Phishing Email Detection Method: Leveraging Data Across Different Organizations

Tanusree Sharma[1], Priscilla Ferronato[2], and Masooda Bashir[3]

University of Illinois at Urbana Champaign, Champaign, IL, 61820, USA
{tsharma6,pf4,mnb}@illinois.edu

Abstract. Phishing attacks are a common tool used by malicious actors to gain access to systems or exploit individuals. Much of the phishing detection schemes focuses on awareness training programs and detection models. While these methods, detection models utilizing machine learning techniques, blacklists, and other email characteristics such as domain names, email addresses, and URLs are helpful in combating the problem, more research is needed in order to detect phishing scams more accurately and precisely where phishing email data can play an important role. Nowadays phishing attacks exploit human vulnerabilities by targeting specific human emotions, such as fear, to trick users into giving up their personal information. Consideration of the emotional exploitation present in phishing emails data combined with current detection schemes could lead to better detection. The goal of this paper is to analyze the emotional related content of a phishing email dataset to see if there is any relation between the emotion exploited and the sender's email domain. In addition, this research demonstrates the need for the collection and analysis of the various types of phishing emails and its related emotional content that can be beneficial in developing better detection methods.

Keywords: Emotions · phishing email · email database

1 Introduction

Phishing emails are a form of fraudulent attempt by malicious actors to gain access to personal and organizational computer systems causing financial and credentials losses [16]. Phishing emails often exploit individuals' emotional and sentimental vulnerabilities to obtain access to people's personal information and sensitive credentials [17]. Despite numerous efforts on the detection and development of filtering systems, phishing emails continue to attract users' attention and pose a serious computer security vulnerability. Cyber-attacks via phishing emails leave a great negative impact and cost on an organization's revenue and further remediation and regulation [14]. Moreover, the loss of user's data can potentially lead to loss of reputation and trust for an organization [13].

Phishing attacks are used to gain access to a system by sending fake emails disguised to be from legitimate sources. These emails could be requesting usernames, passwords, and other sensitive information, or contain links to phishing

websites that imitate real websites, tricking users into entering their personal information. Spear-phishing is a targeted phishing attack where an individual or group of individuals is targeted. The phishing email content is tailored to a person or group in order to gain access to a system that the target has access to. Humans are considered the weakest link in the security world, and phishing attacks take advantage of this by exploiting human vulnerabilities. Phishing is involved in 32% of all confirmed data breaches [3]. Attacker groups are especially fond of phishing as 65% of these groups use spear phishing has their primary infection vector [4]. Since phishing emails are designed to look legitimate, detection remains a much-researched topic. Existing phishing countermeasures include security awareness training programs and various detection models. These models focus on analyzing the senders' domain names, associated email addresses, and URLs to determine if an email is malicious or not [5]. However, some researchers [5] are trying to use natural language processing, sentiment analysis, and other methods to use the content of an email to determine its level of maliciousness to improve existing detection tools.

Majority of the phishing detection schemes focus on awareness training programs and detection models, employing machine learning techniques [15], using blacklists including URL, domain names and email addresses. While all of these approaches have combated phishing emails and have reduced the frequency of exploitation, more research is needed to detect phishing scams more accurately and precisely, especially focusing on human's vulnerabilities. Using social engineering techniques, phishing emails often exploit human vulnerabilities by targeting specific human emotions, such as fear, trust, and anticipation to trick users into giving up their personal information [1]. While previous studies have focused on persuasion techniques adopted in phishing attacks [2], there is still a research gap on how phishing emails target different human emotions and how this strategy can be combined with current detection schemes to achieve better security measures. Therefore, the goal of this paper is to analyze the emotional triggers that appear in phishing email datasets collected from educational and industry organizations, and determine if there are any significant differences in the types of human sentiments that are exploited in these emails. We believe this investigation can guide and influence design considerations for existing detection systems which will be more human centered. To achieve this goal our specific research questions were as follows:

RQ1: Is there a difference in the types of emotions and sentiments that are exploited in the various phishing email datasets that we examined?

RQ2: What are the most frequently used emotions and sentiments exploited in each set of the phishing emails as well as the cumulative database?

We conducted this research with publicly available datasets from two educational institutions and two non-educational datasets. Our analysis process was divided into 4 steps which are coding, pre-processing, analyzing with rule-based sentiment analyzer and extracting emotions score by using tokenization.

