Winter 12-10-2023

Protection Against Phishing Attacks on Social Networks with Use of Selected Machine Learning

Aneta Poniszewska-Marańda
Lodz University of Technology, aneta.poniszewska-maranda@p.lodz.pl

Aleksander Lemiesz
Lodz University of Technology

Witold Marańda
Lodz University of Technology

Protection Against Phishing Attacks on Social Networks with Use of Selected Machine Learning

Aneta Poniszewska-Marańda[1]
Institute of Information Technology
Lodz University of Technology, Poland

Aleksander Lemiesz
Institute of Information Technology
Lodz University of Technology, Poland

Witold Marańda
Department of Microelectronics and Computer Science
Lodz University of Technology, Poland

ABSTRACT

Nowadays, many interactions between people have moved to the Internet, mainly to social media. Due to the huge amount of data, hackers target social media by carrying out cyber-attacks, especially phishing. It focuses on tricking the victim into clicking a link and then providing private information or installing malware on the victim's computer. Phishing attacks are becoming more and more difficult to recognize every year. Therefore, there is a need to support humans in this difficult task and machine learning can be used for this purpose. The paper analyzes the works on phishing recognition by humans and artificial intelligence. Then, the new AlexPhish algorithm for classifying phishing URLs was presented, along with a proposal for its implementation on social media platforms. It is trained on the "Web page phishing detection" dataset and achieves an accuracy of 94.53%.

Keywords: Security, Social Media, Cyberattack, Phishing, Machine Learning.

INTRODUCTION

Nowadays, much of the interaction between people has moved to the Internet. In developed countries, approximately 80-90% of the population are Internet users, and this number is increasing every year. The Internet, and especially the part of it where people communicate with each other in the most accessible way, is social media. Initially, we could primarily post text

[1] Corresponding author. aneta.poniszewska-maranda@p.lodz.pl +48 426312796

or photos on social networks. Nowadays, social media are also used to publish short videos and share and forward content published by others to other people, including links to videos, files such as documents for collaborative work, music. Attackers know many ways to launch attacks on social networks. The most common ones are phishing attacks, malware attacks, identity theft attacks, brute force attacks, social engineering attacks, spam attacks and distributed denial of service (DDoS) attacks. Often, a short video is published on a given topic, containing a link to other content. References and links can appear not only in video materials, but practically in any form of content transmitted on social media, such as a post or private message. This leaves many opportunities for potential phishing attacks (Al-Otaibi and Alsuwat, 2020, Alwanain 2020, Blancaflor et al. 2021).

Phishing attacks aim to steal confidential data, such as data needed to log in to a given website, banking details or sensitive data such as an identification number or ID number. Attackers can launch a successful attack by relying on users' trust in other people they meet on social media. A typical phishing attack involves manipulating the user into using a fake login page where they provide their details to log in to the website. Thanks to this, the cybercriminal gains access to user data. After data theft, the cybercriminal's fake website redirects the user to the authentic website so as not to arouse suspicion among the victim. A phishing attack can also involve sharing a link to a website that installs malware on the victim's machine in the background. Then such a machine can be used remotely by a cybercriminal (Alwanain 2020).

Phishing, although not a new type of attack, still constitutes the vast majority of attacks. However, the number of phishing attacks increases every year. Hackers are constantly refining their phishing attacks to be as sure as possible that they convince the user that the content of the

attacks is authentic and safe. People are definitely bad at recognizing such attacks. Therefore, there is a growing need to develop methods that can effectively recognize phishing attacks.

In recent years, with the development of social networks, machine learning and related algorithms have developed very dynamically. Solutions based on these algorithms are used to ensure cybersecurity. However, there is a lack of solutions that could be used by social networks to filter infected content and at the same time be adaptable to a specific social network. Social networks are reluctant to share data about attacks carried out against them, so solutions independent of them are unable to update their databases based on newly encountered attacks (Al-Otaibi and Alsuwat, 2020, Alwanain 2020, Blancaflor et al. 2021, Dolega et al. 2021, Kizgin et al. 2020, Studen and Tiberius, 2020, Wardati and Mahendrawathi, 2019)

This paper presents the developed solution for classifying phishing content based on machine learning and the evaluation of results achieved by it. Due to the presence of URL link in the vast majority of phishing attacks, recognition of infected content is based on the link and the page beneath it. The developed model was implemented using the Python programming language and PyTorch library dedicated to the development of machine learning algorithms.

