Mateusz Koryciński
Instytut Automatyki i Informatyki Stosowanej, Wydział Elektroniki i Technik Informacyjnych, Politechnika Warszawska
ORCID: 0000-0002-5239-1635
mateusz.korycinski@pw.edu.pl

Konrad A. Ciecierski
Instytut Automatyki i Informatyki Stosowanej, Wydział Elektroniki i Technik Informacyjnych, Politechnika Warszawska
ORCID: 0000-0003-2471-3016
konrad.ciecierski@pw.edu.pl

WYKRYWANIE ATAKÓW PHISHINGOWYCH NA WCZESNYM ETAPIE KAMPANII

Streszczenie

Użytkownicy Internetu codziennie narażeni są na różnorodne zagrożenia cybernetyczne, spośród których szczególnie groźne są ataki phishingowe. Ich celem jest kradzież poufnych informacji – najczęściej danych logowania do kont bankowych – poprzez podszywanie się pod zaufane instytucje i tworzenie stron internetowych łudząco podobnych do portali płatniczych. W przypadku braku odpowiednich zabezpieczeń, skutki takiego ataku mogą być bardzo dotkliwe dla ofiary i wiązać się nawet z utratą wszystkich środków na koncie. Pomimo istnienia szeregu metod ochrony przed tego typu atakami, przestępcy wykorzystują coraz to nowe techniki działania, co utrudnia skuteczne przeciwdziałanie. W niniejszym artykule przybliżamy czytelnikom mechanizmy działania ataków phishingowych, oraz opisujemy potencjalne metody ich wczesnego wykrywania, szczególnie na wczesnym etapie trwania kampanii. Szczególną uwagę poświęcamy wykorzystaniu sztucznej inteligencji, która umożliwia identyfikację potencjalnie niebezpiecznych domen internetowych, już na etapie zapytań do serwera nazw domenowych (DNS, ang. Domain Name Server). Mimo obiecujących wyników, systemy te nadal nie są nieomylne, dlatego ich decyzje powinny podlegać weryfikacji przez wykwalifikowanych analityków. Automatyczne blokowanie fałszywie

zidentyfikowanych jako phishingowe a bezpiecznych stron internetowych wiąże się bowiem z ryzykiem poważnych konsekwencji dla ich właściciela - zarówno wizerunkowych jak i finansowych.

Słowa kluczowe: phishing, detekcja ataków, bezpieczeństwo, wyłudzenia.

WSTĘP

Użytkownicy Internetu narażeni są na różnorakie ataki cybernetyczne, mające na celu kradzież tożsamości, kradzież danych (w tym danych wrażliwych), uzyskanie nieautoryzowanego dostępu do systemu komputerowego, czy też wyłudzenie pieniędzy. Metody ataków stale ewoluują z uwagi na rozwój systemów ich detekcji oraz wzrastającej świadomości samych użytkowników. Mimo to, jest wiele skutecznych ataków paraliżujących działanie firm czy też całych sektorów gospodarki. Przykładami spektakularnych ataków o szerokim zakresie działania są ataki typu ransomware. Atak ten polega na zaszyfrowaniu dysku twardego komputera (osobistego lub serwera) oraz żądaniu okupu (ang. „ransom", skąd nazwa ataku) za jego odszyfrowanie. Najgłośniejszymi przykładami takich ataków jest chociażby WannaCry[1] przeprowadzony w maju 2017 r., czy kampanie Petya/NotPetya[2]. Poza potencjalnym ryzykiem wyłudzenia pieniędzy (w przypadku, gdy ofiara nie posiadała kopii zapasowej zaszyfrowanych danych), firmy ponosiły kolosalne straty materialne wynikające ze sparaliżowania ich systemów wewnętrznych. Niestety ataki dotknęły również placówki medyczne, co skutkowało wyłączeniem z użytkowania między innymi aparatury diagnostycznej. Ataki te trafiły na główne strony gazet z uwagi na swoją niebywałą skalę. Zagrożenie cyberatakiem czyha jednak również na co dzień na każdego z nas i nosi miano ataku phishingowego, a jego celem są zwykle osoby fizyczne. Atakujący tworzą kampanię, w której, aby zmylić i przechytrzyć

[1] Ghafur, S., Kristensen, S., Honeyford, K., Martin, G., Darzi, A., & Aylin, P. (2019). A retrospective impact analysis of the WannaCry cyberattack on the NHS. *Npj Digital Medicine 2019 2:1*, *2*(1), 1–7. https://doi.org/10.1038/s41746-019-0161-6

