technol* 38: 184–197.
13. Ferreira A, Teles S (2019) Persuasion: How phishing emails can influence users and bypass security measures. *Int J Hum-Comput St* 125: 19–31.
14. Seifollahi S, Bagirov A, Layton R, et al. (2017) Optimization Based Clustering Algorithms for Authorship Analysis of Phishing Emails. *Neural Process Lett* 46: 411–425.
15. Canfield CI, Fischhoff B, Davis A (2019) Better beware: comparing metacognition for phishing and legitimate emails. *Metacognition and Learning* 14: 343–362.

16. Nowak J, Korytkowski M, Wozniak M, et al. (2019) URL-based Phishing Attack Detection by Convolutional Neural Networks. *Aust J Intell Inf Process Syst* 15: 60–67.
17. Wei W, Ke Q, Nowak J, et al. (2020) Accurate and fast URL phishing detector: A convolutional neural network approach. *Comput Netw* 178: 107275.
18. The text analytical software TOVEK, 2020. Available from: https://www.tovek.cz.
19. Tobac R (2020) Social Engineer & Ethical Hacker. Live Hacking Demo: Hacking the Human, Sophos Evolve - Cybersecurity Summit, Webinar presentation.
20. DZRO FVT-2, KYBERSILY. Project of faculty research: Cyber forces and resources. University of Defence, Faculty of Military Technologies, Brno, Czech Republic, 2021.

AIMS Press