RELATED WORKS

A review of the literature related to phishing attacks and other solutions used to recognize phishing attacks shows that people are not good at recognizing phishing and that this type of attacks are becoming more and more common and pose a greater threat (Dolega et al. 2021, Kizgin et al. 2020, Studen and Tiberius, 2020, Wardati and Mahendrawathi, 2019). Research was conducted to determine the extent to which people are able to recognize phishing emails on their own, how phishing attacks have changed over the years, and what is the impact of these changes on users' ability to recognize phishing attacks (Butavicius et a. 2022, Carroll et al. 2022). There

are also works on recognizing phishing URL links, but very often when creating this type of solutions, the authors decide to create their own database, which makes the results very difficult to compare with other works (Alani and Tawfik, 2022, Butavicius et al. 2022, Carroll et al. 2022, Hannousse and Yahiouche, 2021, Sánchez-Paniagua et al. 2022). The work of (Hannousse and Yahiouche, 2021) proposes a well-constructed set of data obtained from various sources, which increases its usefulness. This set was created to build a reference or comparison set for various types of solutions for classifying infected URL links. Due to the quality of received data set, it was used to develop a new phishing attack recognition system based on artificial intelligence during our works.

METHOD FOR RECOGNIZING PHISHING IN SOCIAL MEDIA

Undoubtedly, people have a big problem recognizing phishing websites. It is possible to learn how to recognize attacks, but with the constant development of phishing attacks and the very dynamic pace of increasing the frequency of attacks, it is only a matter of time when even a good user will fall victim to an attack. Additionally, phishing attacks are increasingly common on social media. Social media attacks can take place in a way that more or less resembles a typical email attack. As part of our work, we propose the AlexPhish machine learning model to classify URL links into those leading to authentic websites or phishing ones. The classifier can be used as a social media content filter by network administrators to eliminate or reduce phishing attacks on a given platform.

Approaching the problem of phishing in social media

The most used phishing attacks in social media is the one carried out via a text message in the instant messenger built into the social media platform. In this attack, the cybercriminal uses a fake account impersonating a company or a famous person to gain the victim's trust. This

attack resembles classic phishing emails. The infected message is distributed to many users and manipulates them into using the hyperlink contained in the message. In the content of the message, the attacker often also puts time pressure on the victim to do what he or she is asked to do in the message as quickly as possible.

Another way to conduct a phishing attack in social media, which is more different from a typical e-mail attack, is an attack by impersonating a company or institution in order to publish a post in one of the groups or on the bulletin board of an account created for attacks. Cybercriminals decide to use this type of attack because a post once posted can be shared or even appear in users' news feeds if the algorithm of a given social media platform deems it popular, which may arouse less suspicion than a private message. This is because on social media, companies or people share information about their services or products through posts. The attacker can use additional accounts used by bots to create positive comments and a significant number of likes under the post (Dolega et al. 2021, Kizgin et al. 2020, Studen and Tiberius, 2020, Wardati and Mahendrawathi, 2019).

However, if a phishing post is published in a group of people looking for deals, it has a chance to reach almost all of its members and does not require sending messages to each of them, which could additionally arouse their suspicions, unlike a post on one of their favorite groups with promotions. Moreover, if an entry is added to a thematic group, the attacker does not have to impersonate a specific company, but can say that he is a customer and has discovered an error in the system of a given store or wants to share a link he found online, which can be used to shop at a more favorable price. Such entries may also motivate user to click on the link with limited offers or, in the case of the above-mentioned errors in the system, inform user that the pricing error may be fixed at any moment and he should hurry up to use the link so as to take

advantage of the offer before it is too late. If this type of attack is carried out correctly, it can be very difficult for the user to detect. If it also gains sufficient credibility and popularity, it may be shared by accounts unrelated to the attacker and thus spread on its own.