In this paper, we will present some experimentation on the effect of the type of tokenization used on the email data on the results of a sentiment analysis. Fur-

thermore, we will describe our experimentation with finding relations between the emotional content of a phishing email and sender's email domain (educational, non-educational). Through the analyses, we would be able to see if there is any correlation between the domain of the sender email and the sentiment of the email content is trying to exploit.

2 Background

There have been different approaches when it comes to examining and detecting phishing emails such as URL analysis, webpage content analysis, phishing detection schemes, email content and to educate end users to identify phishing content. Common phishing detection schemes fall under two categories: user education and software [6]. User education consists of security training programs where users are trained to be aware of and recognize common signs of a phishing attack. Software solutions focus on detecting phishing attacks on a systems level using machine learning techniques, blacklisting, and visual similarities [6]. Many of the existing detection schemes rely on email metadata such as domain names, associated email addresses, and URLs. While these techniques are useful and help to mitigate phishing attacks, many phishing scams claim victims by exploiting human vulnerabilities. To help improve detection schemes, researchers have started to analyze the content of these phishing emails to better understand how scammers are exploiting human vulnerabilities through social engineering approach.

Email contents or text have high dimensionality and complexity, and thus present a large number of features to be analyzed. Analysis on such a large dataset is slow and leads to poor classifier performance [7], [18]. Researchers use dimensionality reduction techniques such as feature extraction or feature selection to reduce this dimensionality and make it easier to do analysis on email contents. Researchers are also looking into what techniques and language is used in phishing emails to exploit human vulnerabilities. Attackers can exploit certain emotional triggers such as a user's fear and anticipation by using targeted words and subjects [5]. A better understanding of how attackers exploit human emotions would lead to the ability to create better detection schemes.

In this experiment, we aim to build off of the work done by Sharma and Bashir [5] by conducting sentiment analysis on a phishing email dataset to see what types of sentiments can be extracted from email contents and how best to extract them. This information will help further research into using email content to help aide phishing detection. Understanding the way human emotions are exploited in a phishing attack will help build better detection models for the future.

3 Method

3.1 Data Coding and Initial Analysis

We first Coded the emails into email body, email subject, sender domain, and email signature. Our source of email are: 2 public educational databases ("Berkeley Information Security Office" and "SecureIT-Kent State University"), 2 non-educational databases. In this initial stage, we used NLTK Vader Sentiment Analyzer, a lexicon and rule-based sentiment analysis tool for Preliminary Sentiment analysis.

3.2 Polarity Scoring

To start off the experimentation, we wanted to use a simple tool that would allow us to get a rough idea of what sort of sentiments were expressed in the phishing email dataset. This is why we decided to start off using the NLTK Vader Sentiment Analyzer, a lexicon and rule-based sentiment analysis tool [9]. While it is intended for use in social media posts, the tool is fast and does not require any training data, which made it ideal for the initial experimentation. The NLTK Punkt Tokenizer was used because it can differentiate between periods that end sentences and periods used in words like "Mr. Bach" or "Mrs. Smith". Furthermore, it also can recognize that sentences can start with non-capitalized words [8]. This makes it useful in tokenizing email content because some emails do not exhibit formal English language techniques.

NLTK Punkt Tokenizer In order to pre-process the dataset, we used the NLTK Punkt Tokenizer to split up the email contents into sentences. This would allow the Sentiment Intensity Analyzer to analyze each sentence individually rather than as a whole paragraph. This tokenizer take in dataset and put information into data structure called rows and go through each entry row and check to see if there is a sender email. If yes, then separate the results into different lists in the results dictionary based on the email domain of the sender. It then tokenize the email content using english.pickle tokenizer from nltk Once results calculated for each row, print out the average compound sentiment value for each domain and output results. However, it may be useful to run the Sentiment Intensity Analyzer on the full paragraph without tokenization so the email as a whole is analyzed. The results section contains the polarity scores with and without the usage of the English pickle tokenizer.

NLTK Sentiment Intensity Analyzer The NLTK Sentiment Intensity Analyzer was used to score the data. The analyzer gives the input message a positive, negative, neutral, and compound score on a scale from -1 to 1.

3.3 Emotion Extraction

In this process, it takes in dataset and put information into a data structure called rows and then clean message text and simplify it down to a list of words that provide some sort of info on the sentiment of the sentence. After that, it compares each word to the dictionary and see if any of them are in there. If they are, add the corresponding emotion to the emotions list array. Once each row has been processed, it counts up all the emotions that have been extracted to see which one is most prevalent.

