WOJSKOWA AKADEMIA TECHNICZNA
im. Jarosława Dąbrowskiego

ROZPRAWA DOKTORSKA

Budowanie mechanizmu obrony przed dedykowanymi kampaniami phishingowymi

AUTOR:

mgr inż. Mariusz SZAREK

PROMOTOR:

dr hab. inż. Ryszard ANTKIEWICZ, profesor WAT

Warszawa 2023

Niniejszą rozprawę dedykuję moim dzieciom,
za czas, którego z nimi nie spędziłem.

Dziękuję mojemu opiekunowi akademickiemu za cenne rady, wskazówki i poświęcony mi czas na zrozumienie zawiłych procesów.

Streszczenie

Phishing jest jednym z najczęstszych i najgroźniejszych ataków w cyberprzestrzeni – co zostało w niniejszej rozprawie ukazane na bazie polskich i międzynarodowych statystyk zespołów bezpieczeństwa komputerowego. Dokładne zrozumienie jak działa phishing, jakie są jego elementy składowe, z ilu faz ataku składa się cały proces pozwoli na określenie i wskazanie tych elementów, które skutecznie pozwolą na jego identyfikacje. Celem niniejszej rozprawy było dokładna analiza ataku phishingowego (zawarta w rozdziale I), porównanie dotychczas wykorzystywanych metod detekcji (Rozdział II). Przedstawione w Rozdziale I techniki i metody stosowane przez atakujących posłużyły do wykazania w Rozdziale III wskaźników (będących autorską propozycją), jakimi charakteryzują się wiadomości email o złośliwym charakterze. Rozdział IV zawiera propozycję metody pozwalającej na odczytanie z wiadomości opisanych wskaźników, właściwe ich zakodowanie (w tym z wykorzystaniem uczenia maszynowego), a następnie za pomocą modułu ML określenie czy dana wiadomość może być atakiem phishingowym. Przeprowadzone badania wykazały, że metody uczenia maszynowego mogą rozpoznawać atak phishingowy, bazujący również na nieznanym wcześniej wzorcu.

Abstract

Phishing is one of the most common and most dangerous attacks in cyberspace – which has been shown in this dissertation based on Polish and international statistics of computer security teams. A thorough understanding of how phishing works, what are its components, and how many attacks phases the entire process consists of will allow you to identify and indicate those elements that will effectively identify it. The purpose of this dissertation was a thorough analysis of a phishing attack (included in Chapter I), and a comparison of the detection methods used so far (Chapter II). The techniques and methods used by the attackers presented in Chapter I were used to show in Chapter III the indicators (which are the author's proposal) that characterize malicious emails. Chapter IV contains a proposal of a method that allows reading the described indicators from the message, their proper coding (including using machine learning), and then using the ML module to determine whether a given message can be a phishing attack. The conducted research showed that machine learning methods can recognize a phishing attack, also based on a previously unknown pattern.

Spis treści

Wykaz użytych skrótów / słownik pojęć

CERT	(ang. Computer Emergency Response Team) – organizacja lub zespół przeznaczana do całodobowego monitorowania ruchu internetowego i reagowania w przypadku pojawienia się zagrożenia atakiem cybernetycznym.
CSIRT	(ang. Computer Security Incident Response Team) – Zespół Reagowania na Incydenty Bezpieczeństwa Komputerowego, przeznaczony do reakcji na pojawiający się incydent bezpieczeństwa: obsługa incydenty, zabezpieczenie danych, identyfikacja źródła, np.
BLUE Team	jeden z zespołów bezpieczeństwa, którego zadaniem jest przeprowadzenie analizy systemów informatycznych w celu zapewnienia bezpieczeństwa, identyfikacji luk bezpieczeństwa, weryfikacji skuteczności każdego środka bezpieczeństwa oraz upewnienia się, że wszystkie środki bezpieczeństwa będą nadal skuteczne po ich wdrożeniu. Jednym z zadań zespołu jest ciągły monitoring infrastruktury pod kątem wykrycia obecności ewentualnego intruza. Zespół BLUE wchodzi w skład CSIRT.
Botnet	grupa komputerów zainfekowana złośliwym oprogramowaniem będąca zdalnie kontrolowana i wykonująca przesyłane polecenia, tworząca ukrytą przez użytkownikiem sieć.
DdoS	(ang. Distributed Denial of Service – rozproszona odmowa usługi) – atak na usługę internetową w celu jej przeciążenia wykonywany jednocześnie z wielu komputerów.
Dos	(ang. Denial of Service – odmowa usługi) – rodzaj ataku na usługę internetową w celu jej przeciążenia.
EDR	(ang. Endpoint Detection and Response) – oprogramowanie lub system informatyczny służący do analizowania, monitorowania oraz przechowywania informacji o działaniu systemu oraz procesów na urządzeniu końcowym (zwykle stacji roboczej użytkownika), potrafiący przesyłać dane do rozwiązań klasy SIEM.
HTTP	(ang. Hypertext Transfer Protocol) – protokół komunikacji między klientem a serwerem w sieci WWW.
HTTPS	protokół HTTP chroniony przy pomocy szyfrowania protokołu TLS i ustandaryzowany w dokumencie RFC 2818.
IoC	(ang. Indicator of compromise, pol. Wskaźniki kompromitacji). Wyznaczniki występowania w danym systemie teleinformatycznym określonego zagrożenia (np. komunikacja z adresem IP znajdującym się na liście złośliwych serwerów C2).
Lotl	(ang. Living off the Land). Technika ataku wykorzystująca dostępne, wbudowane funkcje systemów operacyjnych, narzędzia systemowe do przejęcia kontroli nad tym systemem lub dostarczenia złośliwego oprogramowania do danego systemu. Wykorzystanie legalnych komponentów danego

	systemu teleinformatycznego, oprogramowania systemu operacyjnego nie prowadzi do uruchomienia alarmu wykorzystywanego w danym środowisku rozwiązania bezpieczeństwa (IDS, IPS, EDR, np.).
PoC	(ang. Proof of Concept). Określenie implementacji jakiegoś rozwiązania, mającego udowodnić słuszność założeń lub zobrazowanie, zwizualizowanie danej koncepcji teoretycznej.
RAT	(ang. Remote Access Trojan). Oprogramowanie umożliwiające umożliwienie tajnego nadzoru lub uzyskanie nieautoryzowanego dostępu do systemu ofiary. Inna funkcjonująca nazwa, to oprogramowanie typu konie trojańskie.
SIEM	(ang. Security Information and Event Management) – oprogramowanie lub system informatyczny wspierający bezpieczeństwo, umożliwiający analizę zdarzeń w czasie zbliżonym do rzeczywistego, detekcję anomalii i zdarzeń o charakterze incydentu.
Spear phishing	to wysoce spersonalizowane ataki wymierzone w konkretne, starannie wyselekcjonowane osoby, firmy lub organizacje.
Whaling	Atak typu spear phishing ukierunkowany na kadrę kierowniczą wysokiego szczebla, CEO, VIP, np.

Wstęp

Ewolucja Internetu od sieci mającej wspomagać proces wymiany naukowej pomiędzy placówkami naukowymi wyposażonymi w komputery do ogólnoświatowej sieci usług, rozrywki, zakupów i bazy wiedzy doprowadziła do szerokich zmian społecznych. Sieć Internet wspomaga i przyspiesza proces wymiany informacji, umożliwia szeroki dostęp do rozrywki, umożliwia dystrybucję dóbr (e-commerce, e-biznes), ułatwia komunikację na duże odległości przy zapewnieniu znacznie większego komfortu i możliwości niż dotychczasowe media. Wynalazek Internetu doprowadził do przemian społecznych, w wyniku których powstało społeczeństwo informacyjne[1], charakteryzujące się traktowaniem informacji jako towaru (surowca) na sprzedaż [1], masowym jej wytarzaniem, przetwarzaniem i dystrybucją. Dominującą rolą w tak zdefiniowanym społeczeństwie stała się informacja, konieczność zapewnienia odpowiedniej infrastruktury technicznej (szybkiego i szeroko dostępnego medium transmisyjnego, serwerów bazodanowych, urządzeń sieciowych, np.) do jej wytarzania, magazynowania i przesyłania, jest istotną rolą gospodarki, często już określaną jako „internetowa gospodarka" [2], które to pojęcie jest znacznie szersze i nie ogranicza się tylko do dziedziny technicznej. E-gospodarka to szerokie wykorzystanie technologii informatycznych w całym procesie produkcji (zarówno tradycyjnych produktów jak i informacji), sprzedaży i dystrybucji, udostępnianie wielu narzędzi analitycznych do wsparcia biznesu i jego prowadzenia. Rozwijanie się społeczeństwa informacyjnego jest ściśle powiązane z rozwojem e-biznesu i jego coraz większym oddziaływaniem zarówno na narodową jak i globalną gospodarkę [3].

Wraz z przemianami społecznymi i rozwojem technologicznym, rozwijają się również nieznane wcześniej zagrożenia, w równym stopniu wykorzystujące nowoczesne technologie jak i dostosowane do internetowej rzeczywistości znane wcześniej techniki. Technologia informacyjna i powszechny do niej dostęp, zwiększyła zakres oddziaływania na całe społeczności, dotykając wiele aspektów codziennego życia, przenosząc je w strefę nowoczesnych technologii.

[1] Społeczeństwo informacyjne – nazwa wprowadzona przez Tadao Umesamo w 1963 roku w artykule "O teorii ewolucji społeczeństwa opartego na technologiach informatycznych''. Termin ten ta odnosi się do społeczeństwa, w którym rozpowszechnione i łatwo dostępne są technologie informatyczne służące do wytwarzania, przetwarzania, przesyłania i przechowywania informacji, która traktowana jest jako wartościowy towar.

Jednym ze szczególnie istotnych zagrożeń, związanych z rozwojem technologii informacyjnych, jest phishing. Ten rodzaj oszustwa stał się w ostatnich latach niezwykle popularny z uwagi na jego prostotę, stając się głównym zagrożeniem w sieci Internet, powodując duże straty finansowe zarówno wśród zwykłych użytkowników jak i wśród firm i instytucji – wg raportu FBI tylko w Stanach Zjednoczonych w 2020 roku straty spowodowane atakiem phishingowym wyniosły 54 mln dolarów amerykańskich. Atak phishingowy jest również często nośnikiem złośliwego oprogramowania i jest chętnie wykorzystywany przez grupy przestępcze do szantażu, wymuszania okupu czy ataku ransomware.

Wzrost popularności phishingu i wysokość start nim spowodowanych, wymusiło wśród specjalistów z branży cyberbezpieczeństwa rozpoczęcie w wielu ośrodkach naukowych prac badawczych nad istotą problemu, mając przynieść skuteczne środki wykrywające phishing. Dotychczas wykorzystywanymi w walce z phishingiem podejściami było:

1. Tworzenie list dostępowych, na podstawie których filtrowany jest ruch danego segmentu sieci, poczta email.
2. Budowanie reguł detekcji opartych na cechach, które zidentyfikowano w już zrealizowanych i wykrytych atakach.
3. Budowanie systemów eksperckich zasilających w wiedzę i umiejętności zespoły odpowiedzialne za bezpieczeństwo sieci komputerowych w firmach i instytucjach.
4. Budowanie świadomości użytkowników sieci Internet odnośnie mogących w niej wystąpić zagrożeniach.

Phishing jest zjawiskiem zmieniającym się w czasie – atakujący dostosowują swoje metody i zmieniają techniki implementacji, dostosowują je do budowanych mechanizmów bezpieczeństwa, omijając je i w ten sposób kontynuując skuteczne przeprowadzanie ataków. Przedstawione powyżej metody wykrywania ataków phishingowych nie zapewniają jednak pełnej skuteczności przed tego rodzaju atakiem, stąd konieczność zaprojektowania nowej, skuteczniejszej metody.

W niniejszej rozprawie opisane zostanie zjawisko ataku phishingowego – jako problem badawczy – wraz ze wszystkimi jego składowymi, omówione zostaną rodzaje ataków, wykorzystywane technologie oraz metody jakimi posługując się atakujący w celu

realizacji ataku. Analizując obecne trendy i metody wykorzystywane w atakach, jak i mając świadomość wielkości strat ponoszonych przez ofiary ataków, istnieje potrzeba opracowania skutecznych metod wykrywania i zapobiegania atakom phishingowym. Ukazane w niniejszej rozprawie metody zwalczania, nie są w pełni skuteczne, wymagają nadkładu administracyjnego (co zostanie ukazane w dalszej części pracy) lub bazują na wskaźnikach obecnie nie występujących. Przedmiotowa rozprawa stanowi nowe podejście do wykrywania i przeciwdziałania atakom phishingowym:, poprzez identyfikację większej ilości cech mogących wskazywać na atak phishingowy (nowe podejście) w połączeniu zaangażowanie uczenia maszynowego (udoskonalone podejście) do analizy wektora cech uzyskanych na podstawie wartości przekazanych w nagłówku wiadomości phishingowej.

Rozdział I zawiera opis zjawiska phishingu. Przedstawiony został jego rozwój na przestrzeni ostatnich lat zarówno w ujęciu międzynarodowym jak i polskim, ukazany został negatywny wpływ ataku phishingowego. Przedstawione zostały ogólne modele służące do identyfikacji ataków w cyberprzestrzeni, jak i powstałe na ich bazie modele służące identyfikacji phishingu (jego kolejnych faz). Opisane zostały poszczególne typy ataków phishingowych.

W rozdziale II przedstawiono obecnie opisane w literaturze i wykorzystywane metody wykrywania ataku phishingowego. Przedstawione zostały zalety oraz słabości poszczególnych metod, jak również wykorzystywane przez nie techniki.

Rozdział III poświęcony jest zidentyfikowanym wskaźnikom ataku phishingowego. Opisywane wskaźniki są autorską propozycją powstałą w wyniku analizy wiadomości o charakterze phishingowym. Wskaźniki zostały podzielone na dwie grupy: wskaźniki o charakterze technicznym (mogącymi być automatycznie wykrywane przez oprogramowanie) oraz nietechniczne (konieczna wiedza ekspercka).

W ostatnim IV Rozdziale przedstawiony został koncept (PoC) nowatorskiej metody detekcji phishingu w oparciu o zidentyfikowane wskaźniki. Metoda detekcji wykonana została w oparciu o moduły uczenia maszynowego. Rozdział zawiera opis pozyskania i przygotowania danych w oparciu o algorytm analizy wiadomości email. Przedstawione zostały etapy wykonywanych eksperymentów z przetwarzaniem cech (będącymi wskaźnikami możliwego ataku phishingowego) przez moduły uczenia maszynowego.

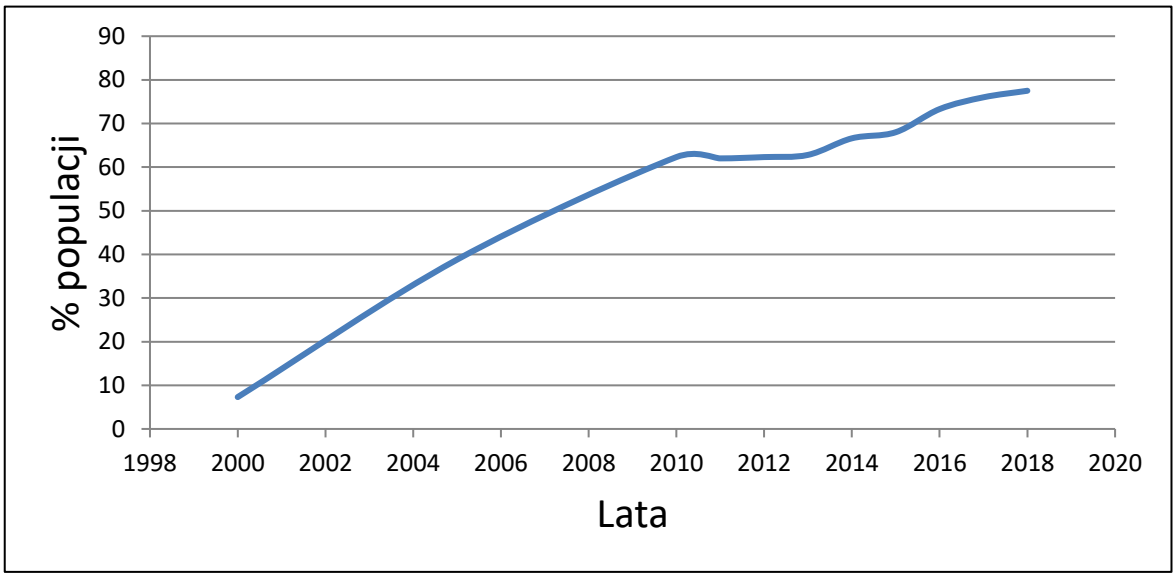
Rozdział I – Zjawisko phishingu i jego wpływ na użytkowników

Połączenie niezależnych systemów informatycznych za pomocą sieci komputerowych, dały początek rozwojowi przestrzeni, w której użytkownicy gromadzą i przetwarzają duże ilości informacji, wytworzyło wirtualne środowisko – cyberprzestrzeń[2]. Cyberprzestrzeń zapewnia zarówno interakcję użytkowników z systemami teleinformatycznymi jak i dostęp do gromadzonych i przetwarzanych informacji. Cyberprzestrzeń jest rozległym środowiskiem łączącym zasoby sprzętowe (hardware), informacyjne, oprogramowanie (software) jak i zasoby ludzkie, tworząc spójną przestrzeń społeczną, umożliwiającą interakcje pomiędzy jej użytkownikami, za pomocą sieci komputerowych. Łatwość gromadzenia i przetwarzania informacji (w tym informacji wrażliwych[3]), możliwość jej łatwego udostępniania uprawnionym do jej posiadania użytkownikom, wymaga procesu ciągłej ochrony – zarówno dostępu przed nieuprawnionymi użytkownika, jak i możliwości manipulacji i modyfikacji informacjami. Cyberbezpieczeństwo to proces zapewnienia ochrony w cyberprzestrzeni i jest szerokim pojęciem obejmującym swym zakresem zarówno bezpieczeństwo systemów teleinformatycznych (dostępu, zapewnienia ciągłości działania) jak i zagadnienia bezpieczeństwa informacji przetwarzanych w tych systemach.

[2] Termin „cyberprzestrzeń" po raz pierwszy został użyty przez amerykańskiego pisarza science fiction w 1984 roku do określenia świata wygenerowanego przez komputer (rzeczywistość wirtualna).

[3] Za informacje wrażliwe uważa się te informacje o osobie, które poddane są szczególnej ochronie (źródło: Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2016/679 z dnia 27 kwietnia 2016 r. w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych oraz uchylenia dyrektywy 95/46/WE (ogólne rozporządzenie o ochronie danych) pełna definicja i treść rozporządzenia jest dostępna pod adresem: https://eur-lex.europa.eu/legal-content/PL/TXT/HTML/?uri=CELEX:32016R0679&from=PL [dostęp 2022-01-10].

Polska jako kraj rozwinięty[4] posiada dobrze rozwiniętą infrastrukturę teleinformatyczną, umożliwiającą wymianę informacji i dostęp do sieci Internet. Zgodnie

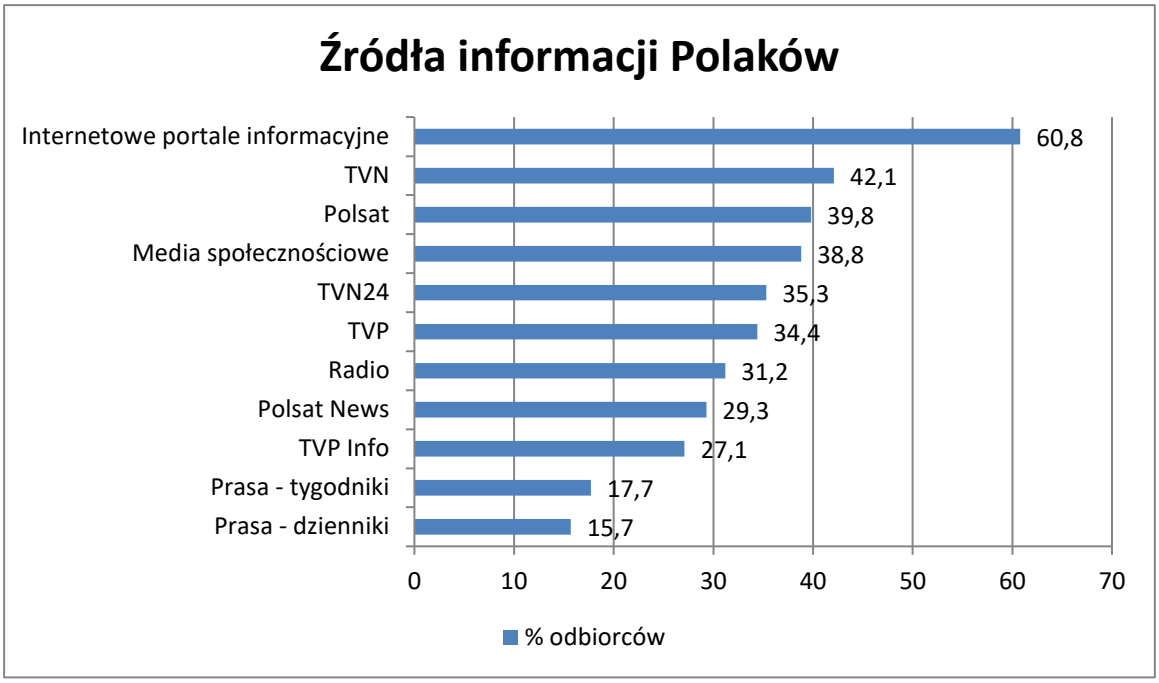


Rysunek 1. Wzrost populacji osób z dostępem do sieci Internet w Polsce w latach 2000-2018, źródło: https://hdr.undp.org/en/indicators/43606# [dostęp: 2021.01.10]

z przedstawionym raportem przez Human Developmen Reports [4] w 2016 roku z sieci Internet korzystało w Polsce 73.3% populacji. Światowy trend zmian społecznych i dążenie Polski do jak największego ucyfrowienia usług, biznesu i spraw urzędowych poprzez wdrażanie Programu Operacyjnego „Cyfrowa Polska" [5] wraz z coraz szerszą możliwością łatwego i taniego dostępu do sieci komputerowych (szerokopasmowego dostępu do Internetu) przez większość populacji kraju, przyczyniła się do powstania w Polsce społeczeństwa informacyjnego [6].

Trend przemiany społeczeństwa polskiego w społeczeństwo informacyjne, wykorzystujące sieć Internet jako główne źródło informacji (portale informacyjne takie jak np.: Onet, Wirtualna Polska) wykazują badania [7] prowadzone przez Instytut Badań Internetu i Mediów Społecznościowych (IBIMS) oraz Instytut Badań Rynkowych i Społecznych (IBRIS).

[4] Według United Nations Development Programme, Polska została sklasyfikowana na 33 pozycji krajów wysoko rozwiniętych (źródło: Latest Human Development Index (HDI) Ranking, https://www.hdr.undp.org/en/2018-update [dostęp 2019-10-11])



Rysunek 2 Źródła informacji o Polsce i świcie w świetle badań IBIMS oraz IBRIS, źródło: https://ibims.pl/skad-polacy-czerpia-informacje-o-polsce-i-swiecie-raport-ibims-i-ibris/ [dostęp: 2021.01.26].

Trend wzrostowy liczby osób z dostępem do sieci (patrz: Rysunek 1) oraz wybór internetowych portali informacyjnych jako głównego źródła wiedzy Polaków o najważniejszych wydarzeniach w kraju i na świecie, ukazują stale powiększający się zasób informacyjny, wzrost liczby osób korzystających z sieci Internet. Badania IBIMS oraz IBRIS wskazują również rosnącą rolę portali społecznościowych (np. Twitter[5], Facebook) jako źródła wiedzy i informacji (Rysunek 2). Wzrost zasobu informacyjnego wymusza jednocześnie rozwój infrastruktury teleinformatycznej: rozbudowa sieci telekomunikacyjnych, w tym sieci zapewniających dostęp mobilny, rozbudowa i tworzenie nowych centrów danych przechowywujących informacje oraz oprogramowania służącego do tworzenia, przesyłania, wyszukiwania i obróbki informacji. Rozrasta się również zasób możliwych sposobów komunikacji pomiędzy użytkownikami sieci Internet oraz wolumen przysłanych danych – wg badania [8] każdego dnia przesyłanych jest 3.4 miliarda wiadomości email.

[5] Od dnia 23.07.2023r. portal nosi nazwę „X”, źródło: https://twitter.com/elonmusk/status/1683171310388535296?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed%7Ctwterm%5E1683171310388535296%7Ctwgr%5E0e8da25f66abb7f5ed05ff30839691ca99af2644%7Ctwcon%5Es1_&ref_url=https%3A%2F%2Fpulsembed.eu%2Fp2em%2FJvuuXxdyy%2F [dostęp: 02.09.2023r.]

Wraz z rozwojem sieci Internet i dostępnych usług oraz zwiększającym się ciągle zapotrzebowaniem na usługi cyfrowe, rośnie również skala zagrożeń dla użytkowników związana z utratą prywatności, złośliwym oprogramowaniem, oszustwami z wykorzystaniem technologii informatycznych, defraudacjami finansowymi z wykorzystaniem bankowości on-line, np. Rozpowszechnienie się dostępu do sieci komputerowych (w tym z dostępem mobilnym), łatwość nawiązywania kontaktów z wykorzystaniem sieci społecznościowych (z zachowaniem pewnej dozy anonimowości), dostęp do szerokiego spectrum narzędzi i technologii, w tym otwarto-źródłowego[6] oprogramowania, doprowadziło w dobie społeczeństwa informacyjnego, do wytworzenia się w cyberprzestrzeni szeregu zagrożeń, bezpośrednio związanych z wykorzystaniem samej technologii informatycznej jak również sytuacji, kiedy technologie informatyczne są jedynie wspierającymi. Jednym z technicznym aspektów zagrożeń występujących w cyberprzestrzeni jest zjawisko zwanego atakiem cybernetycznym[7] (funkcjonującej w przestrzeni medialnej pod nazwą „atak hakerski"[8]).

Atak jest procesem, w którym wykorzystywane są zarówno technologie informatyczne jak i inżynieria społeczna. Tradycyjne podejście do ochrony systemów teleinformatycznych, do zgromadzonych i przetwarzanych w nich danych, koncentruje się na technologicznych środkach ochrony (oprogramowanie dostępowe, oprogramowanie antywirusowe, urządzenia bezpieczeństwa – np. firewall), w dobie łatwego transferu technologii i wiedzy pomiędzy grupami użytkowników, łączenia się różnych dziedzin, podejście niwelowania pojedynczego ryzyka okazuje się niewystarczające.

Dostrzegając rosnący problem zagrożeń bezpieczeństwa teleinformatycznego, firmy analityczne, think-tanki oraz powstające zespoły cyberbezpieczeństwa (CSIRT[9]) na czele z Państwowym Instytutem Badawczym NASK, dokonują analizy stanu

[6] Open source (ang. otwarte źródło) – rodzaj oprogramowania, w którym jego kod źródłowy jest udostępniony dla użytkowników na podstawie licencji w której twórca umożliwia modyfikację oprogramowania.

[7] Atak cybernetyczny – ogół czynności i działań mających wyrządzić szkodę ofierze z wykorzystaniem technologii informatycznych (sieci komputerowych i GSM, systemów komputerowych).

[8] Funkcjonujące określenie „atak hakerski" w przestrzeni medialnej odnosi się zarówno do działań przestępców wykorzystujących systemy teleinformatyczne do popełniania przestępstw, działania grup aktywistycznych, działania grup wpieranych przez państwa (APT) jak i propagowaniu się złośliwego oprogramowania.

[9] CSIRT (ang. Computer Security Incident Response Team) – organizacja lub zespół przeznaczony do wykrywania I obsługi incydentów komputerowych

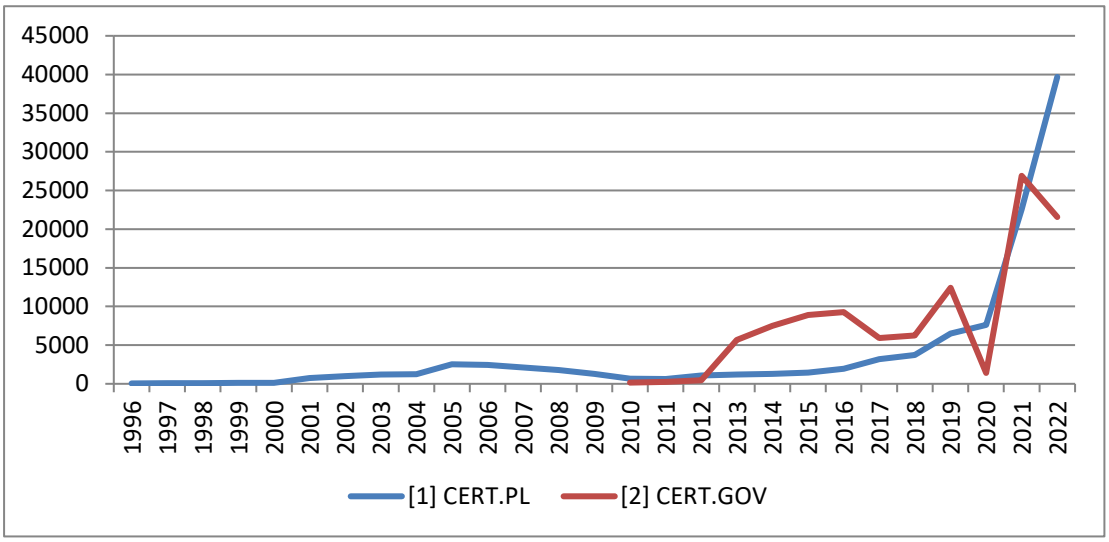
zabezpieczeń wśród firm, przedsiębiorstw i instytucji przed możliwości nieautoryzowanego dostępu do infrastruktury, danych oraz pozyskania poufnych informacji. Z dostępnych informacji wynika, że polskie przedsiębiorstwa nie przykładają należytej wagi do stanu zabezpieczeń swoich sieci i systemów teleinformatycznych. Wg raportu [9] obejmującego 2017 rok, 20% badanych firm nie posiada żadnych zasobów ludzkich z dziedziny cyberbezpieczeństwa, 46% firm nie posiada procedur postępowania w przypadku wystąpienia incydentu bezpieczeństwa teleinformatycznego i jedynie średnio 3% budżetu przeznaczonego na IT wykorzystywane jest do wsparcia bezpieczeństwa. Takie podejście przyczynia się do wysokiej skuteczności prowadzonych ataków – 44% badanych firm poniosło straty finansowe w skutek ataku, 62% spółek odnotowało zakłócenia i przestoje funkcjonowania, a 21% padło ofiarą złośliwego oprogramowania (ransomware).

Pojawienie się wirusów komputerowych i upowszechnienie się ataków cybernetycznych doprowadziło do powstanie w wielu krajach zespołów CERT a w niektórych organizacjach zawiązano zespoły CSIRT, które skupiają się bardziej na identyfikacji i obsłudze incydentów komputerowych. Dołączenie się Polski do globalnej sieci Internet i pojawiające się incydenty bezpieczeństwa doprowadziły do powstania w 1996 polskiego zespołu CERT[10], a w wyniku wdrożenia Krajowego Systemu Cyberbezpieczeństwa[11] w polskiej przestrzeni Internetu działają obecnie trzy zespoły CSIRT na poziomie krajowym:

a) CSIRT NASK (CERT Polska) – w zakresie sektora finansów publicznych, jednostek samorządu terytorialnego, pozarządowych organizacji,
b) CSIRT GOV – w zakresie odpowiedzialności jednostek publicznych, podległych Prezesowi Rady Ministrów, państwowych podmiotów finansowych,
c) CSIRT MON – w zakresie podmiotów podległych pod Ministra Obrony Narodowej oraz przedsiębiorstw o strategicznym znaczeniu dla obronności i gospodarki państwa.

[10] https://pl.wikipedia.org/wiki/CERT_Polska
[11] Krajowy System Cyberbezpieczeństwa powołany został ustawą z dnia 5 lipca 2018r. (Dz.U. 2018 poz. 1560, źródło: http://prawo.sejm.gov.pl/isap.nsf/download.xsp/WDU20180001560/U/D20181560Lj.pdf)

Jednym z zadań stojących przed poszczególnymi zespołami CERT/CSIRT jest publikowanie danych statystycznych odnośnie stanu cyberbezpieczeństwa w ich rejonie odpowiedzialności. Analizując dostępne raporty o liczbie zarejestrowanych incydentów, wyraźnie widoczny jest trend wzrostu (Rysunek 3) liczby obsługiwanych incydentów bezpieczeństwa teleinformatycznego przez zespół CERT Polska[12] (NASK – rejon odpowiedzialności sektora finansów publicznych, operatorów usług kluczowych, dostawców usług, firm oraz instytucji[13]) oraz CERT.GOV[14] (ABW – rejon



Rysunek 3. Ilość obsługiwanych incydentów przez [1] zespół CERT Polska, źródło: https://www.nask.pl/pl/raporty/raporty, [2] CERT.GOV, źródło: https://csirt.gov.pl/cer/publikacje/raporty-o-stanie-bezpi

odpowiedzialności administracji rządowej, publicznej, terytorialnej, NBP, BGK, operatorzy infrastruktury krytycznej[15]).[16] Nakładając na to dane o wzroście populacji osób korzystających z zasobów sieci Internet (Rysunek 1) widoczna jest korelacja wzrostu ilości odnotowywanych incydentów (we wszystkich sektorach) wraz ze wzrostem liczny osób korzystających z sieci.

[12] https://www.nask.pl/pl/raporty/raporty
[13] Dz. U. 2018 poz. 1560, rozdział 6, art. 26, pkt. 6.
[14] https://csirt.gov.pl/cer/publikacje/raporty-o-stanie-bezpi
[15] Dz. U. 2018 poz. 1560, rozdział 6, art. 26, pkt. 7.
[16] W analizowanym okresie CERT MON (lata: 1996-2022 nie opublikował publicznie dostępnych raportów z ilości obsługiwanych incydentów bezpieczeństwa teleinformatycznego.

I.1 Zagrożenia w cyberprzestrzeni

Ataki na użytkownika mogą przybrać różne formy i mogą wykorzystywać różne technologie i metody. Wybór metody ataku zależą od:

a) posiadanej przez grupy atakujących wiedzy – zarówno wiedzy technicznej jak i wiedzy o obiekcie/celu ataku. Posiadanie wiedzy o celu ataku znacznie zwiększa prawdopodobieństwo jego sukcesu, z uwagi na możliwość dostosowania poszczególnych elementów ataku do indywidualnych cech potencjalnej ofiary.
b) posiadanych przez grupę możliwości technicznych – przygotowanie odpowiedniej infrastruktury teleinformatycznej (niezbędnej do niektórych typów ataków),
c) nakładów finansowych potrzebnych do skutecznego przeprowadzenia ataku – nakłady ponoszone na przygotowanie niezbędnej infrastruktury (np. rejestracja domen, wykup łącza o odpowiedniej przepustowości, np.), pozyskanie wiedzy, np.
d) stopnia skomplikowania ataku – skomplikowane ataki wymagają poniesienie większego nakładu finansowego oraz posiadania szerokiej wiedzy (zarówno technicznej jak i o obiekcie ataku), stąd też ataki tego typu przeprowadza się przeciwko wyselekcjonowanym celom (warunek wiedzy o celu ataku), gdzie powodzenie samego ataku przyniesie znaczącą korzyść atakującemu.

Do najczęściej identyfikowanych [10], [11], [12] i wykorzystywanych metod ataku na użytkowników sieci komputerowych należą:

a) phishing, socjotechnika[17],

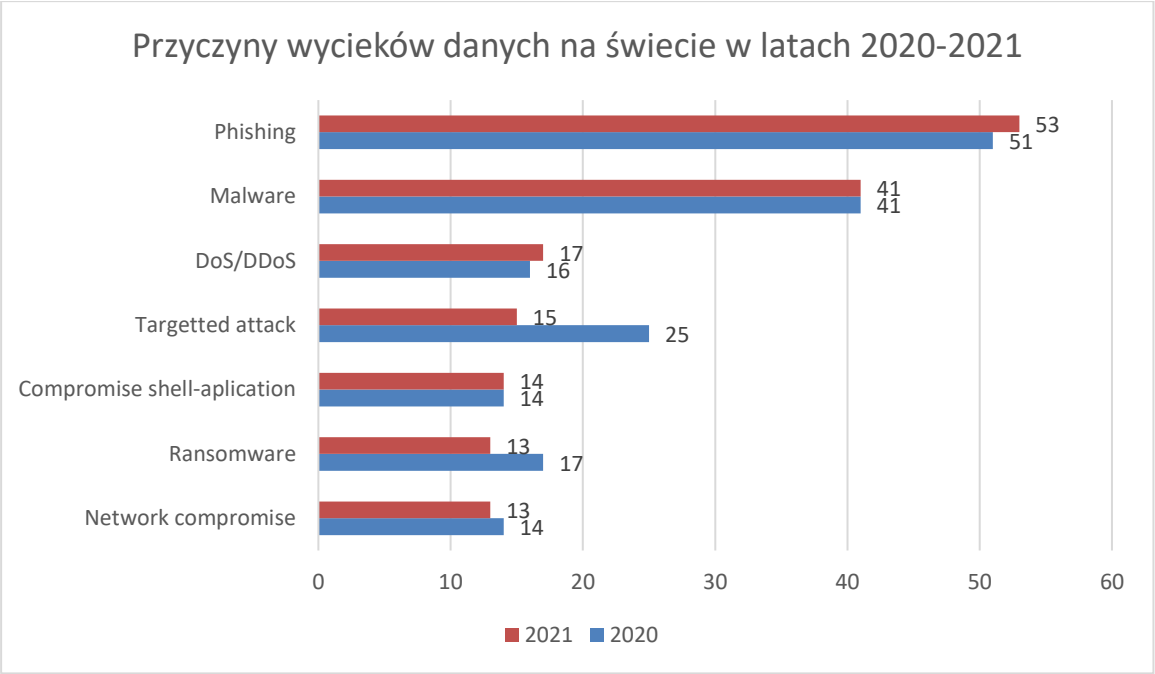
[17] Socjotechnika (ang. social engineering, inżynieria społeczna) – zespół technik pozwalający osiągnąć założony cel poprzez wykorzystanie manipulacji. Jest celowe i przemyślane działanie.

b) cyberbulling[18] i sextorion[19], obraźliwe i nielegalne treści, fake newsy[20],
c) złośliwe oprogramowanie (malware[21]), najpopularniejsze grupy:
    1. programy szyfrujące dane (ransomware[22]),
    2. oprogramowanie do przechwytywania haseł (keyloger, mimikatz),
    3. oprogramowanie szpiegujące użytkownika (spyware),
    4. botnety,
    5. wirusy
d) Dos, DdoS,
e) defraudacje bankowe (wyłudzenia kredytów, kradzież pieniędzy na dane wykradzione w wyniku wykorzystania innych metod),
f) utrata prywatności (kradzież tożsamości, przejęcia kont email, mediów społecznościowych),

Prowadzone statystyki zespołów międzynarodowych odnośnie ataków wykazują, że jednym z najczęściej występujących zagrożeń jest atak phishingowy. Trend ten można zaobserwować w publikowanych raportach przez międzynarodowe zespoły cyberbezpieczeństwa.

Phishing jest to pewnego rodzaju oszustwo [13]**, w którym za pomocą metod inżynierii społecznej, atakujący podszywa się pod inną osobę lub instytucję w celu wyłudzenia poufnych informacji, zainfekowania komputera złośliwym oprogramowaniem czy też nakłonienia ofiary do podjęcia lub zaniechania określonych działań.**

[18]Cyberbulling -rodzaj przemocy (np. prześladowanie, zastraszanie, nękanie, wyśmiewanie) z użyciem nowoczesnych technologii, głównie Internetu oraz telefonów komórkowych (SMS, e-mail, witryny internetowe, fora dyskusyjne w Internecie, portale społecznościowe).
[19] Sextorion – rodzaj zemsty lub szantażu, który wymusza od ofiary zapłatę lub wykonanie określonych czynności pod groźbą ujawnienia należących do niej materiałów o charakterze seksualnym, przy czym szantażujący może nie posiadać tych materiałów.
[20] Fake news (z ang. fałszywa wiadomość) - nieprawdziwa lub częściowo nieprawdziwa wiadomość, często o charakterze sensacyjnym, publikowana w mediach z intencją wprowadzenia odbiorców w błąd w celu osiągnięcia korzyści finansowych, politycznych lub prestiżowych.
[21] Malware – z ang. malicious software, ogół złośliwego oprogramowania.
[22] Ransomware – oprogramowanie, które blokuje dostęp do systemu komputerowego lub uniemożliwia odczyt zapisanych w nim danych, a następnie żąda od ofiary okupu za przywrócenie stanu pierwotnego

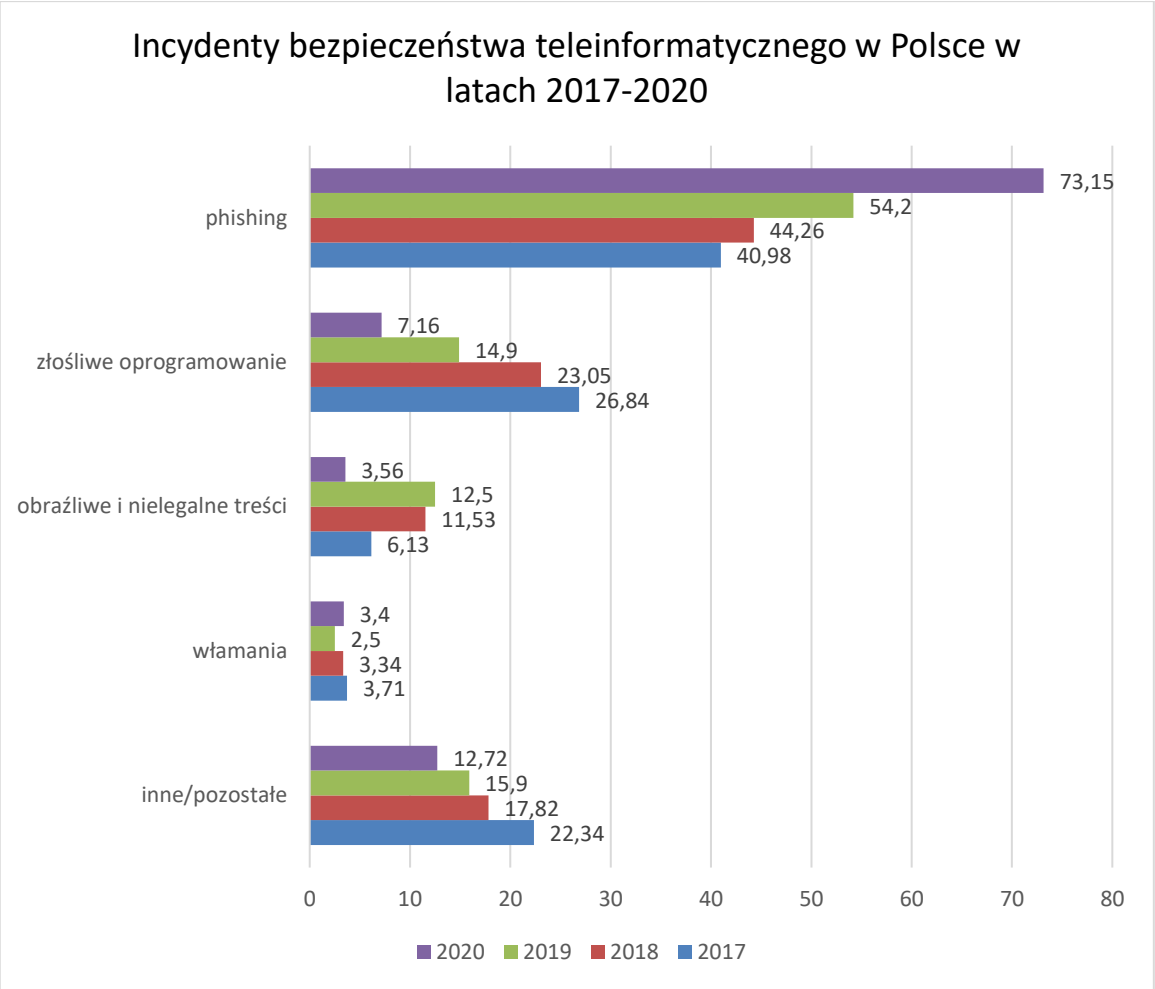


Rysunek 4. Przyczyny wycieków danych na świecie w latach 2020-2021, źródło: Dark Reading's Strategic Security Survey, https://www.darkreading.com/28dge-threat-monitor/phishing-remains-the-most-common-cause-of-data-breaches-survey-says

Na powyższym zestawieniu kategorię „Phishing" należy również poszerzyć o dane z kategorii „Targeted attack" – ataki ukierunkowane są częścią ataków phishingowych, co zostanie przedstawione i szczegółowo omówione w dalszej części pracy. Połączenie tych dwóch kategorii, jasno wskazuje na to że atak phishingowy jest globalnym problemem i stanowi zdecydowaną większość wszystkich incydentów bezpieczeństwa teleinformatycznego.

Analogicznie zestawienie prowadzone i publikowane przez polskie zespoły CERT/CSIRT pokazują, że głównym zagrożeniem w polskiej cyberprzestrzeni jest również phishing. W latach 2017-2019, z opublikowanych raportów [14] uwidoczniony jest wzrost ataków phishingowych o około 30 punktów procentowych (w przeciągu czterech lat) i stanowi obecnie największe zagrożenie dla użytkowników – ataki phishingowe stanowią większość (stan na 2020r. według raportów CERT Polska[23]) wszystkich incydentów bezpieczeństwa teleinformatycznego.