Another attack that can be carried out and is definitely different from the classic attack is placing phishing content in a comment under a post in the news or in a group. For the attack to be effective, the comment should have many likes and contain engaging content on the topic discussed in the post. This method of attack allows to use the popularity of safe content to spread the attacker's content. Thanks to this, the attack can reach many users and also gain their trust as content appearing in the official discussion. To avoid, companies often employ moderators who monitor content related to their employer and remove content that may be harmful. However, without an IT system, this part of the protection of Internet users rests on the shoulders of people who, as it turns out, have an increasing problem with correctly recognizing phishing attacks.

Method for detecting phishing in social media

In order to counteract phishing attacks carried out using posts, comments or messenger messages, it is proposed to use a system that monitors new content on social media and then classifies this content as safe or infected. The classification is based on the link contained in the content to be published. If the link is deemed safe, all content is deemed safe and may be published. However, if the link is considered infected, the new content is not allowed for publication and may be immediately rejected or referred to a moderator for evaluation, depending on the approach preferred by the administrators of a given social media platform. The proposed system is therefore a filter of content published on social media. By knowing which content is phishing, social network administrators can identify accounts used to carry out attacks and therefore send them warnings or remove them. This limits the possibilities of cybercriminals.

Basing the anti-phishing filter on URLs contained in published content carries the risk of not filtering out phishing content without a link. Attackers may choose to separate the content of the attack from the hyperlink they want the victim to click on. However, this means that the separated URL link would be filtered out and this would render the vast majority of attacks harmless. Even if the potential victim can manipulate the content of the phishing, he or she is unable to interact with the hyperlink due to its absence. This makes it impossible to provide data on a cybercriminal's impersonating website or download software using it. This solution to the problem of phishing in social media limits the possibilities of cybercriminals without significantly interfering with the functioning of a given social network.

Social networks have various types of content filters, such as those that prevent the use of hate speech, so an additional filter is only an additional stage in the existing process and does not involve any changes for the users of the social network and does not change the architecture or operation of the social network, which is why the owner or the administrator of the entire network is not forced to redefine the system or redesign it. The only cost that the social network would have to incur is acquiring a solution and adding another filter to the existing ones.

Architecture of developed anti-phishing system

The developed anti-phishing system is based on a neural network. The neural network (NN) prototype was written in Python and using the open-source PyTorch library for programming machine learning algorithms. The NN accepts 86 features from the data set as input, and at the output it performs a binary classification of URL links into safe and phishing ones. Network layers are launched sequentially. This means that the result of the operation from a given layer is passed to the next one. The architecture of the NN is shown in figure 1. The implementation code of created network is shown in figure 2.

The input layer of the network accepts data as a feature vector, and the output determines the label assigned to the data. In the case of phishing detection, it is a binary classification, i.e. a classification of two classes, so the output layer has one output. The model consists of a total of 9 layers. The network architecture consists of 3 linear layers, and 2 of the 3 are followed by ReLU activation function. Additionally, there is also a dropout layer after them. After the second linear layer there is also a normalization layer. This architecture allows to obtain high results while limiting the classification time. Linear layers are very good for training numerical data that are independent features, as is the case with classification based on URL link features. The use of ReLU function is also computationally simple and allows less frequent activation, which further simplifies and speeds up model training. The use of dropout layers can improve results and prepare the model for data different from those encountered during training. In order to avoid too large numbers, a normalization layer was added before the last linear layer and sigmoid layer.