[2] Fayi, S. Y. A. (2018). What Petya/NotPetya Ransomware Is and What Its Remidiations Are. *Advances in Intelligent Systems and Computing*, *738*, 93–100. https://doi.org/10.1007/978-3-319-77028-4_15/FIGURES/10

potencjalne ofiary, podszywają się pod znane podmioty cieszące się powszechnym zaufaniem, takie jak organizacje rządowe, operatorzy pocztowi, firmy kurierskie, operatorzy telekomunikacyjni czy banki. Wykorzystując różne środki komunikacji wyłudzają następnie fałszywe płatności, które stanowią przykrywkę do wykradania danych logowania do kont bankowych, a następnie kradzieży przechowywanych w banku środków.

RODZAJE ATAKÓW PHISHINGOWYCH

Ataki phishingowe przybierają różnorakie formy. Atakujący mogą wykorzystać różne kanały komunikacyjne, takie jak wiadomości e-mail, wiadomości SMS, WhatsApp itp., jak również tradycyjne połączenia telefoniczne. Ich celem jest przekonanie ofiary do dokonania płatności, np. uregulowania należności za fałszywą fakturę czy dopłaty do paczki kurierskiej. Często jednak nie chodzi o zapłacenie rzeczonej kwoty (zwykle kilka złotych) na konto atakujących a wykradnięcie danych logowania do banku poprzez podstawienie fałszywej – łudząco podobnej do oryginalnej - strony internetowej. Atakujący liczą, że nieuważny użytkownik nie sprawdzi poprawności adresu URL, certyfikatu witryny czy szczegółów samej strony, a rozpoznając sam jej wygląd nabierze zaufania i poda wszystkie potrzebne dane. System przygotowany przez przestępców wykorzysta następnie dane logowania do tego, aby przelać znacznie większą sumę na inne konto, w wyniku czego ofiara może stracić wszystkie przechowywane w banku oszczędności. Pieniądze zwykle przelewane są wielokrotnie pomiędzy różnymi kontami, co sprawia, że znalezienie sprawcy staje się niezwykle trudne, a często wręcz niemożliwe.

Ataki phishingowe charakteryzują się określoną strategią. Przygotowywana kampania ukierunkowana jest na konkretne grono odbiorców, dla których przygotowywane są fałszywe strony internetowe. W celu ich udostępnienia cyberprzestępcy muszą zarejestrować w jednym z TLD domenę internetową. Pojedyncza domena wykorzystywana jest tylko do jednej kampanii, a po jej zakończeniu (zwykle kilka godzin) jest usuwana. Z uwagi na stosunkowo krótki czas trwania takich kampanii niezwykle ważnym jest wczesne ich wykrycie i zablokowanie domen phishingowych, aby jak

najbardziej ograniczyć ich zasięg, a tym samym liczbę potencjalnych ofiar. Nazwy domen phishingowych dobierane są przez przestępców w taki sposób, aby na pierwszy rzut oka przypominały oryginalny adres podrabianej strony. URL może być wizualnie bardzo zbliżony, a w przypadku użycia unicode wizualnie identyczny z tym używanym przez oryginalną stronę. Może on również używać stosownego dla podrobionej strony certyfikatu SSL a poprzez to upewniać oszukiwaną osobę, że wszystko jest bezpiecznie i w należytym porządku.

SPOSOBY PRZECIWDZIAŁANIA ATAKOM

Aby przeciwdziałać takim atakom stosowane są między innymi kampanie społeczne. Przygotowywane są one we współpracy z wyspecjalizowanymi zespołami reagowania na incydenty bezpieczeństwa (np. CERT Polska), które informują obywateli o aktualnie odbywających się kampaniach phishingowych. Jednocześnie instruują użytkowników jak można rozpoznać podejrzaną wiadomość, na co należy zwrócić uwagę oraz jak na nie reagować. Zespoły takie często przyjmują też zgłoszenia potencjalnych kampanii phishingowych w celu ich analizy i ostrzeżenia pozostałych użytkowników.