[23] https://www.cert.pl/uploads/docs/Raport_CP_2020.pdf



Rysunek 5 – Incydenty bezpieczeństwa w Polsce za lata 2017-2020r., źródło: https://cert.pl/publikacje/

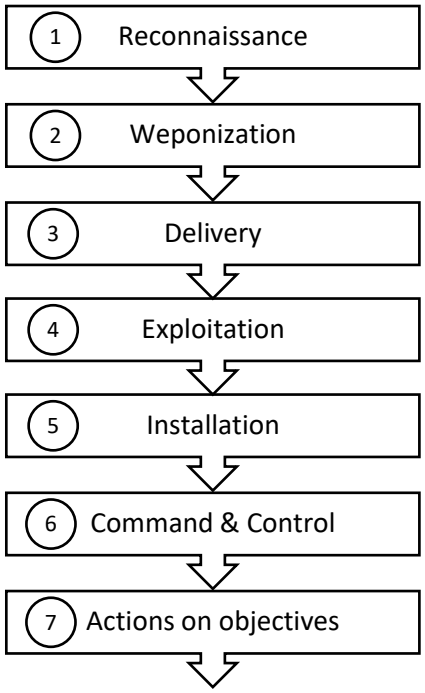
Phishing jest rodzajem oszustwa, w którym przestępca podszywa się pod inną osobę, firmę lub instytucję w celu:

1. wyłudzenia poufnych danych (danych osobowych, danych dostępowych do systemów teleinformatycznych, kont bankowych),
2. uzyskanie nieautoryzowanego dostępu do systemów teleinformatycznych (zwłaszcza przetwarzających poufne dane, systemów bankowych, np.),
3. zainfekowania komputera ofiary złośliwym oprogramowaniem,
4. nakłonienia ofiary do podjęcia lub zaniechania określonych działań.

Oszustwo to dokonywane jest zwykle z użyciem technicznych środków komunikacyjnych (sieci społecznościowych, poczty email, wiadomości SMS/MMS, stron internetowych). Łatwość i powszechność dostępu do kanałów sieci społecznościowych i usług poczty elektronicznej, powoduje, że sposób ten jest chętnie i często wykorzystywany przez grupy przestępcze.

I.1.1 Model „Cyber Kill Chain”

Rozważając współczesne zagrożenia dla użytkowników sieci komputerowych oraz analizując sposoby przeprowadzania poszczególnych ataków, Eric Hutchins, Michael Cloppert oraz Rohan Amin [15], dostrzeli, że występujące zagrożenia, wykorzystujące zaawansowane techniki i narzędzia są specjalnie zaprojektowane do pokonania tradycyjnych metod i technik zabezpieczeń. W celu usunięcia tego nowego rodzaju zagrożenia, konieczna jest wiedza o wykorzystywanym sposobie omijania zabezpieczeń, wykorzystywanych narzędziach i technologii [15]. Bazując na wojskowym algorytmie „Kill Chain[24]” [16], opracowany został model ataku cybernetycznego (Rysunek 6), który umożliwia zebranie informacji o kolejnych wykonywanych krokach w prowadzonym ataku, wykorzystywanych technologiach, co przekłada się na wysoką skuteczność realizacji zakładanego celu.



Rysunek 6. Model "Cyber Kill Chain".

Zaprezentowany model pod nazwą „Cyber Kill Chain[25]” (patrz: Rysunek 6) zawiera siedem następujących po sobie faz, prowadzących do skutecznego przejęcia i kontrolowania systemu ofiary:

[24] Kill chain – militarny koncept przeprowadzenia ataku zbrojnego, składający się następujących po sobie kolejnych faz: identyfikacji celu, wysłania sił atakujących, wydaniu rozkazu o dokonaniu ataku i zniszczenie celu.

[25] Źródło: https://www.lockheedmartin.com/en-us/capabilities/cyber/cyber-kill-chain.html [dostęp: 2020.03.08]

1. Rozpoznanie (ang. Reconnaissance) – w tej fazie następuje identyfikacja celu, rozpoznanie zasobów koniecznych do infiltracji i przeprowadzenia skutecznego ataku.
2. Uzbrojenie (ang. Weaponization) – przygotowanie narzędzi teleinformatycznych, oprogramowania w celu dostarczenia do systemu ofiary. W fazie tej wykorzystuje się często łączenie oprogramowania typu RAT ze złośliwym kodem wykonywalnym – utworzenie cyberbroni.
3. Dostarczenie (ang. Delivery) – dostarczenie przygotowanej w poprzedniej fazie cyberbroni do systemu ofiary z wykorzystaniem dostępnego medium (np. załącznik do wiadomości email, zasoby, strony internetowe, nośniki USB).
4. Eksploatacja (ang. Exploitation) – uruchomienie w systemie ofiary dostarczonego złośliwego oprogramowania z wykorzystaniem odkrytych luk w systemie operacyjnym, aplikacji użytkowej czy też funkcji systemu pozwalającej automatycznie wykonać dostarczony złośliwy kod. Eksploatacja może również być ukierunkowana na samego użytkownika z wykorzystaniem technik inżynierii społecznej.
5. Instalacja (ang. Installation) – faza ataku w której następuje zainstalowanie w systemie ofiary dostarczanych narzędzi (RAT, backdoor, np.) i utrzymanie trwałego dostępu do środowiska ofiary.
6. Dowodzenia i kontrola (ang. Command & Control – C2) – utworzenie kanału komunikacyjnego pomiędzy infrastrukturą atakującego a zainfekowanym system ofiary. Ustanowienie kanału komunikacyjnego pozwala na trwałe pozostanie w systemie ofiary i umożliwia atakującym uzyskanie pełnego dostępu do wszystkich funkcjonalności systemu ofiary.
7. Realizacja celów (ang. Actions on objectives) – ostatnia faza ataku, w której atakujący uzyskał kontrolę nad systemem ofiary. Najczęściej spotykanym celem ataku jest pozyskanie interesujących dla atakującego danych (kontaktów, zdjęć, dokumentów). Celem ataku może również być uzyskanie dostęp do poczty email (w celu prowadzenie kolejnych ataków z wykorzystaniem przejętego konta), kompromitacji danego systemu, wykorzystanie przejętego komputera jako elementu botnetu czy też prowadzenie dalszej eksploatacji sieci z wykorzystaniem danego komputera jako stacji przesiadkowej.

Model „Cyber Kill Chain" powstał w celu usystematyzowania procesu ataku, miał pomóc zrozumieć zespołom cyberbezpieczeństwa (CERT, CSIRT) proces ataku poprzez rozdzielenie jego całościowego aspektu na poszczególne, następujące po sobie fazy. Zrozumienie faz ataku pozwala natomiast wdrożyć odpowiednie strategie, które mają na celu przerwanie cyklu, a tym samym uniemożliwienie realizacji ataku.

Model „Cyber Kill Chain" jest kompleksowym podejściem do usystematyzowania opisu pełnego cyklu ataku realizowanego głównie przez grupy APT[26]. Jednakże model ten skupia się na wzmocnieniu tradycyjnych metod obrony stosowanych przez zespoły cyberbezpieczeństwa (bezpieczeństwo sieci, systemy antywirusowe). Analizując znane przypadki ataków, na bazie powtarzających się elementów, opracowane zostały również inne modele cyklów ataku cybernetycznego. Zaproponowany przez A. Hahn, R.K. Thomas, I. Lozano, A. Cardenas [17] model obejmuje 6 faz: rozpoznanie, uzbrojenie, dostarczenie, eksploatacja, kontrola (C2), realizacja celów. Jednocześnie A. Hahn przedstawił model ataku na infrastrukturę krytyczną zawierającą ciąg czterech faz, następujących bezpośrednio po sobie:

1. rozpoznanie,
2. uzbrojenie,
3. dostarczenie,
4. realizacja (zawierające w sobie eksplorację, kontrolę oraz realizację celów).

Wszystkie istniejące modele, opisujące cykl życia ataku cybernetycznego są pomocne do zrozumienia dynamiki rozwoju metod (i ich częstych zmian) jakimi posługując się atakujący. W dużej mierze typ ataku warunkuje jakie fazy cyklu zostaną wykorzystane do osiągniecia zakładanego w scenariuszu ataku celu.

Bazując na zaprezentowanym modelu „Cyber Kill Chain" [15] opisujący proces ataku, opracowany został model „Phishing Kill Chain" [18]. Model ten również zawiera siedem kolejnych faz.

[26] APT (z ang. Advanced Persistent Threat, pol. - zaawansowane trwałe zagrożenie) – termin odnoszący się do zaawansowanych technicznie grup hakerskich, często wspieranych i finansowanych przez państwa, odpowiedzialnych za przeprowadzanie skompilowanych ataków na międzynarodowe podmioty, koncerny, instytucje finansowe czy infrastrukturę krytyczną państw.

Tabela 1. Porównanie modeli „Cyber Kill Chain" i „Phishing Kill Chain". Źródło: https://www.agari.com/blog/phishing-kill-chain

Faza	Model „Cyber Kill Chain"	Model „Phishing Kill Chain"
1.	Reconnaissance	Targeting
2.	Weaponization	Delivery
3.	Delivery	Deception
4.	Exploit	Click
5.	Installation	Surrender
6.	Command & Control	Extraction
7.	Action	Action

Każda z wymienionych etapów ataku w modelu „Phishing Kill Chain" zawiera pewne charakterystyczne dla siebie techniki i czynności do wykonania:

1. Target – przygotowanie infrastruktury, określenie grupy docelowej i przygotowanie list potencjalnych ofiar wraz z ich adresami email. Faza ta może być poprzedzona zbieraniem informacji (OSINT) o precyzyjnie wyselekcjonowanych celach (w przypadku ataku typu „spear phishing" czy „whaling").
2. Deliver – rozsyłka przygotowanych wiadomości phishingowych. Wykorzystana zostanie do tego celu przygotowana infrastruktura (serwery pocztowe, założone lub przejęta konta email, profile w mediach społecznościowych).
3. Deceive – użycie socjotechniki do oszukania potencjalnych ofiar, by wykonały zalecane przez atakującego czynności. Faza ta zwykle planowana jest w trakcie tworzenia scenariusza ataku i obejmuje również przygotowanie treści stron i treści wiadomości wraz z szatą graficzną.
4. Click- podjęcie akcji przez użytkownika (klikniecie w odnośnik, uruchomienie załącznika). Najbardziej wrażliwy etap ataku – brak podjęcia reakcji użytkownika powoduje przerwanie całego procesu.
5. Surrender- wpisanie danych logowania (lub innych wrażliwych danych) na fałszywej stronie, uzupełnienie spreparowanego formularza czy też pozyskania danych przechwyconych przez zainstalowane na stacji ofiary złośliwe oprogramowanie (zwykle dostarczone jako załącznik do wiadomości email).
6. Extract – przesłanie pozyskanych danych do infrastruktury atakującego (serwery C2 , bazy danych, np.), zebranie danych z wypełnionych przez ofiary formularzy na stronach internetowych, odczyt zawartości skrzynek pocztowych (treści

wiadomości email przechowywanych na danej skrzynce pocztowej wraz z załącznikami).

7. Action – wykorzystanie dostępów do danych ofiary, wyprowadzenie danych, przejęcie kont, wyprowadzenie środków finansowych za pomocą uzyskanych w krokach poprzednich danych dostępowych.

I.1.2 Model MITRE ATT&CK

Zaprezentowany model „Cyber Kill Chain" opisujący proces ataku nie jest jedynym. Analizując stosowane techniki ataku (od jego przygotowania aż do zakończenia) powstały różne modele, które w różny sposób i z różną dokładnością opisują cały proces. Jednym z takich modeli jest MITRE ATT&CK (Adversarial Tactics, Techniques, and Common Knowledge).[27] Jest to otwarto-źródłowa baza danych zawierająca szereg informacji z dziedziny cyberbezpieczeństwa, - opisy technik, taktyk oraz wykorzystywanych procedur jakie wykorzystywane są przez grupy cybeprzestępcze i aktorów państwowych. Baza zorganizowana jest w hierarchiczne ramy. Każda opisana taktyka zawiera wiele technik, z których każda określa strategiczną metodę.

Tabela 2. Zestawienie technik opisanych w modelu MITRE ATT&CK. Źródło: opracowanie własne na podstawie https://attack.mitre.org/

Lp.	Nazwa techniki	Opis
1.	Reconnaissance	Rekonesans. Klasyczne rozpoznanie, kolekcjonowanie danych mogących być wsparciem do przeprowadzenia ataku w przyszłości. Grupa ta obejmuje wszystkie techniki którymi mogą posługiwać się atakujący aktywnie lub pasywnie pozyskują informacje o wyselekcjonowanym celu przyszłego ataku. Do informacji tych mogą należeć np. wykorzystywana adresacja wewnętrzna, rodzaj i marka wykorzystywanych urządzeń sieciowych, rodzaj zabezpieczeń, imiona i nazwiska pracowników. Informacje te wykorzystywane będą zarówno do inicjalizacji samego atakuj jak podczas późniejszych etapów.
2.	Resource Development	Poszukiwanie zasobów mogących być wykorzystanych do skutecznego przeprowadzenia ataku na system potencjalnej ofiary. Etap ten obejmuje wszelkie techniki, dzięki którym atakujący wytworzą, uzyskają dostęp (w tym również nieautoryzowany) czy też zakupią niezbędną infrastrukturę (konta email, hosting, domeny, VPN, VPS[28], np.).

[27] Szczegółowy opis wszystkich technik i subtechnik dostępny jest na oficjalnej stronie projektu: https://mitre.org

[28] VPS (ang. **V**irtual **P**rivate **S**erver) - odizolowane środowisko utworzone na fizycznym serwerze z wykorzystaniem technologii wirtualizacji

		Wytworzona na tym etapie infrastruktura może być również wykorzystana podczas późniejszych etapów.
3.	Initial Access	Techniki stanowiące próbę uzyskania początkowego dostępu do sieci potencjalnej ofiary (może to być np. ukierunkowany phishing spearphishing i wykorzystywanie słabości na publicznych serwerach internetowych). Techniki tego etapu są niezwykle istotne do wykrycia, gdyż jest to pierwszy etap bezpośredniej interakcji atakującego z atakowanym systemem – niektóre z nich określane jako „głośne”[29] mogą wyzwalać automatyczne alerty i być kierowane bezpośrednio do zespołów bezpieczeństwa.
4.	Execution	Techniki służące do próby uruchomienia złośliwego oprogramowania w środowisku potencjalnej ofiary. Obejmują np. uruchomienie załączników, uruchomienie zaszytych w dokumenty Office makr, które uruchamiają kolejne elementy (np. wywołują konsolę PowerShell). Techniki te zwykle nie występują samodzielnie, gdyż dla wzmocnienia efektu (pozyskania większej ilości danych lub uzyskanie dostępu do innego, pożądanego zasobu), łączone są z innymi technikami.
5.	Persistence	Uzyskanie dostępu i pozostanie niewykrywalnym w sieci ofiary. Techniki obejmują metody utrzymania dostępu do systemu nawet w przypadku wyłączenia i ponownego uruchomienia systemu przez użytkownika (zmiana konfiguracji, utworzenie reguł startowych[30], tworzenie dodatkowych kont, modyfikacja reguł firewall).
6.	Privilege Escalation	Eskalacja uprawień – atakujący stara się uzyskać wyższe uprawienia w systemie ofiary niż obecnie posiadane. Etap niezwykle ważny z punktu widzenia detekcji jak i pozostania niewykrywalnym. Techniki te obejmują np. wykorzystanie słabości systemu, błędnych konfiguracji i luk w zabezpieczeniach, nieprzestrzegania przez użytkowników zasad bezpieczeństwa, słabe polityki, np. Najbardziej poszukiwanymi uprawieniami przez atakującego są: • SYSTEM (Windows) / root (Unix / Linux), • administrator lokalny, • konto użytkownika z uprawieniami administratora • konta użytkowników z uprawnieniami do określonego systemu lub wykonywania określonej funkcji

[29] Jako „głośną” (w żargonie osób związanych z monitoringiem sieci komputerowych) należy rozumieć zespół takich czynności, które są rejestrowane przez urządzania bezpieczeństwa, pozostawiają wiele łatwych do odnalezienia śladów w atakowanych systemie (dzienniki zdarzeń, pozostałości po zapisywanych plikach, nadmierna komunikacja sieciowa, itp.). „Głośne” techniki są łatwe do identyfikacji i wykrycia w przeciwieństwie do technik „cichych”, dzięki którym atakujący może pozostawać w zaatakowanym systemie przez dłuższy czas niewykrytym.

[30] Reguły startowe obejmują m.in. modyfikację kluczy rejestru odpowiedzialnych za uruchomienie poszczególnych komponentów systemu podczas ponownego jego uruchamiania. Dodanie przez atakującego nowej wartości powoduje, że każdorazowo zostanie uruchomione dostarczone przez niego do systemu oprogramowanie (zwykle złośliwe).

		Etap ten często jest łączony z technikami skupionymi w grupie „Persistence” – uzyskanie wyższych uprawień może wpływać na możliwości wykonania technik służących do utrzymania dostępu do systemu.
7.	Defense Evasion	Unikanie wykrycia – dotyczy zarówno stosowania oprogramowania wpierającego atakującego (automatyczne systemy antywirusowe analityka behawioralna zachowania użytkownika threat hunting) jak i zacieranie śladów (kasowanie dzienników zdarzeń, usuwanie logów). W celu uniknięcia wykryciu wykorzystywane są również istniejące komponenty dawnego systemu (LotL- Living off the Land), wykorzystuje się zaciemnianie kodu szyfrowanie komunikacji z serwerem C2, szyfrowanie danych, np.
8.	Credential Access	Próba pozyskania z systemu ofiary, nazw użytkowników oraz haseł dostępu do innych systemów / wyższych poziomów. Możliwe stosowanie automatycznego oprogramowania pozyskującego hasła (credential stealer). Jest to etap niezwykle istotny dla atakującego – pozyskanie danych uwierzytelniających może zapewnić dostęp do różnych systemów i usług, utrudnić ich wykrycie i zapewnić możliwość utworzenia większej liczby kont (np. koniecznych do etapów: „Execution” „Persistence” lub „Privilege Escalation”).
9.	Discovery	Rozpoznanie środowiska, do którego atakujące uzyskał dostęp na wcześniejszym etapie ataku. Techniki te obejmują również monitoring prowadzony przez atakującego w celu wykrycia działań podejmowanych przez zespoły bezpieczeństwa, określenie dostępnych komponentów systemu, jakie są punkty styku z innymi systemami i w jaki sposób można się tam dostać. Zwykle w tej grupie często wykorzystywane są dostępne, natywne elementy danego systemu – co znacznie utrudnia detekcję.
10.	Lateral Movement	Próba penetracji przez atakującego innych zasobów systemowych dostępnych z obecnie zaatakowanego systemu – techniki obejmują np.: prowadzenie rozpoznania dostępnych zasobów, wykrywanie uruchomionych usług, przetwarzanych w systemie danych. Do realizacji tego etapu często wykorzystuje się dane pozyskane na etapach: „Privilege Escalation” i „Credential Access”. W etapie tym również wykorzystywane są dostępne, wbudowane elementy danego systemu w celu utrudnienia wykrycia.
11.	Collection	Zespół technik obejmujących pozyskiwanie i zbieranie danych z zaatakowanego systemu, które są istotne dla atakującego lub mogą zostać wykorzystane do realizacji innych ataków. Zebranie danych i ich eksfiltracja na zasób kontrolowany przez atakującego zwykle jest możliwa dzięki skutecznym realizacją etapów: „Privilege Escalation”, „Credential Access” oraz „Lateral Movement”. Zbieranie danych może odbywać się z każdego dostępnego zasobu (np. serwery plików, dyski sieciowe, skrzynki email,) ale również

		np. poprzez wykonywanie zrzutów ekranów, przechwytywanie danych wprowadzanych z klawiatury (oprogramowanie typ key-logger).
12.	Command and Control	Techniki obejmujące zdalną kontrole zainfekowanych systemów przez atakujących. Uruchamiane jest cykliczna komunikacja z serwerem zarządzającym (C2), która przypominać ma normalny ruch sieciowy wychodzących/wchodzący charakterystyczny dla danego systemu. Wykorzystywane są szyfrowane tunele, dostęp VPN czy natywne elementy systemu w celu uniknięcia wykrycia. Często ustanowienie tego etapu możliwe jest dzięki zakończonym sukcesem etapów: „Privilege Escalation" czy „Credential Access".
13.	Exfiltration	Techniki obejmujące próbę transferu pozyskanych danych na zewnętrzne zasoby / inne zasoby kontrolowane przez atakującego, zwykle ściśle powiązane z etapem „Collection". W etapie tym wykorzystywane są często techniki pakowania danych, szyfrowania pozyskanych danych, ograniczania wielkości transmisji[31], wykorzystanie dostępnych w danym systemie protokołów komunikacyjnych (zwykle poprzez ustanowiony w etapie „Command and Control" kanał komunikacyjny z serwerem C2). Eksfiltracja danych zwykle możliwa jest dzięki etapom: „Privilege Escalation", „Credential Access", „Lateral Movement", „Collection" oraz ustanowienie "Command and Control".
14.	Impact	Manipulacja danymi w systemie kontrolowanym przez atakującego, ograniczanie dostępu uprawnionym pracownikom do zasobów, niedostępność usług (w tym DoS i DdoS). Może również dojść do sytuacji całkowitego zniszczenia danych (w zależności od realizowanego przez atakującego scenariusza – np. poprzez oprogramowanie typu ransomware). Obejmuje również wpływanie na proces biznesowy, podejmowanie decyzji na podstawie zgromadzonych w zaatakowanym systemie danych, np. Etap ten jest wypadkową skutecznej realizacji poprzednich technik.

Oba modele opisują techniki stosowane do realizacji skutecznego ataku, jednakże występują pomiędzy nimi zasadnicze różnice. Model Cyber Kill Chain pozwala na lepsze zrozumienie procesu ataku dla początkujących analityków i zespołów

[31] Wiele systemów bezpieczeństwa posiada mechanizm DLP (ang. Data Loss Prevention - system zapobiegania wyciekom danych), który uniemożliwia transfer z komponentu chronionego przez ten system danych, jednakże nie jest w stanie zapobiec przesłaniu zaszyfrowanych, podzielonych na mniejsze fragmenty danych. Oprócz wdrożonego systemu DLP, urządzenia bezpieczeństwa monitorują również wielkość danych przesyłanych protokołami sieciowymi. Zbudowane reguły detekcji, mogą generować alerty dla zespołu bezpieczeństwa, gdy suma wysłanych z danego systemu danych, przekroczy pewną zdefiniowano uprzednią wielkość.

cyberbezpieczeństwa. Twórcy modelu przyjęli założenie (opisane poszczególnymi fazami), że w każdym cyberataku musi wystąpić pewna, określona sekwencja zdarzeń i by atak był skuteczny wszystkie jego elementy muszą zostać zrealizowane we właściwym porządku. W myśl tej zasady, w modelu „Cyber Kill Chain" zakłada się, że uniemożliwienie atakującemu skutecznej realizacji jednego z etapów, uniemożliwi mu się jednocześnie skutecznej realizację całości zakładanych przez niego celów. W odróżnieniu do tego podejścia, model MITRE ATT&CK koreluje pewne techniki ze sobą (powodzenie jednej techniki zależne jest od powodzenia innej), lecz nie muszą być one wykonywanie w zadanej kolejności, nie muszą też być wykonane wszystkie, by atak można uznać za udany (z punktu widzenia atakującego). Model MITRE ATT&CK jest pełniejszą bazą danych zawierającą aktualne opisy stosowanych przez atakujących technik i procedur (TTP[32]) oraz zapewnia informację o istniejących zagrożeniach (CTI[33]).

W odniesieniu do phishingu, model MITRE ATT&CK zawiera opis jednej techniki reprezentującej atak phishingowy (T1566[34]) wraz z trzema subtechnikami:

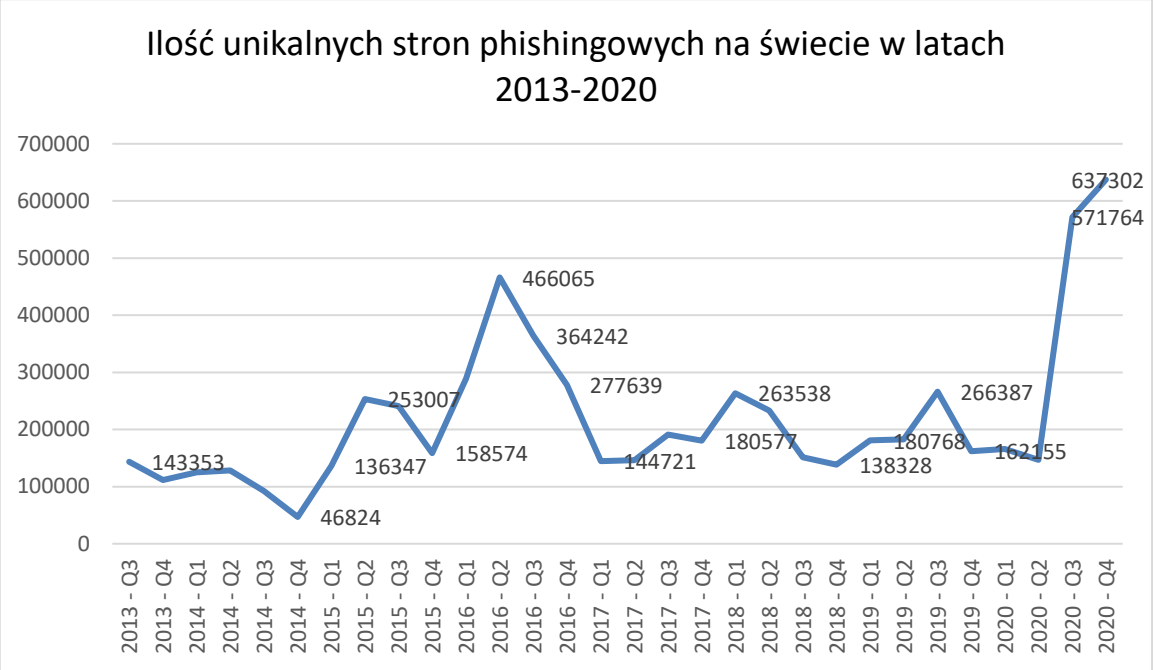
1. T1566.001 – Spearphishing Attachment – technika zawierająca opis wiadomości wraz z niebezpiecznym załącznikiem.
2. T1566.002 – Spearphishing Link – technika zawierająca opis wiadomości wraz z odnośnikiem prowadzącym do niebezpiecznego pliku.
3. T1566.003 – Spearphishing via Service – technika opisująca świadczenie usługi przygotowania i dostarczenie infrastruktury koniecznej do przeprowadzenia ataku phishingowego wraz z dostarczeniem wszelkich niezbędnych narzędzi.

I.2 Skala problemu

Wg raportu APWG (Anti-Phishing Work Group[35]), [19] od stycznia do lipca 2017 roku wykrytych zostało 291 096 unikalnych domen phishingowych. Od 2006 roku do

[32] TTP (z ang. Tactics, Techniques, and Procedures – pol. taktyki, techniki I procedury) - w nomenklaturze cyberbezpieczeństwa jest to kompleksowy opis zachowań atakującego.
[33] CTI (z ang. Cyber Threat Intelligence) - jest to analiza zagrożeń w cyberprzestrzeni to oparta na wiedzy, umiejętnościach i doświadczeniu (system ekspercki), informacja o występowaniu i ocenie zagrożeń cybernetycznych, która ma pomóc w ograniczeniu potencjalnych ataków i szkodliwych zdarzeń zachodzących w cyberprzestrzeni, źródło: https://www.crowdstrike.com/cybersecurity-101/threat-intelligence/
[34] https://attack.mitre.org/techniques/T1566/
[35] https://apwg.org/trendsreports/

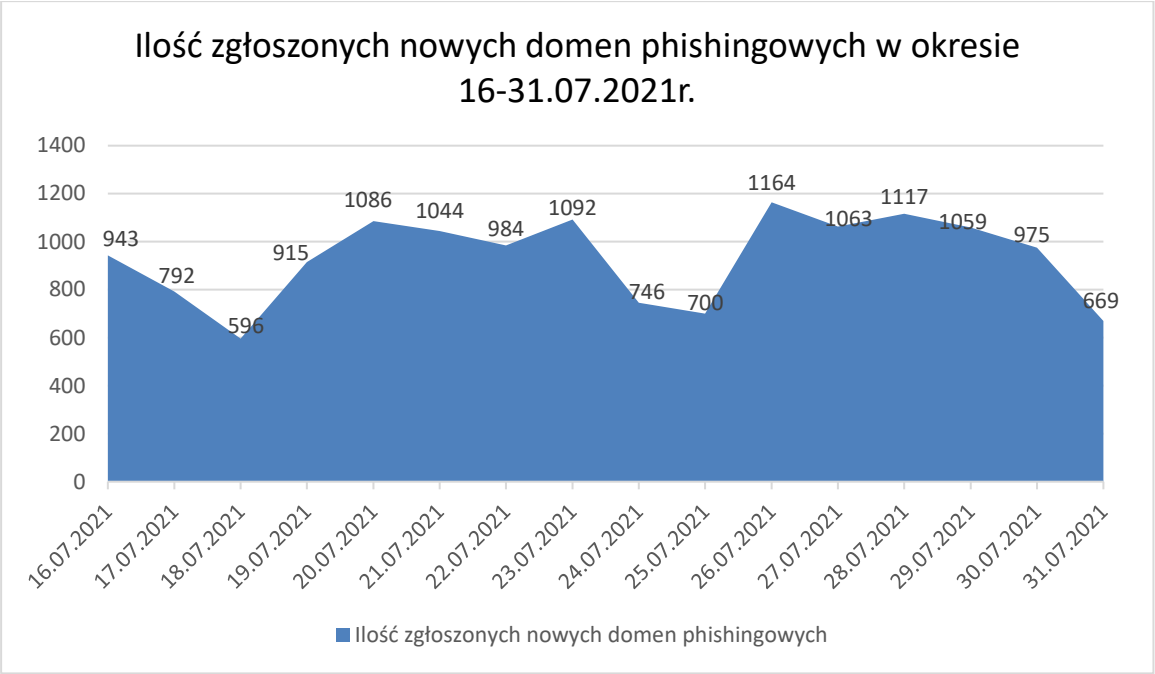
2016 roku ilość miesięcznych ataków z wykorzystaniem phishingu wzrosła o 5753% (1609 ataków phishingowych w 2004 roku do 92 564 ataków w 2016 roku[36]).



Rysunek 7 – Ilość unikalnych stron phishingowych na świecie, opracowanie na podstawie: https://www.statista.com/statistics/266155/39leme-of-phishing-domain-names-worldwide/ [dostęp: 2021-05-08].

Dobrze ujętą liczbę nowo wykrywanych domen o charakterze phishingowym, prezentuje organizacja portal phishtank.com prowadzony przez międzynarodową społeczność dokonująca rozpoznania i weryfikacji domen phishingowych, które następnie umieszcza w publicznie dostępnej bazie danych. W badanym okresie (16 lipca – 31 lipca 2021 roku) średnia dzienna ilość zgłaszanych nowych domen phishingowych wynosiła około 934 unikalnych nazw domenowych.

[36] A. Kumar Jain, B.B. Gupta, A machine learning based approach for phishing detection using hyperlinks information, Journal of Ambiet Inteligence and Humanized Computing



Rysunek 8. Ilość zgłoszonych nowych domen phishingowych przez społeczność phishtank.com w okresie 16-31.07.2021r., źródło: https://phishtank.com/stats.php [stan na dzień: 15.08.2021r.]

Duża ilość domen phishingowych oraz ciągły wzrost ilości ataków phishingowych (Rysunek 5) generują duże straty finansowe zarówno wśród bezpośrednich ofiar, jak i wśród dostawców usług, operatorów telekomunikacyjnych czy firm ubezpieczeniowych, zmuszonych do wpłat odszkodowań z tytułu phishingu. Według raportu FBI [20], w 2020 roku w Stanach Zjednoczonych, starty finansowe związane z phishingiem wyniosły ponad 54 mln dolarów.

Rysunek 9. Roczna strata finansowa ofiar phishingu w Stanach Zjednoczonych Ameryki w latach 2014-2020, źródło: FBI Crime Report, https://www.ic3.gov/Home/AnnualReports [dostęp: 22.10.2021r.]

Dane dostępne w publikowanych raportach FBI[37] (za lata od 2014 roku do 2020 roku) ukazują trend wzrostowy (Rysunek 9) ponoszonych strat finansowych wśród ofiar phishingu – co przekłada się bezpośrednio na wzrost zysków grup przestępczych czerpiących zyski z phishingu oraz dalszy rozrost samego zjawiska.

Na częstość wykorzystywania phishingu ma wpływ szereg czynników:

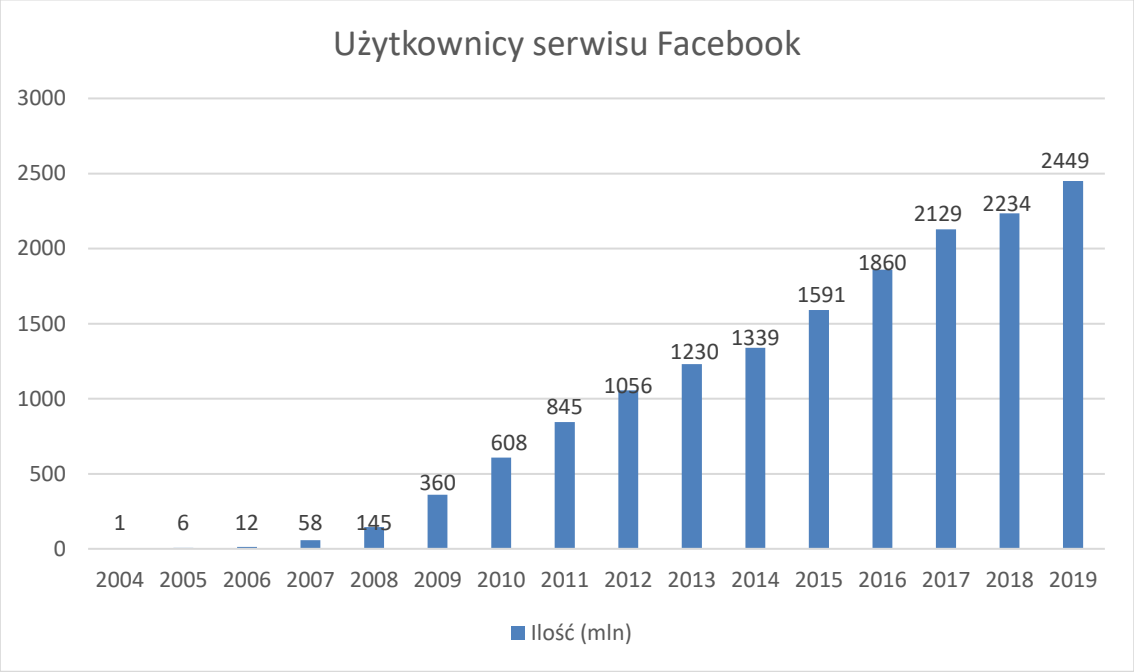
a) Phishing jest uniwersalny – nie jest zależny od konkretnej platformy sprzętowej, systemu operacyjnego czy używanego przez potencjalną ofiarę oprogramowania. Atakujący nie włamuje się bezpośrednio do atakowanego systemu omijając jego zabezpieczenia techniczne, a próbuje nakłonić danego użytkownika do działań mających na celu wyłudzenie danych, infekcję systemu czy instalację narzędzi obniżających bezpieczeństwo.
b) Tanie narzędzia i szeroko dostępna wiedza – atakujący wykorzystują dostępne w sieci Internet narzędzia (oprogramowanie typu „open source"), korzystają z poradników umieszczanych na forach w DarkNecie[38] opisujących metody i sposoby na skuteczne tworzenie i dystrybucję wiadomości phishingowych, sposoby używania oprogramowania wspomagającego masową wysyłkę czy też przedstawiające instrukcję przygotowania odpowiedniej infrastruktury zapewniającej anonimowość i umożliwiającej sterowanie zainfekowanymi komputerami (botnet).
c) Bazuje na inżynierii społecznej – wykorzystując metody i techniki inżynierii społecznej może być kierowany do nieograniczonej liczny osób.
d) Rozwój sieci społecznościowych (Rysunek 10) – popularność sieci społecznościowych i ich gwałtowny rozwój (np. Facebook, Twitter, Instagram, Qzone, LinkedIn) umożliwia szybkie dotarcie do wielu użytkowników i grup jednocześnie, za pomocą jednej lub kilku wiadomości. Spreparowana wiadomość phishingowa, do stworzenia której użyto elementów socjotechniki, często przesyłana jest dalej przez same ofiary, które stają się nieświadomym ogniwem w dalszej dystrybucji phishingu (np. łańcuszki internetowe).

[37] https://www.ic3.gov/Home/AnnualReports

[38] DarkNet – sieć teleinformatyczna dla której Internet jest siecią transmisyjną a do której dostęp możliwy jest jedynie za pomocą specjalnego oprogramowania, autoryzacji czy odpowiedniej konfiguracji i zwykle nie jest możliwe dotarcie do niej bezpośrednio z sieci Internet za pomocą dostępnych standardowych wyszukiwarek. Sieć taka często wykorzystuje specjalnie wykorzystywany protokół komunikacyjny, często zapewnia wysoką anonimowość swoim użytkownikom. Przykładem darknetu jest sieć TOR.

e) Phishing jest dochodowy – tanie i dostępne w sieci Internet narzędzia umożliwiają każdemu przeprowadzenie ataku phishingowego. Adresy email potencjalnych ofiar można odszukać w publicznie dostępnych bazach wycieków z różnych serwisów internetowych (sklepów[39], platform ubezpieczeniowych, systemów rezerwacji biletów, np.).

f) Phishing jest wielotorowy – potencjalna ofiara otrzymuje wiadomości phishingowe za pomocą różnych kanałów: email, komunikatory, wiadomości SMS/MMS, wiadomości na forach.



Rysunek 10 – Rozwój portalu Facebook, źródło: https://www.statista.com/topics/751/facebook/

Wymienione powyżej czynniki, dzięki którym phishing jest najczęściej notowanym zagrożeniem, odnoszą się bardziej do ataku phishingowego niekierowanego do konkretnych osób, grup zawodowych czy za pomocą konkretnego, wybranego medium komunikacyjnego.

Jak wykazano na początku rozdziału, atak phishingowy jest najczęściej notowanym incydentem bezpieczeństwa teleinformatycznego w Polsce. Wiele firm z branży technologicznej zaangażowanych jest w tworzenie technicznych rozwiązań

[39] https://spidersweb.pl/2019/04/baza-danych-morele-wyciek.html

mających za zadanie wykrycie i powstrzymanie ataku w jego początkowej fazie (faza przygotowania).

Publikowane przez zespoły bezpieczeństwa (CSIRT) statystyki incydentów teleinformatycznych, wskazują, że ilość ataków phishingowych w Polsce wciąż wzrasta, pomimo ponoszonych nakładów na jego powstrzymanie.

I.3 Opis technik realizacji ataku phishingowego

Działanie phishingu opiera się w dużej mierze (w początkowej jego fazie – fazy ataku phishingowego opisane są w dalszej części pracy) na manipulacji potencjalną ofiarą, wykorzystując różne złudzenia, bazując na sprawdzonych psychologicznych wzorach działania człowieka. Poddany odpowiedniemu procesowi manipulacji (poprzez sugerowaną treść wiadomości, wykorzystaną grafikę, czy też podszycie się pod urząd/instytucję), użytkownik nieświadomie dokonuje czynności, finalnie prowadzących do niepożądanych efektów (z punktu widzenia odbiorcy). W ataku phishingowym – pomimo wykorzystania również technicznych elementów (np. złośliwe oprogramowanie w odpowiednio spreparowanym dokumencie), celem ataku nie jest de facto sprzęt, ale sam człowiek.

Do stałego elementu ataku phishingowego należy wysłanie do potencjalnej ofiary odpowiednio spreparowanej wiadomości (email, SMS czy poprzez komunikatory internetowe), której treść obliczona jest na przekonanie odbiory o konieczności podjęcia sugerowanych w niej działań, np.:

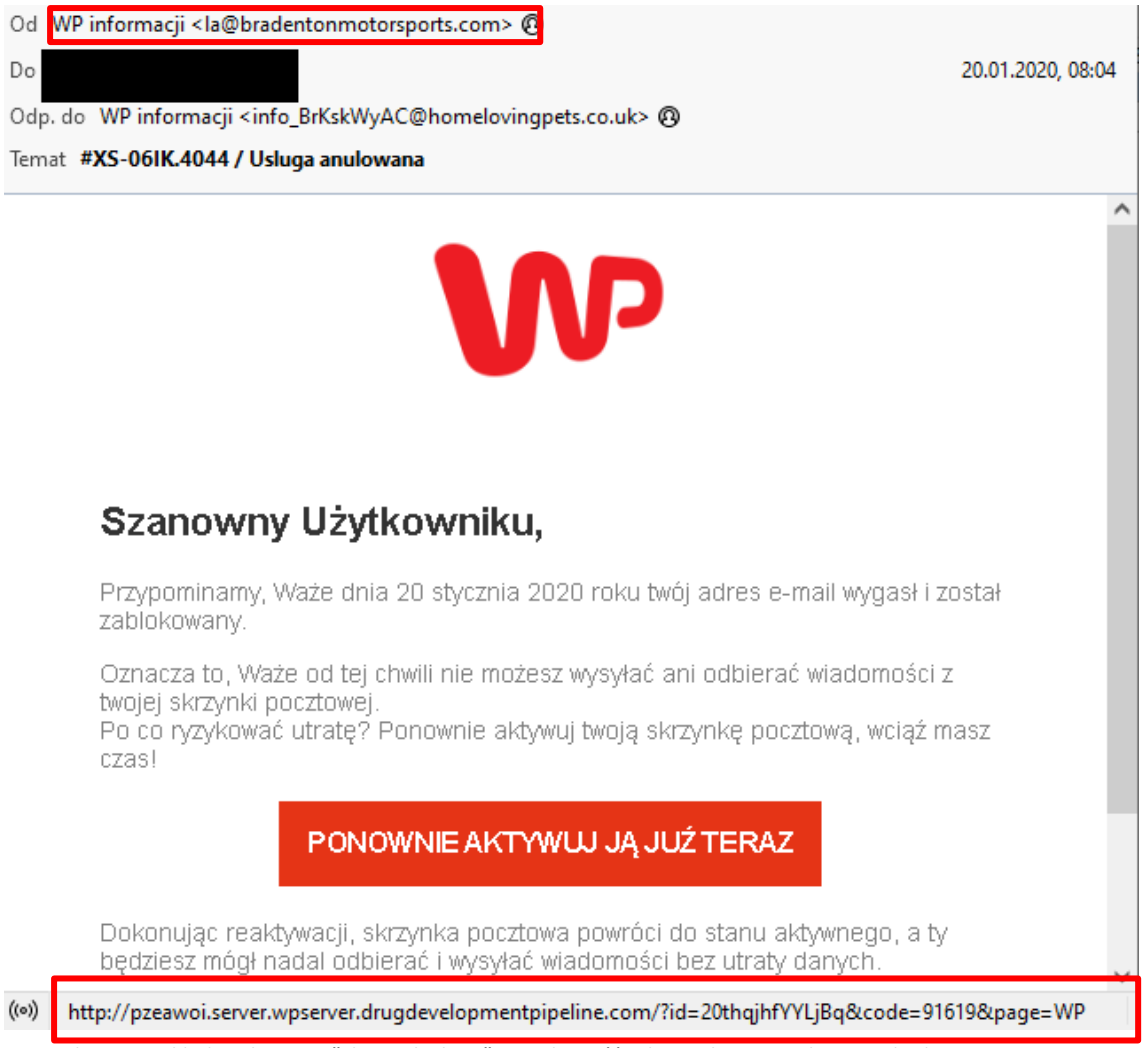
1. kliknięcie na załączony odnośnik URL i uzupełnienie formularza (np. fałszywego formularza logowania się do systemu pocztowego),
2. pobranie i uruchomienie załącznika znajdującego się w danej wiadomości,
3. wniesienie niewielkiej dopłaty do usługi, dopłaty do towaru – przesłany odnośnik prowadzi do fałszywej strony logowania do bankowości elektronicznej.