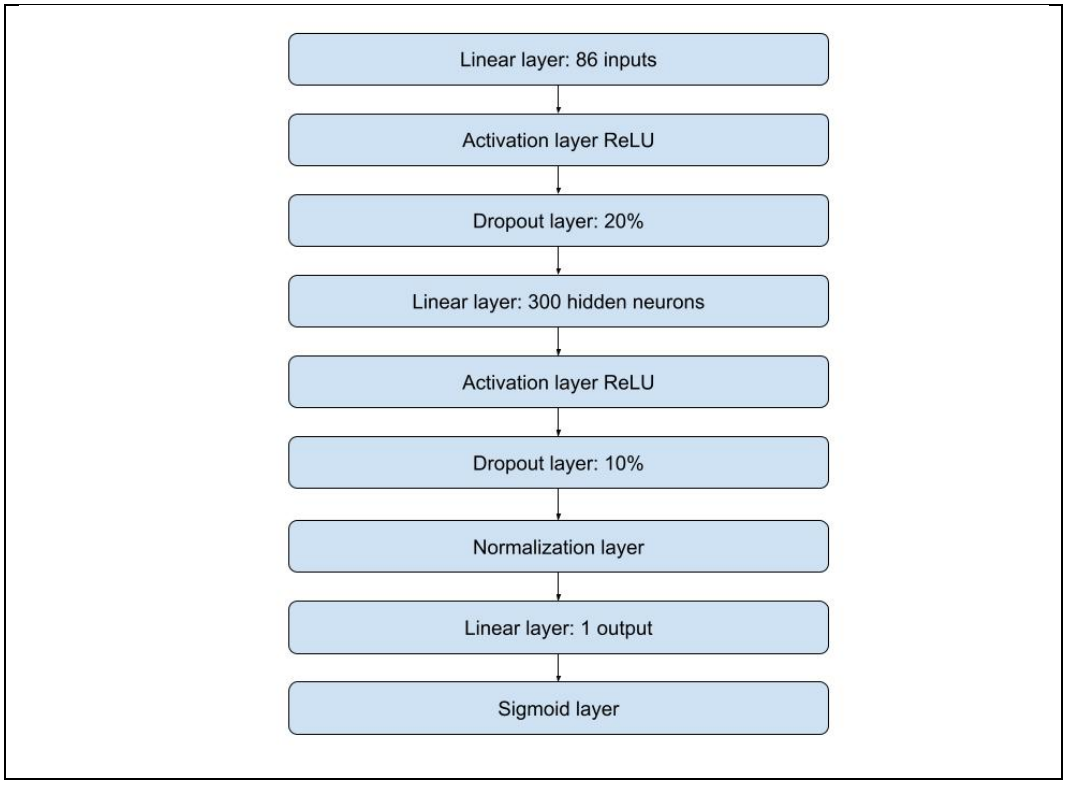

Figure 1. Schema of neural network architecture used to classify phishing URL links

The data set used for training was “Web page phishing detection” set (Hannousse and Yahiouche, 2021). It was selected because it designed as a benchmark dataset for phishing recognition problems. Page features are divided into three categories. These are the features of URL link itself, features obtained through page analysis and features obtained using external websites. Features associated with a link include, among others: the number of characters in the link or the number of dots in it. Features obtained by analyzing the page behind the link are the number of empty hyperlinks inside the page and the number of redirects to external pages. Features of the last group of features included in "Web page phishing detection" include, e.g. the time that has passed since the website was registered or the pagerank result of the analyzed website. The dataset also has a feature extraction algorithm for any given website. This allows for a reliable comparison of the results of proposed solution with the results presented by the authors of the data set. “Web page phishing detection” contains high-quality data because it was collected from many sources to compensate for the shortcomings of individual websites. Additionally, the dataset is balanced, which means it contains the same number of URL links to legitimate sites as to infected ones.

```
model = nn.Sequential(
    nn.Linear(in_features=86,out_features=300),
    nn.ReLU(),
    nn.Dropout(p=0.2),
    nn.Linear(in_features=300,out_features=300),
    nn.ReLU(),
    nn.Dropout(p=0.1),
    nn.BatchNorm1d(num_features=300),
    nn.Linear(in_features=300,out_features=1),
    nn.Sigmoid()
)
criterion = nn.BCEWithLogitsLoss()
optimizer = torch.optim.Adam(model.parameters(), lr=learning_rate)
```

Figure 2. Implementation of NN architecture used to classify phishing URL links using PyTorch

While training the model, the dataset was divided into training and testing subsets. The test subset consists of 20% of the data of entire set. Such a set allows to use a large part of the data for learning and avoid overfitting thanks to verification on the remaining data. The loss function chosen in science was the sigmoidal binary cross-entropy. This function combines the advantages of sigmoid activation and binary cross entropy loss, thereby increasing the efficiency of the training process. Adam (*Adaptive Moment Estimation*) was chosen as the learning optimizer. It is an adaptive optimizer combining the advantages of RMSprop and AdaGrad optimizers. The adaptability of the optimizer is achieved by adjusting the learning speed for each parameter. This allows for more effective training due to the fact that individual parameters converge to the minima individually and therefore do not oscillate around them while waiting for the other parameters to converge. Adam also uses regularization techniques and is resistant to noise in the data, which makes it very versatile tool. It also has parameters such as learning rate. In the training of developed model used to classify phishing content, the coefficient was 0.001 and allows to avoid defects caused by too small or too large values of this coefficient.