Poza informowaniem i kształceniem użytkowników ważne są narzędzia pozwalające oszacować ryzyko związane z odwiedzeniem konkretnej witryny. W przestrzeni Internetu dostępnych jest wiele narzędzi tego typu, jak chociażby LinkChecker[3], VirusTotal[4], czy URLVoid[5]. Wykorzystują one metody heurystyczne, uczenie maszynowe, jak również obszerne bazy danych zawierające adresy URL znanych phishingów, opracowywane przez badaczy bezpieczeństwa analizujących różne kampanie. Przykładem takiej szeroko dostępnej bazy jest lista ostrzeżeń przygotowywana przez CERT Polska[6]. Lista ta, aktualizowana na bieżąco wykorzystywana jest w Polsce przez operatorów telekomunikacyjnych, firmy, organizacje, jak i samych użytkowników do

[3] NordVPN Link Checker, https://nordvpn.com/pl/link-checker/
[4] Virus Total, https://www.virustotal.com/gui/home/url
[5] URL Void, https://www.urlvoid.com/
[6] Lista Ostrzeżeń CERT Polska, https://cert.pl/lista-ostrzezen/

blokowania dostępu do treści, które uważane są za zagrożenie. Stanowi to przykład półautonomicznego systemu blokującego dostęp do potencjalnie niebezpiecznych stron internetowych.

DETEKCJA PHISHINGU

Detekcja phishingu jest procesem dosyć skomplikowanym, wielowarstwowym a również często wieloetapowym. Phishing może być wykrywany na podstawie informacji o podmiocie rejestrującym domenę, na podstawie analizy nazwy domenowej jak również na podstawie analizy statystycznej ruchu kierowanego do badanej domeny. Metody wykrywania są jednak oparte na heurystykach i dają jedynie informację o podejrzeniu, że mamy do czynienia z domeną phishingową. Ostateczna decyzja pozostaje w gestii człowieka, dlatego istotne jest aby mechanizmy wykrywania phishingu nie generowały nadmiernej ilość błędów fałszywie dodatniej detekcji.

Znane są pule adresów IP oraz domeny, które często wykorzystywane są do działalności przestępczej / phishingowej. Jeżeli rejestrowana jest nowa domena i adres IP z którego następuje rejestracja znajduje się w podejrzanej puli, domena podlega dalszej weryfikacji oraz obserwacji, często bywa też niemal natychmiast po rejestracji blokowana.

Nazwa domenowa może być jednym z głównych wskaźników, że mamy do czynienia z domeną phishingową. Jeżeli domena jest łudząco podobna do innej znanej domeny (odległość Levenshteina[7], czy też miara kosinusowa[8]) może to wskazywać, że mamy do czynienia z domeną phishingową. Podobnie wyraźnym sygnałem alarmowym jest używanie w nazwie domenowej znaków spoza ASCII, w szczególności znaków unicode o których wiadomo, że mają wygląd zbliżony bądź identyczny do znaków z puli ASCII. Podejrzane są również domeny zawierające losowe ciągi znaków nie mające odpowiedników w słownikach. Można zdefiniować wiele mechanizmów regułowych pozwalających wychwytywać domeny phishingowe, należy jednak pamiętać,

[7] Levenshtein, V. I., Levenshtein, & I., V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *SPhD*, *10*, 707. https://ui.adsabs.harvard.edu/abs/1966SPhD...10..707L/abstract

[8] Han, J., Kamber, M., & Pei, J. (2012). Getting to Know Your Data. *Data Mining*, 39–82. https://doi.org/10.1016/B978-0-12-381479-1.00002-2

że jest to swoisty wyścig zbrojeń pomiędzy przestępcami a instytucjami zajmującymi się walką z cyberprzestępczością. Tabela 1 przedstawia przykłady domen niebezpiecznych, które zostały stworzone przez przestępców w taki sposób, aby kojarzyć je ze znanymi i popularnymi stronami.

Popularne strony	**Domeny phishingowe**
Allegro.pl	olx.allegrolokolnie.pl; t-allegro.shop; www.allegrogo.shop; www.allegrro-outlets.shop; www.allegroonline.shop; allegrolokalnie.pl-5fe6r.cfd
Amazon.pl	www.amazoon.shop;
Olx.pl	o1x.06481479.xyz; o-lx.7032643.xyz
Interia.pl	www.help-interiamail.biz; www.helpmailinteria.biz; www.interia-pl-eu.icu

Tabela 1 Przykłady domen phishingowych, wykorzystywanych do ataków skierowanych na użytkowników popularnych portali w przestrzeni polskiego Internetu.