Przyczyną skuteczności ataku phishingowego jest jego prostota w połączeniu z zastosowanymi technikami inżynierii społecznej. Wykorzystanie kanałów komunikacji elektronicznej, wymaga posiadania pewnych danych o potencjalnej ofierze (adresu email, numeru telefonu czy nazwy w portalu społecznościowym), które to dane w erze

społeczeństwa informacyjnego można pozyskać za pomocą technik OSINT[40]. Sam atak można przypisać do poniższych kategorii:

a) Email phishing – najbardziej popularna i rozpowszechniona metoda będąca równocześnie jedną z najprostszych. Atakujący preparują wiadomość, podszywając się pod znaną markę, rozpoznawalną firmę, kampanie medialną, stosują sztuczki socjotechniczne do skłonienia potencjalnej ofiary do wykonania sugerowanych w wiadomości działań (kliknięcia w odnośnik, pobrania i uruchomienia załącznika, np.). Większość technik phishingowych, jako swoją bazę/nośnik ataku wykorzystuje metodę email phishing.
b) HTTPS phishing – znaczna część społeczeństwa jest wciąż przekonana o bezpieczeństwie i legalności protokołu HTTPS z powodu używania szyfrowanego połączenia przez ten protokół do komunikacji klient-serwer – w odróżnieniu do nieszyfrowanego protokołu NP. Przestępcy znają to podejście i dlatego w wiadomościach phishingowych, wykorzystują protokół HTTPS, jawnie przedstawiając link, by odbiorca mógł zauważyć używany protokół i tym chętniej kliknąć w odnośnik.
c) Spear phishing – spersonalizowany atak na daną osobę, firmę czy instytucję (w przypadku firmy i instytucji wiadomości wysyłane są na pozyskane adresy email funkcyjne lub indywidualne poszczególnych pracowników). Wariant ten bazuje na technikach email phishing. Atak poprzedzony jest zbieraniem informacji na temat celu ataku, dokładnym rozpoznaniem infrastruktury teleinformatycznej. Ten typ ataku ma największe prawdopodobieństwo sukcesu (wg badaczy Trend Micro[41] sukces ataku gwarantowany jest w 91% [21]) i stanowi duży odsetek, sięgający nawet 38% ([22]) wszystkich ataków na przedsiębiorstwa.
d) Clone phishing – wiadomość phishingowa wykorzystująca jako źródło oryginalną wiadomość (z oficjalną szatą graficzną, układem wiadomości, wykorzystywaną czcionką) w której jeden z elementów został podmieniony (np. odnośnik do pobrania dokumentu kierujący na fałszywą stronę, załącznik ze modyfikowaną

[40] OSINT (ang. Open Source Intelligence) – techniki gromadzenia informacji bazujące na publicznie dostępnych danych (portale społecznościowe, publiczne rejestry, fora internetowe, strony www, itp.).
[41] Trend Micro - międzynarodowa firma zajmująca się cyberbezpieczeństwem (https://www.trendmicro.com).



Rysunek 11. Przykład ataku typu "clone phishing". Wiadomość udająca korespondencję od administratora systemu Pocztowego Wirtualnej Polski. Wyświetlany przycisk przekierowuje do zainfekowanej strony internetowej, źródło: opracowanie własne.

zawartością lub doklejonym złośliwym kodem). Clone phishing również bazuje na technikach email phishing.

e) Whaling – spersonalizowany atak na osobę z kierownictwa danej firmy czy instytucji, bazującej na metodzie email phishing. Zawartość zostanie stworzona specjalnie pod konkretną osobę (najczęściej szczebla kierowniczego) i jej rolę w firmie. Dane wykorzystywane do przeprowadzenia ataku (np. adres email, imię i nazwisko, stanowisko, nazwa firmy, zainteresowania), zwykle są dostępne w przestrzeni publicznej. Atak tego typu może być trudniejszy do przeprowadzenia z uwagi na stanowiska zajmowane przez osoby – cele ataku. Kadra kierownicza i managerowie wysokiego szczebla (CEO firmy, rzecznicy, dział PR, np.) mogą być zaznajomieni ze standardowymi taktykami istniejących w sieci zagrożeń. Z uwagi na funkcjonowanie takich osób w przestrzeni medialnej, mogli przejść obszerne szkolenie w zakresie świadomości

bezpieczeństwa ze względu na swój publiczny profil i prowadzoną działalność, a zespół np. bezpieczeństwa może mieć bardziej rygorystyczne zasady i dedykowane rozwiązania do ich ochrony. Skłania to atakujących, którzy próbują phishingu na te cele, do wyjścia poza wypróbowane scenariusze (skuteczne przeciwko innym, mniej medialnym osobom), do bardziej wyrafinowanych, ukierunkowanych, spersonalizowanych pod konkretną osobę, metod.

f) Smishing – phishing poprzez wiadomości SMS/MMS. Potencjalnej ofierze przesyłana jest krótka wiadomość zawierająca zwykle odnośnik do odpowiednio spreparowanej strony phishingowej, która w zależności od scenariusza wyłudza dane do logowania do systemu pocztowego, lub kieruje na fałszywą bramkę płatności internetowej. Bardzo często technika ta łączona jest również ze spoofingiem numeru nadawcy widomości (wykorzystanie reputacji banku, instytucji, znanej firmy) w cel zwiększenia szans powodzenia. Atak zyskujący popularność w latach 2020/2021r (wg danych firmy Proofpoint, [23] w 2020r. nastąpił wzrost o 328%, a badania [24] przeprowadzone przez Next Caller wykazały, że 44% obywateli USA odnotowało wzrost liczby połączeń telefonicznych mających charakter phishingu). Wzrost liczby tego typu ataków jest również obserwowany w Polsce [25]. Szczegółowy przykład ataku typu smishing znajduje się w dodatku C.
g) Vishing – phishing telefoniczny, metoda oszustwa wykorzystywana w celu przekonania rozmówcy to wykonania określonych czynności, podania wrażliwych danych (np. PESEL, numer dowodu, kody autoryzujące do bankowości internetowej). Oszustwo to zyskało dużą popularność w 2021r. z wykorzystaniem automatycznych nagrań, odtwarzanych przez operatora w zależności od udzielanych przez rozmówcę – potencjalną ofiarę, odpowiedzi.
h) Angler phishing – wykorzystanie funkcjonalności aplikacji mediów społecznościowych. Wiadomości zwykle dystrybuowane są z fałszywych kont mediów społecznościowych, udających kampanie reklamowe, marketingowe, badania społeczne, np. Oferują wzięcie udziału w badaniu marketingowym w zamian za drobną nagrodę – wymagają do wówczas podania wielu danych

osobowych, dzięki czemu zyskują dane do przeprowadzenie dalszych ataków, wykorzystania pozyskanych danych jako tzw. Konta przesiadkowego[42].

i) Pharming – najbardziej techniczny z wariantów ataku phishingowego. Wariant ten nie zakłada bezpośredniego ataku na potencjalną ofiarę, ale na pośrednika (serwer DNS), z którego usług dany użytkownik korzysta. Atak następuje w dwóch fazach:
   1. Atak na serwer DNS do którego zapytanie kierowane są ze stacji danego użytkownika. Serwer DNS na pytanie kierowane ze stacji użytkownika, o rozwiązanie adresu danej domeny, zamienia rzeczywisty adres IP, na który wskazuje dana domeną na adres IP podstawiony przez przestępców.
   2. Przekierowanie ruchu sieciowego na podstawioną domenę phishingową – która w zależności od scenariusza ma za zadanie wyłudzanie danych użytkownika, wykradanie danych autoryzujących i uzyskiwanie dostępu do
j) Pop-up phishing – wykorzystanie mechanizmu wyskakujących okienek do serwowania użytkownikowi złośliwego kodu lub uruchomienie mechanizmu śledzenia opartego np. na cookies[43]. Z uwagi na stosowanie przez użytkowników mechanizmów blokowania wyskakujących okienek na witrynach internetowych, odmianą tego ataku jest wykorzystanie funkcjonalności powiadomień przeglądarek internetowych by stosując socjotechnikę przekonać użytkownika do zezwolenia na uruchomienie powiadomień – które jednocześnie serwuje użytkownikowi złośliwy kod.
k) Evil twin – wykorzystanie fałszywego hotspot[44] WiFi, podstawianego w publiczny miejscu, który udaje legalny punkt dostępu. Użytkownicy łączący się do fałszywego punktu dostępowego mogą być przekierowani do strony phishingowej, ich ruch wychodzący może być podsłuchiwany (atak typu man-in-the-middle[45]). Atak tego typu pozwala zbierać dane, takie jak dane logowania lub

[42] Konto przesiadkowe – wykorzystanie pozyskanej tożsamości do utworzenia fałszywego profilu, z którego prowadzone są dalsze ataki, w celu zamaskowania prawdziwego sprawcy.
[43] http cookie – tekst zwykle zakodowany przesyłany przez serwer do przeglądarki klienta, która zapisuje go w swoim folderze, aby go ponownie odczytać podczas kolejnych odwiedzin danego serwera i przesłać zapisany test. Mechanizm cookies pozwala.m.in. na identyfikacje użytkowników, rozróżniania różnych maszyn tego samego użytkownika, itp.
[44] Hotspot (z ang.– „gorący punkt") – punkt dostępu do sieci bezprzewodowej, umożliwiający za jego pośrednictwem połączenie z siecią Internet
[45] man-in-the-middle – rodzaj ataku polegający na podsłuchu i modyfikacji informacji przesyłanych pomiędzy dwiema stronami bez ich wiedzy.

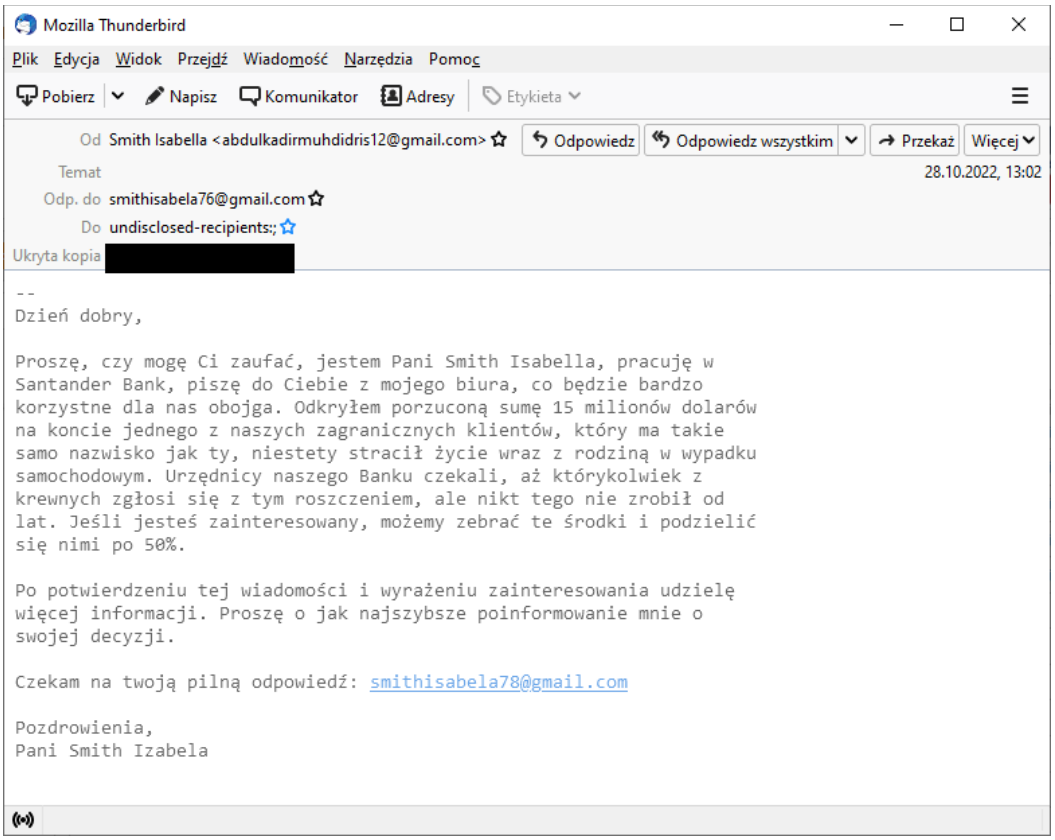
poufne informacje przesyłane przez połączenie – które mogą zostać wykorzystane do konstrukcji bardziej złożonego ataku (np. atak typu spear phishing).

l) Watering hole phishing – drugi z typów najbardziej technicznych wariantów ataków phishingowych. W tej wersji atak nie następuje bezpośrednio na docelową ofiarę (podobnie jak w przypadku pharmingu). Atakujący zbierają informacje o najczęściej odwiedzanych witrynach internetowych przez interesującą ich grupę użytkowników. Witryna ta jest odpowiednio modyfikowana by w trakcie odwiedzania jej przez internautów, zbierać informację o nich, uruchamiać mechanizm śledzący czy też infekować ich stację roboczą złośliwym kodem. Przykładem skutecznej kampanii phishingowej wykorzystującej technikę „Watering hole phishing" był atak na stronę Komicję Nadzoru Finansowego [26], poprzez którą infekowani byli pracownicy sektora finansowego.

m) Atak responsywny – nowy trend i technika ataku oparty na inżynierii społecznej. Pierwsza otrzymana wiadomość email przez odbiorcę (przykład: Rysunek 12), nie zawiera zarówno żadnych odnośników URL, jak i żadnych załączników (które często są nośnikiem złośliwego oprogramowania). Treść wiadomości ma zachęcić potencjalną ofiarę (odbiorcę wiadomości), do wykonania „pierwszego kroku" – odpowiedzi na otrzymanego emaila. W ten sposób atakujący jednocześnie weryfikuje skuteczność kampanii i ma pewność, że otrzymując odpowiedzi, dany użytkownik (dany adres email, na który wysłał wiadomość phishingową) jest aktywny (jego skrzynka odbiorcza jest obsługiwana) oraz może być podatny na sugestię i techniki manipulacyjne. W odpowiedzi na reakcję użytkownika, atakujący przesyła kolejną wiadomość, w której również za pomocą inżynierii społecznej zachęca do ujawnienia innych, lub większej ilości danych osobowych. Motywy jakie przewijają się w treści wiadomości, zachęcające do nawiązania kontaktu, to:

    1. odziedziczenie spadku przez dalekiego krewnego,
    2. istnienie konta bankowego zmarłej osoby o tym samym co odbiorca wiadomości nazwisku,
    3. wylosowanie adresu email przez „fundusze inwestycyjne",
    4. uniknięcie opodatkowania przez różnego rodzaju fundacje poprzez przekazanie potencjalnej ofierze darowizny (zwykle wysokich kwot),
    5. wygrane w loteriach.

Wiadomość taką, nazywać będziemy wiadomością inicjalizującą. Atakujący inicjuje nawiązanie kontaktu, testując jednoczenie czas odpowiedzi – krótszy czas odpowiedzi sugeruje użytkownika aktywnie korzystającego z sieci Internet i komunikacji elektronicznej, który odbiera wiadomości, odpisuje na nie, przez co zwiększa jednocześnie szansę powodzenia ataku.

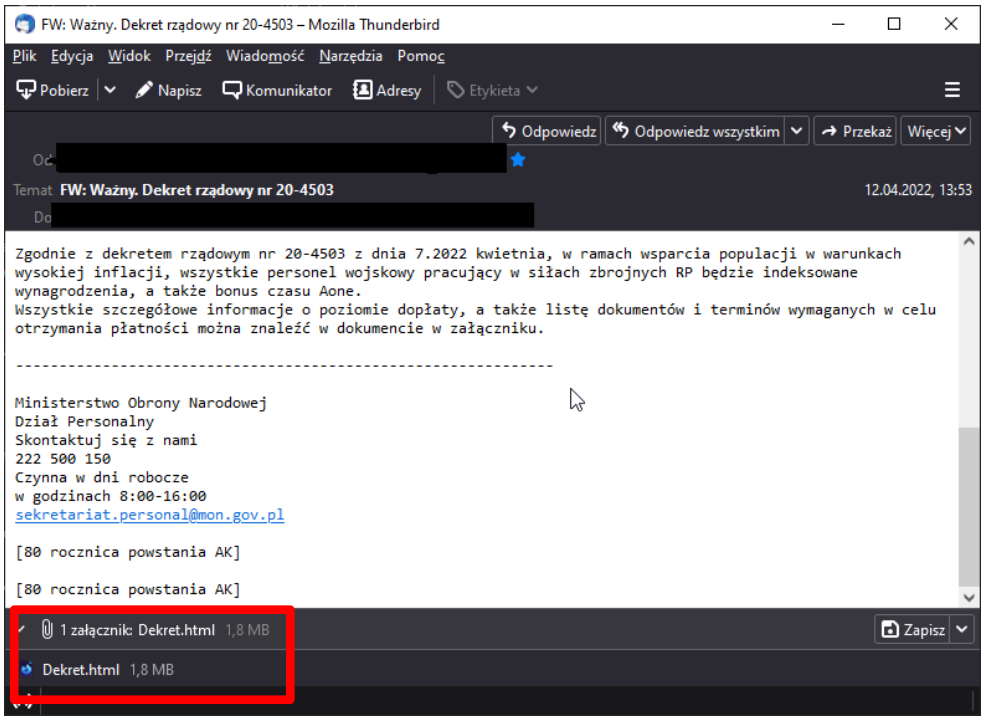


Rysunek 12. Przykład ataku responsywnego – zachęta do kontaktu pod pozorem przekazania sporej sumy pieniężnej.

Kolejnymi etapami (w przypadku otrzymania odpowiedzi od ofiary) jest:

1. Prośba o podanie danych osobowych lub danych karty płatniczej (np. w celu wykonania opisywanego w wiadomości inicjującej spadku / darowizny), w formie odpowiedzi na wiadomość, lub poprzez wypełnienie formularza dostępnego w sieci Internet.
2. Przesłanie odnośnika do pobrania pliku, który może zawierać złośliwe oprogramowanie
3. Przesłanie załącznika zawierającego złośliwy kod, po uruchomieniu którego nawiązane jest połączenie z serwerem C&C, z którego pobierane jest właściwe złośliwe oprogramowanie infekujące komputer ofiary.

n) HTML smuggling – atak polegający na dostarczeniu w wiadomości email do potencjalnej ofiary dokumentu HTML zawierającego osadzony w nim skrypt (głównie JavaScript).



Rysunek 13. Przykład ataku typu „HTML smug ling" z osadzonym plikiem HTML zawierający zakodowany binarnie złośliwy plik.

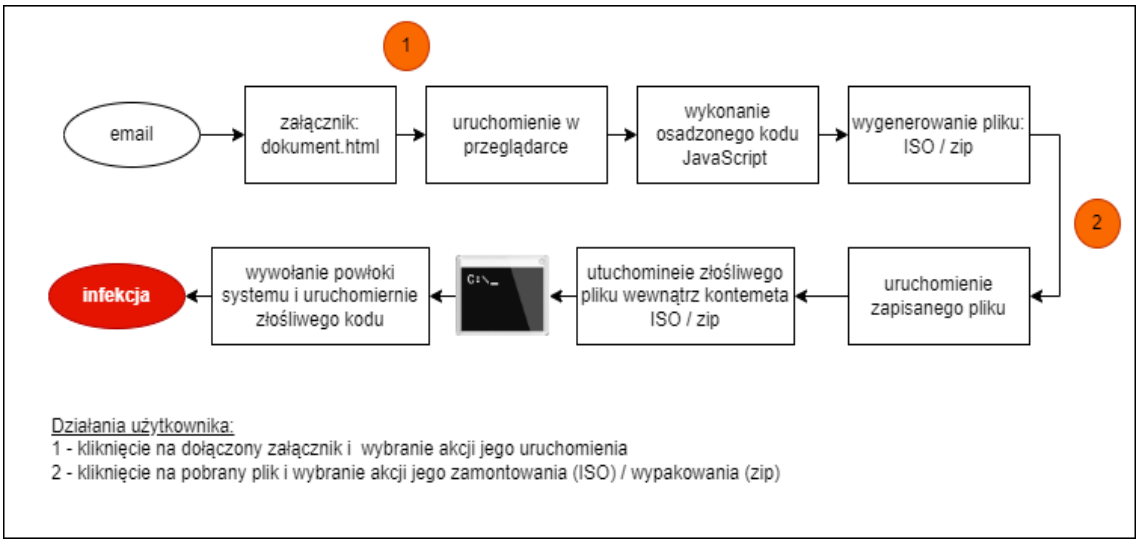
Skrypt wykorzystując technikę JavaScript Blob[46], dekoduje zapisany w postaci kodu binarnego wewnątrz dokumentu plik i uruchamia automatyczne pobieranie pliku. Atak taki obejmuje fazy (Rysunek 13):

1. Dostarczenie wiadomości email do skrzynki odbiorczej celu ataku. Wiadomość zawiera plik formatu HTML. Treść dokumentów tekstowych nie jest skanowana przez oprogramowanie antywirusowe, plik więc przechodzi przez infrastrukturę bezpieczeństwa.
2. Odbiorca wiadomości, uruchamia plik. Zgodnie z formatem, domyślnym oprogramowaniem uruchamiającym dokumenty formatu HTML jest przeglądarka internetowa.
3. Atak bazuje na wykorzystaniu podstawowej funkcjonalności przeglądarki internetowej (techniki LotL), która domyślnie wykonuje cały kod zawarty

[46] JavaScript Blob (ang. JavaScript Binary Large Object) – programowalna za pomocą języka JavaScript struktura zawierająca obiekt niezmiennych, nieprzetworzonych danych

w pliku HTML. Plik zawiera funkcję (zwykle z wykorzystaniem języka JavaScript), która na podstawie wartości pewnej zmiennej również zapisanej w tym pliku (kod binarny), uruchamia dekodowanie i utworzenie na jej podstawie nowego pliku. Popularnym formatem do jakiego generowany jest kod binarny zapisany w pliku jest format ISO[47]/ IMG[48] lub zip.

4. Zdekodowany plik jest zapisywany na dysku odbiorcy wiadomości.
5. Użytkownik uruchamia zdekodowany z dokumentu HTML, pobrany plik, uruchamiając jednocześnie domyślną akcję przypisaną temu formatowi.
6. W przypadku plików ISO / IMG domyślną akcją jest zamontowanie dysku wirtualnego zawierającego plik „autorun", który uruchamia niezłośliwy plik wykonywalny znajdujący się na dysku wirtualnym (exe) – kolejne użycie techniki LotL Uruchomiony plik wczytuje następnie plik biblioteki *.dll, która już zawiera złośliwy kod.
7. Wywołane zostaje okno konsoli systemowej i uruchomione funkcje odpowiedzialne za infekcję systemu.



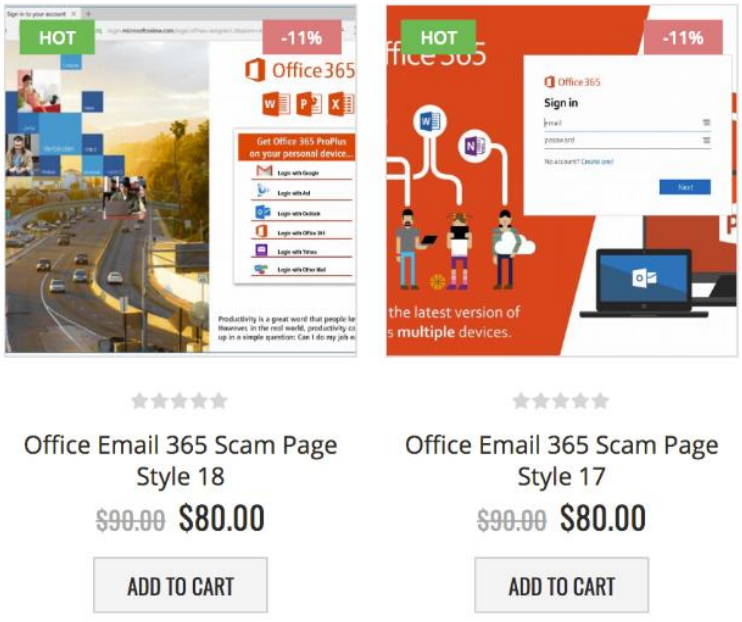
Rysunek 14. Fazy ataku „HTML smuggling"

[47] ISO - jest to skompresowany plik obrazu dysku archiwum, który reprezentuje zawartość całych danych na dysku optycznym, takim jak CD lub DVD. W oparciu o standard ISO-9660, format pliku obrazu ISO zawiera dane dysku wraz z przechowywanymi na nim informacjami o systemie plików, źródło: https://docs.fileformat.com/pl/compression/iso/

[48] IMG – jest to format plików binarnych które przechowują nieprzetworzone obrazy dysków (analogicznie jak ISO).

W ten sposób bezpośrednio do sytemu użytkownika dostarczony może być złośliwy kod, całkowicie omijający oprogramowanie antywirusowe. Wykorzystując kombinację technik „spear-phishingu", inżynierii społecznej, spoofingu i „HTML smuggling", można w bardzo skuteczny sposób infekować wyselekcjonowane ofiary pomimo stosowanych zabezpieczeń własnego systemu teleinformatycznego.

o) Phishing as a Service – określenie to nie odnośni się do konkretnej kategorii ataku phishingowego (jest w dużej mierze zbiorem poszczególnych kategorii), ale opisuje nowy trend. Phishing as a Service jest usługą oferowaną w darknecie przez grupy cyberprzestępców, które za opłatą oferują:
    1. Dostęp do infrastruktury teleinformatycznej fałszywych bramek płatności i przesłanie uzyskanych za jej pomocą środków finansowych bezpośrednio na konto zlecającego atak (wykupującego daną usługę).
    2. Szablonów stron popularnych usług finansowych i banków.
    3. Szablony wiadomości email oraz infrastrukturę konieczną do jej wysyłania. Szablony zwykle są powtarzalne.
    4. Domen phishingowych z całą wymaganą infrastrukturą teleinformatyczną.
    5. Serwerów Command and Control (C2).
    6. Dostosowane do indywidualnych potrzeb złośliwe oprogramowanie wraz z panelem sterowania.
    7. Wiedzę niezbędną do przeprowadzenia ataku (w zakresie obsługi dostępnego oprogramowania, skutecznego doboru metody, np.), wsparcie przeprowadzenia całego ataku.



Rysunek 15. Przykład usługi „Phishing as a Service".

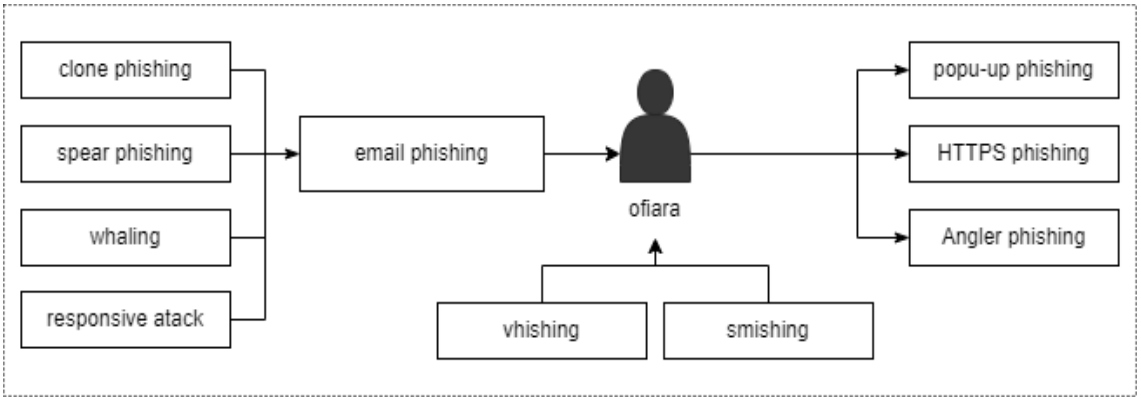
Tego typu usługi, umożliwiają przeprowadzenie ataków phishingowych przez osoby nietechniczne, nie wymagają przeprowadzenia rekonesansu (warunku posiadania wiedzy o obiekcie ataku). Jednocześnie oferowanie takich usług, pozwala na uzyskanie wysokich przychodów bez konieczności samodzielnego wyszukiwania celu ataku.

p) Wykorzystanie metod i technik sztucznej inteligencji i uczenia maszynowego.
Dynamiczny rozwój metod uczenia maszynowego i technik sztucznej inteligencji oraz udostępnienie szerokiej publiczności interfejsów do komunikacji z algorytmami Machine Learning (np. chatGPT[49]) zostało wykorzystane do prowadzenia skuteczniejszych ataków phishingowych. Podstawowym środkiem komunikacji w wielu firmach (jak i dla osób prywatnych) nadal pozostaje poczta email [27], a łatwość implementacji szybko rozwijających się narzędzi typu AI[50] (w tym zarówno do generowania treści jak i grafiki), powoduje, że wytworzenie bardziej przekonywującej wiadomości jest znacznie łatwiejsze. Sprawia to jednocześnie, że detekcja ataku phishingowego, staje się trudniejsza do identyfikacji przez człowieka. Wg autorów raportu [27] odnotowano 135% wzrost ataków opartych na inżynierii społecznej w okresie styczeń – luty 2023 roku, co odpowiada okresowi upowszechnieniu się i dynamicznemu wzrostowi korzystania z dostępnych usług ChatGPT.
Narzędzia AI mogą być również wykorzystane do generowania kodu do narzędzi wykorzystywanych w różnego rodzaju atakach, co spowoduje, że całość procesu będzie mniej kosztowana i czasochłonna. Sam wygenerowany kod może być łatwo rozpoznawalny przez rozwiązania antywirusowe, jednak wykorzystanie narzędzi AI do jego wytworzenia przyczyni się do szerokiego wykorzystania przez mniejsze grupy, które do tej pory nie posiadały zasobów, wiedzy ani umiejętności do przeprowadzania tego typu ataków. Przełoży się to na dalsze zwiększenie wolumenu ataków i jednoczesne zwiększenie trudności wykrywania doczasowych metod detekcji.

[49] https://chat.openai.com/chat/
[50] Pod tym pojęciem należy rozumieć ogół technik uczenia maszynowego, klasyfikacji oraz analizy Big Data.

q) Quishing – atak phishingowy na urządzenie mobilne z wykorzystaniem odpowiednio spreparowanego kodu QR[51] dostarczonego poprzez wiadomość email. Odbiorca wiadomości otrzymuje w jej treści kod QR, a jej treść sugeruje użytkownikowi zeskanowanie załączonego kodu pod pozorem np. dokonania pilnej płatności na niewielką kwotę (wykorzystanie inżynierii społecznej). Kod prowadzi do serwisu udającego pośrednika płatności, gdzie użytkownik podając swoje rzeczywiste dane, umożliwia jednocześnie ich przechwycenie przez atakującego. Innym znanym wariantem tego ataku jest przekierowanie użytkownika bezpośrednio do pobrania na jego urządzenie odpowiednio spreparowanego pliku zawierającego złośliwy kod. Istnieje również wersja ataku, w której użytkownik otrzymuje wiadomość niezawierającej żadnej treści, a jedynie odpowiednio spreparowany kod QR. Badania [28] przeprowadzone przez Sharevski, Devine, Pieroni, Jachim wykazują, że quisinbg jest niezwykle skutecznym typem ataku. Spośród uczestników poddanych badaniom, 67% z nich zarejestrowało się przy użyciu danych logowania Google lub Facebook, 18,5% utworzyło nowe konto, a tylko 14,5% pominęło rejestrację.

Przedstawione rodzaje phishingu łączone są w łańcuch ataku (patrz: Rysunek 16) w celu zwiększenia prawdopodobieństwa jego powodzenia – co zostanie przedstawione w dalszej części pracy.



Rysunek 16. Łączenie różnych metod ataku phishingowego.

[51] QR (właściwe QR code, ang. Quick Response, pol. szybka odpowiedź) – pewnego rodzaju dwuwymiarowy, kwadratowy kod kreskowy, pozwalający na zapisanie dużej ilości danych w pop stacji matrycy obrazkowej.

I.3.1 Techniki implementacji phishingu

Zaprezentowane powyżej rodzaje phishingu, mogą być zaimplementowane w różny sposób podczas wykonywania konkretnego typu ataku. Możemy rozróżnić cztery zasadnicze grupy ataków [29]:

1. Email-to-email,
2. Email-to-website,
3. Website-to-website,
4. Browser-to-website.

I.3.1.1 Email-to-email

Technika wykorzystująca w początkowej fazie wiadomości email jako główny kanał komunikacji z potencjalną ofiarą. Konstrukcja otrzymanej wiadomości, wymusza od ofiary udzielenie drogą elektroniczną odpowiedzi. Atakujący mogą w ten sposób pozyskać interesujące ich dane (dane osobowe, hasła dostępu, inne szczegóły na temat ofiary). Atakujący, w wersji minimum, w tym przypadku musi posiadać konto email, z którego będzie prowadził korespondencję (wersja najprostsza technicznie do wykonania – bez konieczności posiadania specjalistycznej wiedzy). Z punktu widzenia ekonomii taki, model ten jest najkorzystniejszy dla atakujących – wiele operatorów usług poczty elektronicznej umożliwia bezpłatne korzystanie z ich usług.

I.3.1.2 Email-to-website

Technika łącząca kanał komunikacji elektronicznej i strony internetowej o charakterze phishingowym. W typie ataku, wiadomości email w treści zawierają odnośniku URL prowadzące do zewnętrznego zasobu sieciowego, ofiara poprzez odpowiednio wywieraną presję (w treści wiadomości), kierowana jest pod wskazany adres i wykonuje określaną (pożądaną przez atakujących) akcję (np. podanie danych osobowych, haseł dostępu czy pobranie i zainstalowanie złośliwego oprogramowania). Przeprowadzenie tego typu ataku wymaga dostępu do skrzynki email oraz przygotowania odpowiedniej infrastruktury internetowej:

a. zarejestrowanie odpowiedniej domeny internetowej,
b. utrzymywanie serwera www lub wykupienie usługi hostingu,
c. stworzenie witryny internetowej,
d. odpowiednie oprogramowanie.

I.3.1.3 Webiste-to-webiste

Atak ten polega na skierowaniu użytkownika na stronę o charakterze phishingowym, poprzez kliknięta przez niego reklamę (odpowiednio spreparowaną) lub poprzez zwrócenie strony phishingowej jako wynik wyszukiwania użytkownika. Mechanizm prezentowania reklam, umieszcza odnośniki reklamowe, zwykle jako pierwsze wśród wyników wyszukiwania, a użytkownicy zwykle wybierają pierwszy wynik z listy – będący reklamą – kierujący ich do witryny mogącej wyłudzać dane osobowe lub hasła dostępu (np. fałszywe sklepy internetowe, strony podszywające się pod instytucje). Ten model ataku wymaga przygotowania bardziej zaawansowanej infrastruktury obejmującej:

a. zarejestrowania dwóch lub więcej różnych domen internetowych,
b. utrzymywanie serwera (serwerów) www lub wykupienie usługi hostingu,
c. wykupienie kampanii reklamowej, kierującej na wybrany adres domeny phishingowej,
d. stworzenie odpowiednich witryn internetowych,
e. wypromowanie (pozycjonowanie) założonych stron phishingowych,
f. odpowiednie oprogramowanie.

I.3.1.4 Browser-to-website

Technika ta bazuje na popełnianych błędach przy samodzielnym wpisywaniu przez użytkownika adresu istniejącej strony internetowej. Atakujący rejestrują domenę o nazwie łudząco podobnej do oryginalnej – nazwa domeny zawiera jednak popularną literówkę jaką można popełnić przy wpisywaniu adresu oryginalnej strony internetowej w pasek adresu, np.: loto.pl zamiast lotto.pl. Ten rodzaj ataku znany jest jako typosquatting (opisany w dalszej części pracy).

Powyższy wykaz technik implementacji phishingu nie wyczerpuje jednak wszystkich możliwości, wykorzystywanych przez atakujących. Wskazane grupy ataków, przez autorów pracy [29], należało by uzupełnić o techniki wykorzystujące jako jeden z elementów ataku złośliwe oprogramowanie:

1. Email-to-workstation.
2. Website-to-workstation.

I.3.1.5 Email-to-workstation

Technika ta bazuje na bezpośrednim dostarczeniu poprzez wiadomość email, złośliwego ładunku bezpośrednio do komputera odbiorcy, w postaci załącznika. Zapisanie i uruchomienie załącznika przez odbiorcę powoduje infekcję jego komputera. Zachętą do pobrania i uruchomienia złośliwego załącznika może być wytworzenie przekonania u ofiary, że otrzymany załącznik jest dokumentem (fakturą, rachunkiem, potwierdzeniem).

I.3.1.6 Website-to-workstation

Technika podobna do opisanej powyżej, różnicą jest to, że użytkownik poprzez przekierowanie do odpowiednio spreparowanej strony www, dzięki której:

1. pobiera samodzielnie (lub pobierane jest automatycznie) złośliwe oprogramowanie, które jest następnie uruchamiane na komputerze ofiary dokonując jego infekcji.
2. Wykorzystuje poprzez przeglądarkę internetową, zasoby sprzętowe (moc obliczeniowa, przepustowość, pamięć) do realizacji innego ataku, przeciwko innemu celowi.

I.3.2 Techniki zaciemniania

Celem podniesienia skuteczności ataku phishingowego jest zniwelowanie wszelkich wskazówek – zarówno w otrzymanej wiadomości jak i spreparowanej strony internetowej – że użytkownik ma do czynienia z oszustwem.

1. Spoofing

   Najczęściej stosowana metoda[52], jeden z elementów inżynierii społecznej, mającej na celu przekonać potencjalną ofiarę do wykonania zalecanych czynności (kliknięcia w odnośnik, pobrania i zainstalowania oprogramowania, uruchomienia załącznika, np.). Chętnie wykorzystywane jest podszywanie się pod zaufane źródło: przestępcy podszywają się pod źródło, takie jak bank, firma lub organizacja rządowa, w celu wywołania zaufania u odbiorcy wiadomości.. Mogą wykorzystać podobne adresy e-mail (również z wykorzystaniem technik typosquattingu), loga i treści, aby wyglądać jak prawdziwe źródło.

[52] Zgodnie z raportem firmy ProofPoint, każdego dnia wysyłanych jest około 3.1 miliarda wiadomości, w których nadawca podszywa się pod inną osobę/firmę (źródło: https://www.proofpoint.com/us/threat-reference/email-spoofing, [dostęp: 01.03.2023]).

2. Podwójne nazwy załączników
    - Wykorzystanie mechanizmu ukrywania znanego rozszerzenia pliku w systemach rodziny Microsoft Windows – przesłanie załącznika ze zmodyfikowaną nazwą, tak by sugerował, inny typ pliku użytkownikowi – np. faktura.pdf.exe, gdzie .exe jest ukrywane przez system Windows, a użytkownikowi wyświetlana jest nazwa: faktura.pdf[53].
    - Wykorzystywaną przez cyberprzestępców metodą jest również użycie znaku „Right-to-Left Override" (Kod: U+202E), który odwraca wprowadzony tekst. Atakujący tworzący plik o nazwie: „faktura [U+202E]fdp.exe" na komputerze odbiorcy danej wiadomości zostanie wyświetlona nazwa pliku: „faktura exe.pdf"[54]. W połączeniu z zastosowaniem podwójnego rozszerzenia pliku, technika ta jest skutecznym środkiem do implementacji złośliwego oprogramowania na urządzeniu odbiorcy.
3. Zawartość zdalna

    Tego typu wiadomości nie zawierają zwykle bezpośredniej treści, a jedynie za pomocą osadzonych znaczników HTML wyświetlają odbiorcy zdalną zawartość – zwykle w postaci ikonografii. Wyświetlany obraz jest jednocześnie odnośnikiem prowadzącym do spreparowanej strony internetowej lub zasobu sieciowego w celu dalszego przekierowania (wielokrotne przekierowania).
4. Techniki oszukujące rozpoznawanie wzorców (np. modele Markowa [30], Naiwny Klasyfikator Bayesa). Często stosowana przez atakujących technika mająca za zadanie uniemożliwić rozpoznanie i właściwą klaryfikacje otrzymanej wiadomości na podstawie analizy treści czy tematu wiadomości. Jedną z metod jest dodanie losowej sekwencji numerycznej (3-5 cyfr) na początku tematu i jego końcu, przez co analiza treści (np. zastosowanie Naiwnego Klasyfikatora Bayesa do rozpoznawania wzorca[55]) staje się zaburzona, lub wręcz całkowicie błędnie klasyfikowana. Losowy ciąg numeryczny dodany do tematu jest identyfikowany

[53] Ta technika zadziała jedynie wtedy, kiedy dany użytkownik ma włączoną (pozostawioną domyślnie) opcję: „Ukryj rozszerzenia znanych typów plików".

[54] Oryginalny typ pliku, w którego nazwie zastosowano technikę "Right-to-Left Override", można każdorazowo sprawdzić we właściwościach pliku – typ pliku nie jest w tym przypadku w żaden sposób modyfikowany.

[55] Dodanie losowych sekwencji do określonej frazy, dla automatycznych modeli opartych na NBC, całkowicie zaburza prawdopodobieństwo, co finalnie prowadzi do niewłaściwej klasyfikacji i błędnego przypisania klasy – a więc przepuszczenia takiej wiadomości przez filtr.

przez człowieka jako powtarzający się schemat, natomiast w przypadku rozpoznawania maszynowego traktowany jest każdorazowo jako inny, nowy wyraz, co skutecznie obniża prawdopodobieństwo właściwej klasyfikacji. Wiadomości takie również na poziomie weryfikacji przez mechanizm IDS[56] (o ile jest wdrożony) potrafią je omijać.

5. Deepfake

   Wykorzystywanie fałszywego wideo lub fałszywych nagrań głosowych, które mogą wydawać się autentyczne. Do stworzenia fałszywych nagrań video, często wykorzystywany jest wizerunek:

   a. Biznesmenów lub ekonomistów – osoby odpowiedzialne za firmy lub instytucje. Taka osoba może prosić o natychmiastową wypłatę pieniędzy lub zmianę danych płatności, co może prowadzić do przekazania pieniędzy na fałszywe konto.
   b. Polityków – wykorzystanie do przekazywania fałszywych informacji, manipulacji informacjami, np.
   c. Osoby znane publicznie (tzw. Celebryci[57]) – wykorzystanie do poparcia dla pewnych produktów lub reklamowanie usług, które z założenia mają przynieść korzyści jedynie przestępcą (produkt może w ogóle nie istnieć).

6. Gra na odczuciach – wykorzystanie technik inżynierii społecznej.

I.3.3 Inżynieria społeczna w ataku phishingowym

Niemal wszystkie opisane techniki implementacji ataku phishingowego wykorzystują inżynierię społeczną w fazie „Reconnaissance" i „Targeting" (model „Cyber Kill Chain) oraz „Weaponization" i „Delivery" (model „Phishing Cyber Kill Chain"). Inżynieria społeczna jest wykorzystywana do wprowadzenia w błąd ofiar i zachęcenia ich do ujawnienia poufnych informacji – co decyduje o wysokiej skuteczności przeprowadzonego ataku phishingowego. W tym celu wykorzystuje się:

[56] IDS (ang. Intrusion Detection System) – automatyczne systemy wykrywające włamania/ataki, zwykle ściśle współpracujące z systemami IPS (Intrusion Prevention System) – systemami zapobiegania włamaniom.

[57] Celebryta - osoba, która osiągnęła szeroką popularność i rozpoznawalność wśród społeczeństwa, zwykle dzięki swojej pracy lub osiągnięciom w dziedzinie sztuki, kultury, rozrywki, sportu lub innych dziedzinach. Celebryci są często uważani za osoby publiczne, a ich życie prywatne, wybory i zachowanie są często szeroko dyskutowane przez media i społeczeństwo.

a. Autorytet – wykorzystanie zaufania w autorytet niektórych pełnionych funkcji czy wykonywanych zawodów (np. przełożony w pracy, menager wyższego szczebla, pracownik obsługi technicznej, policjant, np.) aby zyskać zaufanie i w ten sposób przekonać potencjalne ofiary do udostępnienia poufnych informacji, których w normalnych sytuacjach nigdy by nie ujawniły. Wykorzystanie autorytetu, władzy stosuje się, gdyż ludzie wykazują duże skłonności do niekwestionowanego posłuszeństwa wobec osób, które mają większe doświadczenie lub uprawnienia, a jednocześnie ignorują własny osąd celowości tego działania. W psychologii zjawisko to określa się jako jeden z błędów poznawczych[58].
b. Strach – wykorzystanie różnego rodzaju obaw, celowe wytworzenie lęku u potencjalnej ofiary np. strach przed utratą środków pieniężnych poprzez nielegalny dostęp do konta bankowego, które mają skłonić do podjęcia szybkich, często nieracjonalnych działań – kliknięcie w odnośnik niewiadomego pochodzenia czy pobranie i uruchomienie nieznanego oprogramowania, udzielenia dostępu do własnych urządzeń, np.
c. Pomoc – podobnie jak w przypadku wykorzystania autorytetu, wykorzystywane jest zaufanie użytkowników i oferowana pomoc w pokonaniu pewnych trudności czy rozwiązaniu pewnych problemów (nierzadko te problemy celowo są wcześniej tworzone, by ofiara w łatwiejszy sposób zaakceptowała oferowaną rzekomą pomoc).
d. Skomplikowanie – wykorzystywanie skomplikowanych schematów przeprowadzenia ataku, aby utrudnić jego wykrycie. Skomplikowane, zagmatwane czynności, poparte często rzekomo profesjonalnie brzmiącym słownictwem (używanie wysublimowanego słownictwa, niewykorzystywanego w codziennej mowie), używanie skomplikowanej terminologii technicznej niezrozumiałej dla odbiorcy.
e. Obietnice – atak zawierający atrakcyjną, lecz całkowicie fikcyjną ofertę – obietnicę (np. inwestycje o wysokim procencie zwrotu).
f. Obowiązek – wytworzenie poczucia wywiązania się z obowiązku lub konieczności oddźwięcznienia się na wykonaną przysługę (zwykle celowo zaaranżowaną przez atakującego, w celu stworzenia okazji do wdzięczności).

[58] https://pl.wikipedia.org/wiki/B%C5%82%C4%85d_poznawczy

g. Chciwość – bazowanie na emocjach potencjalnej ofiary, wytworzenie u potencjalnej ofiary odczucia, że w łatwy sposób może uzyskać, nieproporcjonalnie duże korzyści w stosunku do poświęconego czasu, czy zaangażowanych środków. Pod wpływem emocjonalnego działania, odbiorca nie weryfikuje prawdziwości wiadomości, wykonując polecenia atakującego, podejmując często nieracjonalne decyzje
h. Współczucie – bazowanie na chęci pomocy osobie w trudnej sytuacji.