The model learning takes place iteratively over epochs and is divided into stages relative to the training set. The maximum number of epochs is 100, but a realistic stopping condition to avoid model overfitting during training is that the model accuracy value on examples from the test set cannot fall below the value achieved at the end of training in the previous epoch. Initially, the accuracy increases as the model adapts to the data, but the moment when the accuracy for the test set begins to decline means that the model begins to overfit and over-adapt to the data from the training set, losing the ability to generalize. The training set and the test set were divided into parts consisting of 64 data copies. The used model developed in this way needed to be trained for

four epochs, and then it started to overfit, so learning was stopped. The efficiency achieved in this way reaches 94.53%. The effectiveness after the first epoch was 92.48%.

Classification of new URL link

The developed system is based on the NN model described above, but it is necessary for its operation to provide a simple way to load the features for any link, on the basis of which the classification is made. A simple text interface was created for the prototype. After running the prototype, the user is asked to provide a URL link, and the result of the system is a message notifying about a possible threat or lack thereof. The list of steps the algorithm performs to classify a new URL link, along with their implementation, are as follows:

1. Loading the names of the link features.
2. Waiting for URL link from the user.
3. Feature extraction for the user-provided URL link.
4. Normalization of features according to the proportions used during training.
5. Loading the trained model from the file.
6. Classification based on trained model.

The first step needed to perform classification is to load the names of the features used during model training. Then, feature extraction is performed using algorithms provided by the database creators. The obtained features belong to the set obtained from the link structure, from the set of data obtained from the content of the pages and from the set of features obtained from external websites. To obtain correct results, the features of the provided URL link must be normalized. After loading the features and normalizing them, the system uses the developed and trained model to perform classification. Data during training are additionally normalized during processing to avoid too large values. The classifying a user-supplied link is shown in figure 3.

The database is available at: https://www.kaggle.com/datasets/shashwatwork/web-page-phishing-detection-dataset. Data in used database is saved in CSV file. Single entry looks like:

http://www.crestonwood.com/router.php,37,19,0,3,0,0,0,0,0,0,0,0,0,3,0,1,0,0,0,0,1,0,0,0,1,0.0,0.0,0.0,0,0,0,0,0,0,3,0,0,0,0,0,0,4,4,3,3,3,11,11,6,5.75,7.0,4.5,0,0,0,0,0,0,17,0.529411765,0.470588235,0,0,0,0.875,0,0.5,0,0,80.0,0,100.0,0.0,0,0,0,0,0.0,0,0,0,0,1,0,45,-1,0,1,1,4,legitimate