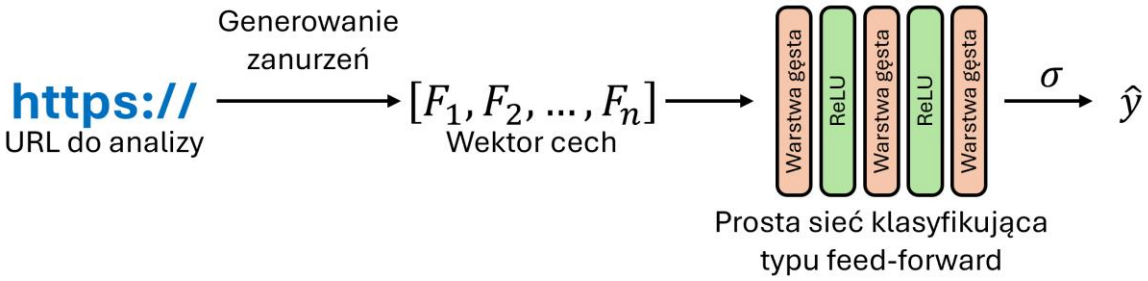
Niezależnie od metod regułowych, nazwy domenowe mogą być również analizowane za pomocą uczenia maszynowego. Przykładowo można zdefiniować szereg cech leksykalnych wynikających z nazwy domenowej (np. najdłuższy podciąg cyfr występujący w nazwie) i na podstawie tych cech zbudować klasyczny klasyfikator jak np. AdaBoost[9]. Konieczne jest w tym miejscu zaznaczenie, że podejścia z użyciem uczenia maszynowego muszą uwzględniać olbrzymią dysproporcję klas tj. ilości domen prawidłowych do domen phishingowych. W ruchu obserwowanym w Internet domeny phishingowe stanowią często ułamki promili. Z tego między innymi względu warto stosować klasyfikatory dobrze nadające się do analizy przypadków z istotnym niezbalansowaniem klas, czyli właśnie np. AdaBoost.

W ostatnich latach dużą popularnością cieszy się zastosowanie głębokich sieci neuronowych (ang. artificial intelligence, AI) do klasyfikacji. Sieci takie mogą być z powodzeniem stosowane do wykrywania phishingu

[9] Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/JCSS.1997.1504

Przykładowa konfiguracja takiego modelu przedstawiona jest na Źródło: Opracowanie własne.

Rys. *1*. Podobnie jak ma to miejsce w innych zastosowaniach sieci neuronowych do analizy danych tekstowych, pierwszym etapem jest uzyskanie tak zwanych zanurzeń[10] tj. wektorowej reprezentacji nazw domenowych w wysokowymiarowych przestrzeniach liczb rzeczywistych. W celu uzyskania zanurzeń można użyć *transfer learning*[11] tj. użycie fragmentów uprzednio przetrenowanych dużych głębokich sieci neuronowych takich jakie np. wykorzystuje się w modelach językowych. Dobór odpowiedniego modelu do generowania zanurzeń jest kluczowy dla samej architektury sieci klasyfikującej, gdyż modele te zwracają zanurzenia o różnym rozmiarze. W większości przypadków, przy założeniu dostatecznie dużej próbki danych treningowych, możliwe jest uzyskanie satysfakcjonującej dokładności klasyfikacji z wykorzystaniem dość prostej sieci typu feed-forward zbudowanej z trzech warst oraz funkcji sigmoidalnej. Sieć taka będzie zwracała wartości z przedziału $\hat{y} \in [0,1]$. Po odpowiedniej analizie modelu na zbiorze testowym możliwe jest następnie określenie progu decyzyjnego dla detekcji domen phishingowych.



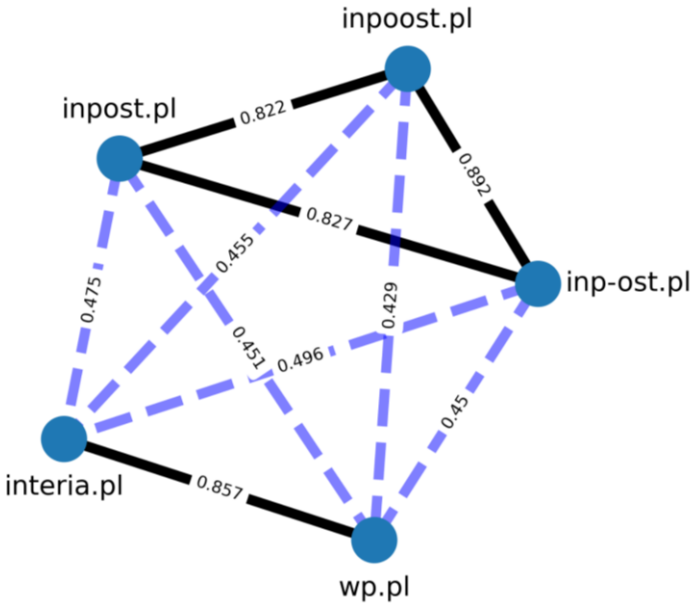
Źródło: Opracowanie własne.