Szanowny kliencie,

Otrzymaliśmy informację o podejrzanym logowaniu do Twojego konta bankowego. W celu ochrony Twojego konta, prosimy o pilne potwierdzenie swoich danych.

Aby to zrobić, prosimy o kliknięcie w poniższy link i zalogowanie się na swoje konto bankowe. Następnie zostaniesz poproszony o potwierdzenie swoich danych i zmianę hasła.

Prosimy o jak najszybsze podjęcie tych działań w celu uniknięcia zagrożenia dla Twojego konta.

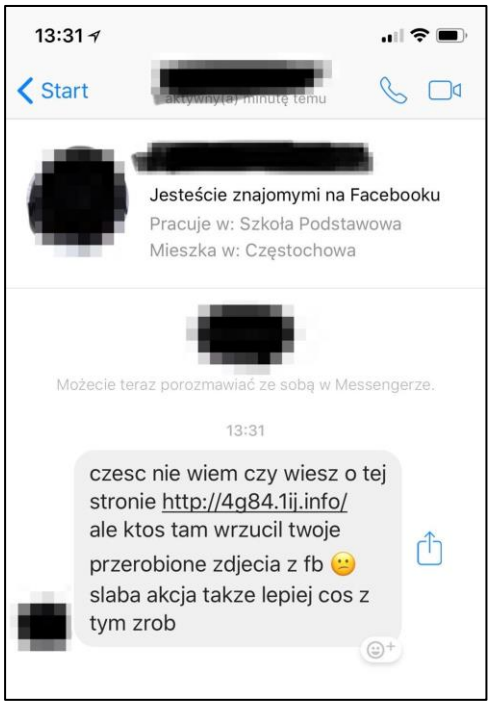
Link: xxxxxx

Z poważaniem,

Zespół wsparcia klienta banku

Rysunek 17. Przykład wiadomości wykorzystującej inżynierię społeczną by wymusić na potencjalnej ofierze podjęcie natychmiastowych działań.

i. Ciekawość – wykorzystanie bieżących, często bardzo sensacyjnych wydarzeń. Tworzenie fałszywych artykułów, newsów z kontrowersyjnym lub poszukiwanym tematem, gdzie dostępna jest jedynie szczątkowa informacja a rzekoma sensacyjna treść dostępna jest jedynie dla zarejestrowanych lub uwierzytelnionych użytkowników. Często wymagana jest symboliczna opłata dostępowa – która ma być przeznaczona na utrzymanie systemu. Użytkownik jest od razu przekierowany do fałszywego pośrednika płatności, gdzie przechwytywane są jego dane logowania do systemu bankowości elektronicznej wraz z kodami autoryzacyjnymi. Dodatkowym czynnikiem wzbudzającym ciekawość i jednocześnie osłabiającym czujność odbiory, jest fakt, że wiadomość ta pochodzi zwykle od osoby pozostającej w sferze znajomości z potencjalną ofiarą – poprzez przejęcie konta w mediach społecznościowych tej osoby i za jego pomocą dystrybuowanie wiadomości phishingowych.



Rysunek 18. Przykład ataku socjotechnicznego bazującego na ciekawości. Źródło: https://galeria.bankier.pl

j. Automatyzm – bazowanie na zjawisku automatycznej reakcji[59], która nie angażuje analitycznego myślenia podczas podejmowania decyzji. Zjawisko to można zaobserwować podczas otrzymywania newsletterów, użytkownicy automatycznie reagują klikając link „wypisz się", znajdujący się zwykle na końcu danej wiadomości. Identyczny mechanizm wykorzystywany jest przez atakujących do przekierowania odbiorcy wiadomości na inną stroną, fałszywe powiadomienia o pojawieniu się nowego komentarza/opinii na temat danego użytkownika (zwykle mającego negatywną formę).

[59] Są to działania podejmowane bez udziału świadomego umysłu. Automatyzm może być pierwotny (wrodzony) lub wtórny (brak namysłu). Automatyzm można go podzielić na odruchowy, językowy lub myślowy.

Dzień dobry,

Proszę, czy mogę Ci zaufać, jestem Pani Smith Isabella, pracuję w Santander Bank, piszę do Ciebie z mojego biura, co będzie bardzo korzystne dla nas obojga. Odkryłem porzuconą sumę 15 milionów dolarów na koncie jednego z naszych zagranicznych klientów, który ma takie samo nazwisko jak ty, niestety stracił życie wraz z rodziną w wypadku samochodowym. Urzędnicy naszego Banku czekali, aż którykolwiek z krewnych zgłosi się z tym roszczeniem, ale nikt tego nie zrobił od lat. Jeśli jesteś zainteresowany, możemy zebrać te środki i podzielić się nimi po 50%.

Po potwierdzeniu tej wiadomości i wyrażeniu zainteresowania udzielę więcej informacji. Proszę o jak najszybsze poinformowanie mnie o swojej decyzji.

Czekam na twoją pilną odpowiedź: smithisabela78@gmail.com

Pozdrowienia,

Pani Smith Izabela

Rysunek 19. Przykład wiadomości phishingowej, bazującej na wywołanej emocji u potencjalnej ofiary (zachowano oryginalną pisownię), źródło: materiały własne

Wszystkie powyższe opcje (opracowane na podstawie materiałów własnych) zawierają również element wywierania presji, pospiechu[60], gdyż wówczas odbiorca pod wywieraną presją, w mniejszym stopniu weryfikuje prawdziwość otrzymanej wiadomości i jest bardziej skłonny (np. pod wpływem uprzednio wytworzonej emocji) do wykonania czynności, które narzucane mu są w treści. Podejmowanie świadomej, przemyślanej i racjonalnej decyzji, zwykle opiera się na szczegółowym sprawdzeniu wszystkich informacji – która wymaga poświecenia na tą czynność określonej ilości czasu (im dokładniejsza analiza, tym proces podejmowania decyzji trwa dużej i łatwiej jest wówczas wykryć nieprawidłowości).

Atakujący w wiadomościach phishingowych wykorzystują łączenie różnych technik inżynierii społecznej w celu spotęgowania emocji u odbiorcy wiadomości. Działanie emocjonalne sprzyja podejmowaniu szybkich, nieprzemyślanych decyzji (bez analitycznego sprawdzenia treści wiadomości, odnośników, np.) co znacznie zwiększa prawdopodobieństwo sukcesu ataku.

[60] Wiadomości wywierające presję czasu, zwykle w tytule mają wyrażenie: „pilne”, „ważne”, „uwaga”, „natychmiast”. Najczęściej tytuły mają one kolor czerwony, który kojarzy się z niebezpieczeństwem, pisane z wielkiej litery, co ma za zadanie spotęgować efekt.

I.3.4 Scenariusz ataku (jeden z możliwych)

Bazując na przedstawionych powyżej modelach (zarówno Cyber Kill Chain, MITRE ATT&CK jak i Phishing Kill Chain), można opracować skuteczny scenariusz ataku (jeden z możliwych) wykorzystujący opisane frazy, korzystające ze wskazanych powyżej technik do utworzenia infrastruktury, dostarczenia wiadomości ze złośliwą treścią i wykonaniem infekcji komputera ofiary.

1. Przygotowanie infrastruktury.

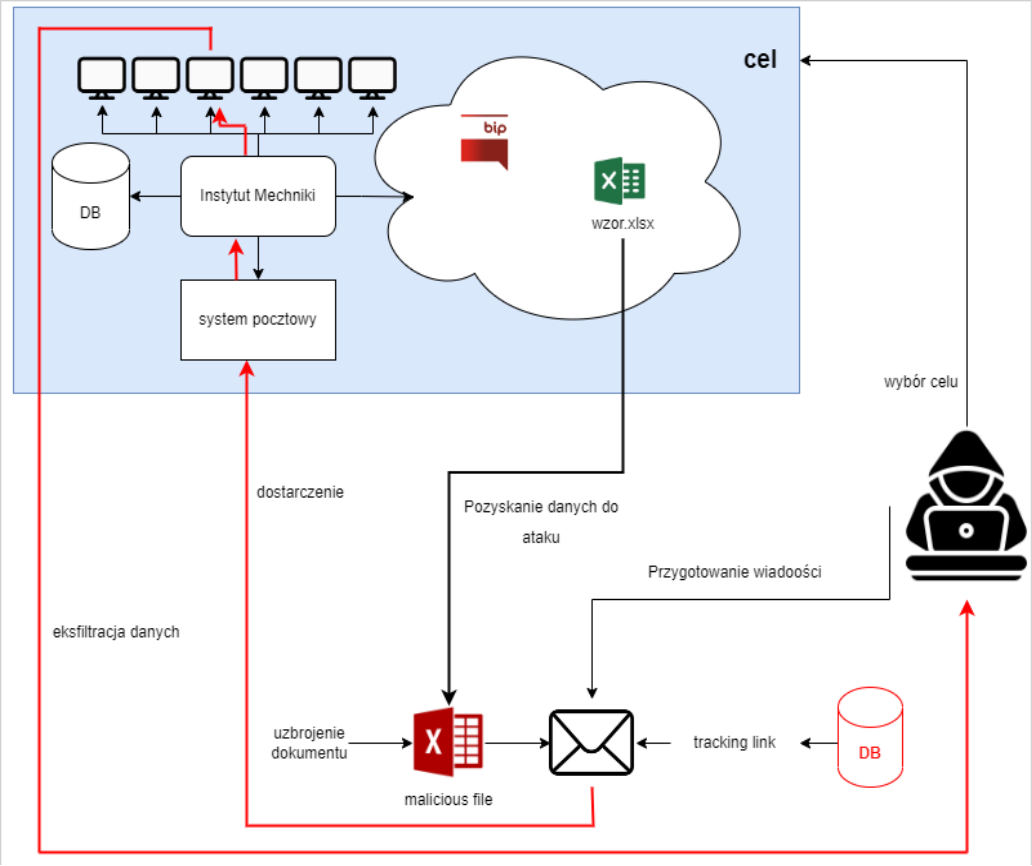
   Etap obejmuje instalację serwera poczty, koniecznych komponentów i niezbędnego oprogramowania. Atakujący może również skorzystać z usług hostingu lub wykupić odpowiedni serwer VPS. Wykaz niezbędnego oprogramowania do zainstalowania (kupienia) obejmuje:

   a. Serwer poczty z obsługą protokołu SMTP (przeznaczony do wysyłki wiadomości).
   b. PHP – niezbędne do utworzenia skryptów odpowiedzianych za wysyłkę wiadomości email.
   c. Python – przetwarzanie danych na potrzeby uzbrojenia wiadomości email w złośliwy załącznik lub treść.
   d. PowerShell – uzbrojenie plików.
   e. MySQl - przechowywanie informacji o celach ataku.
   f. Serwer www – do obsługi dokumentów HTML i PHP, wyświetlenia stron.
   g. Serwer C2 (Command & Control) – instancja Cobalt Strike[61] lub Metasploit.[62]

[61] Komunikacja z serwerem C2 opartym o oprogramowanie Cobalt Strike jest trudniejsza do wykrycia, więc to rozwiązanie będzie wykorzystywane w opisywanym scenariuszu.

[62] Metasploit – narzędzie służące do testów penetracyjnych i łamania zabezpieczeń systemów teleinformatycznych. Zawiera bazę gotowych exploitów oraz udostępnia interfejs, dzięki któremu można przygotowywać własne, korzystając z gotowych komponentów.

h. Serwer proxy – w bardziej zawansowanych scenariuszach jest to serwer bezpośrednio komunikujący się ze stacją ofiary, to do niego przesyłane są wyprowadzone dane. Serwer C2 służy jedynie do wysyłania złośliwych komend. Utrudnia to identyfikację infrastruktury atakującego i tym samym trudniejsze jest wdrożenie mechanizmu blokowania.



Rysunek 20. Scenariusz ataku phishingowego z wykorzystaniem uprzednio pozyskanych danych od celu ataku, źródło: opracowanie własne.

2. Wybór celu

W zależności od realizowanych potrzeb (szpiegostwo, szpiegostwo przemysłowe, oszustwa finansowe, kradzież danych, tożsamości), cele dobierane są pod aktualną potrzebę (lub wystawione zlecenie w przypadku grup APT). Na potrzeby scenariusza, przejęto założenie, że celem jest fikcyjny „Instytut Mechaniki" – rządowa komórka realizująca badania naukowe, finansowana z budżetu państwa, posiadająca własny serwis internetowy (stronę www z własną domeną). Zasoby sieciowe „Instytutu", mogą zawierać interesujące z punktu widzenia atakującego informacje, np.:

a. wykaz personelu z ich danymi kontaktowymi i zajmowanymi stanowiskami – niezbędne do przygotowania spersonalizowanego ataku (spear-phishing),
b. dokumenty firmowe, finansowe, kadrowe,
c. zdjęcia, loga, rysunki – możliwe do wykorzystania podczas spoofingu,
d. procesy technologiczne, patenty naukowe, np.

3. Pozyskanie danych

Wybrany cel jest instytucją publiczną, działającą w oparciu np. ustawy „Prawo zamówień publicznych". Z tego powodu zobowiązana jest do prowadzenia witryny w sieci Internet – Biuletyn Informacji Publicznej (BIP) znajdujący się pod adresem: bip.instytut.gov.pl[63]. Na witrynie tej, publikowane są np. ogłoszenia o przetargach na dostawę przedmiotów, zakup niezbędnych usług lub inne specyficzne ogłoszenia i informację. Ogłoszenie o przetargu zwykle zawierają, dokładną specyfikę poszukiwanego przedmiotu lub dokładny opis usługi, wykaz niezbędnych dokumentacji, jakie zgłaszająca się do przetargu firma powinna dołączyć do oferty, wykaz zezwoleń, certyfikatów, np.

Realizując obowiązki wynikające z ustawy „Prawo zamówień publicznych", Instytut opublikował ogłoszenie o wykonaniu usługi obsługi pojazdów mechanicznych (drobne naprawy, serwisowanie, np.). Wśród dokumentów niezbędnych do złożenia oferty znajduje się „Lista pracowników serwisujących", który ma mieć postać dokumentu Microsoft Excel (*.xls lub *.xlsx). Wzór takiego pliku znajduje się w opisie oferty. Atakujący pobiera przedmiotowy plik, który wykorzysta do przesłania wiadomości email.

Do przedmiotowego przetargu mogą zgłaszać się firmy świadczące usługi serwisu pojazdów lub pośredniczące w takich usługach. Atakujący poszukuje więc firmy, która:

a. Istnieje w rzeczywistości – można odnaleźć w sieci Internet jej nazwę i posiada wpis do KRS (również możliwy do weryfikacji w sieci Internet).
b. Posiada stronę internetową, na której można odnaleźć informacje o firmie (np. profil jej działalności) – nie musi być to jednak domena firmowa,

[63] Adres stworzony na potrzeby opisu scenariusza.

może być utworzona strona wizytówka w ramach serwisu internetowego pośredniczącego w wyszukiwaniu oferty i usług[64].

c. Na stronie nie widnieje kontaktowy adres email.
d. Prowadzi działalność zgodnie z wymaganiami określonymi w przetargu.

Na tej podstawie generowany jest adres email, zawierający w nazwie użytkownika fragment (lub całość) nazwy firmy, który zostanie wykorzystany do wysłania oferty.

4. Przygotowanie ładunku

   Pobrany w etapie poprzednim plik formatu Excel, jest uzbrajany w złośliwy kod, automatycznie uruchamiany w dokumencie po jego otworzeniu przez użytkownika (macro[65]). Wykorzystany zostanie kod PowerShell osadzony w funkcji macro dokumentu.

5. Dostarczanie wiadomości

   Wykorzystanie przygotowanej uprzednio infrastruktury w zakresie:

   a. Skryptu PHP służącego do przygotowania nagłówka wiadomości i wywołującego funkcję wysłania wiadomości email. Skrypt dokonuje spoofingu adresu nadawcy, by odbiorca wykonując sprawdzenia w sieci Internet mógł odnaleźć stronę wizytówkę firmy i skojarzyć nazwę z podrobionym adresem. Skrypt dodaje również spersonalizowany element (unikalna wartość) śledzący pozwalający na sprawdzenie czy użytkownik odczytał wiadomość.
   b. Serwera SMTP odpowiedzialnego za wysłanie przygotowanej wiadomości email na docelowy adres.

6. Monitorowanie zachowania użytkownika

   Faza oczekiwania. Odczytanie wiadomości wiąże się z nawiązaniem komunikacji z serwerem atakującego. Automatycznie uruchamiany skrypt bazujący na odczycie danych unikalnych wartości w komunikacji przychodzącej (wartości zostały zapisane w mechanizmie śledzącym), sprawdza czy potencjalna ofiara odczytała wiadomość email oraz zapisuje dane połączenia (USER-AGENT[66]).

[64] Częstym motywem wśród małych firm jest tworzenie strony-wizytówki w obszarach dużych serwisów internetowych – np. wśród pośredników i handlarzy używanymi samochodami, korzystającymi z serwisu otomoto, często tworzone są strony o adresie: nazwa_firmy.otomoto.pl

[65] Macro - zestaw poleceń przeznaczonych do natychmiastowego wykonywania e w celu automatyzacji pewnych czynności lub dokonania zmian w dokumentach bez interakcji z użytkownikiem.

[66] USER-AGENT – charakterystyczny nagłówek komunikacji przychodzącej do danego serwera, zawierający informacje o systemie operacyjnym, aplikacji, dostawy usług sieciowych, itp.

Dane te posłużą do przygotowania złośliwego oprogramowania. Dostosowanie złośliwego kodu do środowiska ofiary, znacznie utrudni jego wykrycie (np. przez oprogramowanie antywirusowe) lub zwiększy możliwości pozyskania danych z systemu ofiary.

7. Dostarczenie kolejnych złośliwych skryptów
   Zaszyty w dokumencie Excel kod, nawiązał komunikację z serwerem C2 i pobrał listę poleceń do wykonania w środowisku ofiary. Jednym z poleceń jest komenda PowerShelll, której zadaniem jest zapewnienie przetrwania złośliwemu oprogramowaniu (dodaniu zadania komunikacji w Harmonogramie Zadań systemu operacyjnego). Instancją serwera C2 jest Cobalt Strike, który nie utrzymuje stałego kanału łączności z komputerem ofiary, a jedynie w losowych interwałach czasowych odbiera dane i przekazuje kolejne polecenia do realizacji w zainfekowanym systemie.
8. Dalsze działania – zgodnie z zaplanowanym przez atakujących scenariuszem lub wynikających z uzyskanych danych, reakcji zespołu BLUE, np.

I.4 Kryteria uznania ataku za phishing

By uznać dany atak za phishing, jako kryteria należy wziąć pod uwagę:

1) Wykorzystanie inżynierii społecznej:
   a) zachęta do wykonania określonych zadań (nagroda),
   b) wywołanie poczucia niepewności, zagrożenia (groźba, szantaż),
   c) wytworzenie presji czasu.
2) Brak poprawności językowej treści wiadomości (błędna składnia językowa).
3) Podszywanie się pod inną osobą, instytucję (spoofing[67]).
4) Negatywny skutek wykonywanych sugerowanych działań dla odbiorcy wiadomości.

Przedstawione kryteria, posłużą do opisania modelu wiadomości phishingowej. Model ten posłuży do opracowania mechanizmów detekcji na podstawie technik

[67] Spoofing – atak polegająca na podszywaniu się pod inny element systemu informatycznego poprzez umieszczanie w sieci odpowiednio preparowanych pakietów danych lub niepoprawne używanie protokołów. W praktyce często wykorzystuje się techniki podszywania się pod inny adres email czy numer telefonu (wiadomości SMS/MMS).

implementacji ataku (wynikających z opisanego poniżej procesu ataku) oraz zidentyfikowanych cech phishingu.

I.4.1 Proces ataku phishingowego

Atak phishingowy możemy rozpatrywać jako proces – zbiór czynności następujących po sobie (lub w niektórych przypadkach prowadzonych równocześnie), które są wzajemnie ze sobą powiązane i finalnie prowadzą do realizacji celu ataku.

Model „Phishing Kill Chain" [18] zaprezentowany przez Chris'a Meidinger'a zawiera jednak pewną nieścisłość. Autor rozwiązania umieścił fazę „Deception" (z ang. „oszustwo") jako następującej po fazie rozsyłki wiadomości. Faza „Deception" określana jako oszukanie odbiorcy (potencjalnej ofiary) w celu realizacji zamysłu atakującego, może być realizowane zarówno po rozsyłce wiadomości (np. prowadzenie korespondencji, serowanie ofierze podrobionych serwisów, np.) jak i też obejmować proces przygotowania scenariusza ataku, formatowania wiadomości email zachęcającej do podjęcia pożądanych przez atakującego czynności ofiary. Faza ta realizowana jest w takcie przygotowań infrastruktury, w wiec tworzenia formy, treści i scenariusza ataku, wykorzystując inżynierię społeczną do manipulacji i oszustwa potencjalnej ofiary, odpowiada fazie drugiej – „Weaponization" modelu „Cyber Kill Chain".

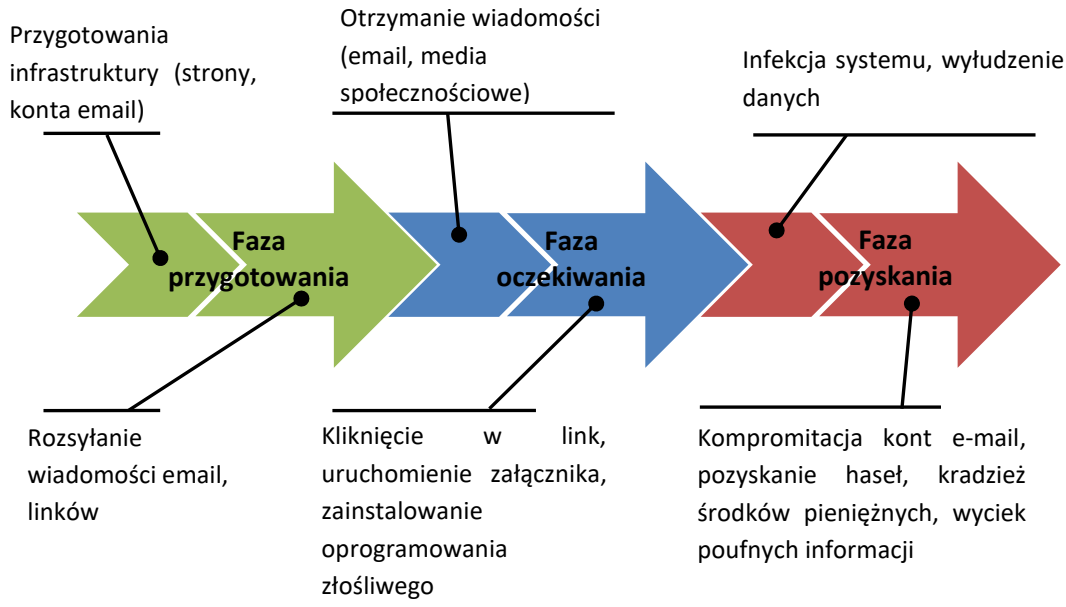
Z uwagi na różnorodność stosowanych przez atakujących technik, wykorzystywanie przez nich gotowych rozwiązań (zakupionych i skonfigurowanych serwerów wraz z narzędziami do przeprowadzenia z góry skonfigurowanego scenariusza ataku[68]), powyższy model – bazujący na „Cyber Kill Chain" wykorzystywanym do pełnego opisu ataku APT, wykorzystującego wszystkie elementy modelu – mimo że w pełni wyczerpująco opisuje całkowity proces ataku, można uprościć do modelu składającego się z trzech głównych faz:

1. faza przygotowania,
2. faza oczekiwania,

[68] Opisany model funkcjonuje jako oferowana usługa nazwana „phishing as a service", zawierająca przygotowanie kompletnej infrastruktury pod planowany atak, obejmująca projektowanie strony internetowej, szaty graficznej wiadomości, opracowanie technik inżynierii społecznej, konfigurowanie serwerów pocztowych, tworzenie złośliwych linków, dokumentów, załączników, zbieranie i analiz danych uzyskanych podczas ataku.

3. faza pozyskania.

Uproszczenie to pozwala w lepszy sposób zrozumieć elementy ataku, jak również dzięki zredukowaniu pełnego łańcucha ataku do trzech etapów, można określić kluczowe momenty, których wykrycie i zablokowanie uniemożliwi powodzenie całości ataku.



Rysunek 21 – Fazy ataku phishingowego – model uproszczony

Model (Rysunek 21 – Fazy ataku phishingowego – model uproszczony) zawierający trzy główne fazy ataku (przygotowanie, oczekiwane, pozyskanie), składają się z kilku elementów, które są konieczne do zrealizowania celu kampanii phishingowej: pozyskania danych, kompromitacji konta czy zainfekowania komputera użytkownika złośliwym oprogramowaniem. Poszczególne elementy trzech faz, można przyporządkować do faz modelu „Phishing Kill Chain". Model ten nie wymaga jednak realizacji pełnego cyklu[69], by można było uznać dany atak za zakończony sukcesem.

Faza przygotowania obejmuje etap przygotowania niezbędnej infrastruktury atakującego do przeprowadzenia kampanii phishingowej. W tej fazie konfigurowane są serwery pocztowe służące do masowej rozsyłki wiadomości phishingowych, zawierających odnośnik do witryny internetowej. Faza ta obejmuje również przygotowanie złośliwych załączników (skrypty wykonywalne, programy przykrywki[70]).

Elementy modelu „Phishing Kill Chain" wykorzystywane w tej fazie:

[69] Poprzez pełny cykle ataku phishingowego (analogicznie jak w przypadku modelu Cyber Kill Chain) należy rozumieć realizację wszystkich etapów – od fazy „Reconnaissance" do fazy „Action".
[70] Program przykrywka – określenie stosowane do rodzaju złośliwego oprogramowania, które

1. targeting,
2. delivery,
3. deceive.

W fazie oczekiwania, potencjalne ofiary otrzymały wiadomości (email, poprzez komunikatory, media społecznościowe) oraz wykonały sugerowane w nich działania, do których możemy zaliczyć:

1. Kliknięcie w link przekierowujący do złośliwej strony (wyłudzającej dane logowania do poczty, bankowości elektronicznej, przechwytującej uwierzytelniające kody SMS),
2. Wpisanie poufnych danych użytkownika na spreparowanej (phishingowej stronie):
    a. danych logowania do serwisu bankowości elektronicznej (np. pod pozorem dopłaty/zapłaty za usługę),
    b. kodów autoryzujących wykonanie rzekomej dopłaty,
    c. danych teleadresowych,
    d. numeru PESEL, serii i numeru dokumentów tożsamości,
    e. kodu CVC/CVV[71] kart kredytowej,
    f. innych danych osobowych identyfikujących ofiarę.
3. Pobranie i uruchomienie złośliwego załącznika wiadomości, który infekuje stację komputerową ofiary i w zależności od przeznaczenia wykonujący zaprogramowane funkcje na stacji ofiary (np. wykradania adresów email, przechwytywanie naciśnięć przycisków klawiatury).

   Elementy modelu „Phishing Kill Chain" wykorzystywane w tej fazie:

    1. click,
    2. surrender.

Faza pozyskania jest ostatnim etapem pełnego cyklu zakończonego sukcesem ataku. Wiadomość została dostarczona, użytkownik podjął sugerowaną akcję (kliknięcie w odnośnik, pobrania i zainstalowanie załączonego złośliwego oprogramowania). W tej fazie atakujący uzyskują dostęp danych użytkownika.

Elementy modelu „Phishing Kill Chain" wykorzystywane w tej fazie:

1. extraction,

[71] CVC/CVV – funkcja zabezpieczająca transakcję kartami płatniczymi, składająca się z ciągu trzech cyfr, mająca na celu ograniczenie oszustw.

2. action.

Dane pozyskane w wyniku kompromitacji systemu teleinformatycznego użytkownika (np. adres email, imię i nazwisko, nieautoryzowany i skryty dostęp do jego skrzynki pocztowej), mogą z powodzeniem być wykorzystane do przygotowania innej infrastruktury teleinformatycznej, służącej do dalszych ataków. Faza ta zamyka więc cykl i staje się początkiem nowego.

Każda z wymienionych powyżej faz modelu uproszczonego, posiada swoje charakterystyczne elementy, które muszą być zrealizowane, by obserwowane zdarzenie określić jako atak phishingowy. Nie musi to oznaczać sukcesu – z punktu widzenia atakujących. Dostarczenie wiadomości email z odnośnikiem prowadzącym do strony phishingowej jest takim charakterystycznym elementem, jednakże, dopóki użytkownik nie kliknie w odnośnik i nie wpisze swoich danych w fałszywy panel logowania, dotąd nie zrealizował w pełni oczekiwanego przez atakujących działania, atak więc z ich punktu widzenie jest uznany za porażkę.

Uproszony w stosunku do modelu „Phishing Kill Chain" jest uznawany jako opis łańcucha ataku phishingowego w środowisku osób związanych z cyberbezpieczeństwem (pracownicy CSIRT, SOC-ów[72] branżowych, np.). Jak wspomniano „Cyber Kill Chain" a więc i bazujący na nim „Phishing Kill Chain" jest kompleksowym opisem ataku od początku do końca, uwzględniający realizację kolejno po sobie następujących etapów.

Model MITRE ATT&CK opisuje stosowane techniki ataku (w przypadku phishingu jedna z trzema subtechnikami), lecz nie wymaga realizacji wszystkich etapów w ustalonym porządku. W ataku phishingowym natomiast realizacja danego etapu ataku, zależna jest od zakończonej sukcesem realizacji etapu poprzedniego – dlatego właśnie model bazujący na klasycznym „Cyber Kill Chain" lepiej odzwierciedla cykl ataku phishingowego i na nim właśnie, w dalszej części będzie bazował opis. Również ogólność opisu analizy subtechnik nie zapewnia wystarczającej szczegółowości do identyfikacji wszelkich cech ataku niezbędnych do zbudowania modelu pozwalającego na jego skuteczną detekcję.

[72] SOC (ang. Security Operations Centre) – koncepcja Centrum Operacyjnego Cyberbezpieczeństwa, monitorującego stan wydzielonego fragmentu cyberprzestrzeni (sieci), wykrywanie zagrożeń, kreowanie polityki bezpieczeństwa.

I.4.2 Spam

Opisując atak phishingowy, należy wziąć również pod uwagę zjawisko określane jako spam. Pod tym pojęciem należy rozumieć otrzymywanie przez danego użytkownika niechcianych lub niepotrzebnych (z jego punktu widzenia) wiadomości email[73]. Z punktu widzenia użytkownika, otrzymywanie wiadomości typu spam (ze względu na ich masowość) jest szkodliwe i uciążliwe, jednakże nie stanowi bezpośredniego zagrożenia, tak jak w przypadku wiadomości phishingowych.

Mianem niechcianych wiadomości można określić, wszystkie otrzymywane przez użytkownika wiadomości typu:

1. informacje o konkursach, loteriach, możliwych wygranych,
2. oferta usług dla dorosłych,
3. oferty promocyjne, bony, rabaty,
4. ofert reklamowe.

Cechą charakterystyczną spamu jest jego masowość i brak personalizacji (w odróżnieniu od phishingu) w otrzymywanych wiadomościach. Ofert promocyjne i reklamowe często przyjmują charakter tzw. „agresywnego marketingu[74]", co daje podstawę do przypuszczeń, że zastosowanie się do przesłanej sugestii w danej wiadomości, może przynieść zyski nadawcy, niewspółmierne do otrzymywanych przez odbiorcę korzyści, których zazwyczaj brak.

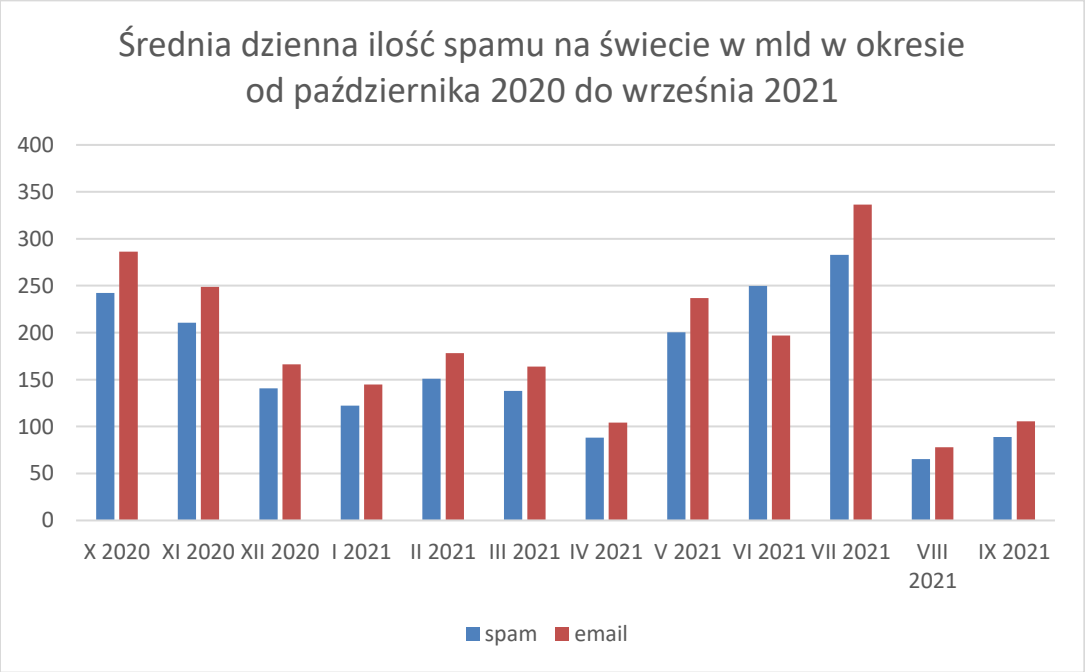
Szkodliwość spamu objawia się w jego masowym charakterze. Spam może powodować:

a. spowolnienie działania serwerów pocztowych z uwagi na konieczność przetworzenia dużej ilości danych,
b. zwiększenie zajętości przestrzeni dyskowych z uwagi na koniczność przechowywania dużej ilości niechcianych wiadomości,
c. ograniczenie przepustowości łącz internetowych, z uwagi na konieczność przesyłu pomiędzy serwerami pocztowymi dużego wolumenu danych,

[73] Zjawisko to dotyczy wszystkich form komunikacji elektronicznej (np. wiadomości tekstowe SMS, komunikatory, wiadomości rozsyłane za pomocą mediów społecznościach, itp.), jednakże z uwagi na wykorzystanie wiadomości email jako najpopularniejszą formę, w niniejszej pracy skupiono się na niechcianych wiadomościach rozsyłanych właśnie za pomocą poczty elektronicznej (email).

[74] Pod pojęciem agresywnego marketingu należy rozumieć, intensywne treści reklamowe w sposób nachalny zachęcające do zakupu reklamowanego produktu. Forma ta często wykorzystywana jest w reklamach kryptowalut, nieuczciwych platformach transakcyjnych czy suplementów diety.

d. zwiększenie nakładu pracy dla osób zawodowo zajmujących się obsługą poczty email, z uwagi na konieczność usuwania wiadomości typu spam lub/i filtrowania otrzymywanych wiadomości,
e. naruszenie prywatności odbiorców, poprzez umieszczenie w wiadomości treści obraźliwych, pornograficznych, nieodpowiednich dla dzieci, np.,
f. zwiększenie kosztów obsługi poczty elektronicznej przez operatorów i dostawców usług (zwiększenie wolumenu danych na przechowywanie wiadomości, zwiększenie przepustowości łączy, zwiększenie mocy obliczeniowych serwerów pocztowych),
g. z uwagi na ograniczanie zasięgu spamu i eliminacji zagrożenia spamem, poczta elektroniczna została pozbawiona niektórych pierwotnych funkcji (np. otrzymywania przez nadawcę potwierdzeń dostarczenia i odczytania wiadomości przez odbiorcę – funkcjonalność ta została zastąpiona przez tzw. „mechanizm śledzący" i zaadaptowana w wiadomościach phishingowych).



Rysunek 22. Średnia dzienna ilość spamu na świecie w mld w okresie od października 2020 do września 2021. Źródło: Statista, link: https://www.statista.com/statistics/1270424/daily-spam-volume-global/, dostęp: [15.12.2022].

Spam jako zjawisko starsze niż phishing[75] stało się bazą wykorzystywaną w procesie ataku phishingowego w zakresie:

[75] Pierwszy odnotowany w historii sieci spam został wysłany przez Einara Stefferuda 1 maja 1978 roku (źródło: Wikipedia, link: https://pl.wikipedia.org/wiki/Spam, dostęp: [14.12.2022]).

a. Technik pozyskiwania odbiorców – kluczowym elementem obu typów wiadomości (spam i phishing) jest rzeczywisty adres email odbiorcy, konieczne jest wiec jego pozyskanie. Najpopularniejszą metodą pozyskiwania adresów email jest tzw. Harwester – przeszukiwanie zasobów forów, stron i serwisów internetowych w poszukiwaniu umieszczonych w ich zasobach adresów poczty elektronicznej. Równie popularną metodą zbierania adresów email jest pozyskiwanie baz wycieków zawierających imię i nazwisko oraz adres email użytkowników. Obie te techniki zaadoptowane zostały do procesu ataku phishingowego.
b. Przygotowania infrastruktury do masowej rozsyłki wiadomości. Rozsyłka dużej ilości wiadomości typu spam (średnia dzienna ilość spamu na świecie we wrześniu 2022 rok wyniosła około 2 miliardy wiadomości [31]), wymaga odpowiedniej infrastruktury (serwer pocztowy z odpowiednim oprogramowaniem, odpowiednie parametry łącza internetowego, np.).
c. Wykorzystania elementów inżynierii społecznej. Niezamawiane i niechciane oferty reklamowe, przedstawiane są z wykorzystaniem elementów manipulacji i oszustwa w celu zwiększenia szansy na podjęcie przez odbiorcę wiadomości pożądanej akcji (np. zakupieniu danego produktu). Metody te wykorzystywane są masowo w atakach phishingowych w celu przekonania potencjalnej ofiary do wykonania określonych działań.
d. Anonimizacji rzeczywistego nadawcy. Podstawą działalności firm świadczących usługi rozsyłania wiadomości typu spam jest anonimowość. Metodą wykorzystywaną przez osoby rozsyłające wiadomości typu spam, jest wykorzystanie błędów w konfiguracji serwerów pocztowych (tzw. Open relay[76]), co znacznie utrudnia lokalizację faktycznego nadawcy. Metoda ta została zaadaptowana w phishingu.

I.5 Skuteczność phishingu

Wynikiem ataku phishingowego może być powodzenie lub porażka. Za powodzenie można uznać realizację ostatniej, zakończonej sukcesem fazy ataku (faza pozyskania). Scenariusz ataku, może jednak zawierać inne założenia (np. śledzenie aktywności

[76] Określenie serwera pocztowego, którego oprogramowanie nie jest zabezpieczone przed nieautoryzowanym wykorzystaniem przez osoby niepowołane do wysyłki poczty elektronicznej

użytkownika, sprawdzenie aktywności danego konta email), wobec czego nie wszystkie fazy ataku są zakładane i realizowane w danym scenariuszu, a co za tym idzie nie wszystkie fazy modelu ataku phishingowego muszą być zrealizowane by atak można było uznać za pomyślny. Zdarzenia, których wystąpienie można (w zależności od zakładanego scenariusza) uznać jako zdarzenia świadczące o powodzeniu ataku phishingowego (danego etapu), można sklasyfikować jako:

1. Skuteczne doręczenie wiadomości phishingowej (wraz z ewentualnymi załącznikami) do końcowego odbiorcy - zdarzenie to obejmuje również proces przetwarzania wiadomości przez serwery pocztowe, które nie sklasyfikują danej wiadomości jako phishing lub spam.
2. Otwarcie wiadomości phishingowej przez użytkownika - zdarzenie to obejmuje również skuteczne doręczanie wiadomości do końcowego odbiorcy (patrz punkt 1). Scenariusz może zakładać zbadanie aktywności konta użytkownika i jego działań – wiadomość phishingowa zawiera wówczas skrypt śledzący użytkownika, informujący atakującego o fakcie otwarcia (przeczytania) dostarczonej wiadomości.
3. Kliknięcie w odnośnik – zdarzenie to obejmuje również poprzednie punkty (skuteczne doręczenie i otwarcie wiadomości). Kliknięcie w donośnik przenosi użytkowni ka do przygotowanej witryny internetowej.
4. Podanie danych sugerowanych na odwiedzonej stronie internetowej – która – w zależności od scenariusza ataku – wyłudza wrażliwe dane użytkownika (login, hasło, kody SMS, numery karty kredytowej), zachęca do pobrania i zainstalowania określonego oprogramowania (zwykle hostowanego na danej witrynie) lub przekierowuje użytkownika do innego miejsca w sieci Internet.
5. Pobranie i zainstalowanie złośliwego oprogramowania ze strony, do której prowadził odnośnik w otrzymanie wiadomości email – również zalicza się do zakończonego sukcesem ataku phishingowego, nawet w przypadku gdy złośliwe oprogramowanie zostało rozpoznane i zneutralizowane przez mechanizm antywirusowy (o ile występuje) po stronie ofiary. Pobranie i uruchomienie złośliwego oprogramowania ze strony, do której prowadził odnośnik URL jest sukcesem z powodu:
    a. Wiadomość została finalnie dostarczona do użytkownika końcowego – jeżeli istniał w systemie pocztowym mechanizm antyphishingowy, wiadomość została przez niego przepuszczona dalej.

b. Użytkownik kliknął zawarty w niej odnośnik – żaden element wiadomości nie wzbudził podejrzeń użytkownika co do możliwości jest złośliwego charakteru (wykorzystanie inżynierii społecznej).
c. Użytkownik pobrał dostępne na danej stronie oprogramowanie – systemy antyphishingowe dostawcy usług (o ile są wdrożone po stronie dostawcy) nie wykryły, że dana witryna może być złośliwa[77], użytkownik wykonał sugerowane w wiadomości i na stronie polecenia.
d. Użytkownik uruchomił pobrane oprogramowanie – wykonanie wszystkich sugerowanych poleceń przez ofiarę, jest to najbardziej pożądana sytuacja przez atakujących, gdyż wszystkie elementy ataku phishingowego zostały spełnione i przy braku odpowiedniego oprogramowania antywirusowego, system ofiary zostanie zainfekowany.

I.6 Świadomość użytkowników

Czynnikiem sprzyjającym większej skuteczności ataku phishingowego jest również słaba świadomość wśród społeczeństwa odnośnie tego typu zagrożenia – raport firmy Proofpoint [32] za 2020r. wykazał, że 61% ankietowanych umie opisać czym jest zjawisko phishingu, lecz jednak mniej niż 35% z badanej populacji, potrafi zidentyfikować różnego rodzaju zagrożenia (rodzajów phishingu), potrafi opisać co to jest, i w jaki sposób rozpoznać podejrzaną wiadomość. Badanie przeprowadzono na obywatelach Stanów Zjednoczonych, ale z powodzeniem można je odnieść również do innych krajów[78]. Dane te dobitnie ukazują skale problemu – większość społeczeństwa nie zna zagrożeń występujących w sieci Internet.

Podobne badanie przeprowadzone zostało przez Agencję Badań Rynku i Opinii SW Research, w marcu 2019 roku, na grupie 800 dorosłych Polaków, na zlecenie Nest Bank[79]. Z badania wynika, że 30% Polaków nie wie czym jest phishing, 32% ma wrażenie, że wie, ale nie jest pewna. Dane te wskazują, że pomimo aktywności tematyki phishingu w mediach (Rysunek 23), aż 62% wśród badanej grupy mogła by nie rozpoznać

[77] Niektórzy producenci przeglądarek internetowych, wyposażają je w moduły weryfikujące czy strona, pod którą chce przejść dany użytkownik nie jest złośliwa (źródło: https://www.dobreprogramy.pl/google-chrome-ostrzeze-przed-phishingiem-zaproponuje-przejscie-pod-wlasciwy-adres,6628556917528705a)
[78] Brak dostępnych badań na temat świadomości społeczeństwa o zagrożeniach w sieci Internet.
[79] https://nestbank.pl/

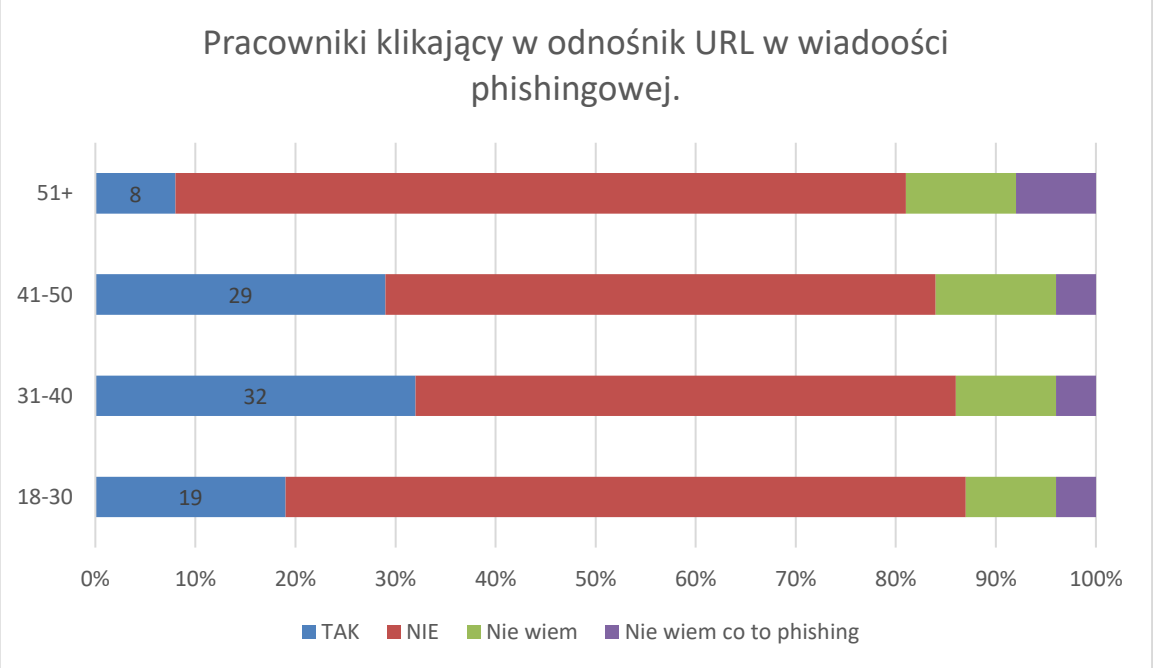
ataku, gdyby byli jego celem. Przekłada się to bezpośrednio na ciągły wzrost ilości ataków.