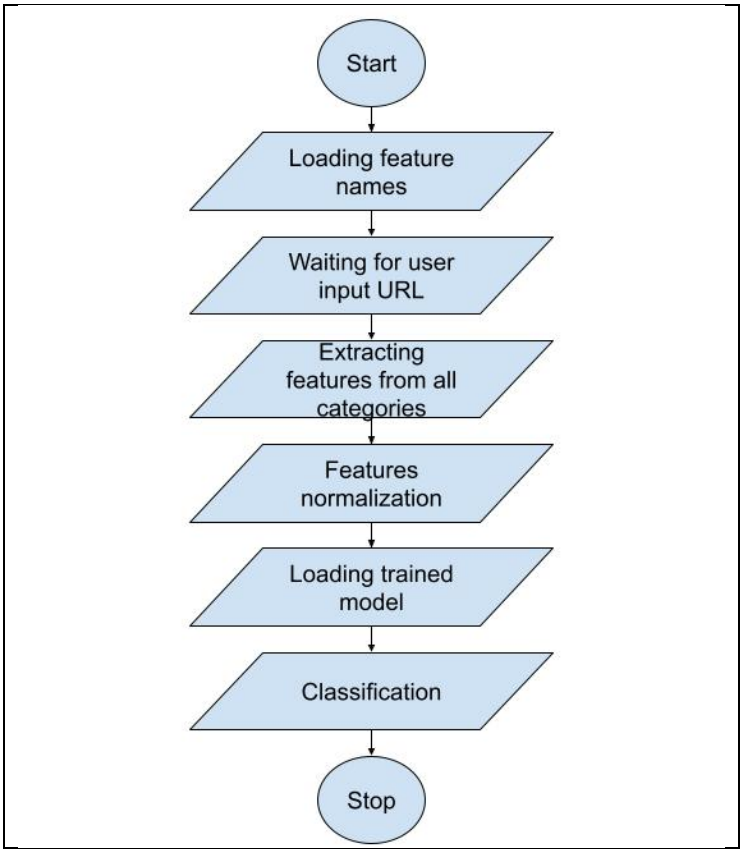

Figure 3. Algorithm for classifying a new link provided by the user

Possible use of the model

The developed anti-phishing model can be used in social networks as a filter used when publishing new content. Unlike similar solutions, such as “PhishNot” (Alani and Tawfik, 2022), which is an external service (social media do not want to send links to an external service, so as not to reveal how often their site is used as a platform to carry out the attacks), the proposed model is not made available via API. Instead, the prototype is released in source code form. This

is the preferred method because sharing the social network solution code allows administrators to allocate the appropriate amount of resources and does not require sending all URL links published on social media to an external site. Social networks may not want to share with others data about the links their users publish and the attacks that are carried out using them.