Rys. 1 Przykład systemu do detekcji domen phsihingowych, który wykorzystuje metody przetwarzania języka naturalnego oraz prostą sieć neuronową typu feed-forward.

[10] Selva Birunda, S., & Kanniga Devi, R. (2021). A Review on Word Embedding Techniques for Text Classification. *Lecture Notes on Data Engineering and Communications Technologies*, *59*, 267–281. https://doi.org/10.1007/978-981-15-9651-3_23

[11] George Karimpanal, T., & Bouffanais, R. (2019). Self-organizing maps for storage and transfer of knowledge in reinforcement learning. *Adaptive Behavior*, *27*(2), 111–126. https://doi.org/10.1177/1059712318818568

Generowanie zanurzeń dla nazw domenowych ma dodatkowe zalety. Zanurzenia stanowią wektory wartości, które da się interpretować w przestrzeniach wielowymiarowych. O ile problematycznym jest zwizualizowanie takich przestrzeni, to istnieją miary pozwalające na mierzenie odległości (podobieństwa) pomiędzy dwoma wektorami. Przykładem jest stosowana często miara kosinusowa, definiowana jako znormalizowany iloczyn skalarny dwóch wektorów. We wspomnianej przestrzeni wielowymiarowej reprezentuje on kosinus kąta pomiędzy wektorami reprezentującymi nazwy domenowe. Czym wyższa wartość, tym bliższe sobie są dwie domeny, co ilustruje przykład widoczny na Źródło: Opracowanie własne przy wykorzystaniu biblioteki NetworkX. **Rys. *2***. Domeny bliskie domenie „inpost.pl" to te, które starają się ją naśladować i zmylić tym samym potencjalną ofiarę. Pozostałe, są od niej daleko, choć zastanawiającym jest ich własna bliskość.



Źródło: Opracowanie własne przy wykorzystaniu biblioteki NetworkX[12].

Rys. 2 Graf podobieństw różnych nazw domenowych z miarą kosinusową jako wagą krawędzi.

[12] Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, "Exploring network structure, dynamics, and function using NetworkX", in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008

Analiza podobieństwa bliskości nazw domenowych może okazać się pomocna w ręcznej analizie mającej na celu oszacowanie charakteru domeny, ale nie powinna być wykorzystywana do blokowania zautomatyzowanego.

STRATEGIE BLOKOWANIA TREŚCI PHISHINGOWYCH

Blokowanie stron phishingowych wykonywane jest na poziomie organizacji hostującej daną podrobioną stronę bądź na poziomie instytucji zarządzającej domeną używaną do popełniania przestępstwa. Pierwsza metoda jest trudniejsza w realizacji, ponieważ wymaga kontaktu z organizacją hostującą stronę a trzeba pamiętać, że owa organizacja może być w działalność podrobionej strony bezpośrednio zaangażowana. Podejście drugie jest znacznie szybsze i pewniejsze, ale wymaga współpracy w operatorem domeny będącej nadrzędną w stosunku do domeny phishingowej. Operator w takim przypadku dokonuje przejęcia przestępczej domeny i ruch do niej kierowany jest przekierowywany do strony informującej, że domena została zablokowana.

W obu przypadkach system może działać automatycznie lub półautomatycznie. System działający automatycznie decydowałby sam czy daną domenę należy zablokować. Jednak wiąże się to z potencjalnie dużym ryzykiem zablokowania bezpiecznej strony, co generowałoby straty zarówno wizerunkowe jak i finansowe jej właścicieli. Należałoby zatem zastosować system półautomatyczny, w którym modele i algorytmy jedynie wskazują potencjalnie niebezpieczne domeny. Dopiero analiza dokonana przez przeszkolonego analityka pozwalałaby na oszacowanie charakteru stronu i jej potencjalne zablokowanie. Rozwiązanie to jest dużo bezpieczniejsze zważywszy na potencjalne pomyłki w szacowaniu przez modele.