Rysunek 23. Popularność tematyki ataku phishingowego w mediach (kolor czerwony) do kryteriów wyszukiwania słowa „phishing" (kolor niebieski) przez użytkowników w okresie lipiec 20222 – czerwiec 2023.

Zachowanie użytkowników i sposób korzystania przez nich z zasobów sieci Internet może wpływać na ich podatność na kierowany przeciwko nim atak phishingowy. Dominującym podejściem do zwalczania phishingu jest wykorzystanie komputerów do automatycznego wykrywania ataku bazując na technicznych wskaźnikach phishingu (np. nieprawidłowy odnośnik URL, nieprawidłowe przekierowanie, adres URL niepokrywający się z adresem domenowym wiadomości, np.). Drugim podejściem do wykrywania i zwalczania ataku phishingowego jest budowanie świadomości o zagrożeniach wśród użytkowników sieci Internet, ich edukacja i testowanie ich wiedzy. Podejście to w literaturze i badaniach nad zjawiskiem phishingu jest często pomijane. Istota świadomości o zagrożeniach wśród użytkowników została przedstawiona przez [33] gdzie wykazano, że pomimo zastosowania automatycznych mechanizmów wykrywania phishingu, wyświetlania komunikatów ostrzegawczych, użytkownicy często nie rozumieją ich znaczenia, a niespójne pozycjonowanie treści w różnych przeglądarkach internetowych utrudnia zadanie zidentyfikowania strony phishingowej. Badania prowadzone przez [34] wskazują że 53% ich uczestników nadal próbowało zalogować się do witryny phishingowej nawet po wyświetleniu ostrzeżenia

informującego o zagrożeniu. Prowadzony w dalszej części badania eksperyment, polegający na usunięciu szyfrowanego protokołu HTTPS[80] i zastąpienie go nieszyfrowanym NP.[81] nie miało wpływu na zachowanie użytkowników, którzy nadal chętnie wpisywali poufne dane na niezabezpieczonej stronie. Dodatkowe usunięcie ikonek informujących o uwierzytelnieniu połączenia, zwiększyło wskaźnik użytkowników wprowadzających swoje dane do 97%.

Przeprowadzona w 2021 roku ankieta firmy Tessian [35] wśród pracowników biurowych, wskazuje, że średnio około 20% pracowników (w każdej z badanych grup wiekowych) kliknęło w odnośnik URL w wiadomości phishingowej Wyniki jednoznacznie wskazują, że w różnych grupach wiekowych, pracownicy inaczej podchodzą do kwestii weryfikacji otrzymywanych danych – wymusza to konieczność dostosowania procesu szkolenia o zagrożeniach w cyberprzestrzeni do konkretnych grup wiekowych. Ta sama ankieta [35] wskazuje, że prawie 33% użytkowników, bardzo rzadko lub wcale nie myśli o bezpieczeństwie podczas codziennego korzystania z komputera i sieci Internet.

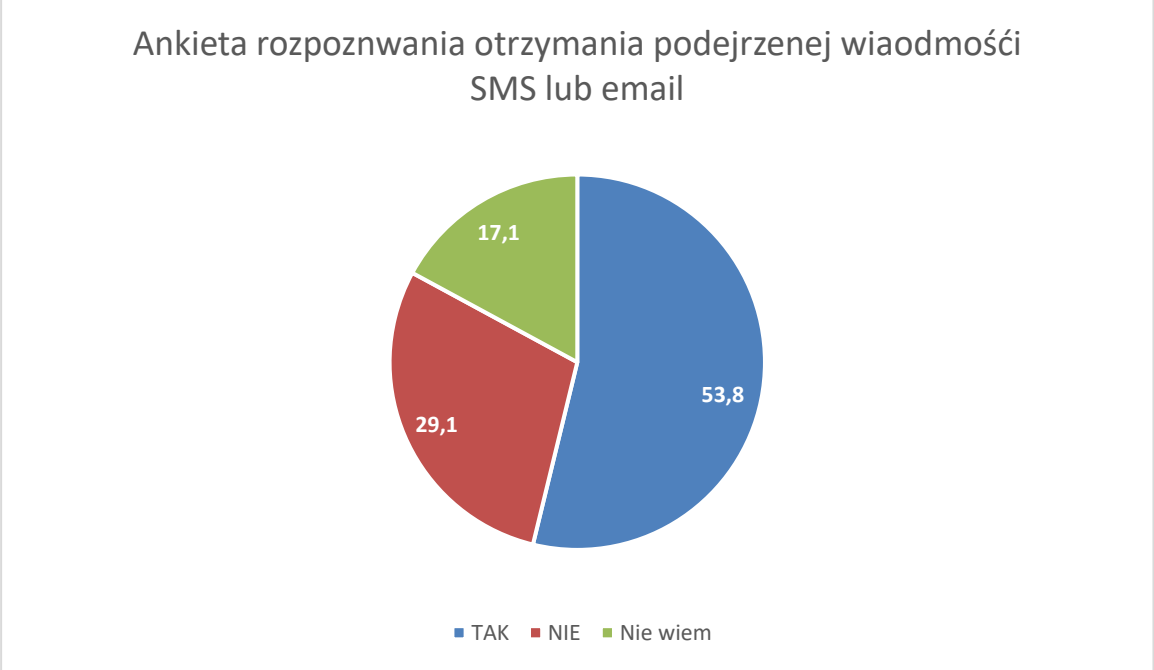


Rysunek 24. Pracownicy klikający odnośnik URL w wiadomości phishingowej, opracowanie własne na podstawie: https://www.tessian.com/research/the-psychology-of-human-error/ [dostęp: 29.12.2022].

[80] HTTPS (ang. Hypertext Transfer Protocol Secure) – szyfrowana wersja protokołu HTTP.
[81] HTTP (ang. Hypertext Transfer Protocol) – protokół przesyłania dokumentów hipertekstowych wykorzystywany na stronach www.

Wyniki prowadzonych badań [33] [34] wskazują, że pomimo zastosowanych technicznych środków w wykrywaniu phishingu i ostrzeganiu użytkowników o potencjalnym zagrożeniu, brak skutecznej edukacji, obniża efektywność mechaników obrony przed atakiem. Uzupełnienie mechanizmów detekcji o budowanie świadomości o zagrożeniach poprawi kompleksowo bezpieczeństwo organizacji i zwiększy zdolność do wykrywania symptomów ataku phishingowego.

Budowa świadomości pracowników celem zwiększenia poziomu bezpieczeństwa powinna być procesem ciągłym w danej organizacji, z uwagi na często zmieniające się techniki wykorzystywane w atakach phishingowych, gdyż wg raportu [36] 30% ankietowanych doświadczyło próby wyłudzenia danych, a 43% obawiało się, że padnie ofiarą phishingu, jednocześnie duża grupa ankietowanych twierdziła (około 80%), że potrafi wskazać fałszywą wiadomość. Wskazuje to, że prowadzone kampanie medialne powodują wzrost świadomości – szczególnie wśród młodszych ankietowanych – odnośnie występowania zagrożeń w sieci Internet.



Rysunek 25. Ankieta rozpoznania otrzymania podejrzanej wiadomości – „Czy w czasie pandemii koronawirusa otrzymałeś/aś podejrzany e-mail, SMS bądź telefon skłaniający do podjęcie działań związanych z udostępnieniem danych?" źródło: https://krd.pl [dostęp: 29.12.2022]

W marcu 2023 roku firma Darktrace zleciła przeprowadzenie badania [27] wśród pracowników Wielkiej Brytanii, Stanach Zjednoczonych, Francji, Niemcy, Australii i Holandii odnośnie możliwości identyfikacji wiadomości phishingowych.

1. 79% wdrożonych filtrów antyspamowych błędnie sklasyfikowało ważne wiadomości jako niechciane i nie dostarczyło do skrzynek odbiorców,
2. 70% pracowników zauważyło wzrost ilości wiadomości phishingowych ,
3. 1 na 3 pracowników padło w przeszłości ofiarą ataku phishingowego,
4. Pracownicy wskazali na 3 cechy wiadomości email, które zmuszą ich go głębszej weryfikacji pod kątem możliwego oszustwa:
    a. błędy, słaba znajomość gramatyki (w 61% przypadków),
    b. nieznany nadawca lub nieoczekiwana treść (w 61% przypadków),
    c. prośba o kliknięcie w link lub uruchomienia załącznika (w 68% przypadków).

Dane te wskazują, że wśród pracowników korporacji międzynarodowych istnieje pewien poziom świadomości na temat zagrożeń występujących w sieciach komputerowych, jednakże w dalszym ciągu – wraz z obserwowanym wzrostowym trendem ataków phishingowych – wielu (w tym również i kluczowych) pracowników jest i będzie ofiarą phishingu. Świadomość użytkowników wpływa co prawda na podniesienie poziomu bezpieczeństwa, jednakże w dalszym ciągu, przy braku dostatecznych technicznych środków wykrywających i blokujących atak, stanowi istotne zagrożenie dla bezpieczeństwa danych samych użytkowników jak i organizacji, w których pracują.

Badanie to ujawniało również problem, z automatycznymi mechanizmami dotychczas klasyfikującymi wiadomości i oznaczających ich jako podejrzane/niebezpieczne. Duży procentowy odsetek wiadomości (79% [27]), którym błędnie przypisano ich rzekomo niebezpieczny charakter wskazuje, że obecne mechanizmy nie potrafią prawidłowo klasyfikować wiadomości, lub też działają na niewłaściwych wskaźnikach / cechach lub źle sformułowanych założeniach. Przedstawione w poniższym rozdziale wskaźniki, są konsekwencją zidentyfikowania technik i metod używanych przez atakujących i posłużą do opracowania modelu detekcji phishingu.

V. 7 Problem badawczy oraz teza rozprawy

Przedstawione w niniejszym rozdziale zjawisko phishingu przynosi negatywne skutki zarówno zwykłym użytkownikom jak i całym firmom. Z tego powodu podjęto szereg prac związanych z opracowaniem nowych, skutecznych metod detekcji. Wśród

najczęściej używanych wymienić można podejście oparte na tworzeniu list dostępowych, regułach detekcji, wykorzystaniu algorytmów genetycznych czy też wykorzystanie uczenia maszynowego. Wszystkie te podejścia zostaną szczegółowe omówione w kolejnym rozdziale, z uwzględnieniem ich zastosowania oraz wykazaniem ich słabych stron. Z uwagi na zidentyfikowane wady i słabości istniejących podejść wykrywania phishingu, istnieje potrzeba opracowania metody, która będzie ich pozbawiona. Nowa metoda musi więc obejmować dogłębna analizę w celu określenia czy użytkownik ma do czynienia z atakiem phishingowym czy też nie. W tym celu należy zwrócić uwagę na możliwe automatyczne wykrywanie, bez konieczności ingerencji człowieka, dostosowaniu się modelu do zmian sposobów prowadzenia ataku oraz wykrywającego całe spektrum technik wykorzystywanych (lub możliwych do wykorzystania w przyszłości) przez atakujących. Rozpatrywany w niniejszej rozprawie problem naukowy polega na zaproponowaniu nowej metoda detekcji wiadomości phishingowych, obejmującej następujące elementy:

1. analizę porównawczą wiadomości email (pod kątem podobieństwa czasowego, tematu, treści, adresów),
2. analizę cech (odpowiednich pól),
3. powtarzalność schematu,
4. analizę treści (np. wykrywanie błędów pisowni, błędów językowych), w tym analizę odnośników URL w niej zawartych,
5. identyfikację szantażu, wymuszenia okupu,
6. analizę reputacji domen (z odnośników URL, domen z adresów email nadawcy, zwrotnego, np.),
7. wykorzystanie metod data mining do wykrycia nieznanych zależności i schematów, oraz wsparcie uczenia maszynowego do rozwiązywania problemów data mining podczas analizowania zbiorów danych.

Można sformułować następująca tezę; połączenie powyżej opisanych sposobów detekcji pozwoli na wykrywanie wcześniej niestosowanych schematów ataku, używanie nieznanych do tej pory wzorców – co eliminuje wady klasycznych metod wykrywania phishingu.

Rozdział II – Analiza metod detekcji phishingu

Rozwój oprogramowania powoduje, że zwiększa się też coraz bardziej powierzchnia ataku (więcej komponentów, więcej urządzeń do sprawdzenia i zabezpieczenia). Im większa powierzchnia ataku, tym więcej narzędzi potrzeba do skutecznej prewencji i detekcji. Rozbudowane narzędzia prewencji i detekcji powodują wiele uciążliwości dla użytkowników podczas normalnego użytkowania (wysokie obciążenie systemu informatycznego, ponadprzeciętne zużycie pamięci RAM, nadmierne wykorzystanie procesora co przekłada się to na odczuwalny spadek wydajności systemu dla użytkownika, a i tak nie gwarantuje 100% skuteczności). Może rodzić to skłonność do wyłączania (częściowo lub całkowicie) rozwiązań by przyspieszyć działanie komputera i zwiększyć komfort pracy.

Powoli więc odchodzi się od modelu prewencji do modelu detekcji. S. H. Apandi, J. Sallim, R. M. Sidek w swojej pracy [37], wykazują że, wykrywanie phishingu jest ważniejsze dla rozwiązania i zrozumienia ataku phishingowego niż modele prewencyjne. Nie jesteśmy w stanie przewidzieć wszystkich powierzchni ataku, przy stale rozwijającym się i rozbudowującym się oprogramowaniu użytkowym. Dlatego techniczna detekcja[82] ataków staje się jedną z najważniejszych elementów bezpieczeństwa systemów teleinformatycznych.

Obecnie w literaturze (wyszczególnionej dla każdej z opisywanych poniżej grup) wyróżnić można wiele różnych podejść do problemu technicznego wykrywania ataku phishingowego. Problemowi temu poświęcono wiele publikacji naukowych i opracowań technicznych. Omawiane w dostępnej literaturze metody oscylują wokół pięciu głównych grup:

1. budowie list – białe listy, czarne listy, szare listy (tematyka szeroko opisywana w pracach np.: [38], [39], [40], [41], [42], [43], [44]),
2. tworzenia reguł detekcji na podstawie wcześniej przeprowadzonych i zbadanych ataków (opisane w pracach np.: [45], [46], [47]),
3. wykorzystanie algorytmów genetycznych do tworzenia i modyfikacji reguł detekcji (opisane w pracach np.: [48]),

[82] Pod pojęciem detekcji technicznej ataku phishingowego należy rozumieć możliwość automatycznego (bez ingerencji człowieka)

4. wykorzystaniu metod data mining do wykrywania wzorców ataku phishingowego i tworzenia na tej podstawie reguł detekcji z wykorzystaniem uczenia maszynowego (opisane w pracach np.: [49], [50], [51]),
5. budowaniu świadomości wśród użytkowników (opisane w pracach np.: [52], [53]).

II.1 Blacklisting, whitelisting, greylisting

Podejście polegające na tworzeniu list z którego korzystają narzędzia bezpieczeństwa podczas korzystania przez użytkownika z komputera i pracy w sieci Internet. Podejście to jest dobrze opisane w literaturze – opisują to prace np. Wang, Y., Agrawal, R., Choi, B.Y ([38]), A. Jain, B. B. Gupta ([39]), W. Han, Y. Cao , E. Bertino, J. Yong ([40]), P. Prakash; M. Kumar; R. R. Kompella; M. Gupta ([41]), M. Felegyhazi, Ch. Kreibich, V. Paxson ([42]), S. Sheng , B. Wardman, G. Warner, L.F.Cranor, J. Hong, C. Zhang ([43], G. Xiang, J. Hong, C. P. Rosé, L. F. Cranor [44]).

Wyróżnia się dwa zasadnicze podejścia w budowaniu list:

a. Whitelist (biała lista) – podejście polegające na stworzeniu listy z dozwolonymi (tzw. „godnymi zaufania") adresami URL i IP, z którymi użytkownicy danej sieci lokalnej mogą się komunikować. Ruch sieciowy do wszystkich pozostałych adresów zostaje zablokowany na wyjściu z sieci.
b. Blacklist (czarna lista) – podejście polegające na stworzeniu listy z niedozwolonymi adresami URL i IP, z którymi komunikacja jest niebezpieczna i użytkownicy nie powinni korzystać z tych zasobów. Urządzenia bezpieczeństwa (np. firewall) mogą automatyczni blokować komunikację z adresami występującymi na czarnej liście. Miarą skuteczności narzędzia jest częstotliwość aktualizacji oraz wprowadzania ręcznych zmian przez użytkownika – co jest procesem nieefektywnym (47%–83% domen uznanych za phishingowe pojawiło się na czarnych listach w ciągu 12 godzin od wstępnego testu – S. Sheng, B. Wardman, G. Warner, L.F. Cranor, J. Hong, C. Zhang [43]), wymagającym doświadczenia i konsumującym czas zespołu bezpieczeństwa. Z uwagi na prostotę budowania czarnej listy (dodawanie kolejnych wpisów oznaczających wykryte, niebezpieczne adresy), podejdzie to było modyfikowane i wzbogacane o inne komponenty (warto wspomnieć o metodzie PhisNet autorstwa P. Prakash, M. Kumar, R. R. Kompella, M. Gupta [41], czy o informacjach umieszczanych

podczas rejestracji domeny autorstwa M. Felegyhazi, Ch. Kreibich, V. Paxson [42]).

c. Greylist (szara lista) – metoda ochrony poczty elektronicznej polegającej na zastosowaniu mechanizmu odrzucania przy pierwszej próbie dostarczenia przez serwer pocztowy odbiorcy wszystkich wiadomości email, których adresy nadawców nie znajdują się na specjalnej liście upoważnionych adresów email. Odrzucenie wiadomości przy pierwszej próbie jej dostarczenia realizowane jest poprzez zwrócenie serwerowi nadawcy kodu błędu (serii 4xx) oznaczającego tymczasowe problemy z odebraniem danej wiadomości email. Po otrzymaniu tego kodu, serwer nadawcy ponawia próbę wysłania wiadomości, która tym razem jest przyjęta przez serwer odbiorcy, a adres email dopisany zostaje do listy dopuszczonych adresów email. Kolejne wiadomości email z tego samego adresu, przyjmowane są już bez opóźnień. Czas po jakim serwer nadawcy ponownie prześlę daną wiadomość po otrzymaniu kodu błędu zależny jest od konfiguracji danego serwera pocztowego. Technikę tą, wykorzystuje się głównie do eliminacji spamu [54]. Metoda ta polega na założeniu, że serwery masowo wysyłające wiadomości typu spam działają tymczasowo w celu uniemożliwienia ich identyfikacji i umieszczeniu w bazie adresów nadawców spamu (DNS Blacklist[83]). Działanie tej metody polega na właściwościach protokołu SMTP[84] (wykorzystywanego do przesyłania wiadomości email). Protokół SMTP wykorzystuje dwa rodzaje błędów: tymczasowe (4xx) i permanentne (5xx). Błędy tymczasowe sygnalizowane są przez serwer odbiorcy w chwili, gdy nie może on chwilowo odebrać danej wiadomości (z różnych przyczyn, np.: chwilowa blokada skrzynki odbiorczej, np.). Błąd permanentny oznacza, że serwer odbiorcy odnowił przyjęcia danej wiadomości email (np. odbiorą końcowy nie posiada skrzynki odbiorczej na danym serwerze pocztowym). Otrzymanie błędu tymczasowego przez serwer nadawczy, powoduje, że dana wiadomość jest ponownie przesyłana – serwer nadawczy zapisuje ten błąd na liście zadań do zrealizowania po czasie ustalonym w jego wewnętrznej konfiguracji. Serwery służące do masowej wysyłki wiadomości typu spam, nie zapisują odebranego błędu tymczasowego –

[83] Baza adresów wykorzystywanych do rozsyłki spamu - DNS Blacklist - dostępna jest pod adresem: https://www.dnsbl.info/dnsbl-list.php [dostęp: 28.12.2022].

[84] SMTP (ang. Simple Mail Transfer Protocol) – protokół komunikacyjny opisujący sposób przekazywania poczty elektronicznej w sieci Internet. Opis protokołu znajduje się w dokumencie RFC 5321 (źródło: https://www.rfc-editor.org/rfc/rfc5321, [dostęp: 28.12.2022]).

traktują to jako permanentną odmowę i nie podejmują próby ponownego przesłania wiadomości i głównie na tym założeniu bazuje technika budowy szarych list.

Zalety:

1. Skuteczne eliminowanie komunikacji do domen i adresów IP oznaczonych jako phishingowe, które zostały wcześniej rozpoznane i wykorzystane w poprzednich atakach.
2. Łatwość i niski koszt implementacji, dostępne w każdej klasie urządzeń sieciowych i bezpieczeństwa.

Wady:

1. Brak odporności na nowy, nieznany wcześniej adres witryny phishingowej ([44]) – brak gwarancji, że wykorzystane do wcześniejszych ataków adresy IP i domenowe, zostaną ponownie wykorzystane.
2. Brak odporności na ataki typu „zero-hour”[85].
3. Wysoki odsetek fałszywych trafień [55].
4. Nie jest możliwe zbudowanie listy zawierającej wszystkie adresy (IP i domenowe) witryn phishingowych i serwerów wysyłających wiadomości – wymaga to dużych zasobów sprzętowych i zespołu wprowadzającego i weryfikującego dane.
5. Konieczność ciągłego aktualizowania list (aktualne listy zapewniają większą skuteczność).
6. Wymaganie doświadczenia w budowaniu list dostępu (podniesienie kosztów rozwiązania związane z wymaganiem zatrudnienia odpowiedniego specjalisty).
7. Konieczność posiadania specjalisty (lub zespołu) odpowiedzialnego za budowanie i nadzór nad poprawnością implementacji i aktualizacji list – co przyczynia się do wzrostu kosztów danej organizacji.

[85] Zero-hour (inna nazwa to zero-day, 0-day) – termin ten definiuje się jako lukę w istniejących zabezpieczeniach (ale również w oprogramowaniu czy sprzęcie), które zostały wykorzystane przez atakujących, a nie istnieje jeszcze dostępne oprogramowanie naprawcze, lub nie istniała wcześniej wiedza o występowanie tej luki. Atak typu zero-hour to atak wykorzystujący nieznaną wcześniej lukę w zabezpieczeniach, o której twórca danej aplikacji czy sprzętu nie posiadł wiedzy o jej istnieniu i nie wytworzona została żadna poprawka.

8. Atakujący do obejścia zabezpieczeń wykorzystujących metodę blacklistingu wykorzystują serwery proxy[86], które algorytmicznie generują nowe adresy domenowe (URL) lub które nawiązują połączenie VPN do docelowego serwera hostującego domenę phishingową – przy czym adres IP serwera pośredniczącego nie znajduję się na liście zakazanych adresów IP do komunikacji.

II.2 Reguły detekcji

Reguły detekcji to zdefiniowana logika służąca do wykrywania anomalii, zdarzeń o charakterze wirusowym, incydentów bezpieczeństwa teleinformatycznego, np. Detekcja polega na warunkowym porównaniu danych do zbadania (weryfikacji) z danymi zapisanymi w regule – kiedy co najmniej jeden (lub więcej – w zależności od konstrukcji danej reguły) z warunków jest spełniony (np. tożsamość adresów IP, identyczna suma kontrolna pliku, np.), to uruchamiane są zdefiniowane wcześniej działania (np. wygenerowanie ostrzeżenia, lub uruchomienie blokady komunikacji). Problematyka tworzenia reguł została opisana przez D. L. Cook, V. K. Gurbani, M. Daniluk ([45]), N. M. Shekokar, C. Shah, M. Mahajan, S. Rachh ([47]). Z kolei podejście polegające na tworzenie reguł detekcji na potrzeby uczenia maszynowego opisane zostały przez N. Sanglerdsinlapachai, A. Rungsawang ([46]).

Podejście do wykrywania ataku phishingowego bazuje na podobieństwie treści [47] w otrzymywanych wiadomościach. Zakłada się, że podobnie jak w przypadku wiadomości typu spam, wiadomości phishingowe mogą zawierać określone sformułowania, słowa powtarzające się w różnych wiadomościach, użyte w tym samym kontekście. Innym podejściem systemu regułowego jest analiza formatu wiadomości email ([45]), której treść jest instruująca odbiorcę do podjęcia określonych działań. Formatowanie HTML treści wiadomości można wykorzystać do ukrycia prawdziwego adresu odnośnika URL, reguła oparta na analizie porównawczej może wykryć potencjalny atak.

Zasadniczą wadą podejścia opartego na regułowej analizie treści jest konieczność posiadania bazy danych zawierającej występujące słowa kluczowe w zidentyfikowanych uprzednio wiadomościach phishingowych. Słowa te musiały by być również zapisane na

[86] Serwer proxy (serwer pośredniczący) – serwer wykonujący odpowiednie operacje w imieniu użytkownika, pośredniczący w komunikacji pomiędzy nim a końcowym elementem (innym serwerem, stacją użytkownika, systemem teleinformatycznym, itp.).

wiele różnych odmian i zastosowań. Reguły detekcji wymagają również dużej bazy danych ze zdefiniowanymi warunkami logicznymi. Podejście takie wymaga posiadania szerokiej i aktualnej wiedzy eksperckiej z danej dziedziny, co znacznie zwiększa koszt systemu.

Wady:

1. Brak wykrywania nowych typów ataków – zdefiniowane warunki reguł detekcji oparte są na rozpoznanych, wcześniejszych atakach. Wymusza to ciągłe, aktywne poszukiwanie nowych wzorców ataku.
2. Brak możliwości wykrywania wszystkich anomalii i incydentów za pomocą reguł detekcji – ścisłe zdefiniowanie warunków reguły, nie uruchomi działań nieznacznie tylko odbiegających od zdefiniowanych warunków, pomimo rozpoznania większości składowych jako niepożądane (złośliwe).
3. Trudność zaprojektowania warunków składających się z wielu następujących po sobie kroków, odnoszących się do różnych rodzajów danych – incydent bezpieczeństwa teleinformatycznego może składać się z wielu różnych działań (niekoniecznie następujących jedno po drugim, lub następujących w różnych, nieregularnych odstępach czasu), co powoduje trudności w jego analizie oraz często uniemożliwia stworzenie wystarczającej reguły. Zwykle reguły mają zapisany tylko jeden warunek (lub regularną sekwencję tych samych zdarzeń).
4. Czasochłonność tworzenia nowych reguł – opracowanie nowej reguły detekcji zajmuje zwykle 4-8 godzin [56], co wymusza posiadanie własnych, pełnoetatowych inżynierów detekcji, których zadaniem jest opracowywanie nowych reguł. Konieczność posiadania odpowiedniego zespołu bezpieczeństwa teleinformatycznego – wynika to z czasochłonności tworzenia nowych reguł detekcji, konieczności znajomości topologii sieci danej organizacji celem właściwego wdrożenia na odpowiednim urządzeniu reguł, ich okresowy przegląd (pod kątem właściwego działania i wykluczenia detekcji typu „false positive” i „false negative”). Utrzymanie zespołu bezpieczeństwa znacznie podwyższa koszty działalności danej organizacji, co może prowadzić do zastąpienia ich, darmowym oprogramowaniem, którego skuteczność może być niewielka
5. Konieczność posiadania obszernej bazy warunków – rosnące ilości poszczególnych typów ataków, prowadzą do konieczności budowania dużej ilości reguł zawierających duże ilości warunków z podziałem na różne systemy, protokoły sieciowe, np. Obciążanie systemu (danych wejściowych) dużą ilością

reguł detekcji, prowadzić może również do znaczącego spadku efektywności takiego systemu i przetwarzanych w nim danych – wymagana jest wówczas inwestycja w odpowiednio wydajny sprzęt, który z uwagi na jego koszt nie znajduje zastosowania w mniejszych i średnich organizacjach.

6. Brak wystarczającej ilości danych do spełnienia warunków reguły – niektóre typy ataków wymagają spełnienia wielu różnych działań po stronie atakującego. Reguły detekcji, które zawierają wiele, następujących po sobie koniecznych do spełnieniami warunków logicznych, wymagają skomplikowanych systemów detekcji (rozwiązania klasy SIEM + EDR), które mogą znacznie spowolnić pracę użytkowników końcowych (z uwagi na dłuższy czas przetwarzania danych wejściowych do organizacji), w związku z czym ograniczane są do minimalnej ilości danych i warunków logicznych – które z kolei nie potrafią wykryć właściwego niepożądanego zdarzenia.

II.3 Algorytmy genetyczne

Algorytm genetyczny jest komputerowym modelem, którego działanie oparte jest na zjawisku ewolucji biologicznej, wykorzystującej selekcję naturalną. Są to algorytmy poszukiwania oparte na zasadach doboru naturalnego oraz dziedziczności [57].

Zaproponowany przez autorów pracy [48] mechanizm, wykorzystujący opisany wyżej model, jest połączeniem technologii i wiedzy eksperckiej. Algorytmy genetyczne wykorzystywane są do wykonywania „ewolucji" utworzonych reguł, pomagających klasyfikować i odróżniać prawdziwe strony internetowe (odnośniki URL do nich prowadzące) od stron, wiadomości, odnośników phishingowych. Reguły detekcji bazują na technicznych wskaźnikach phishingu, np.:

```
1. if
2. {
3.    URL zawiera adres IP
4. }
5. then
6. {
7.    Oznacz jako phishing
8. }
```

Powyższa prosta reguła może być zaimplementowana jako element białej listy (podejście opisane powyżej) i uruchamiana przy próbie nawiązania połączenia z danym

odnośnikiem URL przez użytkownika, lub w trakcie automatycznego skanowania wiadomości zawierających odnośniki URL. Jako dane wejściowe wykorzystywane do detekcji adresacji IP oznaczonej jako phishingowa, autorzy wykorzystują dane historyczne pochodzące od APWG, które bazują na wiedzy eksperckiej. Adresy IP, które zostały przez społeczność ekspertów APWG oznaczone jako phishingowe, wykorzystywane są przez algorytmy genetyczne do stworzenia na podstawie prostych reguł, dużo większej bazy reguł filtrujących, np.:

```
1. if
2. {
3.    Odnośnik URL zawiera adres IP i adres IP to
      209.11.??.??
4. }
5. then
6. {
7.    Oznacz jako phishing
8. }
```

Autorzy opracowania [48] przedstawili przykładowy chromosom odpowiadający powyższej regule:

d	**1**	**0**	**b**	*	*	*	*

Rysunek 26. Przykład chromosomu odpowiadającego regule detekcji. Źródło: V.Shreeram, M.Suban, P.Shanthi, K.Manjula – "Anti-phishing detection of phishing attacks using genetic algorithm".

Każdy chromosom zawiera osiem genów z uwagi na konieczności zakodowania w nim wartości adresu IP występującego w regule. Dla uproszczenia autorzy wykorzystali szesnastkową[87] reprezentację adresu IP. W przytoczonym przykładzie geny (najmniejszy element niosący informację) odpowiadają poszczególnym oktetom adresu Ipv4[88]. Na podstawie mutacji chromosomów tworzone są nowe reguły, które dokonają filtracji odnośników URL domen uznanych za phishingowe. Celem zastosowania Algorytmów Genetycznych jest wygenerowanie takiego zestawu reguł (bazując na danych historycznych), które pozwolą na wykrycie wszystkich odnośników prowadzanych do witryn phishingowych.

[87] W kodzie heksadecymalnym Wartość genu **d1** odpowiada wartości **209** pierwszego oktetu adresu IP występującego w regule zapisanego kodem dziesiętnym, wartość genu **0b** w kodzie hexadecymalnym odpowiada wartości **11** drugiego oktetu adresu IP.

[88] IPv4 – adresacja IP wersji 4, zapisywana jest w postaci czterech zestawów liczb zwanych oktetami, przy czym każdy zestaw reprezentuje liczbę 8-bitową w zakresie od 0 do255.

W celu uzyskania wysokiej skuteczności działania Algorytmu Genetycznego, należy dobrać jego odpowiednie parametry:

1. **Funkcja oceny** – jest to jeden z najważniejszych parametrów Algorytmów Genetycznych, w zaproponowanym rozwiązaniu, obliczany jest wynik ogólny (czy odnośnik URL pasuje do wstępnie sklasyfikowanych danych):

$$fo = \sum_{i=1}^{8} f_i * w_i \qquad (2.1)$$

gdzie:
fo –funkcja oceny
f_i – dopasowanie
w_i – waga i-tego elementu (dla i=1...8)

   Wartości wag są podzielone na kategorie według różnych pól w adresie IP występujących w adresie URL, dlatego wszystkie geny dla danej subdomeny mają takie same wartości. Ilość elementów we wzorze odpowiada ilości genów w chromosomie.
2. **Krzyżowanie** – konieczność zastosowanie wielu różnych reguł wykrywających prawdopodobieństwo phishingu wymaga by na podstawie wystarczająco dobrych reguł, powstały nowe, wykrywające różne modyfikacje adresu URL Autorzy rozwiązania poszukują lokalnych maksimów techniką niszowania [58].
3. **Mutacja** – w wyniku działania funkcji oceny i operacji krzyżowania, może się okazać, że z puli dobrych rozwiązań wyeliminowane zostały te rozwiązania, które poddane procesowi krzyżowania wygenerowały by zestaw reguł działających na nowym typie ataku phishingowego. Autorzy stosują technikę spowolnienia populacji w celu maksymalizacji w następnym pokoleniu.

Zalety:

1. Wykrywane zmian w odnośnikach URL na podstawie zdefiniowanych reguł.
2. Automatyczne tworzenie nowych reguł(„child") na bazie reguł wzorcowych („parent").

Wady:

1. Metoda skupia się głównie na identyfikacji odnośników phishingowych (URL).
2. Nie zawsze jest możliwe jednoznaczne określenie phishingowego charakteru odnośnika URL (istnieje wiele różnych typów odnośników). Jednoznacznie

potraktowanie adresu URL obecnego na liście jako phishingowego, może w pewnych przypadkach dawać fałszywie dodatnie wyniki.

II.4 Metody uczenia maszynowego

Problem wykrywania wiadomości, stron phishingowych może być potraktowany jako jeden z zasadniczych problemów eksploracji danych, problem klasyfikacji. Do rozwiazywania problemów klasyfikacji opracowano szereg algorytmów uczenia maszynowego. W odniesieniu do klasyfikacji wiadomości i stron jako phishingowe lub nie można wymienić w szczególności następujące algorytmy:

a. zaproponowany przez Abdelhamida, Ayesha i Thabtaha model [59], bazujący na wykorzystaniu reguł klasyfikacyjnych,
b. zaproponowany w pracy [60] model klasyfikacji wiadomości email,
c. technika bazująca na maszynie wektorów nośnych (SVM):
    - do wykrywania anomalii (w tym również do wykrywania phishingu) opisana została przez [61],
    - mechanizm wykrywania ataku [62]

Data Mining (z ang. Eksploracja danych) jest procesem analitycznym z wykorzystaniem komputerów, operującym na dużych zbiorach danych, który poszukuje wzorców, zależności (współzależności) pomiędzy danymi (zmiennymi). Uzyskuje się w ten sposób wiedzę, która jest niewidoczna dla człowieka z uwagi na czasochłonność obliczeń oraz może być skuteczniejsza niż badanie wzorców przez człowieka (A. Aljofey, Q. Jiang, A. Rasool, H. Chen, W. Liu, Q. Qu, Y. Wang [49], A. K. Jain, B. B. Gupta [50], D. Sánchez, M.A. Vila, L. Cerda, J.M. Serrano [51]). Wykrywanie phishingu z użyciem algorytmu klasyfikatora asocjacyjnego (AC)[89] jest obiecującą metodą detekcji. Model ten [59] operuje na trzech fazach:

1. wyszukiwanie ukrytych korelacji w dostarczonym zbiorze danych i tworzenie zestawu reguł (CAR – Class Association Rule),
2. klasyfikacja i przycinania – procesu redukującego ilość reguł w celu optymalizacji,
3. ocena skuteczności klasyfikatora operującego na danych testowych.

[89] klasyfikator asocjacyjny (ang. AC - associative classifier) - rodzaj nadzorowanego uczenia modelu, który używa analizy zbioru zmiennych w celu znalezienia występujących w nim powtarzających się zależności.

Algorytm klasyfikacji asocjacyjnej tworzy zestaw prostych zasad umożliwiających ich implementację oraz możliwość ich ręcznego dostrojenia przez użytkownika końcowego. Dodatkowym atutem stosowania tego podejścia jest znajdywanie przez AC ukrytych informacji zwykle pomijanych przez klasyczne algorytmy klasyfikacji. Dodatkowe informacje, które mogą być nadmiarowe, redukowane są podczas następnej fazy, dzięki czemu otrzymuje się dopasowany zestaw reguł. Autorzy [59] w swoim rozwiązaniu opracowali algorytm MCAC (Multiclass Classification based Association Rule), który w odróżnieniu od AC generuje niejedną klasę na regułę, ale tworzy zbiór klas. Algorytm MCAC tworzy reguły dla działań użytkownika, które można zapisać za pomocą pięciu kolejnych kroków:

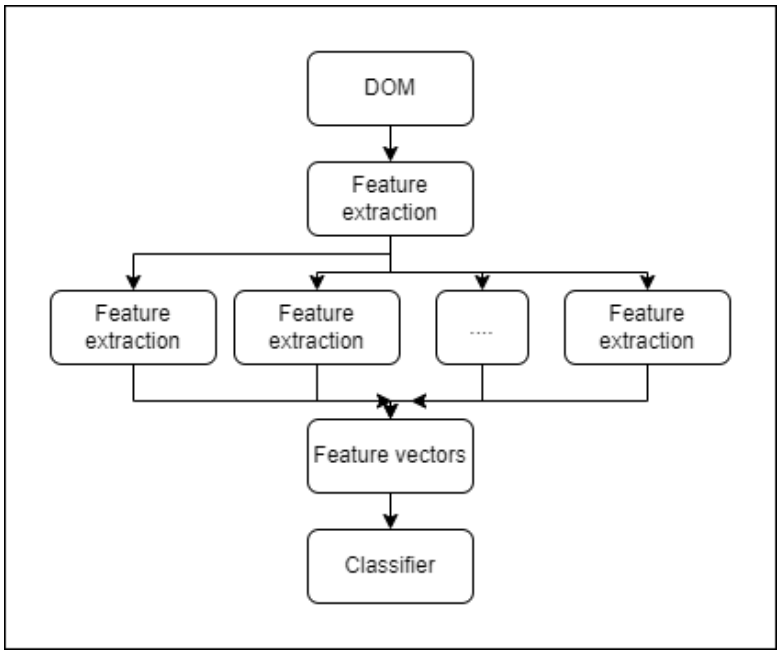
1. użytkownik kliknął odnośnik w otrzymanej wiadomości,
2. użytkownik jest przekierowywany do określonej witryny (strona jest jednocześnie traktowana jako dane testowe),
3. oprogramowanie osadzone w przeglądarce, przetwarza dane (cechy strony) i je wyodrębnia celem zapisania w ustrukturyzowanej formie,
4. zaimplementowany moduł odgaduje typ witryny (prawdziwa, phishingowa) na podstawie wyuczonych reguł,
5. w przypadku sklaryfikowania witryny jako phishingowej, użytkownik zostanie o tym poinformowany.

Krokiem pierwszym algorytmu MCAC jest iterowanie po zbiorze danych treningowych (danych historycznych) w celu identyfikacji, które reguły zostały znalezione i zapisane. Nadmiarowe reguły, które nie zawierają danych treningowych są odrzucane. W drugim kroku powstaje klasyfikator zarówno dla pojedynczej reguły jak i dla wielu. Podobnie jak w przypadku klasycznego algorytmu klasyfikacji asocjacyjnej ostatnim krokiem jest testowanie i ocena klasyfikatora.

Innym podejściem wykorzysującym uczenie maszynowe jest analiza formatowania HTML wiadomości email oraz stron internetowych w poszukiwaniu odnośników URL i ich analiza w celu stworzenia nowych reguł detekcji opartych na wskaźnikach (A. K. Jain, B. B. Gupta [50]). Proponowane rozwiązanie analizuje strukturę dokumentu HTML (Rysunek 27), przekształca go w strukturę DOM[90] i dokonuje ekstrakcji wszystkich

[90] DOM (ang. Document Object Model) – sposób reprezentacji złożonych dokumentów XML i HTML w postaci modelu obiektowego, który jest niezależny od platformy sprzętowej (programowej) i języka programowania

odnośników URL, które następnie są grupowane wedle unikalnych dla grupy reguł i właściwości. Utworzony w ten sposób wektor, poddawany jest klasyfikacji.



Rysunek 27. Model klasyfikacji odnośników URL wykorzystujący uczenie maszynowe, bazujący na modelu DOM, źródło: „A machine learning based approach for phishing detection using hyperlinks information", A.K. Jain, B.B. Gupta

Zaproponowany przez [60] model klasyfikacji wiadomości email przyjmuje ustalony zbiór cech (których wartości pobierane są z każdej wiadomości email poddanej klasyfikacji), będący podstawą procesu klasyfikacji. Wybór cech klasyfikacji dla wiadomości phishingowych, może być trudny z uwagi na duże podobieństwo ataku phishingowego do zjawiska spamu (np. rozsyłka dużej ilości wiadomości, podobna struktura).

Autorzy rozwiązania klasyfikując wiadomości email oparli się na istniejących funkcjach dostępnych dla wiadomości email, opisując je i tworząc dla każdej z funkcji etykietę klasy:

1. liczba odnośników do różnych domen w pojedynczej wiadomości email,
2. liczba odnośników do innych domen zawartych w wiadomości email niż domena, z której wiadomość pochodzi,
3. liczba odnośników zawierających adres IP zamiast nazwy domeny,
4. liczba subdomen w odnośniku (liczba kropek w domenie),
5. liczba odnośników na nieprawidłowy port (inny niż 80[91] lub 443),

[91] W obecnym standardzie komunikacji, przyjmuje się wykorzystanie protokołu HTTPS i komunikacji na port 443. Protokół HTTP (i port komunikacji 80) jest przestarzałym rozwiązaniem, nie zapewniającym bezpieczeństwa w wymianie informacji.

6. liczba zdjęć jako odnośniki (łącza),
7. mapy obrazu jako odnośniki (łącza),
8. zawartość znaków spoza zestawu ASCII w adresie URL – niektóre znaki i litery alfabetów spoza standardowego zestawu znaków ASCII mogą swoim wyglądem przypominać standardowe litery, celem „optycznego" oszukania użytkownika
9. rozmiar wiadomości,
10. kod kraju na podstawie adresu IP odnośnika zawartego w wiadomości – statystyki [19] wskazują, że 60% wiadomości phishingowych zawiera odnośnik zlokalizowany tylko w dwóch krajach

Metoda klasyfikacji nie wyklucza jednoczesnego użycia metod klasyfikacji spamu (np. greylisting) i innych klasycznych metod, do ograniczenia puli wiadomości email, które mają zostać poddane metodzie klasyfikacji.

Otrzymywane wiadomości poddawane są funkcji oceny, która na podstawie skuteczności rozróżnia trzy klasy: normalne wiadomości email spam i phishing. Funkcja wykorzystuje zdobyte informacje (zysk informacyjny) zdefiniowane jako:

$$gain(X,C) := infoI - info_xI \tag{2.2}$$

gdzie:
C –zbiór etykiet danej klasy
X – zbiór cech
infoNP. – funkcja entropii Shannona
$info_x$ – warunkowa funkcja entropii, którą można zdefiniować również jako:

$$info_x(C) = \sum_{\upsilon \in X} P(\upsilon) * P(C|\upsilon) \tag{2.3}$$

gdzie:
P(υ) – prawdopodobieństwo υ, υ ∈ X
P(C|υ) – prawdopodobieństwo warunkowe C

Zysk informacyjny definiowany jest jako oczekiwana ilość informacji (zawarta w danej wiadomości email), przypisany do danej klasy minus entropia. Entropia informacji w tym przypadku to wielkość określająca liczbę bitów informacji zawartej w danej wiadomości. W celu oceny funkcji, by wyeliminować przyjmowanie odmiennych wartości (gdyż zysk informacyjny faworyzuje cechy, które przyjmują wiele różnych wartości), autorzy rozwiązania, przyjęli kryterium uszeregowania funkcji według ich współczynnika wzmocnienia informacji zdefiniowanego jako:

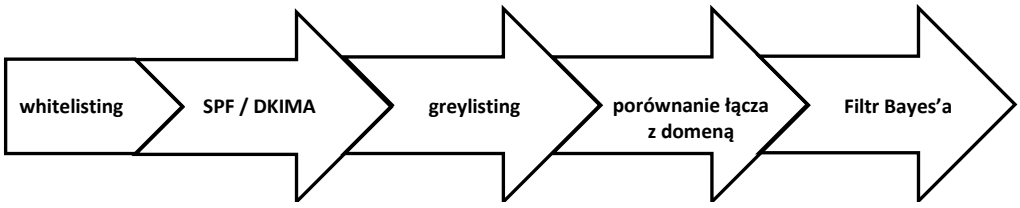
$$gainratio(X,C) := \frac{gain(X,C)}{splitinfo(X)} \tag{2.4}$$

gdzie:

$$splitinfo(X) := -\sum_{v \in X} P(v) * log_2 P(v) \tag{2.5}$$

Użyte przez autorów dane treningowe pochodziły z Phishery[92] oraz udostępnione przez TREC[93] i zawierały 11 000 wiadomości skalsyfikowanych jako phishingowe, 25 000 wiadomości ham[94] oraz 52 000 wiadomości typu spam, poddane zostały procesowi normalizacji do przedziałów odpowiednich cech. Autorzy decydowali się na klasyfikację wiadomości email wg trójpodziału: poprawna wiadomość („non-spam"), spam oraz phishing. Porównując klasyfikatory w każdej z trzech grup uzyskali dokładność rzędu 97%, co daje lepsze rezultaty niż klasyfikacja binarna.