Text Cleaning Following the method used by Attreya Bhatt [10], we cleaned the text data by converting the text to lowercase and removing punctuation so that only the words remained.

Tokenization Once the text was cleaned, the text was tokenized using a word tokenizer that simply created a list of all the words in the cleaned text.

Stop Words The stop words were then removed from the tokenized text using a stop words list from NLTK. If the list contained a stop word found in the NLTK list, it was removed from the final list of tokenized words.

Emotion List To extract emotion data from the tokenized words, we used a dictionary of words matched with their related emotion. Most of the lexicons we found only provided positive or negative sentiments associated with different words, such as SentiWordNet [11] and TextBlob [13], but the one we ended up using, provided by Bhatt [10], had words mapped to their emotion rather than a positive or negative value. This dictionary is by no means an exhaustive list. The emotions expressed are then tallied using a counter module and outputted as a result.

4 Data and Results

4.1 Dataset

Dataset 1 containing data on 500 separate phishing emails from educational institution. This dataset was provided by [5]. It contained the following columns: $Email_Subject$, $Email_Content$, $Sending_Date$, $Sending_Time$, Day, URL_Title, Coined.Word, $Sender_Name$, $Sender_Title$, $Closing_Remarks$, $Sender_Email$, Logo, and To. Dataset 2 and 3 contain non education data from kaggle [1] and data containing both ham and spam email from kaggle [2].

[1] kaggle ernon data. https://www.kaggle.com/wcukierski/enron-email-dataset
[2] https://www.kaggle.com/balakishan77/spam-or-ham-email-classification

4.2 Findings from Dataset-1

Overall Polarity Scores After scoring all of the phishing email data with the Vader Sentiment Analyzer, we took the averages of each category (compound, positive, negative, and neutral) over all of the email content data, shown in Table 1. The positive, negative, and neutral values represent a percentage of the text that fall into those categories. The sum of these values should be 1. The compound score represents a normalized value that falls between -1 (most negative) and 1 (most positive). Figures 1-4 are histograms of all the datapoints for each category (compound, neutral, positive, and negative). We can see that most of the compound values fall between -0.2 and 0, most of them being 0. The positive and negative values sit mostly between 0.0 and 0.2, while the neutral values sit between 0.6 and 1.0, depending on tokenization. In this paper, if the data is tokenized, that means that the NLTK Tokenizer was used to break the email contents into sentences instead of running the sentiment analyzer on each email content as a whole.

type	Tokenization	w/out Tokenization
Comp	0.114	0.128
Pos	0.131	0.122
Neg	0.038	0.046
Neu	0.831	0.832

Table 1: Overall Avg. Scores for data-points by sentiment analyzer

Polarity Scores By Domain The polarity scores were calculated in the same fashion as above, but in this section, averages were taken based on sender email domain. For example, the email "records@dol.gov" would be grouped with other emails ending in .gov. This was done to see if there was any correlation between the email domain, and the sentiment of the email. Due to the small amount of datapoints containing email domain information in this dataset, no conclusions can adequately be drawn by these results. Table 2 shows the results without tokenizing the email contents (meaning analyzing each email as a whole), while Table 3 shows the results with tokenization (analyzing each email sentence by sentence).

Emotion Extraction Phase II was our attempt at further analyzing the email contents after realizing polarity scores did not give as much insight into the data as we wanted. In this phase, we tokenized each email content data point into words, removed those words that did not provide any insight into the emotional content of the email, and then compared the resulting list of words with a dictionary matching words to the respective emotion they represent. Each time an emotion was tallied, it was added to a counter. In the end, each emotion represented was tallied up, and the results are shown in Figure 2 .

Fig. 1: With/out tokenized polarity scores

type	Comp	Pos	Neg	Neu
.com	0.134	0.091	0.112	0.797
.gov	0.517	0.126	0.0	0.875
.net	0.465	0.179	0.0	0.821
.edu	0.651	0.197	0.007	0.796
.org	0.36	0.125	0.011	0.864
none	0.058	0.116	0.049	0.835

type	Comp	Pos	Neg	Neu
.com	0.064	0.139	0.069	0.792
.gov	0.258	0.076	0.0	0.924
.net	0.26	0.216	0.0	0.784
.edu	0.308	0.178	0.005	0.817
.org	0.203	0.164	0.01	0.826
none	0.059	0.119	0.046	0.836

Table 2: Polarity score by domain: a) shows the results without tokenizing the email contents (meaning analyzing each email as a whole), while b) shows the results with tokenization (analyzing each email sentence by sentence).