ZAKOŃCZENIE

Pomimo przeprowadzanych kampanii informacyjnych, jak również rozwoju algorytmów służących wykrywaniu ataków phishingowych, ataki te nadal stanowią realne zagrożenie dla wszystkich użytkowników Internetu. Cyberprzestępcy nieustannie udoskonalają swoje techniki mające na celu kradzież środków finansowych swoich ofiar. Wykrywanie i blokowanie kampanii phishingowych na wczesnym etapie ich trwania rodzi nadzieję na

ograniczenie liczby okradzionych osób. Obiecującym kierunkiem rozwoju systemów przeciwdziałania takim atakom jest wykorzystanie sztucznej inteligencji i technik przetwarzania języka naturalnego. Odpowiednie modele pozwoliłyby wykrywać nazwy domenowe wykorzystywane w kampaniach phishingowych, a co za tym idzie, ich skuteczna identyfikacja umożliwiłaby blokowanie ruchu do złośliwej strony internetowej już na poziome zapytań do serwera nazw domenowych bez konieczności dokonywania wnikliwej analizy całej kampanii.

Bibliografia

Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, "Exploring network structure, dynamics, and function using NetworkX", in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008

Fayi, S. Y. A. (2018). What Petya/NotPetya Ransomware Is and What Its Remidiations Are. *Advances in Intelligent Systems and Computing*, *738*, 93–100. https://doi.org/10.1007/978-3-319-77028-4_15/FIGURES/10

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, *55*(1), 119–139. https://doi.org/10.1006/JCSS.1997.1504

George Karimpanal, T., & Bouffanais, R. (2019). Self-organizing maps for storage and transfer of knowledge in reinforcement learning. *Adaptive Behavior*, *27*(2), 111–126. https://doi.org/10.1177/1059712318818568

Ghafur, S., Kristensen, S., Honeyford, K., Martin, G., Darzi, A., & Aylin, P. (2019). A retrospective impact analysis of the WannaCry cyberattack on the NHS. *Npj Digital Medicine 2019 2:1*, *2*(1), 1–7. https://doi.org/10.1038/s41746-019-0161-6

Han, J., Kamber, M., & Pei, J. (2012). Getting to Know Your Data. *Data Mining*, 39–82. https://doi.org/10.1016/B978-0-12-381479-1.00002-2

Lista Ostrzeżeń CERT Polska, https://cert.pl/lista-ostrzezen/ [dostęp: 13.05.2025]

Levenshtein, V. I., Levenshtein, & I., V. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *SPhD*, *10*, 707.

NordVPN Link Checker, https://nordvpn.com/pl/link-checker/ [dostęp: 13.05.2025]

Selva Birunda, S., & Kanniga Devi, R. (2021). A Review on Word Embedding Techniques for Text Classification. *Lecture Notes on Data Engineering and Communications Technologies*, *59*, 267–281. https://doi.org/10.1007/978-981-15-9651-3_23

URLVoid, https://www.urlvoid.com/ [dostęp: 13.05.2025]

VirusTotal, https://www.virustotal.com/gui/home/url [dostęp: 13.05.2025]

EARLY DETECTION OF PHISHING ATTACKS IN CAMPAIGNS

Abstract

Internet users are exposed to a variety of cyber threats on a daily basis; phishing attacks are of particular concern in this regard. The objective of these actors is to procure confidential information, most frequently comprising bank account login credentials, by impersonating trusted institutions and creating websites that are deceptively similar to payment portals. In the absence of adequate safeguards, the consequences of such an attack can be very severe for the victim and even involve the loss of all funds in the account. Despite the existence of a number of methods to protect against this type of attack, attackers are using new techniques of operation, which makes effective counteraction difficult. In this article, the mechanisms of phishing attacks are introduced, and potential methods of detecting them early, especially in the early stages of the campaign, are discussed. The utilisation of artificial intelligence is of particular significance in the identification of potentially hazardous Internet domains. This process can be initiated at the stage of DNS (Domain Name Server) queries. Despite the fact that these systems have hitherto yielded promising results, it is important to note that they are not infallible. Consequently, it is imperative that their decisions are reviewed by qualified analysts. It is important to note that the automatic blocking of secure websites can pose significant risks to their proprietors, both in terms of their reputation and financial well-being.

Keywords: phishing, attack detection, security, phishing scams.