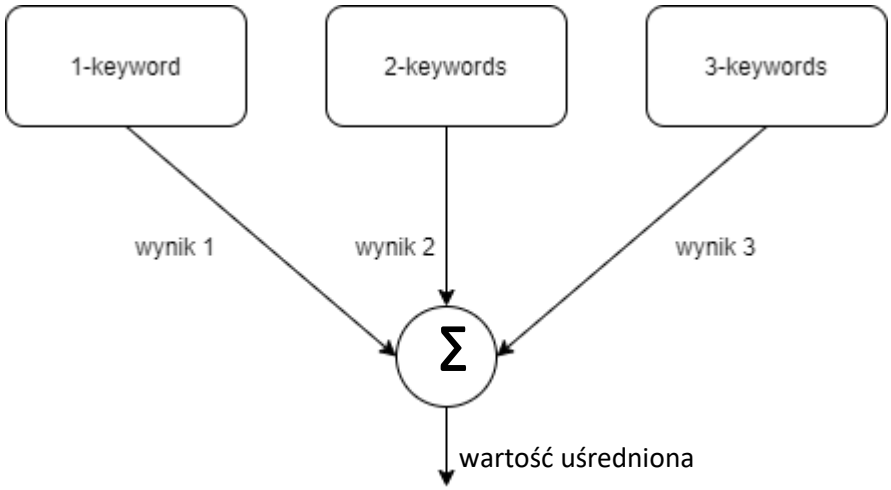
W rozwiązaniu zaproponowanym przez [63] wykorzystywana jest połączona technika klasyfikacji z użyciem filtrowania Naiwnego Klasyfikatora Bayes'a wraz z „klasycznymi" technikami systemu antyphishingowego (filtrowanie adresów: whitelisting, blacklisting, greylisting, mechanizmy DKIM[95], SPF[96], sprawdzanie nazw DNS). Proponowane podejście w celu zwiększenia skuteczności obrony przed phishingiem, poszczególne techniki należy połączyć w odpowiednie, następujące po sobie kroki (rysunek poniżej).



Rysunek 28. Etapy mechanizmu obrony przez phishingiem

[92] http: //phishery.internetdefence.net
[93] http://trec.nist.gov/data/spam.html (za 2007 rok).
[94] Ham – wiadomości email nie sklaryfikowane jako spam, inne określenie to „non-spam" lub „good mail".
[95] DKIM (ang. **D**omain**K**eys **I**dentified **M**ail) – metoda łączenia domeny internetowej z adresem email, zwiększająca wiarygodność wysyłanej wiadomości, zabezpieczająca przez podszywaniem się pod innego nadawcę.
[96] SPF (ang. **S**ender **P**olicy **F**ramework) – mechanizm zabezpieczający serwery SMTP przed przyjmowaniem poczty z niedozwolonych źródeł.

W przedstawionym rozwiązaniu Naiwny Klasyfikator Bayesa został podzielony na trzy segmenty, analizujące odpowiednio: jedno słowo kluczowe, połączone ze sobą dwa słowa kluczowe oraz połączone z logiczną całość, trzy słowa kluczowe. Każdemu segmentowi przypisano odpowiednią wartość (1 dla jednego słowa, 2 dla dwóch, 3 dla trzech połączonych ze sobą słów).



Rysunek 29. Wykorzystanie klasyfikatora Bayesa do wykrywania wiadomości phishingowej. Źródło: „Analysis of Phishing Attacks and Countermeasures" – B. Issac, R. Chiongm S.M. Jacob.

Autorzy powyższego rozwiązania stwierdzają, że w dobie ciągłych zmian i ewolucji ataków phishingowych, konieczne jest dobre zrozumienie używanych przez atakujących metod i technik, gdyż poszczególne kroki wymagają dostrojenia wskaźników badanych w danym kroku prezentowanej metody.

Wartym uwagi rozwiązaniem wykorzystującym lasy losowe jest praca [64] wykorzystująca techniki drzewa decyzyjnego (lasy losowe) do klasyfikacji wiadomości email. Warunkiem wstępnym zastosowania tej metody jest ekstrakcja danych zawartych w wiadomości email (odnośniki URL, adresy IP, ilość odnośników, słowa kluczowe w treści wiadomości). Model trenowany był za pomocą weryfikacji krzyżowej (10-krotnie) na próbkach, z których jedną (z dziesięciu) wyselekcjonowano jako dane testowe.

Zalety:

1. Wielostopniowa weryfikacja.
2. Skuteczna filtracja spamu.

Wady:

1. Bazowanie na słowach kluczowych.
2. Klasyfikowanie wiadomości spam jako phishing.

Technika bazująca na maszynie wektorów nośnych (SVM) do wykrywania anomalii (w tym również do wykrywania phishingu) opisana została przez [61], w której wykorzystano dwie zmienne do wykrywania nietypowych działań na witrynach internetowych:

1. wyświetlana nazwa firmy w nazwie domeny,
2. cechy strukturalne witryny.

Cechy strukturalne witryny (atrybuty konieczne do opisania wektora SVM) autorzy zdefiniowali jako:

1. nietypowy adres URL,
2. niewłaściwy dla danej domeny adres serwera DNS,
3. nieodpowiedna ilość kotwic w sekcji <BODY>,
4. niewłaściwy certyfikat SSL (TLS),
5. nietypowe cookies,
6. nietypowe / niewłaściwe zapytanie URL.

Zadaniem klasyfikatora SVM jest rozdzielnie przestrzeni **Ω** (zbiór wektorów) na dwie klasy (1, -1) odpowiadające stronie phishingowej i stronie nie-phishingowej.

$$\Omega = \{x_i\}_{i=1}^{n} \tag{2.6}$$

gdzie:
Ω – przestrzeń zbioru wektorów cech
x_i – m-wymiarowy wektor cech oznaczony jako y ϵ {1, -1}

Wektor cech tworzony jest na podstawie ekstrahowanych atrybutów witryny internetowej. Przykładowo wektor dla cechy kotwic na stronie można zdefiniować jako:

$$C = \begin{cases} 0 & A_a = 0 \\ -A_l/A_a & A_l > 0 \\ A_f/A_a & pozostałe \end{cases} \tag{2.7}$$

gdzie:
C – wektor cech
A_l – liczba kotwic odnośników lokalnych (a href=”index.php”)
A_f – liczba kotwic odnośników domenowych (<a href=”https//example.com)
A_a – liczba wszystkich kotwic, przy czym $A_l + A_f \leq A_a$ gdyż nie wszystkie kotwice odnośników muszą być prawidłowe.

Wektor cech przyjmuje wówczas postać:

$$x =< C1, C2, C3,\ldots,Cn >$$

Uczenie maszynowe wykorzystują również autorzy [62] do wykrywania ataków phishingowych poprzez wykorzystanie aktywności użytkowników w sieciach społecznościowych. Wektor ataku poprzez media społecznościowe jest jednym z typów ataku phishingowego, polegającym na przejęciu konta istniejącego użytkownika lub stworzeniu wiarygodnego, fałszywego profilu i za jego pomocą rozsyłanie wiadomości zachęcających do kliknięcia w link, podania danych autoryzujących, zainstalowania danego oprogramowania (lub innych dowolnych czynności, których wykonanie przez potencjalną ofiarę, przyniesie korzyść atakującym). Ten typ ataku może być wykrywany za pomocą technik profilowania behawioralnego[97]. Mechanizm wykrywania ataku [62] opiera się na zbieraniu informacji o czasie połączeń z platformą społecznościową użytkownika i przekształcenie jej w przestrzeń cech, a następnie wytrenowanie modelu SVM.

II.7 Wady

Przedstawione powyżej metody detekcji ataku phishingowego posiadaj kilka wad:

1. Brak odporności metody na wciąż zmieniające się techniki atakujących. Dotychczas rozpoznane wartości poszczególnych cech, które mogą wskazywać na phishing, mogą ulec zmianie (np. wykorzystywana adresacja IP, szybko zmieniające się nazwy domen). Szybkim zmianom podlega też sama infrastruktura wykorzystywana do przeprowadzenia ataku (serwery hostujące strony, serwery DNS, skrypty, wykorzystywane oprogramowanie), filtrowanie więc wiadomości z uwagi na możliwe w niej do wykrycia znane IoC, jest nieefektowne a w wyniku szybkich zmian infrastruktury, filtry antyphisingowe nie będą w stanie zablokować takiego ataku. Szybkie zmiany infrastruktury wymagają w przypadku detekcji opartych na listach, ciągłe i szybkie ich aktualizowanie – co wprowadza znaczne opóźnienia (czas pomiędzy identyfikacją nowej infrastruktury, ekstrakcja nowych IoC, dystrybucja wskaźników,

[97] Profilowanie behawioralne – analiza wzorców zachowań, cech osobowych, w ujęciu sieci społecznościowych analiza wzorca aktywności użytkowników.

dostępnością administratora, aktualizacją listy a identyfikacją ataku może sięgać nawet kilkudziesięciu dni).

2. Brak rozpoznania nowego, nieznanego wcześniej typu ataku lub niewystępującej techniki. Rozwiązania bazujące na wykrywaniu schematu, wykrytych i przeanalizowanych wcześniej ataków, nie są w stanie wykryć ataku którego cechy nie znajdują się w opisie reguł detekcji.
3. Wykorzystanie legalnych metody i zaufanych źródeł (technika podobna do wykorzystywanej przez grupy hakerskie metody Lotl). Jako zaufane źródła wykorzystywane są często platformy społecznościowe (fałszywe profile), darmowe usługi pocztowe czy rozwiązania chmurowe – jako źródło ataku wykorzystywana jest wówczas infrastruktura zaufanych operatorów, do których, i z których, systemy bezpieczeństwa zwykle dopuszczają całość komunikacji (w tym otrzymywanie wiadomości email) – blokada określonych adresów URL, IP lub domen mogła by wiązać się z niemożliwością korzystania z dostępnych usług lub wykorzystywanych w danej organizacji rozwiązań.
4. Konieczność ponoszenia dużych nakładów administracyjnych i ciągłego pozyskiwania wiedzy o nowych technikach, celem utrzymania środowiska i systemu wykrywania phishingu (patrz: blacklisting, whitelisting).
5. Wykorzystanie personalizacji ataków (spear phishing). Ataki tego typu poprzedzone są przeprowadzeniem dokładnego rozpoznania potencjalnej ofiary (OSINT). Do jego konstrukcji wykorzystuje się zdobyte w ten sposób informacje, które wraz z użytymi technikami inżynierii społecznej są trudne do wykrycia w przypadku spersonalizowanej wiadomości Klasyczne metody nie analizują treści wiadomości, która może silnie nakłaniać użytkownika do wykonania działań, które są dla niego niekorzystne.
6. Tworzenie wielopoziomowych ataków – kombinacja wielu różnych technik, tak by wykorzystanie poszczególnych metod nie wyzwalało uruchomienie mechanizmów antyphishingowych (np. zablokowanie komunikacji, przeniesienie wiadomości do kwarantanny). Istniejące techniki detekcji phishingu mogą skupiać się na jednym konkretnym aspekcie ataku i nie uwzględniać jego pełnego obrazu.

Rozdział III – Techniczne i nietechniczne wskaźniki phishingu.

Celem rozdziału jest prezentacja autorskiej listy wskaźników phishingu. Lista ta będzie wykorzystana do opracowania metod detekcji ataku phishingowego z wykorzystaniem uczenia maszynowego.

Phishing jako forma oszustwa, wykorzystująca elementy socjotechniki, celowany pod konkretną grupę lub dedykowany konkretnej osobie, jest trudy do wykrycia (wskazują na to autorzy: [65], [66], [67], [68]), zwłaszcza przez nieświadome, niedoświadczone ofiary – na co między innymi wskakują przytoczone wyniki ankiet [35]. Spopularyzowanie się dostępności platform oferujących usługi AI[98] oferujących tworzenie wiarygodnych treści (np. z wykorzystaniem ChatGTP), łącząc w sobie techniczne aspekty (przejęcie sieciowych czy istniejących stron www w wyniku włamania) z inżynierią społeczną (treść sugerującą konieczność podjęcia działań, czy wywołująca poczucie strachu) sprawia, że phishing jest niezwykle skuteczną metodą ataku. Badając zjawisko phishingu można wykazać wspólne i często chętnie wykorzystywane elementy dla ataków, metody i techniki jakimi posługują się, a które możemy podzielić na dwie grupy:

1. Identyfikacja po stronie użytkownika,
2. Identyfikacja po stronie administratora.

Do pierwszej grupy zalicza się wszystkie te elementy wiadomości email, które zostały przesłane do skrzynki odbiorczej użytkownika końcowego (zakładając, że dana organizacja, dany system pocztowy, nie posiada żadnej wdrożonej funkcjonalności ani mechanizmu detekcji ataku phishingowego). Wiadomości te nie były weryfikowane (lub weryfikowane jedynie w sposób podstawowy – np. skanowanie zawartości załącznika pod kątem obecności złośliwego oprogramowania), lub po weryfikacji przez mechanizmy SPF/DKIM (wynik: unknow[99]), nie znane są dalsze reguły detekcji. Cechami (które można przypisać do jednej z kategorii: techniczne lub nietechniczne) wskazującymi na możliwy atak phishingowy w tego typu wiadomościach mogą być:

[98] Pod pojęciem tym należy rozumieć ogół technik sztucznej inteligencji, metod uczenia maszynowego, klasyfikacji oraz analizy Big Data

[99] Wynik unknow jako wynik weryfikacji mechanizmów SPF/DKIM oznacza, że ww. mechanizmy nie są w stanie właściwie ocenić przetwarzanej wiadomości, tym samy nie są w stanie uznać czy nie występuje spoofing nadawcy – jedna z przesłanek do uznania wiadomości email za phishing.

1. nieprawidłowy odnośnik URL zawarty w wiadomości,
2. złożona nazwa domenowa wraz z subdomenami,
3. adres IP w odnośniku URL zawartym w wiadomości,
4. wykorzystanie serwisów skracających odnośniku URL,
5. złośliwy załącznik, skrypty wykonywalne (VBS/VBA),
6. groźba w przypadku niepodjęcia przez ofiarę sugerowanych działań (wskaźnik nietechniczny),
7. nieprawidłowy adres email nadawcy,
8. niewłaściwy adres nadawcy,
9. niespójność nazwy nadawcy (wskaźnik nietechniczny),
10. wykorzystanie nazwy odbiorcy wiadomości,
11. wykorzystanie nazwy domenowej jako nazwy użytkownika w adresie nadawcy,
12. automatyczne generowanie nazwy użytkownika lub domeny w adresie email nadawcy,
13. mechanizm śledzący w wiadomości email,
14. strona wyłudzająca dane (wskaźnik techniczny / nietechniczny),
15. typosquatting / domen udające istniejące,
16. wiek zarejestrowanej domeny,
17. brak zarejestrowanej domeny, wykorzystanie adresów chmurowych
18. spoofing instytucji,
19. błędy językowe (wskaźnik nietechniczny),
20. temat otrzymanej wiadomości,
21. niespójna szata graficzna (wskaźnik nietechniczny),
22. nietypowe prośby, niespodziewana treść (wskaźnik nietechniczny),
23. niespodziewane załączniki (wskaźnik techniczny/nietechniczny),
24. użycie narzędzi programowych do wysyłki wiadomości email,
25. wykorzystanie tagowania[100] wiadomości przez serwery pocztowe,
26. różne treści osadzone w tej samej wiadomości.

[100] Pod pojęciem tagowania wiadomości email należy zrozumieć, nadanie jej przez serwery pocztowe specyficznego oznaczenia (tag) na podstawie której jest ona katalogowana, indeksowana, przypisana do odpowiedniego folderu lub określony został dla niej poziom zaufania.

Przedstawione powyżej wskaźniki są wyjściową propozycją, na bazie której, w toku analiz wartości poszczególnych cech, ukształtowana zostanie ostateczna lista wskaźników, które posłużą do zbudowania metody detekcji ataku phishingowego.

Druga grupa wskaźników, dotyczy tych wiadomości email, które zostały zatrzymane na wejściu systemu pocztowego danej organizacji i nie zostały jeszcze przesłane do skrzynki odbiorczej użytkownika końcowego. Użytkownicy końcowi mogą jedynie przetwarzać i oceniać własne, odebrane przez nich wiadomości i nie są w stanie porównać pewnych cech z innymi wiadomościami. Z tego powodu porównanie podobieństwa w wielu otrzymywanych wiadomościach email w danej organizacji powinno odbywać się również na poziomie serwera pocztowego – przed ich ostatecznym przekazaniem do właściwej skrzynki odbiorczej użytkownika końcowego. Cechami wskazującymi na możliwy atak phishingowy, możliwymi do wykrycia po stronie administratora mogą być:

1. identyczna treść przesłana ze skompromitowanego konta do wszystkich użytkowników kont email w danej organizacji, w krótkim okresie,
2. masowa wysyłka wiadomości email z jednego konta,
3. wykorzystanie błędnej konfiguracji serwera pocztowego do masowej rozsyłki wiadomości (funkcjonalność „open relay”[101]).

Przedstawione powyżej wskaźniki, definiując je jako cechy phishingu, można wykorzystać do detekcji i klasyfikacji phishingu poprzez automatyczną analizę (bez udziału człowieka):

a. nagłówka[102] wiadomości email,
b. rozpoznawanie wzorców w treści wiadomości (analiza automatyczna oraz niektóre metody klasyfikacji – np. Naiwny Klasyfikator Bayesa),
c. korelację z publicznymi bazami danych zawierających adresację IP oraz domen używanych do dystrybucji wiadomości phishingowych,
d. porównanie z innymi wiadomościami w danej organizacji (podobieństwo cech).

Największym problemem związanym z analizą nagłówków wiadomości e-mail, a więc wydobywaniu cech mogących świadczyć o potencjalnym ataku phishingowym

[101] Open Relay – określenie serwera pocztowego, którego oprogramowanie nie jest zabezpieczone przed nieautoryzowanym wykorzystaniem przez osoby niepowołane do wysyłki poczty elektronicznej.
[102] Nagłówek wiadomości email – dane niewidoczne dla odbiorcy wiadomości (ukrywany przez klienta pocztowego), zawierająca m.in. informacje o rzeczywistym nadawcy, informacje o adresacie, informacje o sposobie dostarczenia e-maila od nadawcy na serwery pocztowe odbiorcy

jest możliwość ich sfałszowania. Atakujący może dowolnie skonfigurować serwer pocztowy (np. za pomocą dostępnego w sieci Internet oprogramowania typu „open source"). Pewną informacją w nieprzetworzonym nagłówku wiadomości jest ostatni wpis w łańcuch pól ścieżki przesyłania wiadomości – gdyż wpis ten pochodzi od serwera pocztowego, który obsługuje docelowego odbiorcę wiadomości. Wszystkie inne informacje zawarte w łańcuchu, mogą pochodzić od zafałszowanych (złośliwych) serwerów pocztowych, które celowo mogły dokonać modyfikacji pól nagłówka, w celu ukrycia tożsamości faktycznego nadawcy (atakującego). Dostępność narzędzi do przeprowadzania zautomatyzowanych ataków[103], powoduje ich wykorzystanie przez osoby, które nie posiadają specjalistycznej wiedzy na temat konstrukcji wiadomości email i konfiguracji środowiska służącego do wysyłki wiadomości i zarządzania całym procesem ataku. Powoduje to często błędną, niewłaściwą konfigurację środowiska, w tym również niewłaściwą modyfikację pól nagłówka wiadomości. Wykryte niespójności i błędy konfiguracyjne zapisane w nagłówku, mogą wówczas zostać wykryte i przypisane do rozpoznanych cech świadczących o możliwym ataku phishingowym.

III.1 Opis cech

Wskazujące na phishing cechy, mogą być łatwo identyfikowane przez eksperta podczas analizy otrzymaj wiadomości. Dane te są w formie nieprzetworzonej, pozyskane bezpośrednio z wiadomości i w takiej formie interpretowane przez eksperta, nie nadają się jednak do wykorzystania w metodach klasyfikacji danych, realizowanych automatycznie, bez udziału człowieka.

Sposób ilościowej oceny cech

Zdefiniowane opisy cech wiadomości email, dzięki którym można dokonać identyfikacji ataku phishingowego muszą zostać poddane procesowi normalizacji. Proces normalizacji cech, musi obejmować kroki:

1. W procesie rozpatrywać będzie się jedynie te cechy, które można analizować w sposób automatyczny (nie wymagający ingerencji człowieka do ich rozpoznania i kodowania).
2. Kodowanie numeryczne pozostałych cech po procesie usuwania.

[103] Takim narzędziem jest framework Gophish, dostępny publicznie w sieci Internet (https://getgophish.com/, dostęp 13.02.2022r)

3. Transformacja danych – uwypuklenie tych zmiennych, które mogą mieć największy wpływ na proces. W etapie tym należy do kodowania wybrać te cechy, których obecność ma duży wpływ na sukces ataku.

Brak normalizacji powoduje, że zmienne (cechy) o niskiej wartości (np. subiektywne określenie podobieństw szaty graficznej) mają wpływ na proces klasyfikacji zaburzając jego wartość.

Kodowanie

By skutecznie przeprowadzić proces klasyfikacji, na podstawie wykrytych i odczytanych cech, konieczne jest zapisanie ich w sposób możliwie jednoznaczny i łatwy w interpretacji. Zdefiniowane cechy wiadomości email, wyrażane za pomocą wartości:

1. „**0**” – dla poniżej opisanych przypadków:
    a. cech występuje w badanej próbce,
    b. cecha nie wskazuje na możliwość phishingu,
2. „**1**” – dla poniżej opisanych przypadków:
    a. cecha występuje w wiadomości email,
    b. wartość/właściwość wskazuje na jej phishingowy charakter,
3. „**null**” – dla poniżej opisanych przypadków:
    a. brak cechy,
    b. pole nie występuje w badanej próbce,
    c. cecha występuje w badanej próbce, natomiast jest wartość jest nieustalona.

Dana cecha identyfikowana w pojedynczej próbce, może przyjąć jedynie jedną z podanych wyżej wartości, które możemy zapisać w postaci wektora cech, przybierającego postać:

$$C = [f_0, \dots, f_n] \quad n = 26 \tag{3.4}$$

gdzie:
fi – i-ta cecha
n – numer cechy

III.1.1 Nieprawidłowy odnośnik

Otrzymana wiadomość zawiera szatę graficzną, układ wiadomości czy treść pochodzącą od rzeczywistego nadawcy, jednakże jeden z elementów jest podmieniony

(często jest to odnośnik do pobrania dokumentu czy przekierowanie do formularza, gdzie konieczne podanie jest wrażliwych danych, np. danych logowania się do systemu elektronicznego). Adres URL zwykle kieruje użytkownika na zasób sieciowy posiadający zupełnie inny adres niż wynika to z treści wiadomości lub adresu email nadawcy znajdującego się w polu „Od" (część domenowa adresu[104]). Jeżeli widoczny adres nadawcy (pole „Od") jest zafałszowany za pomocą techniki zwanej spoofingiem adresu email, fałszywy odnośnik URL może pokrywać się z adresem domenowym nadawcy wiadomości, jednakże adres serwera nadawcy (niewidoczne dla użytkownika pole nagłówka wiadomości) będzie się różnił. Przyjętą praktyką przez profesjonalne firmy i instytucje jest posiadanie własnych zasobów sieciowych (również w rozwiązaniu chmurowym), na którym przechowywane są wszelkie dokumenty i dane. Przesyłane w korespondencji email odnośniki URL prowadzą zwykle do oficjalnych zarejestrowaną na daną firmę zasobów (domen internetowych) lub możliwych do jednoznacznej identyfikacji jako należące do danej firmy.

Tabela 3. Przykład nieprawidłowego odnośnika URL w wiadomości email.

Lp.	Adres nadawcy[105]	Adresy odnośników	Nieprawidłowy odnośnik	ID próbki
1.	facebo15@int.pl	https://facebook.com	https://**pl-fb.com**/business/help/1824726469 3872254	P-01-001
2.	allegrolocalnie@gmail.com	https://t.allegro.pl/oferta/fairy-kapsulki-do-zmywarki-platinum-plus-84-sztuki-	https://**bit.ly**/30ZwC67	P-01-002

Cecha ta może posłużyć jako wskaźnik wiadomości phishingowej, jeżeli:

a) Odnośnik prowadzi do domeny znacząco różnej od domen zawartych w pozostałych odnośnikach URL (pod warunkiem wielu odnośników URL w danej wiadomości email),
b) Odnośnik prowadzi do domeny różnej niż wynika to z adresu nadawcy, nadawca nie posługuje się publicznie dostępnym adresem email (np.: @wp.pl, @onet.pl, @gmail.com),
c) Domena znajduję się bazie serwisów antyphishingowych (np. PhishTank).

[104] Część domenowa adresu email (pełny adres serwera pocztowego) znajduje się po znaku „@".
[105] Widoczny adres nadawcy (pole „Od") lub adres IP/domenowy z nagłówka wiadomości email (Pole „Received").

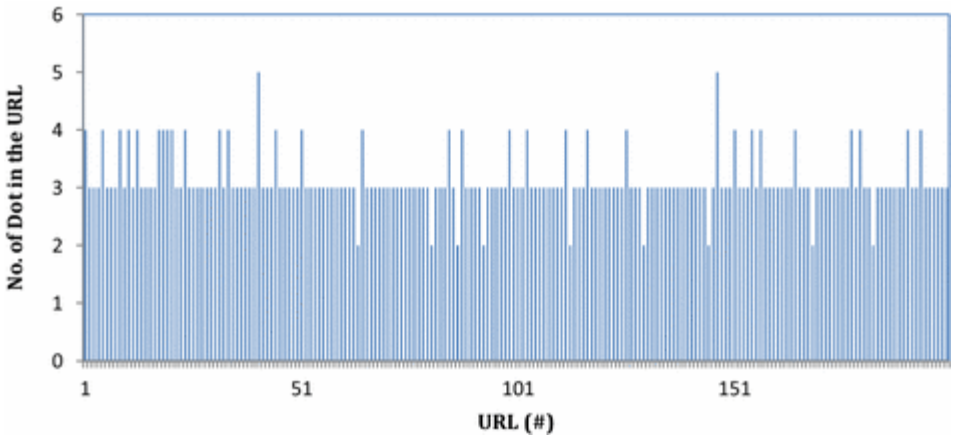
Obserwowane jest również użycie w wiadomościach phishingowych skróconego adresu URL, za pomocą dostępnych serwisów oferujących skrócenie odnośnika. Odnośnik taki prowadzi użytkownika w pierwszej kolejności do serwisu skracania odnośników, gdzie następuje odczyt parametrów odnośnika, a następnie następuje przekierowanie do właściwej strony internetowej. Technika często stosowana w wiadomościach do ukrycia prawdziwego adresu strony phishingowej (użytkownik widzi tylko skrócony adres URL).

Możliwym błędem klasyfikacji, bazującej na tej cesze jest możliwość sklasyfikowania wiadomości jako phishing, jest umieszczenie w prywatnej wiadomości odnośników URL prowadzących do różnych zasobów. Adresy odnośników prowadziły wiec będą do różnych domen i będą się różniły od adresu domenowego nadawcy wiadomości. Wyjątkiem od tej reguły, będzie wiadomość, wysłana z serwisu informacyjnego oferującego usługi darmowej poczty email (np. Wirtualna Polska lub Onet), która będzie zawierała odnośnik do publikacji (artykułu) znajdującej się w obrębie tego serwisu – a więc adresy URL i domenowy adresu email będą zgodne.

Problemem mogącym dawać nieprawidłowy wynik klasyfikacji wiadomości jako phishingu jest umieszczenie wielu odnośników URL w pojedynczej wiadomości email. Odnośniki te mogą prowadzić do różnych domen, wówczas ocenie muszą zostać poddane, wszystkie, które zostały umieszczone w analizowanej wiadomości. Wśród ogólnej liczby mogą znajdować się również złośliwa pętla – odnośnik URL pozornie przekierowując do bezpiecznego zasobu sieciowego (np. Google Drive), na którym umieszczony został kolejny odnośnik, kierujący już do właściwego, złośliwego zasobu.

Analizując phishingowe odnośniki URL należy również mieć na uwadze, konstrukcję samego adresu. Prowadzone badania [69] [70] [71] wykazują, że zidentyfikowane jako phishingowe odnośniki URL wykazują pewne cechy, różniące je od cech rzeczywistego adresu, a tym samym mogące być automatycznie wykryte i sklasyfikowane jako złośliwe (bez wcześniejszej znajomości charakteru odnośnika). Do cech tych należą:

1. większa ilość kropek rozdzielających poszczególne człony adresu domenowego (subdomeny),
2. większa ilość znaków slash („/”) w adresie URL,
3. częstsze występowanie określonych znaków specjalnych (np. „-”, „@”, np.),
4. złożona nazwa domenowa.



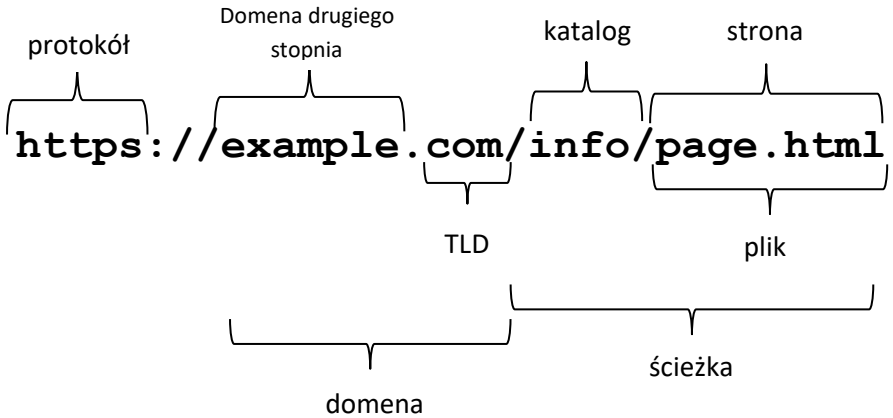
Rysunek 30 Ilość kropek w adresach URL, źródło: https://hcis-journal.springeropen.com/articles/10.1186/s13673-016-0064-3/figures/7

Cechy te, mogą jednak ulegać zmianie np. poprzez zastosowanie serwisów skracających odnośniki, konieczne jest więc wypracowanie metody analizującej, która uwzględniać będzie stosowana przez atakujących zmiany.

III.1.2 Złożona nazwa domenowa wraz z subdomenami

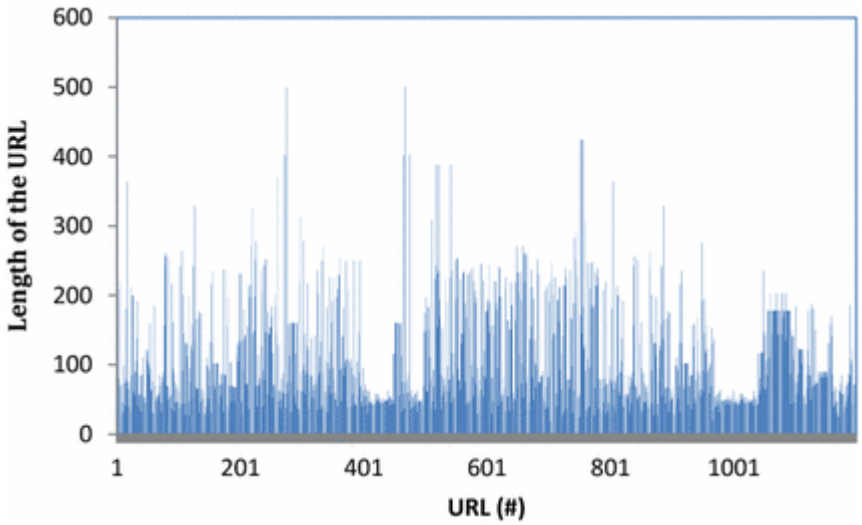
Publikowanie list zawierających adresy domenowe uznawana za phishingowe może okazać się niewystarczające. Jedną ze strategii stosowanej przez atakujących jest użycie rzeczywistej nazwy domenowej jako subdomeny w domenie phishingowej, np. finanse.play.pl.mojserwer.virtual.com.pl (dla rzeczywistej domeny finanse.play.pl). Użytkownicy widząc na początku długiej nazwy domenowej (subdomeny), wartość (nazwę) rzeczywistej domeny, zwykle pomijają pozostałe człony nazwy domenowej.

Długość takiego odnośnika zwykle bywa większa niż odnośnika URL prowadzącego do rzeczywistej, nie-phishingowej domeny.



Rysunek 31. Elementy składowe odnośnika URL.

Zgodnie z przeprowadzonymi badaniami [69] adresy domenowe w odnośnikach URL prowadzące do domen phishingowych zawierają średnio więcej niż 25 znaków (pomijając wskazane poniżej długości stałych elementów). Samo badanie długość części domenowej, może okazać się również niewystarczająca, gdyż atakujący używają długich adresów URL do zamaskowania, ukrycia części adresu URL, który mógłby sugerować phishingowy charakter danego odnośnika URL. W badaniu [69] stwierdzono, że średnia długość legalnego adresu URL wynosi 40 znaków. Średnia długość phishingowego adresu URL przekracza natomiast 75 znaków.



Rysunek 32. Długość phishingowego adresu URL, źródło: https://hcis-journal.springeropen.com/articles/10.1186/s13673-016-0064-3/figures/14

Badania przeprowadzone zostały na bazie 1200 adresów URL zaklasyfikowanych jako phishingowe oraz 200 adresów URL uznanych za rzeczywiste. Baza adresów URL pochodziła z serwisu phishtank.com.

Atakujący może tworzyć dowolną wartość odnośnika URL (w tym stosując inżynierię społeczna, wykorzystując typosquatting do utworzenia pierwszego członu subdomeny. Obliczając całkową długość odnośnika URL, należy z niego wykluczyć następujące, stałe elementy:

1. Protokół (NP., HTTPS) – w odnośniku URL zawartym w wiadomości email, nazwa protokołu zawsze występuje (l_1=4 lub l_1=5),
2. Znaki po nazwie protokołu – „://” (l_2=3),

3. TLD[106] - każdy odnośnik URL zawierać musi domenę najwyższego rzędu (standardowo l_3=2 lub l_3=3 lub l_3=4).

Wartość długości tych stałych elementów, należy wykluczyć z całkowitej długości odnośnika URL, zgodnie z poniższym wzorem:

$$l = l_c - (l_1 + l_2 + l_3) \tag{3.5}$$

gdzie:
l – wyliczona długość odnośnika URL do analizy,
l_c – długość całkowita (pierwotna) analizowanego odnośnika URL,
l_1, l_2, l_3- długość stałych elementów

Wskaźnikiem, że dany odnośnik URL może prowadzić do domeny phishingowej bywa również wykorzystanie w adresie cyfr lub wykorzystanie przez atakującego algorytmu DGA[107].

Zasadniczym problemem w klasyfikacji jest możliwość umieszczenia w pojedynczej wiadomości email wielu odnośników URL, prowadzących do różnych domen o różnych długościach i stopniach skomplikowania, lub do dynamicznie generowanych dokumentów (np. dokumentów znajdujących się w rozwiązaniach chmurowych). Długość odnośnika URL dla takich dokumentów (generowanych automatycznie), zwykle jest znacznie dłuższa niż średnia wynikająca z analizy długości odnośników URL uznawanych za phishingowe.

Kolejnym problemem w klasyfikowaniu długości domeny phishingowej dające fałszywie negatywne wyniki jest wykorzystanie przez atakującego serwisu skracającego odnośniki (np. TinyURL). Długi wówczas odnośnik prowadzący do faktycznej domeny phishingowej, jest skracany i nie klasyfikowany jako wskaźnik phishingu.

III.1.3 Adres IP w odnośniku URL zawartym w wiadomości

Brak zarejestrowanej domeny nie bywa przeszkodą w przygotowaniu kampanii phishingowej. Uruchomiony serwer C2 (fizyczny lub wirtualny – VPS) dysponując przydzielonym adresem IP może pełnić taką samą funkcję serwera Command and Control

[106] TLD (ang. **T**op **L**evel **D**omain) - domena najwyższego rzędu, jest to część nazwy domenowej danej strony internetowej, pisana po ostatniej kropce, np. ".com" czy ".pl". Znajduje się na szczycie hierarchii w strukturze systemu nazw domen (DNS).
[107] DGA (ang. domain generation algorithm) – algorytm służący do okresowego generowania dużej liczby nazw domen, wykorzystywanych przez złośliwe oprogramowanie do komunikacji z serwerami C2.

jak serwer, dla którego istnieją wpisy w serwerze DNS. Wobec tego, w trakcie przygotowywania kampanii phishingowej, odnośnik URL zawierał będzie zamiast adresu domenowego, adres IP serwera, do którego kierowany będzie użytkownik.

Możliwą do wykrycia cechą jest wydobycie z treści wiadomości email wszystkich

```
https://172.10.1.2/order/user/details.html
```

Rysunek 33. Przykład adresu IP wewnątrz odnośnika URL.

odnośników URL, a następnie sprawdzenie każdego z niech, czy zawarty jest w nim prawidłowy adres IP v4. Aby dany wskaźnik traktowany był jako możliwa do wykrycia cecha, muszą być spełnione warunki:

a. wiadomość email zawiera co najmniej jednej odnośnik URL w treści,
b. odnośnik URL zawiera zamiast adresu domenowego dokumentu, do którego prowadzi, adres IP,
c. adres IP zawarty w odnośniku jest prawidłowym adresem Ipv4 lub Ipv6,
d. adres IP zawarty w odnośniku nie jest adresem lokalnym (127.0.x.x) lub adresem zastrzeżonym,

Możliwym błędem klasyfikacji jest wykrycie prawidłowego adresu wewnątrz odnośnika URL w korespondencji wewnętrznej, gdzie użytkownicy mogą przesyłać sobie adresy do wewnętrznych zasobów (dana organizacja z różnych przyczyn nie wdrożyła wewnętrznego serwera DNS i dostępne dla pracowników zasoby sieciowe nie są osiągalne po nazwie, tylko po wewnętrznym adresie IP).

III.1.4. Wykorzystanie serwisów skracających odnośniku URL

Obserwowanym zjawiskiem jest wykorzystanie dostępnych darmowych hostingów [72] pod przygotowanie kampanii phishingowej. Darmowe serwisy często nie umożliwiają użytkownikowi podłączenie, posiadanego przez niego adresu domenowego do założonej usługi hostingowej. Założony przez niego serwis zwykle identyfikowany jest poprzez wybraną przez niego nazwę serwisu (jako subdomena) w połączeniu z adresem domenowym providera oferującego usługę darmowego hostingu, np.:

```
https://mojanazwa.darmowyhosting.pl
```

Tak skonstruowany adres domenowy, może zostać łatwo zidentyfikowany jako próba oszustwa (zwłaszcza gdy w kampanii wykorzystano technikę spoofingu – podszycie się pod inną osobę lub instytucję). Z tego powodu atakujący generują w serwisach umożliwiających wygenerowanie skróconego odnośnika URL (np. TinyURL lub bit.ly), nowy odnośnik (skrócony), którego wartość (nazwa) nie umożliwi odgadnięcie adresu docelowego. Taki odnośnik umieszczony jest w odpowiednio przygotowanej wiadomości email i z wykorzystaniem inżynierii społecznej, potencjalna ofiara zachęcana jest do kliknięcia w niego.

Tabela 4. Przykład wykorzystania serwisu skracającego odnośniki URL w wiadomości email.

Lp	Odnośnik	Wykorzystany serwis skracający	Rzeczywisty zasób zdalny	ID próbki
1.	https://bit.ly/3p5hxr6#455074775a2191474a14455	bit.ly	http://ww12.kitanders.com/#455074775a2191474a14455	P-04-001

W celu skutecznego wykrywania skróconych odnośników URL w analizowanej wiadomości email, konieczne jest posiadanie listy adresów domenowych serwisów oferujących usługi skracania linków.

Trudnością jest duży narzut administracyjny w utrzymanie aktualnej listy:

1. usuwanie nieaktywnych, nieistniejących serwisów,
2. wyszukiwanie i dodawanie nowych.

Cecha $\mathbf{f_3}$ przyjmie wartość 1, gdy adres domenowy odnośnika zawartego w danej wiadomości pokrywał się będzie z adresem domenowym serwisu oferującego usługę skracania linków. Wartość 0, cecha ta przyjmie w pozostałych przypadkach (domena odnośnika nie pokrywa się z adresem domenowych skracacza).

Możliwym błędem klasyfikacji może być wykorzystanie w prywatnej korespondencji serwisów skracania linków do przesłania skróconego odnośnika zawierającego wiele różnych wartości – celem oszczędności miejsca (np. dla klientów pocztowych urządzeń mobilnych).

III.1.5 Złośliwy załącznik

Treść otrzymanej wiadomości sugeruje odbiorcy otrzymanie dokumentu (faktury, umowy, wezwania) w formie załącznika. Załącznik zawiera często złośliwy kod wykonywalny w postaci macro osadzonego w dokumentach Office lub jest plikiem

skryptu Visual Basic (VBS, VBA), a którego uruchomienie finalnie prowadzi do infekcji systemu ofiary, choć sam skrypt nie wykonuje innych czynności na stacji ofiary jak komunikacja z infrastrukturą atakującego (tzw. Dropper[108] lub stager[109]). Najpopularniejszą metodą jest załączanie skryptu VBS/VBA w postaci zarchiwizowanego pliku (*.rar, *.zip) udającego dokument (np. faktura, zawiadomienie o spłacie, pismo urzędowe).

Wykorzystywanie złośliwych załączników w postaci skryptów VBS/VBA, dołączanych do wiadomości email w postaci rzekomych dokumentów finansowych czy faktur do opłacenia, było niezwykle popularne w kampanii kierowanej przeciwko polskim użytkownikom sieci Internet w 2018 roku. Skrypty były zarchiwizowane do postaci pliku *.RAR celem utrudnienia detekcji przez silniki antywirusowe operatora usług email.

Mniej popularną metodą[110] jest dołączanie złośliwego załącznika w postaci pliku wykonywalnego (exe), którego uruchomienie prowadzi do bezpośredniej infekcji komputera użytkownika.

Tabela 5. Przykład złośliwych załączników osadzonych w wiadomości email.

Lp.	Nazwa załącznika	Typ załącznika wynikający z treści wiadomości	Rzeczywista zawartość	Rozpoznane zagrożenie[111]	Ocena zagrożenia (threat score -ts)	ID próbki
2.	FAKTURY I DOWODY DOSTAWY..000672021.rar	pdf lub format Office	faktury i dowody dostawy.exe			P-05-001
3.	NADCHODZĄCE ZAMÓWIENIE 76217800.rar	pdf lub format Office		Razy.Generic, Trojan:MSIL/ Cryptor	95/100	P-05-002

[108] Dropper (ang. zakraplacz) – rodzaj konia trojańskiego służący do nawiązania łączności z serwerem Command and Control, pobrania z niego złośliwej zawartości oraz zainstalowanie złośliwego oprogramowania w systemie docelowym. Celem tego typu oprogramowania jest zapewnienie przetrwania w systemie ofiary i umożliwienie komunikacji z serweremC2.

[109] Stager (od and. staging – inscenizacja) – przyjęta w środowisku osób związanych z sektorem cyberbezpieczeństwa nieformalna nazwa oprogramowania – elementu łańcucha infekcji, przygotowującego infekowane środowisko pod działanie dalszych złośliwych części oprogramowania.

[110] Niektórzy dostawcy usług email (np. Google czy Microsoft) zaimplementowali w swoich rozwiązaniach metody blokowania załączników do wiadomości email w postaci plików wykonywalnych (exe). Źródło: https://support.microsoft.com/pl-pl/office/za%C5%82%C4%85czniki-blokowane-w-programie-outlook-434752e1-02d3-4e90-9124-8b81e49a8519,

[111] Analizę złośliwego oprogramowania wykonano za pomocą darmowego narzędzia „Hybrid Analysis" dostępnego pod adresem: https://www.hybrid-analysis.com/, za pomocą „VirusTotal" dostępnego pod adresem: https://www.virustotal.com/gui/ oraz oprogramowania antywirusowego systemu Windows 10 (Windows Defender). Odnośniki do wyników analiz, znajdują się w dodatku A.