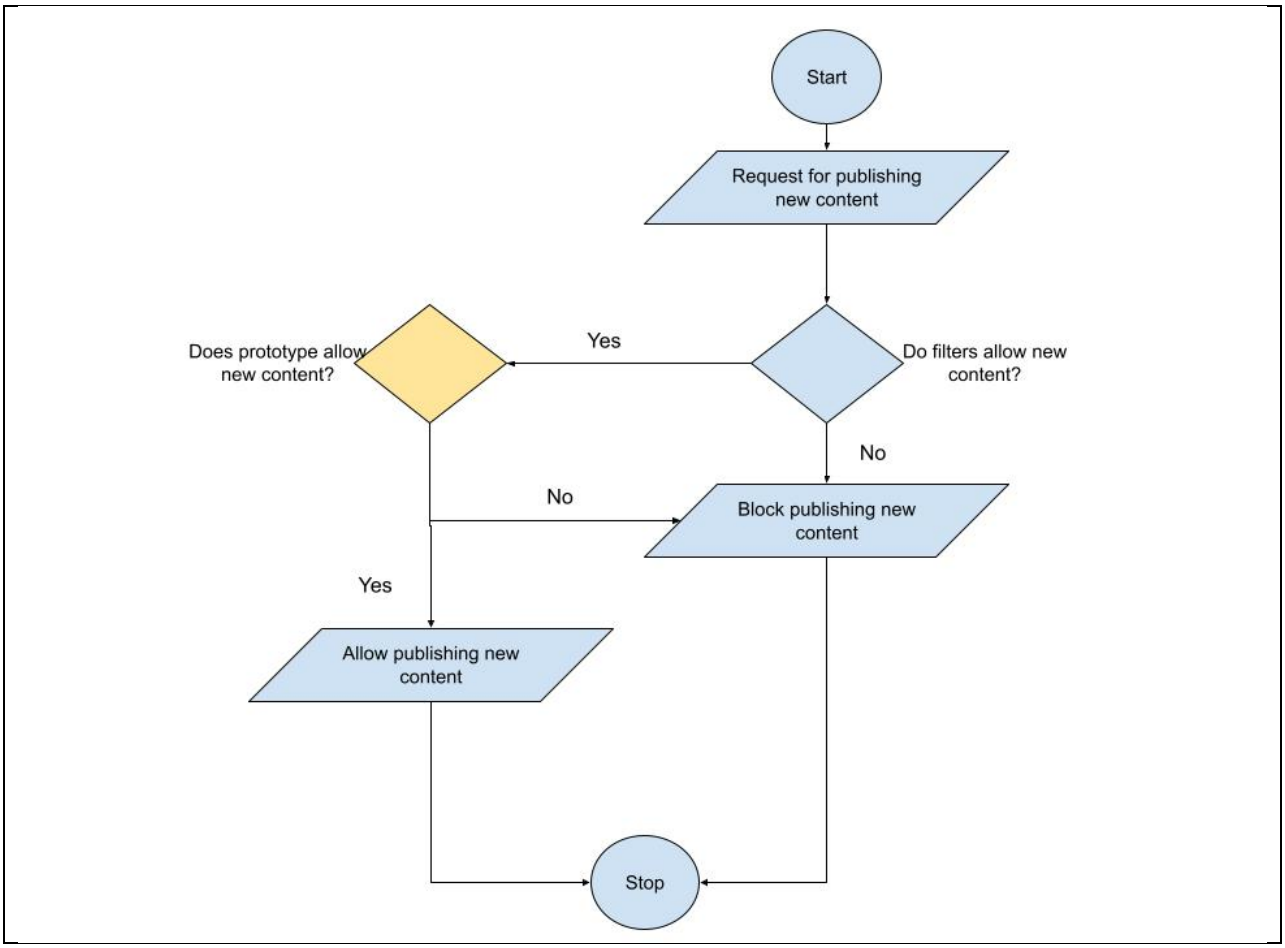

Figure 4. Filtering algorithm for published content, the existing part of social networking systems is marked in blue, the required extension is marked in yellow

Social networks already use content filters. Existing filters prevent the publication of content aimed at spreading hate speech. This means that using the proposed anti-phishing solution does not require changing the architecture of the social networking system. To effectively implement the prototype, simply run it, pass new URL links to the model, and then read the classification result. Having the code and model locally gives us the additional ability to

add new infected links to the dataset, so that the model learns on ongoing basis based on new data. Figure 4 presents the content filtering algorithm with the proposed anti-phishing filter added. It shows how simple it is to add another filter to an existing architecture. In the code of filters used by a social network, the use of the proposed model may be limited to adding another condition for content publication.

PROTOTYPE CLASSIFICATION RESULTS AND THEIR ANALYSIS

The developed and trained NN model, named AlexPhish achieved an efficiency of 94.53% on all types of features available in the data set (Tab. 1). A total of 86 features from all three groups were used. This means that this algorithm achieved better results than 4 out of 5 algorithms used by the authors of the database (Hannousse and Yahiouche, 2021). It turned out to be only less accurate than the random forest algorithm, and the difference between the effectiveness of the two algorithms differed by 2.08% points. Genuine sites are recognized with an accuracy of 96%, and phishing sites with an accuracy of 91%. This means that authentic websites are recognized with greater accuracy. Infected sites are considered authentic in 8% of cases, and honest sites are considered phishing only in 4%. Based on this, it can be concluded that AlexPhish correctly recognizes the vast majority of data regardless of its class.

Table 1. Distribution of true positive (TP), false negative (FN) and false positive (FP) values for trained AlexPhish model

Type of pages	TP	FN	FP
Authentic	96%	8%	4%
Infected	91%	4%	8%

The developed AlexPhish model has a macro F1 score of 93.97%. This result is close to the accuracy of the model. The authors of the database for their best solution obtained an index value of 96.60%, which is also very close to the accuracy of their model, which is 96.61%. The macro F1 value for AlexPhish is unfortunately not that good compared to the compared

algorithms and is only better than the two tested algorithms, SVM and Naive Bayes. However, its value does not differ significantly from that achieved by decision tree and logistic regression.