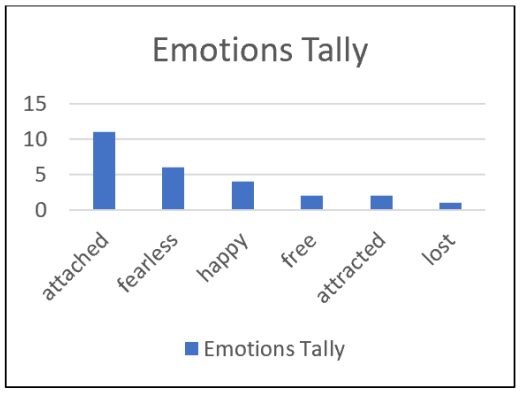

Fig. 2: Emotion Extraction from dataset 1

4.3 Findings from Dataset 2 and 3

For the dataset 2 and 3, we have followed the similar method, after scoring all of the email data with the Vader Sentiment Analyzer, we took the averages of each category (compound, positive, negative, and neutral) over all of the email content data. The positive, negative, and neutral values represent a percentage of the text that fall into those categories. The sum of these values should be 1. The compound score represents a normalized value that falls between -1 (most negative) and 1 (most positive). Figures 3 are histograms of all the datapoints for each category (compound, neutral, positive, and negative). We can see that most of the Enron compound values fall between 0.8 and 1. The negative values sit mostly between 0.0 and 0.2. The neutral values mostly sit between 0.8 and 1. The positive values sit mostly between 0.0 and 0.2. The overall polarity scores are: compound: 0.581, neg: 0.022, neu: 0.878, pos: 0.099

Similarly for dataset 3, after scoring all of the spam email data with the Vader Sentiment Analyzer, we took the averages of each category (compound, positive, negative, and neutral) over all of the email content data. The positive, negative, and neutral values represent a percentage of the text that fall into those categories. The sum of these values should be 1. The compound score represents a normalized value that falls between -1 (most negative) and 1 (most positive). Figures 4 are histograms of all the datapoints for each category (compound, neutral, positive, and negative). We can see that most of the ham compound values fall between -0.8 and 1, most of them being 0. Most of the spam compound values fall between -0.8 and 1, most of them being 1. The ham and spam positive and negative values sit mostly between 0.0 and 0.2. The ham neutral values mostly sit between 0.8 and 1, while the spam neutral values sit between 0.6 and 0.8.