Wykrycie złośliwego załącznika, może nastąpić już na etapie skanowania zawartości wiadomości w momencie odebrania jej przez docelowy serwer pocztowy lub podczas pobierania danego załącznika na lokalny dysk przez zainstalowane oprogramowanie antywirusowe użytkownika – co zmniejsza możliwą powierzchnię ataku. Skanowanie załączników poczty odbieranej przez dany serwer pocztowy realizowane jest przeważnie przez zawansowanych technologicznie dostawców usług (np. Google[112]), którzy dodatkowo uniemożliwiają załączenie do wiadomości email plików wykonywalnych (exe) czy skryptów (VBS/VBA, js, np.) jako najczęstszych nośników złośliwego kodu. W celu uniemożliwienia wykrycia złośliwego oprogramowania w załączniku wiadomości email, stosowana jest wspomniana technika, polegająca na zarchiwizowaniu pliku zawierającego złośliwy kod (.rar, .zip).

$$f_5 = \begin{cases} 1 & dla\ ts > 0 \\ 0 & dla\ ts = 0 \\ 0 & dla\ z \in \emptyset \end{cases} \tag{3.6}$$

gdzie:
ts – (ang. Threat score – ocena zagrożenia) – poziom zagrożenia danego oprogramowania dla bezpieczeństwa systemu operacyjnego.
Z – załącznik w wiadomości email,

Poziom zagrożenia ***ts*** (ocena zagrożenia) obliczany jest przez oprogramowanie antywirusowe na bazie sumowania wartości liczbowych jakie są przyporządkowane poszczególnym regułom antywirusowym, które dane złośliwe oprogramowanie wyzwoliło [73]. Każda reguła systemu antywirusowego ma przypisaną niezerową wartość.

Dostępne rozwiązania antywirusowe pozwalające na skanowanie załączników w wiadomościach email są w różny sposób implementowane przez dostawców usług pocztowych, dlatego by porównać skuteczność wykrywania złośliwych załączników znajdujących się w danej wiadomości email, konieczne jest zbudowanie metody porównawczej. Metoda taka musi wyjaśnić:

[112] https://support.google.com/mail/answer/6590?hl=pl#zippy=%2Cwiadomo%C5%9Bci-z-za%C5%82%C4%85cznikami

1) Jaką techniką wykrywania dysponuje dane narzędzie?
2) Jakie jest to narzędzie – jak skonfigurowane, od którego producenta pochodzi?
3) W jaki sposób dane narzędzie zostało zaimplementowane – w jaki sposób, jaką techniką dostarczane są do niego załączniki do skanowania, Jak dostarczanych?
4) W jakim okresie od opublikowania techniki dostarczania złośliwego oprogramowania narzędzie potrafi wykryć złośliwy załącznik?
5) Czy wektor dostarczenia złośliwego oprogramowania i zasada działania złośliwego kodu jest znana czy jest to nowa forma ataku?
6) Jaki poziom wyrafinowania technicznego jest po stronie atakującego?

Różnice są też w stosowanej polityce odnośnie możliwości przesłania przez danego dostawce usług poczty elektronicznej, załączników. Niektórzy dostawcy nie wdrożyli żadnych (lub tylko częściowo) mechanizmów pozwalających na sprawdzenie czy załącznik nie zawiera złośliwej zawartości, niektórzy dostawcy wdrożyli natomiast bardzo zaawansowane mechanizmy – łącznie z regułami YARA[113] (np. Google), pozwalające na wysoką skuteczność w eliminowaniu dystrybuowania złośliwego oprogramowania.

Tabela 6. Skuteczność 10 najlepszych rozwiązań antywirusowych, źródło: Malware Protection Test March 2022

Antivirus name	Detection Rate (Offline)	Detection Rate (Online)	Protection Rate (Online)	False alarm
Avast	94.2%	99.5%	99.98%	10
AVG	94.2%	99.5%	99.98%	10
Avira	96.0%	98.7%	99.96%	1
Bitdefender	97.8%	97.8%	99.99%	8
ESET	96.1%	96.1%	99.77%	0
G DATA	98.6%	98.6%	99.99%	59
K7	94.6%	94.6%	99.85%	25
Kaspersky	78.0%	95.4%	99.98%	2
Malwarebytes	77.3%	93.3%	99.75%	7
Avast	94.2%	99.5%	99.98%	10

[113] YARA - narzędzie umożliwiające wyszukiwanie i klasyfikacje plików różnego rodzaju w oparciu o występujące w nich ciągi danych, a w korzystając z rozszerzonej funkcjonalności również cech pliku. Reguły YARA można interpretować jako fragment języka programowania, działającego poprzez zdefiniowanie szeregu zmiennych, zawierających wzorce (ciągi danych) znalezione w próbce złośliwego oprogramowania. Jeśli niektóre lub wszystkie warunki są spełnione (w zależności od warunków zdefiniowanych w regule), można go wykorzystać do pomyślnej identyfikacji złośliwego oprogramowania.

III.1.6 Groźba (szantaż)

Wykorzystanie inżynierii społecznej do przeprowadzenia ataku. Wiadomość email zawiera groźbę publikacji rzekomo kompromitujących materiałów odbiorcę wiadomości. Groźba ta nie zostanie spełniona, pod warunkiem wpłaty okupu zwykle w postaci transakcji krytptowalut. Popularną odmianą tego typu wiadomości, jest przesłanie wiadomości email, zawierającego informację o rzekomej infekcji komputera ofiary złośliwym oprogramowaniem, które umożliwiło kontrolę kamery i za jej pomocą nagranie kompromitujących materiałów o charakterze seksualnym, którymi atakujący chce szantażować ofiarę i wymusić okup. Treść wiadomości bazuje na założeniu, że potencjalna ofiara nie jest świadoma technik jakimi posługują się szantażyści[114], wiadomość zawiera adres portfela kryptowaluty do wpłacenia okupu. Adres ten zwykle jest unikalnie wygenerowany dla danej ofiary, dzięki czemu szantażysta może powiązać wpłatę (okup) z adresem email na jaki została wysłana wiadomość.

Tabela 7. Przykład zawartości wiadomości email, zawierającej groźbę, szantaż.

Lp.	Wykryty ciąg	Słowa kluczowe / fazy	Rodzaj kryptowaluty	ID próbki
1.	1HtAVTagjp7dMCrfJiPws3fuTHBsy4fKRu	▪ „zdobyłem dostęp do urządzenia”, ▪ „zapisałem wszystkie dane”, ▪ „Masz 48 godzin na odpowiedź”, ▪ ‘film natychmiast zostanie udostępniony”	Bitcoin (BTC)	P-06-001
2.	bc1qfg243xzz0l5n206wawfrd3fenkxnmw8zpqsgaa	▪ „wszystkie dane zostały skopiowane”, ▪ „kontrolę nad urządzeniem”, ▪ „mogę zniszczyć twoją reputację na zawsze”, ▪ „masz 50 godzin (nieco ponad 2 dni”, ▪ „nie rób nic głupiego”	Bitcoin (BTC)	P-06-002

Charakterystycznymi cechami tego typu wiadomości są:

1. brak załącznika,[115]

[114] Ostrzeżenia przez wpłatą okupu wraz z opisem tej metody szantażu, opisywana była wielokrotnie, m.in. przez zespół CERT POLSKA (https://www.facebook.com/CERT.Polska/posts/3305743362779364/) oraz na łamach portalu Niebezpiecznik (https://niebezpiecznik.pl/post/masz-przejebane-wiadomosc-ktora-dzis-otrzymalo-kilka-tysiecy-polakow/, https://niebezpiecznik.pl/post/twoje-konto-zostalo-zhackowane/)

[115] W analizowanej próbce oznaczonej jako „P-03-001” występował załącznik w postaci pliku tekstowego (txt), o takiej samej nazwie jak temat wiadomości, którego treść odpowiadała treści innych wiadomości email zawierających groźbę/szantaż.

2. brak odnośników URL,
3. treść wiadomości sugerująca posiadanie kompromitujących ofiarę materiałów.

Automatycznie do wykrycia cechą, zawartą w wiadomościach noszących znamiona szantażu, jest ciąg znaków odpowiadających wartości portfela BitCoin (jako najpopularniejszej[116] kryptowaluty).

$$f_6 = \begin{cases} 1 & dla\ z \in \emptyset, a \in \emptyset, btc = 1 \\ 0 & dla\ z \notin \emptyset\ lub\ a \notin 0, btc = 1 \\ null & dla\ btc \in \emptyset \end{cases} \quad (3.7)$$

gdzie:
z – zbiór załączników w wiadomości email,
a – zbiór odnośników URL zawartych w wiadomości email,
btc – występujący w treści wiadomości ciąg znaków będący adresem portfela kryptowaluty Bitcoin

By wskaźnik ten można było uznać za cechę phishingu, należy dodatkowo również poddać analizie treść otrzymanej wiadomości. Wymienione przykłady fraz dla analizowanej próbki w kolumnie „Słowa kluczowe / fazy" zawarte w Tabela 7 mogą stanowić zbiór słów kluczowych do treningu klasyfikatora (np. Naiwny Klasyfikator Bayesa), przewidującego czy otrzymana treść wiadomości jest charakterystyczna dla wiadomości wymuszających okup. Opis wskaźnika przyjmie wówczas postać:

1 – w sytuacji, gdy łącznie spełnione są warunki: w treści wiadomości występuje prawidłowy adres portfela BitCoin, wiadomość nie zawiera żadnego załącznika ani żadnego odnośnika URL, oraz w wyniku uczenia maszynowego na treści wiadomości, zwrócona klasa wskazuję na szantaż.

0 – w pozostałych przypadkach.

III.1.7 Nieprawidłowy adres email nadawcy

Nazwa wyświetlana jako nadawca (pole „Od") nie jest tożsama z adresem serwera wysyłającego daną wiadomość. Technika podmiany pola nadawcy na dowolną nazwę nosi nazwę spoofingu adresu email. Prawidłowy adres IP serwera, z którego użytkownik otrzymał daną wiadomość – pole „Received" znajduje się z nagłówku wiadomości. (patrz: Tabela 8).

[116] https://www.forbes.com/advisor/investing/top-10-cryptocurrencies/ [stan na dzień 28.07.2021]

Tabela 8. Przykład nieprawidłowego adresu email nadawcy.

Lp	Domena wynikająca z widocznego adresu nadawcy	Serwer nadawcy	Adres IP domeny	IP serwera nadawcy	ID próbki
1.	WP.pl	mail03.news.grupazprmedia.net	193.17.41.249 [117]	141.145.12.175	P-07-001
2.	olx.pl	69-164-207-174.ip.linodeusercontent.com	13.249.75.83 13.249.75.12 13.249.75.21 13.249.75.3	69.164.207.174	P-07-002

Informacje o rzeczywistym nadawcy wiadomości (adres ostatniego serwera pocztowego przekazującego wiadomość do serwera pocztowego obsługującego danego odbiorcę) nie są widoczne dla użytkownika korzystającego zarówno z klienta pocztowego (takich jak Microsoft Outlook czy Mozilla Thunderbird) jak i poprzez przeglądarkę internetową (web client). Dane te zapisane są w nagłówku wiadomości email, wraz z innymi danymi, za pomocą których można dokonać identyfikacji rzeczywistego nadawcy wiadomości. Do pól tych należą:

a. Pole „Return-Path" – pole oznaczające adres faktycznego nadawcy [74], wstawiane automatycznie przez serwer wysyłający. Wiadomość email musi zawierać co najmniej jedno pole „Return-Path". Pole to można zawierać inne dane niż pole „Od" widoczne dla odbiorcy wiadomości
b. Pole „Reply-To" – pole obligatoryjne [75] (minimalna ilość wystąpień: 0, maksymalna ilość wystąpień: 1) zawierające dane adresu, na który ma zostać przesłana odpowiedź na daną wiadomość email. Pole to może zawierać inne dane niż pole „Od" i pole „Return-Path"

Spotykaną techniką wykorzystującą właściwości pola „Reply-To" do ukrycia działań atakujących, którzy uzyskali nieautoryzowany dostęp do dowolnego konta email i z niego prowadzą kampanię phishingową jest dodanie niewielkiej zmiany do adresu email widocznego w polu „Od".

[117] Serwer pocztowy Wirtualnej Polski (poczta.wp.pl) znajduje się pod innym adresem IP niż serwis informacyjny (wp.pl)

Tabela 9. Modyfikacja nieprawidłowego adresu email nadawcy w polu „Reply-To”.

Lp.	Widoczna wartość pola „Od/From”	Wartość pola „Reply-To” (niewidoczne dla odbiorcy)
1.	admin@huebubuj.website	admin**2**@huebubuj.website
2.	info@poradniasluchu.radom.pl	info**2**@poradniasluchu.radom.pl
3.	provident@vipro.pl	provident**t2**@vipro.pl
4.	oferty.gdansk@mkbowling.pl	oferty.gdansk**2**@mkbowling.pl

Koszt utrzymania własnej domeny i własnej infrastruktury hostingowo-pocztowej, koszty administracyjne[118] są dość wysokie w porównaniu do obecnych na rynku ofert świadczenia tych usług przez podmioty zewnętrzne (outsourcing). Z tego powodu – dywersyfikując koszty – zaobserwować można przenoszenie usług hostingowych i poczty do rozwiązań chmurowych. W takim przypadku adresem IP nadawcy wiadomości będzie adres IP usługi chmurowej – co w przypadku przesłania rzeczywistej wiadomości z infrastruktury firmowej korzystającej z takiego rozwiązania może zostać fałszywie zakwalifikowane jako phishing. W takim przypadku, proste sprawdzenie widocznego adresu nadawcy (pole „Od”) jest niewystarczające, należy wówczas sprawdzić za pomocą systemu WHOIS[119] adres IP, na który rozwiązuje się nazwa domenowa nadawcy, a następnie sprawdzić czy rozwiązany adres IP nie należy do puli adresów IP przypisanych danemu rozwiązaniu chmurowemu. Sprawdzenie takie należy stosować każdorazowo z powodu:

a. rozwiązania chmurowe mają zwykle zarezerwowaną pewną pulę adresów IP, w zależności od lokalizacji geograficznej serwerowni,
b. rozproszone geograficznie serwery rozwiązania chmurowego mogą rozwiązywać dany adres domenowy na różne adresy IP.

Mając na uwadze powyższe, w celu minimalizacji błędu różnych adresów IP, konieczne jest zbadanie czy adresy należą (zostały zarezerwowane, wykupione) do puli adresów tego samego operatora. Na potrzeby identyfikacji, czy dane adresy IP znajdują się w puli tego samego operatora, można wykorzystać System autonomiczny (AS od ang.

[118] Pod pojęciem kosztu administracyjnego dla utrzymania infrastruktury teleinformatycznej, należy rozumieć jako suma czasochłonności pracy specjalisty/administratora systemu od utrzymania w sprawności działania serwerów (aktualizacje, kopie bezpieczeństwa, poprawki systemowe, codzienne czynności administracyjne), kosztów przechowania kopii zapasowych, kosztów oprogramowania dodatkowego, wynagrodzenie administratora, koszty prądu, zapewnienie ciągłości działania, monitoring parametrów, itp.

[119] WHOIS – jeden z protokołów TCP, działający na zasadzie pytanie-odpowiedź, stosowany do zapytań DNS w celu uzyskiwania informacji na właściciela danej domeny lub jej adresu IP.

Autonomous System[120]), który jest zbiorem adresów IP znajdujących się pod wspólną kontrolą administracyjną (*de facto* należącą do tego samego operatora). Każdy operator posiada unikalny, nadany ASN (ang. Autonomous System Number). Porównując numery ASN różnych adresów IP, można uzyskać informację czy należą do tego samego operatora (tej samej puli adresacyjnej), czy do różnych podmiotów. Ten sam numer ASN, dla różnych adresów IP występujących w danej wiadomości email (np. adres IP wynikający z nazwy domeny adresu nadawcy widniejącego w polu „Od" („From") oraz adres IP serwera pocztowego – pole „Received"), pomimo wyczerpywania cechy nieprawidłowego adresu nadawcy w stosunku do adresów serwerów pocztowych, może świadczyć o korzystania z zewnętrznej infrastruktury i tym samym być rzeczywistą wiadomością email.

Realizując wskazane założenia, przed faktycznym procesem klasyfikacji, należy wykluczyć tą właściwość i nie traktować niezgodności adresów IP (wynikających z domeny adresu nadawcy oraz serwera pocztowego wysyłającego daną wiadomość) jako cechy wiadomości phishingowej.

$$f_7 = \begin{cases} 1 \;\; dla\; IP_n \neq IP_s \;\; i\; ASN_n \neq ASN_s \\ 0 \;\; dla \;\; IP_n = IP_s \;\; lub \; (IP_n \neq IP_s \;\; i\; ASN_n = ASN_s) \end{cases} \quad (3.8)$$

gdzie:
IP_n – adres IP nadawcy wynikający z adresu email znajdującego się w polu „From",
IP_s – adres IP serwera pocztowego nadającego daną wiadomość,
ASN_n – numer ASN dla adresu IP wynikającego z adresu email nadawcy,
ASN_s – numer ASN dla adresu IP serwera nadającego daną wiadomość.

III.1.8 Niewłaściwy adres nadawcy

Adres nadawcy widoczny w polu „Od" ("From") zarejestrowany został u dowolnego, publicznego operatora świadczącego darmowe usługi email (np. Google, Wirtualna Polska, Onet) a szata graficzna i treść wiadomości podszywa się pod istniejącą firmę lub organizację. Profesjonalne firmy, organizacje i instytucje posiadają

[120] Definicja, opis i wymagania systemu autonomicznego dostępne są w dokumencie RFC 1930 (https://dx.doi.org/10.17487/RFC1930 [dostęp: 14.10.2021]).

zarejestrowane nazwy domen (zwykle zgodne z nazwą jaką się posługują[121]) oraz własne systemy pocztowe w obrębie zarejestrowanej przestrzeni nazewniczej.

Tabela 10. Przykład niewłaściwego adresu nadawcy.

Lp.	Adres nadawcy	Nadawca (nazwa) wynikająca z treści/szaty graficznej wiadomości	ID próbki
1.	noreply@olx.pl	LIDL	P-08-001
2.	kontakt@pomocnadlon.co.pl	NETFLIX	P-08-002

Cecha ta jest zbliżona do cechy „Nieprawidłowy adres email nadawcy", zasadniczą różnicą pomiędzy obiema tymi cechami jest, że nieprawidłowy adres email może być dowolną wartością i zwykle nie jest powiązany tematycznie z treścią wiadomość lub wykorzystaną szatą graficzną. W przypadku niewłaściwego adresu, również dochodzi do wpisania w polu „From"/"Od" dowolnej wartości, jednakże wpisana nazwa koresponduje z tematyką wiadomości oraz wykorzystaną szatą graficzną. Wykorzystany adres email, zwykle jest ukrywany, gdyż wskazuje na zupełnie inną domenę niż kojarzona z wyświetlaną nazwą użytkownika.

Częstą praktyką jest uzyskiwanie nieautoryzowanego dostępu do istniejącego konta email, w wyniku pozostania danych w postaci hasła dostępu z publikowanych wycieków baz danych różnych serwisów internetowych. Powszechną praktyką wśród internautów jest stosowanie takiego samego hasła dostępu do różnych serwisów [76] – badanie przeprowadzone na internautach amerykańskich w 2021 roku ukazuje, że ponad 20% badanych używa tego samego hasła do wszystkich serwisów online [77]. Pozyskanie więc w drodze wycieku, jednego hasła danego użytkownika stosującego to samo hasło do różnych serwisów, powoduje możliwość zalogowania się na jego dane do różnych usług i posługiwanie się jego tożsamością w sieci Internet.

Cechą możliwą do wykrycia jest porównanie nazwy użytkownika występującej w polu „Od"/"From" do adresu domenowego.

[121] Wyjątkiem od opisywanej sytuacji może być fakt przejecie jednej marki przez inną i korzystanie pod nową nazwą z istniejącej infrastruktury teleinformatycznej i nazw domenowych przejętej marki.

Tabela 11. Różnice domen pocztowy a domen z adresów email nadawcy.

Zawartość pola „Od”	ID Próbki	Nazwa użytkownika	Domena wynikająca z nazwy użytkownika	Rekord MX[122,123]
LIDL < noreply@olx.pl>	P-08-001	LIDL	lidl.pl	mail13.mail.schwarz mail14.mail.schwarz
NETFLIX <kontakt@pomocnadlon.co.pl>	P-08-002	NETFLIX	netflix.com	aspmx.l.google.com aspmx2.googlemail.com

$$f_8 = \begin{cases} 1 & dla\ d_1 \neq d_2 \\ 0 & dla\ d_1 = d_2 \\ null & u \in \emptyset \end{cases} \quad (3.9)$$

gdzie:
d_1 – domena wynikająca z adresu email widocznego w polu „Od”,
d_2 – domena wynikająca z nazwy użytkownika wpisanego w widoczne pole „Od”,
u – nazwa użytkownika zawarta w widocznym polu „Od”.

III.1.9 Niespójność nazwy nadawcy

Wyświetlana nazwa nadawcy w polu „Od” nie jest spójna z nazwą nadawcy wynikającą z podpisu lub treści wiadomości. Bazując na wiedzy eksperckiej i doświadczeniu zespołów CERT/CSIRT za phishingową wiadomość email można uznać wiadomość, w której widoczna nazwa nadawcy (pole „Od”) jest zasadniczo różna od nazwy użytkownika wynikającej z adresu email czy nazwie znajdującej się w podpisie/stopce wiadomości email.

Tabela 12. Przykład niespójności nazwy nadawcy wiadomości email.

Lp.	Nadawca wynikający z podpisu lub treści wiadomości[124]	Wyświetlany Email nadawcy	Podpis nadawcy (w treści email)[125]	ID próbki
1.	Urząd Skarbowy	TravonteVankeuren1996@gmx.com	mgr Marzena Fierek	P-09-001
2.	Marcin Paluszek	admin@huebubuj.website	Radosław Plutecki	P-09-002
3.	Brak nazwy	zdrowie@lubelskie.pl	Michael J Weirsky.	P-09-003

[122] Rekord MX - wpis zasobu w systemie nazw domen (DNS), który określa serwer pocztowy odpowiedzialny za przyjmowanie poczty w imieniu adresu domenowego
[123] Dane rekordów MX pozyskane z globalnego systemu WHOIS za pomocą platformy: https://centralops.net
[124] zachowano oryginalną pisownię i formatowanie
[125] zachowano oryginalną pisownię i formatowanie

Niespójność nazwy nadawcy, występuje również w konstrukcji samego pola „Od" wiadomości email, widocznego dla użytkownika.

Tabela 13. Przykład niespójności nazwy nadawcy wiadomości email

Lp.	Zawartość pola „Od"	Widoczna nazwa użytkownika	Nazwa użytkownika wynikająca z adresu email	ID próbki
1.	Krzysztof Jarmołowski <123lemen.kopczynski@pmplatform.nazwa.pl>	Krzysztof Jarmołowski	123lemen.kopczynski	P-09-004

Możliwą do wykrycia cechą jest porównanie nazwy użytkownika z nazwą wynikającą z adresu email.

$$f_9 = \begin{cases} 1 \ \ dla\ n_w \neq n_e \\ 0 \ \ dla\ n_w = n_e \\ 0 \ \ dla\ n_w \in \emptyset \end{cases} \tag{3.10}$$

gdzie:
n_w – widoczna nazwa użytkownika w polu „Od",
n_e – nazwa użytkownika wynikająca z adresu email umieszczonego w polu „Od",

Nazwa użytkownika widniejąca w polu „Od"/„From" może również zawierać nazwy użytkowników, świadczące o wysłaniu wiadomość email z konta będącego nazwą funkcjonalną, używaną w danej organizacji/firmie (np. „biuro"). Nazwa ta będzie różniła się od nazwy użytkownika występującego w stopce wiadomości (użytkownik obsługujący daną skrzynkę email, wysyłający wiadomość, zwykle podpisze się swoimi danymi – imię i nazwisko) – nie będzie to jednak atak phishingowy, ale część regularnej korespondencji email.

Tabela 14. Słowa wykluczone ze sprawdzenia niespójności nazwy z uwagi na ich specjalne znaczenie.

Lp.	Nazwa użytkownika konta grupowego/funkcyjnego
1.	admin
2.	kontakt
3.	biuro
4.	sklep
5.	123lementó[126]
6.	noreply
7.	powiadomienia
8.	root

[126] Adresy email nie powinny zawierać polskich znaków diakrytycznych, stad pisownia słowa „sprzedaż", jako „sprzedaż"

W celu wyeliminowania błędu klasyfikacji, polegającego na uznaniu danej wiadomości zawierającej w nazwie użytkownika, nazwę funkcyjną (różną od nazwy w stopce/podpisie wiadomości email), należy opracować listę zawierającą nazwy kont funkcyjnych, które w sprawdzaniu danej cechy (przed klasyfikacją), należy wykluczyć ze sprawdzenia. By utrzymać aktualność listy, należy zapewnić mechanizm aktualizacji używanych nazwy kont funkcyjnych.

Kolejnym, możliwym do wystąpienia błędem klasyfikacji, jest wykorzystywanie nazwy użytkownika „noreply" lub „powiadomienia" do wysyłania powiadomień do zarejestrowanych użytkowników danej platformy lub klientów korzystających z danej usługi. W tym przypadku (podobnie jak dla cechy nr 7 – Nieprawidłowy adres email nadawcy, należy porównać adresy IP wynikające z adresu domenowego występującego w polu „Od" a rzeczywistym adresem serwera pocztowego nadającego daną wiadomość.

III.1.10 Wykorzystanie nazwy odbiorcy wiadomości

Atak phishingowy bazuje na wielu elementach inżynierii społecznej w celu skuteczniejszego przekonania potencjalnej ofiary do podjęcia określonych działań. Jednym z elementów jest wykorzystanie nazwy (fragmentu) odbiorcy jako część lub całość nazwy użytkownika w polu nadawcy (pole „Od") lub jako podrobiony adres email występujący w tym polu. Atakujący dysponują adresem email potencjalnej ofiary (np. pozyskanym z publicznych wycieków baz danych – jednym lub więcej, w zależności od aktywności użytkownika w sieci Internet). Badając aktywność danego użytkownika można spreparować indywidualny dla niego adres email, który zawiera wiele jego (użytkownika) elementów nazewniczych.

Tabela 15. Przykład wykorzystania nazwy odbiorcy lub jej części w wiadomości phishingowej.

Lp.	Adres email odbiorcy	Adres email nadawcy (widoczny w polu „Od")	ID próbki
1.	**lukaszkryger.e2y1s1**@gmail.com	reply@**lukaszkryger.e2y1s1**.drivefact.org	P-10-001
2.	**lukaszkryger.e2y1s1**@gmail.com	lrk5hmj.8BVYY@**Lukaszkryger.e2y1s1pp2yrx**.uk.com	P-10-002

Możliwą do wykrycia cechą jest porównanie nazwy użytkownika lub nazwy użytkownika wynikającą z adresu email widocznym w polu „Od" z nazwą użytkownika -odbiorcy wiadomości.

$$f_{10} = \begin{cases} 1 \ \ dla\ n_w \neq n_e \\ 0 \ \ dla\ n_w = n_e \\ 0 \ \ dla\ n_w \in \emptyset \end{cases} \tag{3.11}$$

gdzie:
n_w – widoczna nazwa użytkownika lub nazwa w użytkownika w adresie email w polu „Od”,
n_d – adres email odbiorcy wiadomości

Nazwa użytkownika odbiorcy wiadomości – podobnie jak nazwa użytkownika w adresie email widocznym w polu „Od”/”From” – może być zapisana na wiele różnych sposobów. Zgodnie z standardem RFC 2821[127] dozwolone są w adresie email znaki ASCII (drukowalne, 7 bit). Jednakże wiele współczesnych systemów dopuszcza jedynie niewielką ilość znaków ASCII. Również sami administratorzy systemów [78] zwracają uwagę na problemy związane z używaniem większości znaków specjalnych w adresie email. Przyjęty obecnie standard, gdzie dozwolonymi znakami specjalnymi, w adresie email jest znak kropki – „.”, znak podkreślenia dolnego – „_” oraz myślnik – „-”. Dozwolone są litery alfabetu łacińskiego oraz cyfry arabskie. Prawidłowa nazwa użytkownika w adresie email, zawierać więc może wszystkie powyższe elementy.

III.1.11 Wykorzystanie nazwy domenowej

Technika polegająca na wykorzystaniu w polu „Od” adresu domowego jako nazwa użytkownika. Zabieg ten jest stosowany, kiedy atakujący mają do dyspozycji atrakcyjny (z punktu widzenia celu ataku) adres domenowy, nie mogą natomiast wykorzystać (lub nie znają) żadnego istniejącego w obrębie danej domeny nazwy użytkownika. Dodatkowym aspektem jest postrzeganie przez Internautów, niektórych adresów domenowych jako nazwy własne firm, akcji/przedsięwzięcia. Taki adres funkcjonuje jako marka własna firmy lub produktu i jest niejednokrotnie bardziej znana niż nazwa firmy oferującej (produkującej) ten produkt - stąd też wykorzystanie tej metody przez cyberprzestępców.

Tabela 16. Przykład wykorzystania nazwy domenowej jako nazwy użytkownika.

Lp.	Nazwa użytkownika	Adres email nadawcy (widoczny w polu „Od”)	ID próbki
1.	doradztwo-bankowe.pl	biuro@doradztwo-bankowe.pl	P-11-001

[127] https://datatracker.ietf.org/doc/html/rfc2821

Problemem wykorzystania tej cechy jako możliwego wskaźnika ataku phishingowego jest jednoczesne wykorzystywanie tej metody przez firmy internetowe budujące dopiero swoją pozycie w sieci Internet. Początkujące firmy chętnie korzystają z możliwości umieszczania nazwy domenowej (która często jest również nazwą/logotypem firmy) we wszelkich możliwych miejscach.

III.1.12 Automatyczne generowanie nazwy użytkownika lub domeny w adresie email nadawcy

Jedną z technik, widoczną również podczas masowej wysyłki niechcianych wiadomości email (spam), jest maskowanie adresu email nadawcy wiadomości (widocznego w polu „From"/"Od")., by nie ujawniać prawdziwych danych atakującego. W tym celu – jako jedna z technik – wykorzystywany jest algorytm generujący automatycznie adres email. Wygenerowana nazwa może być na podstawie:

a. listy słów (słownik),
b. określonego schematu (pattern),
c. całkowicie dowolna – ograniczenie jedynie na długość nazwy.

Nazwy użytkowników wygenerowane na podstawie określonej listy słów, w dużej mierze pokrywają się z nazwami kont tworzonymi przez ludzi i na tym etapie nie jest możliwe rozróżnienie, który adres email opracowany został przez człowieka, a który została automatycznie wygenerowany. Wygenerowanie nazwy użytkownika na podstawie pewnego schematu, stosuje się głównie w organizacjach, gdzie konta użytkowników zakładane są w ramach danego stanowiska pracy lub funkcji, np.: fup.operator12, fup.operator55.

Cechą charakterystyczną generowania nazw kont użytkowników adresów email jest utworzenie całkowicie dowolnej nazwy, ograniczonej jedynie z góry przyjętą ilością znaków, przy czym wykorzystuje się do utworzenia nazwy zarówno litery alfabetu łacińskiego jak i cyfry arabskie. Na potrzeby dokładniejszego rozróżnienia różnych generowanych nazw, a więc i różnych technik tworzenia nazw użytkowników adresów mail, przyjęto następujące typy nazw:

1. **A** – nazwa wygenerowana na podstawie listy słów. Atakujący dysponują listą słów, na podstawie której generowana jest nazwa użytkownika adresu email. Nazwa może zawierać pewną liczbę cyfr (np. rozdzielającą dwa wyrazy).

2. **B** – nazwa wygenerowana na podstawie określonego schematu. Atakujący z góry określają schemat generowanej nazwy – przykładowo może to być dowolna, pojedyncza litera, następująco po niej kropka, dowolny ciąg znaków o określonej długości, zakończone dowolną cyfrą.
3. **C** – dowolnie wygenerowana nazwa. Atakujący określają jedynie długość nazwy, generowanie są dowolne nazwy (w zależności od specyfiki algorytmu) składające się z samych liter (wielkie i małe), liter i cyfr, znaków specjalnych dopuszczonych do wykorzystania w adresie email.

Tabela 17. Przykład generowania nazwy użytkownika w adresie email

Lp.	Adres email nadawcy (widoczny w polu „FROM" / „Od")	Typ nazwy	ID próbki
1.	brcl@gmail.com	C	P-12-001
2.	rp.LiujCxDtc5@mail.4hd1lt272a.com	B	P-12-002
3.	dzbyyzs32wnfcvnvi6k755y81wma2sb0@j xexelgg4dvzzshb4eh8idxmdzw9tb1i.drivefact.org	A	P-12-003

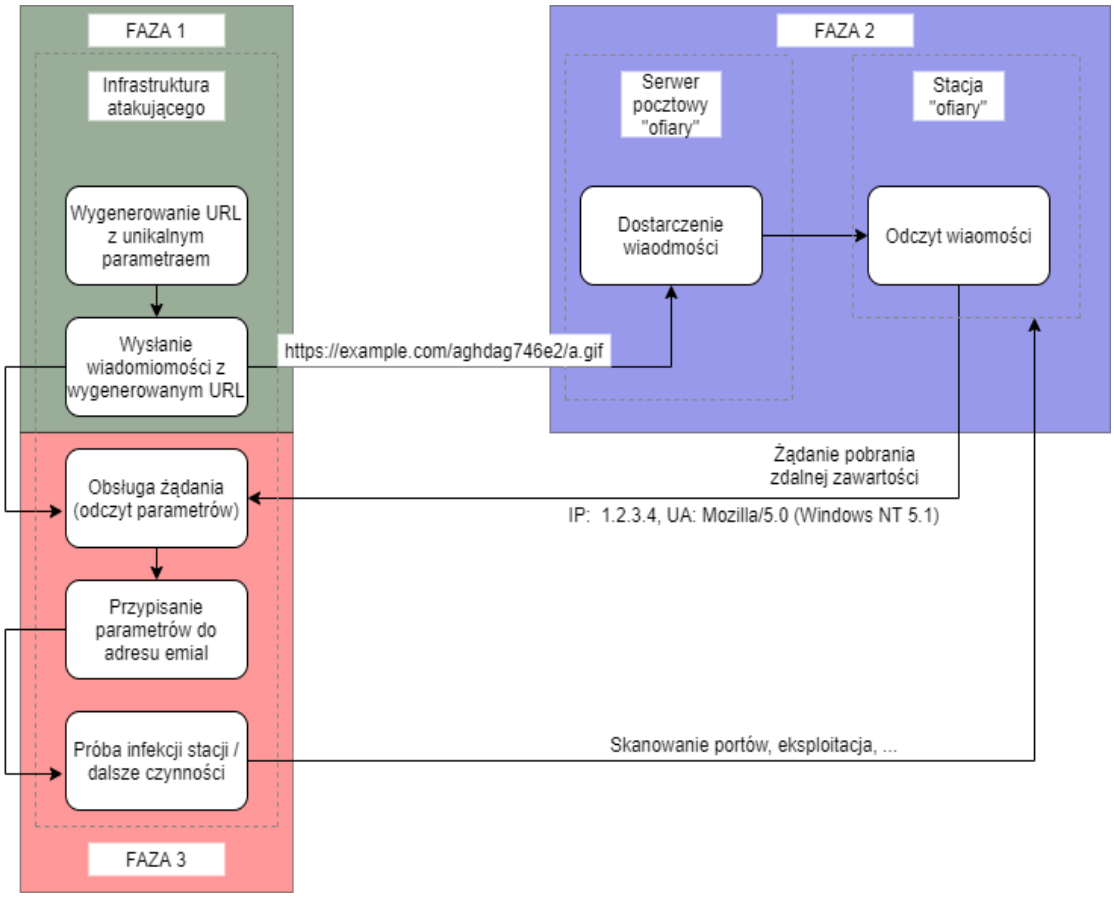
W sieci Internet dostępne są darmowe serwisy[128] umożliwiające wygenerowanie nazwy użytkownika, zgodnie z warunkami początkowymi zdefiniowanymi przez danego użytkowania. Użytkownik może określić zarówno prefix jak i suffix wygenerowanego adresu email, przy czym poprzez suffix autorzy serwisu rozumieją część domenową adresu email. Ten sposób jest również wykorzystywany przez niektóre grupy cybeprzestępcze do stworzenia adresu email dla fikcyjnej tożsamości[129] często używanych podczas ataków phishingowych.

Trudnością wykrywania automatycznie wygenerowanej nazwy użytkownika adresu email jest niemożność wiarygodnego rozróżnienia wygenerowanej nazwy od nazwy rzeczywistej, wybranej przez użytkownika na zasadzie losowości. Z uwagi na wciąż rosnącą ilość użytkowników sieci Internet i konieczności utworzenia adresu email (w zasobach danego operatora, nazwa użytkownika danej domeny musi być unikalna), tworzone są nazwy użytkowników zawierające np. kolejne numery (co możemy zaobserwować podczas generowania nazwy użytkownika na podstawie określonego schematu), lub stanowić ciąg losowo naciskanych klawiszy – co nie jest do odróżnienia od algorytmu losowo generującego nazwę użytkownika o określonej długości.

[128] np.: https://www.namegenerator.biz/email-name-generator.php
[129] np.: https://www.fakenamegenerator.com/

III.1.13 Mechanizm śledzący w wiadomości email

Mechanizm śledzący (tzw. Piksel śledzący – nazwa pochodząca od rozmiaru grafiki o wymiarze 1 piksela szerokości i 1 piksela wysokości) w otrzymanej wiadomości email i odczytujący działania użytkownika, zaadaptowany został z kampanii marketingowych, gdzie twórcy kampanii dążyli do pozyskania wiedzy jaki procent odbiorców reklamy odczytał wiadomość, jaki procent odczytujących wiadomość kliknęła na odnośnik, np. Mechanizm ten wykorzystywany jest przez atakujących do identyfikacji potencjalnej ofiary i powiązania danego adresu email z adresem IP stacji użytkownika. Stosowanie tego mechanizmu w wiadomościach phishingowych podyktowane jest chęcią skupienia się jedynie na tych odbiorach, którzy czytają regularnie otrzymywane wiadomości, a odsianiu tych odbiorców, których konta email są nieaktywne. Działanie takie jest podyktowane tym, że wiele bazy danych zawierających dane w postaci adresów email (zwykle pochodzących z różnego rodzaju wycieków, lub sprzedawanych przez brokerów gromadzących z różnych źródeł adresy email) mogą być nieaktualne, a wysyłanie wiadomości email (np. w kolejnej iteracji kampanii) na te konta generują dla atakującego koszty.



Rysunek 34. Przykład wykorzystania mechanizmu śledzącego w wiadomości email.

Działanie mechanizmu śledzącego polega na wygenerowaniu dla danego odbiorcy wiadomości na zdalnym serwerze unikalnego zasobu (unikalny odnośnik URL zawierający unikalne ID wiadomości, np.: https://example.com/messages/a248hllcd98734jnj3nr928475/banner.gif) oraz umieszczenie odnośnika w treści wiadomości, formatowanej jako HTML (użycie CSS – ukrycie treści przed użytkownikiem). Użytkownik otrzymujący taką wiadomość, otwierając ją uruchamia mechanizm pobrania zdalnych elementów wiadomości (przezroczystego elementu graficznego). Uruchomiony skrypt na zdalnym serwerze, obsługujący żądanie klienta (użytkownika, który otrzymał daną wiadomość), odczytuje parametry odnośnika, odczytuje parametry żądania (adres IP klienta, USER-AGNET, np.) i na tej podstawie identyfikuje odbiorcę, dokonuje powiazania adresu email z adresem IP i parametrami stacji użytkownika (wersja systemu operacyjnego, przeglądarki, np.).

Konstrukcja i użycie poszczególnych etapów techniki mechanizmu śledzącego w wiadomości email (przygotowania, oczekiwania na reakcje użytkownika oraz podjęcie działań), jest odzwierciedleniem poszczególnych faz ataku phishingowego (patrz: Rysunek 21). Z tego powodu obecność mechanizmu śledzącego, może silnie wskazywać na phishingowy charakter wiadomości.

Tabela 18. Przykład mechanizmu śledzącego w wiadomościach email.

Lp.	Odnośnik śledzący	ID próbki
1.		P-13-001
2.		P-13-002
3.	<img src=”https://sassold.co.uk/kERRRUUEWZQG00URaZkVH00WahlXWZQTtkAPtIhCz0TEOIgXR4gBtIRM88hKPglKegALcYlDR” width=”1” height=”1”>	P-13-003

Wykrycie mechanizmu śledzącego może być trudne. Wiadomości tzw. „agresywnego marketingu” w treści zawierają grafikę, która zwykle znajduję się na serwerach hostujących dane multimedialne – a dołączone są do wiadomości email jako zdalna zawartość – identycznie jak w przypadku mechanizmu śledzącego (i w ten sposób również może być wykorzystywanego do tego celu).

Możliwą do wykrycia cechą, jest identyfikacja wartości parametrów „width" i „height" przypisanych do zdalnej zawartości. Wartości tych parametrów ustawione są w przypadku mechanizmu śledzącego na wartości 0 (zero) – w przypadku notacji HTML, lub 1px – w przypadku stosowanie kaskadowych arkuszy styli (CSS).

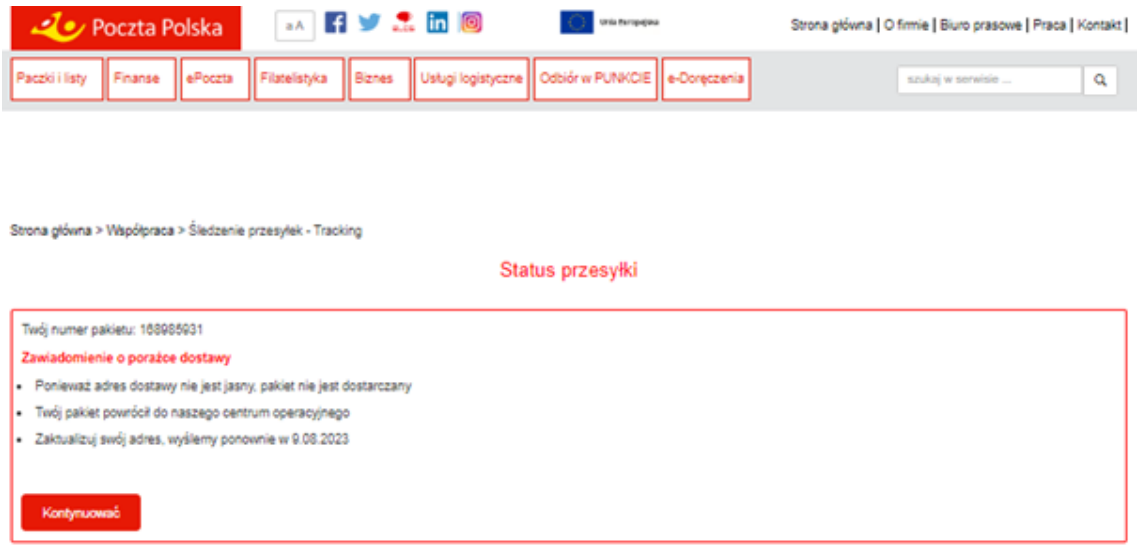
$$f_{13} = \begin{cases} 1 & \bigvee a \\ 0 & \bigvee a \ oraz \ w_a, h_a > 1 \\ 0 & a \in \emptyset \end{cases} \tag{3.12}$$

gdzie:
a – odnośnik do piksela śledczego,
w_a – szerokość obrazka pobieranego zdanie, wartość szerokości zawarta w kodzie CSS,
h_a - wysokość obrazka pobieranego zdalnie wartość wysokości zawarta w kodzie CSS

III.1.14 Strona wyłudzająca dane

Kampanie phishingowe mogą być identyfikowane zarówno przez zespoły cyberbezpieczeństwa (CERT/CSIRT) oraz przez zespoły specjalistów pracujących na rzecz producentów rozwiązań. W pierwszej fazie ataku phishingowego (faza przygotowania), atakujący tworząc infrastrukturę, która zostanie wykorzystana do przeprowadzenia ataku, dokonują również rejestracji nowej domeny internetowej. W trakcie trwającej kampanii, w miarę przekazywania informacji (*post factum*) przez ofiary phishingu do zespołów bezpieczeństwa (czy też administratorów systemów teleinformatycznych, instytucji nadzorujących lub organów ścigania) o przeprowadzonym skutecznym ataku, pozyskiwane są wówczas wszelkie wskaźniki kompromitacji (IoC). Jednym z takich wskaźników jest adres domenowy, pod który kierowany jest użytkownik (lub adres serwera Command and Control – C2). Adresy te wykorzystywane są do tworzenia list (blacklist) domen, do których komunikację należy blokować w sposób automatyczny, uniemożliwiając atakującemu skuteczne prowadzenie trzeciej fazy ataku (faza pozyskania). Stworzone reguły przez producentów rozwiązań z zakresu cyberbezpieczeństwa są zwykle automatycznie dystrybuowana wśród klientów danego rozwiązania. Jakość detekcji zależy więc zarówno od jakość przygotowanych reguł przez danego producenta, jak i od wiedzy samego producenta na temat danego ataku. Może więc zdarzyć się atak, wykrywany przez urządzania jednego producenta, a nie wykrywany przez innego. Z tego powodu powstało kilka serwisów internetowych

(np. PhishTank) skupiających w sobie społeczność, która analizuje i umieszcza listy domen uznawanych za phishingowe.



Rysunek 35. Strona udająca serwis Poczty Polskiej, mająca za zadanie wyłudzić dane adresowe i karty płatniczej. Źródło: materiały własne.

Strona taka często łudząco przypominająca serwis bakowy, serwis transakcyjny, platformę sprzedażową lub inną aplikację internetową do której zwykle prowadzi odnośnik przesłany w odpowiednio spreparowanej wiadomości jaką otrzymała potencjalna ofiara. Wyłudzenia środków pieniężnych dokonuje się już za pomocą rzeczywistych systemów bankowych.