Table 2. Indicator values for each class for the trained AlexPhish model

Class	F1 score	Recall	Precision
Authentic	94,12%	92,31%	96%
Infected	93,82%	95,79%	91,92%

The advantage of AlexPhish is that it is easy to modify. Further research into the optimal network architecture may lead to improved results, and the random forest algorithm leaves no room for improvement. The algorithms that turned out to be worse than the proposed NN model were Decision Tree, Logistic Regression, Naive Bayes and SVM algorithm. This means that the developed NN is well suited to classifying phishing URL links. Data extraction for a new link takes between 40 and 90 seconds. This means that in the case of an instant messenger implemented on a social networking site, it is not an optimal solution because it will significantly delay communication, but in the case of publishing a post or comment, it should not have a significant impact on the operation of the network. Additionally, this result was obtained on a mid-range machine owned by a private individual and with a standard Internet connection. If feature extraction was performed on a machine with greater computing power and very fast Internet access, it would undoubtedly shorten the time needed to obtain features. Additionally, feature acquisition could be performed in parallel, which would shorten the waiting time for a response. Another method of accelerating the algorithm, but reducing the effectiveness of the filter, would be to use only the features that the database authors marked as the most significant. Fewer features would certainly result in obtaining them faster. Combining this solution with the introduction of parallel obtaining them would shorten this process even more. However, limiting the features would certainly have an impact on the results achieved. AlexPhish does a good work

of recognizing infected URL links. The effectiveness achieved is high and the algorithm can be used in social media to effectively recognize phishing websites. Further research should focus on improving effectiveness of AlexPhish model and improving the recognition of phishing attacks.

REFERENCES

Alani, M.M., and Tawfik, H. 2022. "PhishNot: A Cloud-Based Machine-Learning Approach to Phishing URL Detection," *Computer Networks*, (218).

Al-Otaibi, A.F., and Alsuwat, E.S. 2020. "Study On Social Engineering Attacks: Phishing Attack," *International Journal of Recent Advances in Multidisciplinary Research*, (70).

Alwanain, M.I. 2020. "Phishing Awareness and Elderly Users in Social Media," *International Journal of Computer Science and Network Security*, (20:9).

Blancaflor, E.B., Alfonso, A.B., Banganay, K.N.U., Dela Cruz, G.A.B., Fernandez, K.F., and Santos, S.A.M. 2021. "Let's Go Phishing: A Phishing Awareness Campaign Using Smishing, Email Phishing, and Social Media Phishing Tools," *Proceedings of the 11th Annual International Conference on Industrial Engineering and Operations Management*, Singapore, March 7-11.

Butavicius, M., Taib, R., and Han, S.J. 2022, "Why people keep falling for phishing scams: The effects of time pressure and deception cues on the detection of phishing emails," *Computers & Security*, (123).

Carroll, F., Adejobi, J.A., and Montasari, R. 2022. "How Good Are We at Detecting a Phishing Attack? Investigating the Evolving Phishing Attack Email and Why It Continues to Successfully Deceive Society," *SN COMPUT. SCI*. (3:170).

Dolega,, L., Rowe, F., and Branagan, E. 2021. "Going digital? The impact of social media marketing on retail website traffic, orders and sales", *Journal of Retailing and Consumer Services*, (60).

Hannousse, A., and Yahiouche, S. 2021. "Towards benchmark datasets for machine learning based website phishing detection: An experimental study," *Engineering Applications of Artificial Intelligence*, (104).

Kizgin, H., Dey, B.L., Dwivedi, Y.K., Hughes, L., Jamal, A., Jones, P., Kronemann, B., Laroche, M., Peñaloza, L., Richard, M.O., Rana, N.P., Romer, R., Tamilmani, K., and Williams, M.D. 2020. "The impact of social media on consumer acculturation: Current challenges, opportunities, and an agenda for research and practice," *International Journal of Information Management*, (51).

Sánchez-Paniagua, M., Fidalgo, E., Alegre, E., and Alaiz-Rodríguez, R. 2022. "Phishing websites detection using a novel multipurpose dataset and web technologies features," *Expert Systems with Applications*, (207).

Studen, L., and Tiberius, V. 2020. "Social Media, Quo Vadis? Prospective Development and Implications," *Future Internet*, (12), p. 146.

Wardati, M.K., and Mahendrawathi, E.R. 2019. "The Impact of Social Media Usage on the Sales Process in Small and Medium Enterprises (SMEs): A Systematic Literature Review," *Procedia Computer Science*, (161).