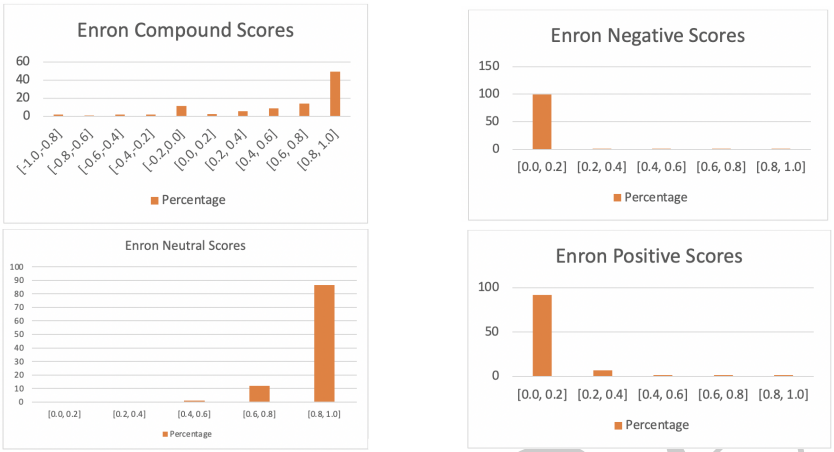

Fig. 3: With/out tokenized polarity scores for dataset2

5 Discussion

Results of these analyses show a clear difference in the types of emotions and sentiments that are included among the three datasets. For example, educational institution's phishing emails are leaning more towards trust, anticipation, and fear, while the other datasets that include industries show different emotions like attachment, fear, and attraction. Although our study results are exploratory and more research is needed to confirm these findings, we believe that our methodology can provide a step forward towards better detection mechanisms that include human vulnerabilities that can easily be exploited. In addition, our research demonstrates the need for a universal database that can collect and share all types of phishing emails across different kinds of institutions in order to capture the various emotional triggers and sentiments that attackers may use to exploit via phishing attacks. This will also improve individual organizations efforts in defending against phishing email on their own which may fail due to the lack of a comprehensive set of emotions and sentiments that can be utilized. Therefore, our study presents a preliminary step towards building a comprehensive and consolidated emotional/sentimental dataset that can help in the detection and deterring of phishing email which continues to be a costly endeavor for many organizations.

Overall, results may vary based on the type of dataset tested. In order to test this question more in-depth, we experimented sender email domain information across pure phishing datasets as well as spam and ham datasets in dataset3. This would allow us to see the consistent correlation between sender email domain and the emotional content of the email. Our analysis of ham emails as well show that values differ or hold true whether or not the email is legitimate or not in respect to negative and neutral polarity score.

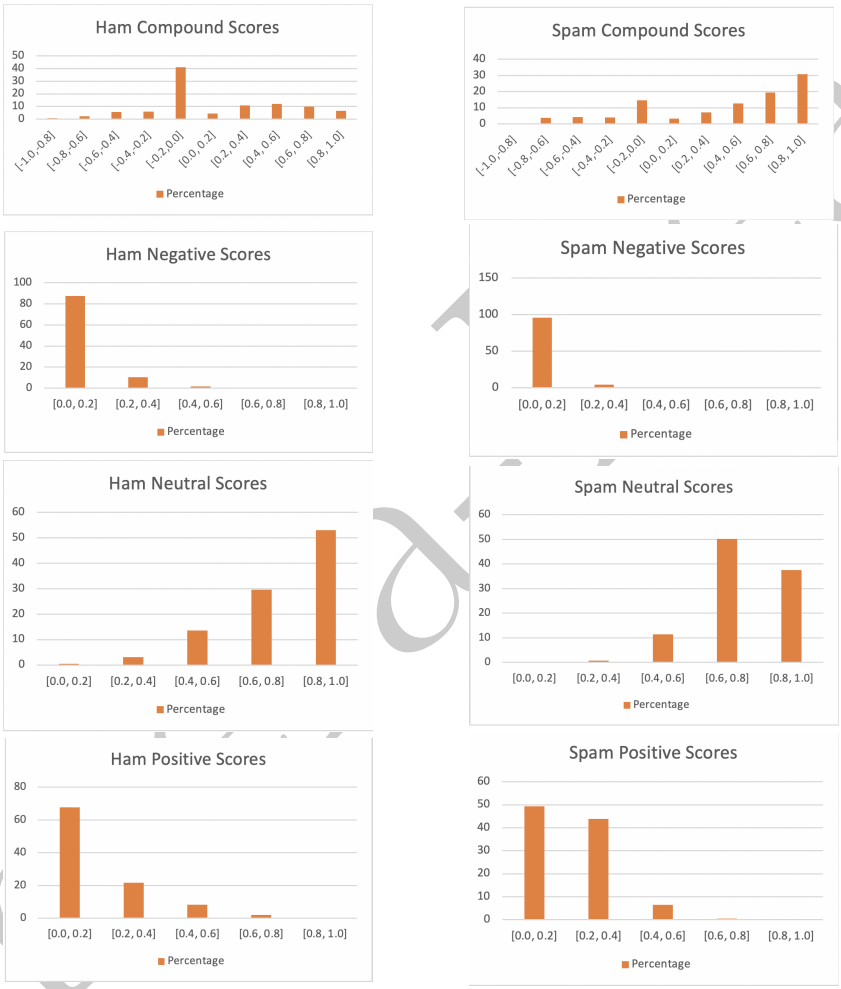

Fig. 4: polarity scores for dataset 3 (comparison of ham and spam email)

Due to the limitations of this experiment, the results do not give generalized recommendation which can draw strong conclusions for all types of phishing email data. The initial dataset was small, having only 500 entries, and of those, only 32 of those data points had email domain information. Datasets Two and Three were significantly larger. This makes the results from Phase 1 noninformative in determining the relation between sender email domain and positive and negative emotion. If tested on a much larger dataset with sender domain data, we would be able to see if there is any correlation between the domain and the emotional content of the email. In terms of the effect of tokenization on the polarity results, we do see that the tokenized averages lend more towards positive polarity. However, if we look at the histograms in figures 1 and 2 there does not seem to be a large difference in clusters of the compound, negative, and positive polarity values between tokenized and untokenized data for this specific dataset. We do see a difference in the neutral polarity values. The untokenized values a higher percentage of neutral scores in the 0.6-0.8 range than the tokenized values.

The results of Phase 2 of emotion extraction of the experimentation may also be slightly unreliable due to the limitations of the dictionary word. Certain words such as “urgently”, “expiring”, and “disregard”, which provide some sense of the type of emotions present in the email, are not in the dictionary. Thus, some words that are important to understanding the emotional content of the text are left unanalyzed.

6 Conclusion

Though this experiment had a few limitations including dataset size and proper emotion mapping techniques, further experimentation would be useful to investigate the relation between sender email domain and emotional content, if any. Future studies would benefit from running analytics on a larger dataset of both spam and ham data types, allowing for any relations between the emotions exploited and the maliciousness of an email. In order to more accurately gauge the emotional content of the email data, machine learning techniques could be used to help classify the overall emotional content of the data. This would be a more accurate method of emotion extraction.

References

1. Gupta, S., Singhal, A., Kapoor, A. (2016, April). A literature survey on social engineering attacks: Phishing attack. In 2016 international conference on computing, communication and automation (ICCCA) (pp. 537-540). IEEE.
2. Ferreira, A., Coventry, L., Lenzini, G. (2015, August). Principles of persuasion in social engineering and their use in phishing. In International Conference on Human Aspects of Information Security, Privacy, and Trust (pp. 36-47). Springer, Cham.
3. Symantec, ”Internet Security Threat Report,” 2019.
4. Verizon, ”Data Breach Investigations Report,” 2019.

5. Sharma, T., Bashir, M. (2020, July). An analysis of phishing emails and how the human vulnerabilities are exploited. In International Conference on Applied Human Factors and Ergonomics (pp. 49-55). Springer, Cham.
6. A. K. Jain and B. B. Gupta, "Phishing Detection: Analysis of Visual Similarity Based Approaches," Hindawi, vol. 2017, p. 20, 2017.
7. M. Zareapoor and S. K. R., "Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection," I.J. Information Engineering and Electronic Business, 2015.
8. W. S. Bird, E. Loper, J. Nothman and A. Darcet, "Source code for nltk.tokenize.punkt," NLTK Project, 2001-2020. [Online]. Available: http://nltk.org/. [Accessed 05 June 2020].
9. C. . G. E. Hutto, "VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text," in Eighth International Conference on Weblogs and Social Media, Ann Arbor, 2014.
10. A. Bhatt, "attreyabhatt/Sentiment-Analysis," GitHub, 2020 Mar 2020. [Online]. Available: https://github.com/attreyabhatt/Sentiment-Analysis. [Accessed 26 May 2020].
11. A. Esuli and F. Sebastiani, "SENTIWORDNET: A Publicly Available Lexical Resource".
12. S. Loria, "TextBlob," [Online]. Available: https://textblob.readthedocs.io/. [Accessed 05 June 2020].
13. Sharma, T., Dyer, H. A., Bashir, M. (2021). Enabling User-centered Privacy Controls for Mobile Applications: COVID-19 Perspective. ACM Transactions on Internet Technology (TOIT), 21(1), 1-24.
14. Sharma, T., Bashir, M. (2020). Use of apps in the COVID-19 response and the loss of privacy protection. Nature Medicine, 26(8), 1165-1167.
15. Arif, M. H., Li, J., Iqbal, M., Liu, K. (2018). Sentiment analysis and spam detection in short informal text using learning classifier systems. Soft Computing, 22(21), 7281-7291.
16. Almomani, A., Gupta, B. B., Atawneh, S., Meulenberg, A., Almomani, E. (2013). A survey of phishing email filtering techniques. IEEE communications surveys tutorials, 15(4), 2070-2090.
17. Caldwell, T. (2013). Spear-phishing: how to spot and mitigate the menace. Computer Fraud Security, 2013(1), 11-16.
18. Moghimi, M., Varjani, A. Y. (2016). New rule-based phishing detection method. Expert systems with applications, 53, 231-242.