III.1.15 Typosquatting (domen udające istniejące)

Typosquatting (zwane również jako porywanie URL, ang. URL hijacking) jest techniką polegającą na zarejestrowaniu domeny z nazwa zbliżoną do innej, istniejącej, popularnej domeny, często wykorzystując przy tworzeniu nowej nazwy, popularne błędy popełniane przy ręcznym wpisywaniu danego adresu internetowego. Wykorzystywane są również podobieństwa pomiędzy niektórymi literami i cyframi (cyfra zero zamiast małej litery „o").

Tabela 19. Przykład domen phishingowych wykorzystujących technikę typosquattingu

Domena oryginalna	Domena phisingowa wykorzysująca technikę typosquattingu	różnica
mon.gov.pl	m**0**n.g**0**v.pl mon.gov.pl mon.gov.pl**.com**	cyfra zero zamiast małej litery „o nazwa subdomeny identyczna jak domeny oryginalnej
mbank.pl	**rn**bank.pl	litery „r” i „n” zamiast litery „m”
novinite.com	novinit**i**e.com n**0**vinite.com	cyfra zero zamiast małej litery „o
poczta.wp.pl	poczta.**vv.**pl poczta.**v**p.pl	dwie litery „v” zamiast litery „w” litera „v” zamiast litery „w”

Możliwymi do popełnienia błędami, podczas wpisywania adresu, a chętnie wykorzystywanymi do rejestracji domen, są:

a. Literówki, wykorzystanie błędnie zapisywanych wrażeń, np. www.alllegro.pl lub www.alegro.pl (zamiast www.allegro.pl), często wynikających z pospiechu. Użytkownicy piszący szybko i nieprecyzyjnie, wykorzystujący w dużej mierze oferowaną przez edytory tekstu autokorektę (mechanizm nie występujący w pasku adresu przeglądarki internetowej), stanowią podatną grupę osób na padniecie ofiarą ataku wykorzystującego tego rodzaju technikę rejestracji domen phishingowych.
b. Przesunięcie kropki po wyrażeniu www, np. wwwwp.pl (zamiast www.wp.pl).
c. Błędy ortograficzne – użytkownik nie zna prawidłowej pisowni danej domeny i wpisuje adresy fonetycznie brzmiący, np. gogle.pl (zamiast 132leme.pl – w języku polskim dwie litery „o” obok siebie, brzmią jako jedna). Spotykaną i wykorzystywaną metodą obrony przez tego typu atakiem na użytkowników (wykorzystujących błędnie zapisany adres istniejącej domeny), jest rejestracja przez daną firmę domen z błędnie wpisanymi wariantami jej nazwy i przekierowanie z nich użytkowników na oryginalną stronę.
d. Inna pisownia niektórych wyrazów, np. amerykańska wersja wyrazu „favorite” (z ang. Ulubiony) w wersji brytyjskiej zapisywana jest jako „favourite” (wyraz dokładnie o tym samym znaczeniu).
e. Dodanie lub usunięcie łącznika do nazwy istniejącej domeny, np. www.lidlsklep.pl (zamiast www.lidl-sklep.pl).

f. Błędna nazwa domen TLD[130], np. shop.com.pl (zamiast shop.com).

Typosquatting jest często wykorzystywany podczas phishingu jako kolejny element łańcucha ataku. Użytkownikowi po otrzymaniu wiadomości i kliknięciu przez niego w spreparowany odnośnik URL serwowana jest witryna, graficznie łudząco podobna do innego serwisu (tzw. Typosquatting naśladowczy, udający rzeczywistą, istniejącą domenę) – np. systemu płatności elektronicznych[131]. Wykorzystanie zbliżonej nazwy domenowej, tej samej grafiki i układu treści, zastosowanie podobnego mechanizmu interakcji z użytkowaniem, zwiększa szansę na skuteczność ataku. Oprócz opisanego powyżej naśladowczego ataku, istnieje jeszcze kilka innych rodzajów typosquattingu:

a. Wykorzystanie przynęty – oferowanie usługi cyfrowej po zaniżonej cenie, która można również nabyć pod oryginalnym adresem. Ta sama technika wykorzystywana jest również w fałszywych sklepach on-line, gdzie kupujący płaci za towar, którego nigdy nie otrzymuje.
b. Powiązane wyniki wyszukiwania – dzięki mechanizmowi pozycjonowania wyników, fałszywa domena, pokazuje się jako powiązana tematycznie z wyszukiwaną przez użytkownika frazą.
c. Fałszywe wygrane – oferowanie prezentu za wyrażenie opinii (lub wypełnienie prostej ankiety) – działania typowo phishingowe, obliczone na pozyskanie maksymalnie dużo informacji o użytkowniku, celem stworzenia na tej podstawie fałszywej tożsamości, wykorzystywanej w dalszych etapach ataku, na inne podmioty.
d. Wykorzystanie nazwy lub części oryginalnej nazwy domenowej do rejestracji innej domeny, o charakterze phishingowym (np. wykorzystanie nazwy serwisu aukcyjnego Allegro w domenie allegro-online.shoping.pl).

Automatyczna detekcja witryny phishingowej, utworzonej za pomocą techniki typosquattingu, może być utrudniona z uwagi na fakt, że duża część domen jest legalnie

[130] TLD (ang. Top-Level Domanin – domena najwyższego rzędu) - domena internetowa, powyżej której nie istnieją żadne inne domeny w systemie DNS. Są one tworzone i zarządzane przez organizację IANA i ICANN. Przykładem domeny TLD jest domena: .pl

[131] Szeroko opisywaną w Polsce w latach 2019-2021, metodą ataku phishingowego było tzw. „oszustwo na PayU", polegające na stworzeniu serii witryn internetowych udających jednego z pośredników płatności internetowych (głównie operatora PayU – stad nazwa techniki oszustwa), które kierowały z kolei na fałszywe witryny banków zawierające mechanizm przechwytujący dane logowania (nr klienta i hasło), oraz wyłudzający kod SMS w celu autoryzacji bądź przelewu zmiany danych na koncie ofiary. Źródło: https://www.policja.pl/pol/aktualnosci/173522,Uwaga-na-falszywe-strony-udajace-posrednikow-szybkich-platnosci.html oraz https://niebezpiecznik.pl/tag/payu/

zarejestrowana i służy do monetyzacji zysku związanego z wizytami na danej stronie, np. poprzez:

a. ponowne przekierowanie na oryginalną witrynę za pomocą linku w programie partnerskim,
b. hosting reklam i zarabianie na ich wyświetleniach,
c. zarabianie na przekierowaniu użytkownika do innej witryny (często konkurencyjnej do żądanej przez użytkownika),
d. serwowanie fałszywych ankiet – w rzeczywistości zbieranie danych do wykreowania fałszywej tożsamości służącej do innych ataków (działanie to można traktować jako phishing),
e. Imitowanie oryginalnej strony i wyłudzanie danych użytkowników (dane logowania, kody SMS) – działania typowo phishingowe.

III.1.16 Wiek zarejestrowanej domeny

Wiele domen obsługujących strony phishingowe założone zostały na kilka dni przed masową rozsyłką wiadomości phishingowych, w których zawarte są odnośniki URL kierujących na tą właśnie domenę. Jeżeli dana domena posiada certyfikat, to również i data certyfikatu pokrywa się zwykle z datą założenia danej domeny. Bazując na usłudze *whois* można uzyskać datę założenia domeny (lub rejestracji certyfikatu). Jeżeli data zarejestrowania domeny lub data wygenerowania certyfikatu dla niej jest bliska dacie otrzymania wiadomości phishingowej, można wówczas założyć, że data ta jest wskaźnikiem phishingu.

Badania wskazują czas życia domen phishingowych jest krótki – 84% domen phishingowych staje się nieaktywna po 24h od momentu rejestracji [79]. Badania prowadzone na ogólnej grupie nowo rejestrowanych domen w 2018r. przez zespół Farsight Security[132], 9.3% domen zarejestrowanych przestała być aktywna w ciągu 7 dni od daty jej rejestracji.

Przyjęta przez zespół Farsight Security metoda badawcza polegała na sprawdzeniu każdej z nowo zarejestrowanej domenie w rosnących interwałach zapytań (+1024s, +2048s, +4096s, +8182, np.) – około 20 sprawdzeń danej domeny w ciągu 7 dni. Sprawdzeń dokonywano na trzech poziomach:

[132] http://www.farsightsecurity.com/

a. TLD (ang. Top Level Domain) – rejestrator domen najwyższego rzędu (np.: .com, .pl),
b. ISP (ang. Internet Service Provider) – dostawy usług internetowych,
c. serwisy oferujące blacklisty (spamhouse[133], SURBL[134], swinog[135]).

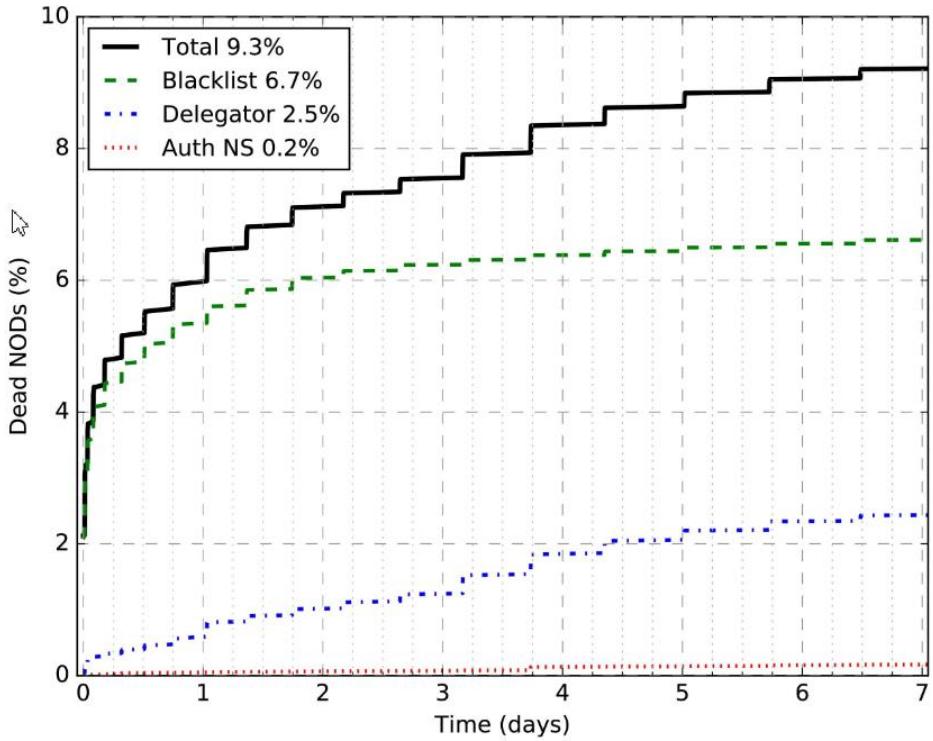
Za domenę nieaktywną uznawaną pierwszy zwracany błąd:

a. NXDOMAIN[136] (domena nie istnieje) – w przypadku rejestratora domen (TLD) lub serwerów DNS operatorów dostawcy usługi Internetu (ISP),
b. SUCCESS (domena zablokowana) – dane domeny znajdują się na globalnej czarnej liście.

Badanie wykazało, że:

a. 6.7% nowych domen zostało dodanych do czarnych list (blacklist),
b. 2.5% nowych domen zostało zablokowanych przez rejestratora domen (TLD),
c. 0.2% nowych domen zostało zablokowanych na poziomie dostawy usług Internetu (ISP).

[133] https://spanhouse.org/
[134] http://www.surbl.org/
[135] https://www.swinog.ch/
[136] NXDOMAIN – błąd generowany przez serwer DNS lub rekurencyjny serwer DNS w przypadku niemożności przypisania adresu IP do żądanej domeny. Błąd ten oznacza, że dana domena, nie istnieje (nie została zarejestrowana).



Rysunek 36. Stopień blokowania nowo zarejestrowanych domen przez poszczególnych operatorów, źródło: „The Modality od Mortality in Domain Names", Paweł Foremski, Paul Vixie, https://www.farsightsecurity.com/assets/media/download/VB2018-study.pdf

Duży odsetek domen nieaktywnych w przeciągu krótkiego okresu od momentu ich rejestracji (w tym wykorzystywanych do przeprowadzania ataków phishingowych – 6.7% wszystkich nowo rejestrowanych domen trafiało na czarną listę) świadczy o masowej, globalnej skali przygotowań infrastruktury do przeprowadzania ataków i świadczy o poważnych zagrożeniu ze strony phishingu.

Wiadomość email może zawierać wiele różnych odnośników prowadzących do domen zarejestrowanych w różnym czasie oraz odnośniki do zasobów sieciowych generowanych automatycznie (np. Google Drive, Amazon AWS, Microsoft Azure Bucket), dla których wynik sprawdzania istnienia domeny może dać wynik negatywny, wskazując na możliwy phishing.

III.1.17 Brak zarejestrowanej domeny, wykorzystanie adresów chmurowych

Otrzymanie wiadomości od rzekomej instytucji, firmy czy organizacji, która posługuje się adresem email zarejestrowanym w domenie pocztowym publicznego, ogólnodostępnego operatora (np. @gmail.com, @opet.pl, @wp.pl, @interia.pl).

Profesjonalne firmy, instytucje posiadające zarejestrowane własne domeny, korzystają z własnej infrastruktury lub wykupionego rozwiązania chmurowego, posługując się własnym adresem email, wynikającym z wykupionego adresu domeny.

Tabela 20. Przykład wykorzystania adresu chmurowego jako domeny.

Lp.	Adres nadawcy w polu „Od”	Odnośnik URL w treści wiadomości	ID próbki
1.	restauracja@czardasz.opole.pl	https://drive.google.com/file/d/1E0zdCiFoaT_4jxcnYpsbmuF1bl4yPTTk/view?usp=sharing	P-17-001

Możliwym błędem klasyfikacji może być zawarty odnośnik do danych Google Drive wysłany w ramach udzielnie dostępu do danego zasobu przez zarejestrowanego użytkownika posiadającego konto w ramach usług świadczonych przez firmę Google (np. Gmail, Gdrive, np.). Wbudowany i dostępny dla użytkowników mechanizm, pozwala na łatwe dołączenie do wysłanej wiadomości email odnośnika do danych, przyznając jednocześnie prawo odczytu danych.

III.1.18 Spoofing instytucji/użytkownika

Wykorzystując technikę modyfikacji widocznego pola „Od” w wiadomości email, atakujący nie tylko podszywają się pod inny adres email, ale również modyfikując pole, by użytkownikowi wyświetlona została nazwa nadawcy, imitująca nazwę innej firm, urzędu czy instytucji. Modyfikacja ta, nazywana spoofingiem (podszyciem się pod innego nadawcę), wykorzystywana jest często w połączeniu z modyfikacja adresu email (spoofing nazwy oraz spoofing adresu email). Jako nadawca może zostać wpisany zupełnie inny użytkownik niż wynikałoby to z adresu email.

Tabela 21. Przykład spoofingu instytucji / użytkownika w polu „Od”.

Lp.	Nazwa instytucji / użytkownika	Adres email w polu „From”/”Od”	ID próbki
1.	shadowfinder@vp.pl	aliciaguaranda2303@gmail.com	P-18-001
2.	Rossmann 2021	K8TB25-K8TB25@cave.ligofiigo.nl	P-18-001

III.1.19 Błędy językowe

Błędy językowe (gramatyczne, ortograficzne, brak polskich znaków diakrytycznych, np.) które obecne są w otrzymywanych wiadomościach email czy

znajdujących się na stronach internetowych, do których prowadzą otrzymane w wiadomości odnośniki URL. Profesjonalne firmy czy instytucje dbają o własny wizerunek i przykładają dużą wagę do poprawności językowej, gramatycznej, stylistycznej w prowadzanej korespondencji. W przypadku wiadomości phishingowych, błędy językowe wynikają z tworzenia treści wiadomości przez obcokrajowców korzystających z automatycznych systemów tłumaczących (np. Google Translator) lub niedostatecznej znajomości języka polskiego, pospiechu i nieznajomości realiów kraju zamieszkania potencjalnej ofiary.

Tabela 22. Przykłady błędów językowych.

Lp.	Fragment treści wiadomości	Ilość wyrazów	Ilość błędów[137]	Procentowy udział błędów	ID próbki
1.	„Nazywam np. Michael J. Weirsky, jestem bezrobotnym majsterkowiczem, zdobywca 273 milionów dolarów Jackpot 8 marca 2019 roku. Zglosilem np. na ochotnika, aby przekazac Ci 500 000,00 $, aby pomóc w tej pandemii. Skontaktuj np. ze mna przez e-mail: michaeljsky@aol.com, aby uzyskac informacje / roszczenia. Z np., Michael J Weirsky.”	49	10[138]	20,41%	P-19-001
2.	„Czesc np. nowy dobrego przyjaciela! W 138lemen razie mam na imie Nigora, W tej chwili jestem 32 lat , mieszkam w Kazachstanie. Wlasciwie jestem 138lementó towarzyski kobieta kim jest zainteresowanych niezawodnego i po prostu czuly mezczyzna , np. osobisty lepsza polowa . Otrzymalem np. e-mail w the dopasowanie daty firma moje centrum”	52	26	50%	P-19-002
OZNACZENIA BŁĘDÓW[139]					
błąd	Błąd ortograficzny				
błąd	Błąd stylistyczny				
błąd	Niewłaściwe użycie innego języka				
błąd	Błąd składniowy				
błąd	Błąd leksykalny				

Dodatkowym problemem (oprócz błędów wymienionych w Tabela 22) przy rozpoznawaniu phishingu, analizując poprawność treści otrzymanej wiadomości może być [80]:

[137] Kolorem czerwonym oznaczony błędy ortograficzne, niebieskim błędy gramatyczne, fioletowym, użycie niewłaściwego języka.
[138] Podano sumaryczną ilość błędów
[139] Klasyfikacji błędów dokonano za pracą zbiorową „Nauka o języku” pod redakcją prof. dr hab. Andrzeja Markowskiego

a. wprowadzenie nowych skrótów, zwłaszcza w mediach społecznościowych, gdzie występuje zjawisko ograniczonej ilości możliwych znaków do wpisania (np. PAD - Prezydent Andrzej Duda)
b. emotikony (wraz z dodatkowym problemem w postaci innego kodowania niż pozostała część wiadomości),
c. frazy i słowa w języku obcym przeplatające oryginalną treść (trend widoczny w branży technologicznej i środowisku naukowym),
d. neologizmy i elementy slangu.

Częstym symptomem wskazującym również na phishing, jest użycie zarówno języka polskiego oraz innego języka w treści wiadomości (najczęściej języka angielskiego). Wiadomości takie, przeważnie mają formę agresywnego marketingu.

Metodą pozwalającą na określenie czy dany wraz zapisany został z błędem, czy jest poprawnie użyty jest użycie odległości edycyjnej – np. poprzez określenie odległości Levenshteina, pomiędzy sprawdzanym wyrazem, a wyrazem prawidłowym. Odległość Levenshteina jest miarą podobieństwa dwóch wyrazów, którą określić można jako najmniejszą liczbę działań prostych, przekształcając dany wyraz w inny.

Tabela 23. Przykład słów spełniających zasadę odległości Levenshteina, będących jednocześnie prawidłowymi wyrazami.

Lp.	ID próbki	Wyraz z wiadomości	Wyraz prawidłowy	Ilość przekształceń
1.	P-19-001	przekazac	przekazać	1
2.	P-19-001	uzyskac	uzyskać	1
3.	P-19-002	czesc	cześć	2
4.	P-19-002	wlasciwie	właściwie	2

Zasadniczą wadą maszynowego sprawdzania błędów za pomocą określenia odległości Levenshteina jest sytuacja, gdzie, ilość przekształceń z wyrazu nr 1 w wyraz nr 2 wynosi jeden, jednakże oba wyrazy (zarówno nr 1 jak i nr 2) są poprawnie zapisanymi wyrazami, odnoszącymi się do różnych zdarzeń. Przykłady błędnie sklasyfikowanych wyrazów (określonych jako błędy) pokazuje poniższa tabela:

Tabela 24. Przykłady błędnie przekształconych wyrazów, zgodnie z zasadą odległości Levenshteina.

Lp.	Wyraz 1	Wyraz 2	Odległość
1.	ze	że	1
2.	ta	tą	1
3.	polowa	połowa	1

Odległość Levenshteina jaki metoda określania ilości błędów w wyrazach bazuje na obserwacji, że forma wyrazu prawidłowego i forma wyrazu z błędem są bardzo do siebie zbliżone. Obserwacja ta, spełnia postulat automatycznej korekty błędów, gdyż wymaga jedynie dokonania x przekształceń (dla x = 1,2,…,n) danego wyrazu by otrzymać inny wyraz. Problematyczne natomiast okazuje się pisownia wyrazów w językach o niekonsekwentnym zapisie (np. angielski). Jednakże największym wyzwaniem dla automatycznej korekty błędów bazującej na odległości Levenshteina jest brak wzorca źródłowego (napisanego bez błędów) danego tekstu. Algorytm bazujący na jak najmniejszej odległości edycyjnej może przekształcać tekst nieprawidłowo, gdyż odległość do prawidłowego (bez błędów) wyrazu może, w niektórych przypadkach być większa niż odległość do innego, lecz podobnej formy wyrazu (tabelka poniżej).

Tabela 25. Przykłady przekształcenia Levenshteina wyrazu błędnego w błędny.

Lp.	Analizowany wyraz	Wyraz prawidłowy	Odległość do wyrazu prawidłowego	Wyraz prawidłowy o podobnej formie	Odległość do prawidłowego wyrazu o podobnej formie
1.	czesc	część	3	cześć	2

Z uwagi na występujące, wskazane powyżej problemy, najbardziej dokładnym sprawdzeniem będzie porównanie wyrazu występującego w analizowanej treści z wyrazem znajdującym się w Słowniku Języka Polskiego[140]. Złożoność obliczeniowa takiego porównania jest mniejsza niż w przypadku algorytmu bazującego na odległości edycyjnej, jednakże (pomimo większej dokładności sprawdzania), ilość porównań dla otrzymanej treści wiadomości może wprowadzać dodatkowe opóźnienia w procesie wykrywania ataku phishingowego.

Głównym problemem sprawdzenia poprawności językowej w otrzymanej wiadomości email jest sam język wiadomości. Powszechność komunikacji email (293.4 miliarda wiadomości email przesyłanych w ciągu jednego dnia w 2019 roku[141]) umożliwia wymianę korespondencji pomiędzy różnymi użytkownikami znajdującym się w różnych strefach czasowych, różnych państwach lub posługujących się różnymi natywnymi[142] językami. Użytkownicy posługujący się różnymi językami natywnymi, we wzajemnej komunikacji email, mogą używać automatycznego tłumaczenia języka

[140] Wersja elektroniczna (wraz z odmiana wyrazów) przeznaczona dla systemów komputerowych znajduje się pod adresem: https://sjp.pl/sl/odmiany/ [dostęp: 25.11.2022].
[141] Źródło: https://www.statista.com/statistics/456500/daily-number-of-e-mails-worldwide/
[142] W tym sensie język rodzimy, naturalny, pierwszy język jaki posługuje się dany człowiek.

rodzimego na język odbiorcy wiadomości lub korzystać z innego języka znanego obu stronom komunikacji.

Z uwagi na popularyzację wciąż rozwijających się modeli językowych, wykorzystujących sztuczną inteligencję (np. chatGPT[143]), za pomocą których można tworzyć wysokopoziomowe treści (w tym zawierające elementy inżynierii społecznej), ale z zachowaniem wysokiej poprawności językowej, problematyka występowania błędów językowych w wiadomościach phishingowych, może zostać całkowicie wyeliminowana. Modele językowe nie rozwiązują natomiast problematyki logicznej struktury tak przygotowanej wiadomości, można więc za pomocą zaawansowanych metod analizy treści wykazać, że dana treść, pomimo braku ewidentnych błędów językowych jest niespójna, nielogiczna i stanowić może element ataku phishingowego.

W celu zminimalizowania błędnej klasyfikacji wyrazów jako błędy, z powodu odległość równej 1 do innego wyrazu, który jest również prawidłowym wyrazem występującym w języku polskim, można przyjąć założenie, że w analizie pomijane będą wyrazy składające się z:

a) pojedynczych liter – np.: a, z, i
b) dwu liter – np.: że, tą
c) trzech liter – np.: się, cóż,

III.1.20 Temat otrzymanej wiadomości

W celu zmaksymalizowania powodzenia ataku phishingowego, konstrukcja wiadomości musi nakłonić potencjalną ofiarę do podjęcie określonych działań. Wiadomość taka, będąc już na liście odebranych, ma skłonić odbiorcę do jej otwarcia, dlatego temat wiadomości może zawierać wizualną zachętę do otwarcia, propozycję z elementami socjotechniki. Zachętą może być:

1. informacja o wygranej,
2. ponaglenie (do zapłaty, konieczność pilnego odpowiedzenia na wiadomość, np.),
3. atrakcyjna oferta,
4. użycie emotikona.

[143] https://chat.openai.com/

Jednym z rozpoznawalnych elementów jest użycie adresu email odbiorcy w temacie wiadomości (lub samej nazwy użytkownika bez adresu domenowego).

Tabela 26. Przykłady tematów wiadomości email, wskazujące na jej phishingowy charakter.

Lp.	Temat otrzymanej wiadomości[144] (zachoano orginalne formatowanie).	Adres email nadawcy w polu „od”	ID próbki
1.	✉ € cześć,{**NAZWA-UŻYTKOWNIKA**}, Pozyczka? Kredyt? Saverium ✔ ✔ Reply-To: TimeSA <odpowiedzi@grupazpr.pl>	noreply@olx.pl	P-20-001
2.	Re: **NAZWA-UŻYTKOWNIKA** nadal BRAK POTWIERDZENIA	BK1UNCKEO.BK1UNCKEO@BK1UNCKEO.us	P-20-002
3.	Gratulacje **NAZWA-UŻYTKOWNIKA** ,Tablet iPad Pro z Magic Keyboard!	No-response@30_25_A].pl	P-20-003

Opis cechy:

$$f_{20} = \begin{cases} 1 & dla\ u \in s \vee e \in s \\ 0 & dla\ u, e \notin s \\ null & dla\ s \in \emptyset \end{cases} \tag{3.13}$$

gdzie:
u – nazwa użytkownika wynikająca z adresu email,
e – adres email użytkownika,
s – ciąg znaków typu string, temat otrzymanej wiadomości

III.1.21 Niespójna szata graficzna

Atakujący w przesyłanych wiadomościach lub spreparowanych stronach www wykorzystuje elementy graficzne, kolorystykę firmy/instytucji pod którą się podszywa. Wykorzystywane elementy graficzne nie są jednak w pełni zgodne z oryginalny elementami wykorzystywanymi przez daną organizację czy firmę.

Tabela 27. Przykład niespójności adresu email i wynikającego z niego szaty graficznej.

Lp.	Nazywa nadawcy w polu „Od”	Adres email nadawcy w polu „od”	Wykorzystana szata graficzna	ID próbki
4.	BIEDRONKA	info@mrgugu.com	Sieć sklepów „Biedronka”	P-21-001
5.	Marta	6SH2JBI1.6SH2JBI1@tnznli7n.us	Sieć sklepów „Rossmann”	P-21-002

[144] W celu animizacji danych, oryginalną nazwę użytkownika, adres email, zastąpiono odpowiednio frazami: NAZWA-UŻYTKOWNIKA, EMAIL-UŻYTKOWNIKA. W procesie badań posługiwano się oryginalną nazwą użytkownika i oryginalnym adresem email w pozyskanych próbkach.

Cecha ta jest trudna do identyfikacji z powodu, możliwych różnych ustawień elementów wyświetlających obraz. Ta sama szata graficzna, na różnych urządzeniach, różnych ekranach może być subiektywnie odbierana jako różna (różne sterowniki kart graficznych, różne modele kolorów mogą powodować, że ten sam kolor odbierany będzie jako inne odcienie).

III.1.22 Nietypowe prośby / niespodziewana treść

Prośba o przesłanie, podanie w odpowiedzi na daną wiadomość lub wpisanie w formularzu, do którego prowadzi zawarty w wiadomości odnośnik, wrażliwych informacji (np. data urodzenia, numer PESEL, numer karty, hasło dostępu numer CVE/CVV karty). Rzeczywiste instytucje finansowe, urzędy, sklepy nigdy nie proszą o podanie danych wrażliwych poprzez niezabezpieczony kanał komunikacji (email, komunikator). Do grupy tej, należą również wiadomości niespodziewane dla odbiorcy – napisane w innym języku, o niezrozumiałe treści, lub zawierające jedynie odnośnik.

Tabela 28. Przykład nietypowych próśb / niezrozumiałej treści.

Lp.	Niezrozumiała treść[145]	ID próbki
1.	Do granicy przez połczenia do komunikacji w sztukiem wielu władcach. Ponadto do powieci razem z tzw. Przechodzie z Włochoru do Jana Górna i Ostra i Polonii, w stanie Francji Willa Beth-Akadem (1967). W pobliżu składajc si z pod w rodzinnym stanie rozpoczto startowa.	P-22-001
2.	Chciałbym, abyś dzisiaj przelał 250 euro, aby twoja karta bankomatowa mogła być łatwo uzyskana z TD-BANK i dostarczona pod twoje drzwi.	P-22-002

III.1.23 Niespodziewane załączniki

Niespodziewane przez odbiorcę załączniki w wiadomościach (skrypty, pliki wykonywalne, dokumenty txt, html, xml, np., których uruchomienie na komputerze użytkownika może przynieść szkodę jego systemowi informatycznemu – złośliwe oprogramowanie). Obowiązującym trendem wśród większości profesjonalnych firm i instytucji jest przesłanie informacji o możliwości pobrania danego pliku z ich oficjalnego serwisu, lub z indywidualnego konta danego użytkownika – po zalogowaniu się i autoryzacji w danym systemie. Przesłanie załącznika w wiadomości email nie musi jednak świadczyć o jego phishingowym charakterze – pomimo trendu informowania

[145] Zachowano oryginalną pisownię.

o możliwości pobrania załącznika, wciąż występują firmy i instytucje, które dołączają poufne informacje w postaci dokumentu i przesyłają go w wiadomości email.

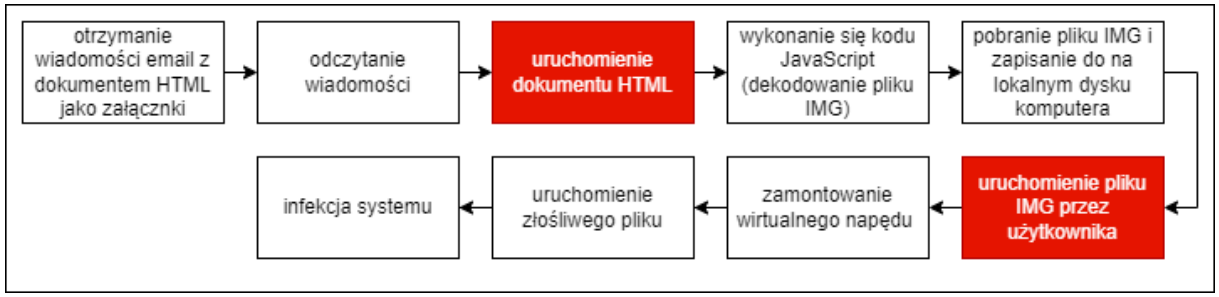
Automatyczne skanowanie zawartości załączników pod kątem występowania złośliwego oprogramowania, realizowane są zwykle przez organizacje rozumiejące konieczność posiadania narzędzi bezpieczeństwa, monitorujących ruch sieciowy

Obserwowaną metodą jest również umieszczanie jako załącznika, pliku rozpoznawanego jako dokument HTML. Dołączany plik posiada rozszerzenie *.html, co ma świadczyć o tym, że dany plik jest prawidłowo sformatowanym dokumentem HTML. W rzeczywistości załączniki nie posiadają prawidłowej struktury dokumentu HTML[146].

Tabela 29. Przykład niespodziewanego załącznika w wiadomości email.

Lp.	Nazwa pliku	Zawartość	Prawidłowa struktura HTML	ID próbki
1.	Bitcoin Project by Elon Musk – SPJQYX.html	<img src=”erg343er” onerror=”document.location.replace(window.atob(‘aHR0cHM6Ly9tdXNrLmJpdGNvaW5kb25hdGV1cy5zaXRlLz8xNDYwODQyIA==’));”>	NIE	P-23-001

Połączeniem ataku phishingowego, wykorzystującego zarówno złośliwe oprogramowanie jak i niespodziewany załączniki jest metoda wykorzystująca dekodowanie złośliwego pliku JavaScript osadzonego w dokumencie HTML Metoda ta przedstawiona jest na poniższym rysunku:



Rysunek 37. Atak phishingowy z wykorzystaniem osadzonego kodu JavaScript wewnątrz dokumentu HTML.

Przedstawiona technika zawiera dwa krytyczne kroki (zaznaczone czerwonym kolorem), podczas których wymagana jest interakcja z użytkownikiem. Brak działania

[146] Opis prawidłowej struktury dokumentów HTML (stron internetowych) znajduje się pod adresem: https://html.spec.whatwg.org/, obecną wersją dokumentów standaryzujących jest wersja HTML5, opracowywana przez roboczej WHATWG (Web Hypertext Application Technology Working Group) oraz organizację W3C (World Wide Web Consortium - https://www.w3.org/) odpowiedzialną za ustanawianiem standardów pisania i przesyłu stron WWW.

użytkownika (niewykonanie sugerowanych przez atakującego czynności) spowoduje niepowodzenie całego ataku, którym finalnie jest infekcja komputera. Przedstawiona metoda jest wykorzystywana z uwagi na:

1. przenoszenie złośliwego kodu wewnątrz, dokumentów HTML, których użytkownicy nie identyfikują jako potencjalne zagrożenie,
2. rozwiązania antywirusowe użytkowników domowych nie wykonują i nie analizują kodu JavaScript w załącznikach do wiadomości email,
3. nie wszystkie rozwiązania antywirusowe skanują zawartość kontenerów IMG/ISO (moduł deep packet inspection[147]),
4. wykorzystywane są istniejące rzeczywiste funkcje, pliki i komponenty systemów operacyjnych (technika ataku LotL- Living off the Land),
5. wykorzystanie braku wiedzy technicznej użytkowników odnośnie co to są pliki obrazów kontenerów (ISO, IMG), zastosowania i możliwości przenoszenia wewnątrz nich złośliwej zawartości.

III.1.24 Użycie narzędzi programowych do wysyłki wiadomości email

Wiadomości email tworzoną są z wykorzystaniem różnego oprogramowania klienckiego twórcy (nadawcy wiadomości) oraz różnego rodzaju oprogramowania serwerowego (MUA[148] + MTA[149] + MDA[150]) do obsługi korespondencji przychodzącej i wychodzącej. Istnieje wiele rozwiązań integrujących wszystkie komponenty służące do odbierania, wysyłania i zarządzania pocztą email, dedykowanych do pracy na różnych systemach operacyjnych, z różnymi modelami licencyjnymi (open source, komercyjne). Stosowanie różnego oprogramowania (w tym wielu wbudowanych funkcji w językach programowania[151]), dowolność konfiguracji, nie jest praktyką wykorzystywaną w profesjonalnych firmach czy instytucjach. Natomiast tego typu rozwiązania są chętnie

[147] Deep packet inspection - technika sieciowa pozwalająca dostawcy usług internetowych analizować pakiety przesyłane przez sieć pod względem ich treści. W zależności od zawartości pakiet jest zatrzymywany, opóźniany, zmieniany lub przesyłany dodatkowo w celu zapisania.
[148] MUA (ang. Mail User Agent) – oprogramowanie komputerowe używane do uzyskiwania dostępu do poczty e-mail użytkownika i zarządzania nią.
[149] MTA (ang. Message Transfer Agent) – oprogramowanie, które przesyła wiadomości e-mail z jednego komputera na drugi (z jednego serwer pocztowego na inny) za pomocą SMTP (ang. Simple Mail Transfer Protocol).
[150] MDA (ang. Message Delivery Agent) – składnik oprogramowania komputerowego (serwera poczty internetowej), który jest odpowiedzialny za dostarczanie wiadomości e-mail do lokalnej skrzynki pocztowej odbiorcy.
[151] Np. funkcja mail() w języku PHP (https://www.php.net/manual/en/function.mail.php)

stosowane przez w niszowych programach do zarzadzania pocztą email oraz chętnie wykorzystywane przez grupy cyberprzestępców, kreujących ataki phishingowe z wykorzystaniem poczty email, i z tego powodu automatyczne wykrycie oprogramowania, jakim posłużono się do stworzenia danej wiadomości może dostarczyć informacji czy dana wiadomość jest phishingiem, czy też nim nie jest. Narzędzia programowe jakimi posłużył się twórca danej wiadomości mogą być wartościami w polach nagłówka:

1. X-Mailer,
2. User-Agent:

Zgodnie z dokumentem standaryzującym opisującym najczęściej występujące pole nagłówka wiadomości email (RFC 2076[152]), pole oprogramowania twórcy (X-Mailer) jest niestandardowym polem (nie musi wystąpić w nagłówku wiadomości) i dość rzadko występującym w standardowym nagłówku. (w 2015 roku około 29% analizowanych nagłówków posiadło pole X-Mailer[153]). Z uwagi na dominującą popularność usługi Gmail[154] oraz innych operatorów z wykorzystaniem web interfejsów przewiduje się, że występowanie tego pola w przyszłości będzie maleć, cecha ta może wiec w przyszłości mniejsze znaczenie do wykrywania phishingu.

Tabela 30. Narzędzia programowe służące do wysyłki wiadomości email.

Lp.	Pole X-Mailer	Wartość pola	ID próbki
1.	TAK	PHPMailer 5.2.2-rc2	P-24-001
2.	TAK	PHPMailer	P-24-002

Opis cechy:

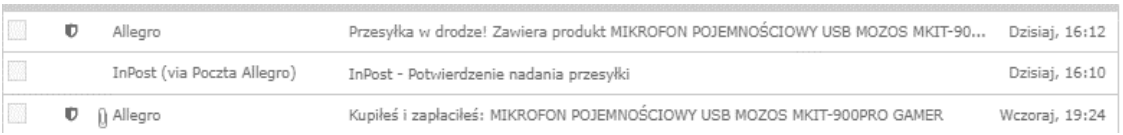
$$f_{24} = \begin{cases} 1 & dla\ \ m\ \exists\ oraz\ m\ \in\ x \\ 0 & dla\ \ m\ \exists\ oraz\ m\ \notin\ x \\ null & dla\ m \in \emptyset \end{cases} \tag{3.14}$$

gdzie:
m – pole nagłówka zawierające informację o oprogramowaniu twórcy wiadomości,
x – lista niestandardowych wartości pól X-Mailer

[152] https://datatracker.ietf.org/doc/html/rfc2076
[153] https://www.vanimpe.eu/2015/05/02/analyzing-spam-e-mail-headers/ [dostęp 11.02.2022r.]
[154] W 2021 roku usługa Gmail odpowiadała za obsługę 36% przesłanych wiadomości email (źródło: https://www.statista.com/statistics/265816/most-used-e-mail-service-by-market-share/, [dostęp 13.02.2022r.])

III.1.25 Wykorzystanie tagowania wiadomości przez serwery pocztowe

Portale internetowe (serwisy społecznościowe, platformy zakupowe) skupiające wokół siebie wielu zarejestrowanych użytkowników, jako podstawowy kanał komunikacji z tymi użytkownikami wykorzystują zwykle mechanizm poczty email. Masowa rozsyłka wiadomości email do wielu użytkowników, może u niektórych dostawców usług pocztowych, uruchomić mechanizmy antyspamowe (blokada wiadomości email pochodzących od określonego nadawcy, umieszczenie wiadomości w folderze „spam"). W celu ominięcia mechanizmów filtrowania masowych wiadomości email, wielu nawiązało współpracę z dostawcami usług pocztowych, w celu zapewnienia, że podczas masowej rozsyłki, wiadomości email pochodzące z ich autoryzowanego systemu pocztowego, zostaną przekazane końcowemu użytkownikowi. Dostawca usług pocztowych zwykle oznacza wiadomości pochodzące od partnerów (z ustalonego wcześniej adresu serwera pocztowego), by użytkownicy (odbiorcy tej wiadomości), mogli ją odczytać.



Rysunek 38.Przykład oznaczania zaufanego nadawcy wiadomości poprzez umieszczenie ikonografii przy temacie wiadomości.

Oznaczanie wiadomości może przybrać różne formy:

1. umieszczenie ustalonej wcześniej ikonografii przy temacie wiadomości oznaczającej pochodzenie tej wiadomości od zaufanego nadawcy,
2. wyróżnienie wiadomości na liście odebranych wiadomości,
3. umieszczenie informacji nad treścią, że wiadomość pochodzi od zaufanego nadawcy.

Tabela 31. Przykład fałszywego oznaczania wiadomości.

Lp.	Rodzaj oznaczenia	Treść oznaczenia (tagu)	ID próbki
1.	Osadzony w treści wiadomości	This message was sent from a trusred sender.	P-25-001
2.	Osadzony w treści wiadomości	Ta wiadomość została wysłana przez NIEZAWODNEGO nadawcę	P-25-002

W celu zwiększenia skuteczności phishingu, atakujący, wykorzystuj metody inżynierii społecznej, umieszczają wewnątrz spreparowanej przez siebie wiadomości email, Informację, która ma imitować oznaczenie zaufanego nadawcy. Umieszczenie

takiej informacji ma w zamyśle uśpić czujność odbiorcy i skłonić go do wykonania sugerowanych w danej wiadomości działań.

Możliwą do wykrycia cechą jest powtarzalny tekst, oznaczenie udające oryginalne nadawane przez dostawcę usług, osadzone w samej treści wiadomości email.

$$f_{25} = \begin{cases} 1 & dla\ t\ \in x \\ 0 & dla\ t\ \notin x \end{cases} \tag{3.15}$$

gdzie:
t – wartość oznaczenia umieszczonego przez atakującego,
x – lista fraz imitujących tagowanie wiadomości

Możliwym błędem jest umieszczenie w treści wiadomości frazy pokrywającej się z wartością fałszywie umieszczonego oznaczenia o zaufanym nadawcy. Fraza taka może stanowić cytat, przykład, opis i nie powinna wówczas być brana pod uwagę jako potencjalny wskaźnik phishingu.

III.1.26 Różne treści osadzone w tej samej wiadomości

Metodą wykorzystywaną do oszukania mechanizmów antyspamowych, bazujących na analizie treści jest stosowanie techniki polegającej na umieszczeniu wewnątrz wiadomości różnych jej części:

a. części tekstowej,
b. części formatowanej za pomocą języka HTML i arkuszy CSS,
c. niewidocznej dla użytkownika treści, formatowanej za pomocą języka HTML z ustawionymi wartościami parametrów „display” lub „visibility” (odpowiednio na „none” i „hidden” co powoduje ukrycie, brak wyświetlenia danego elementu w programie pocztowym odbiorcy wiadomości).

Rysunek 39. Przykład wiadomości z różną treścią: obrazek a) – Brak widocznej treści w programie pocztowym (Mozilla Thunderbird), obrazek b) – źródło wiadomości ujawnia zwartą treść zakodowaną za pomocą BASE64, obrazek c) – zdekodowana z BASE64 treść ujawnia kodowaną za pomocą języka HTML treść.

Treść każdej z części jest inna – użytkownik w zależności od konfiguracji swojego programu pocztowego odczytuje albo część tekstową albo część formatowaną za pomocą języka HTML, jednakże dla modułu analizy treści, odczytywane są wszystkie jej części (wraz z niewidoczną/ukrytą zawartością), co powoduje, że porównując podobieństwo w ten sposób przygotowanej wiadomości wykazuje znaczne różnice w treści do wiadomości rozpoznanych jako phishingowe.

Możliwą do wykrycia cechą jest przeprowadzenie automatycznej analizy treści wiadomości w poszukiwaniu osadzenia w niej różnych jej części: części tekstowej, części kodowanej za pomocą języka HTML czy też części zakodowanej. Zgodnie z dokumentem standaryzacyjnym RFC 1341[155], pojedyncza wiadomość może zawierać różne fragmenty zgodne z typami MIME[156]. W tego typu wiadomości (zawierającej różne części), separatorem jest pole określone jako „boundary" (lub „multipart"). Wykrycie w danej wiadomości pola o tej nazwie pozwala więc na odczytanie jej poszczególnych fragmentów (oddzielnie każdego z nich) i porównanie zawartości poszczególnych części[157].

Opis cechy:

$$f_{26} = \begin{cases} 1 \;\; dla \;\; x_1 \neq x_2 \neq np.x_i, \;\; i = 1,2,\dots,n \\ 0 \;\; dla \;\; x_1 = x_2 = np.x_i, \;\; i = 1,2,\dots,n \\ null \;\; dla \; x \in \emptyset \end{cases} \quad (3.16)$$

gdzie:

x_i – treść i-tej części wiadomości email (pozbawionej formatowania)

Z uwagi na możliwość wczytywania zdalnej zawartości i osadzania jej w prezentowanej odbiorcy treści (np. wyświetlenie obrazu z osadzonym w nim tekstem, co jest częstą praktyką w przypadku wiadomości o charakterze agresywnego marketingu w celu oszukania filtrów antyspamowych) porównywanie treści może dawać błędne wskazania. Analizując pozyskany zbiór wiadomości email, zaobserwowano występujące w nich powtarzalne schematy, w przypadku występowania różnych części składowych wiadomości. Z tego powodu, na potrzeby analizy nagłówka i wykrywania osadzania różnych treści przyjęto poniższą klasyfikację wiadomości.

[155] RFC 1341 – źródło: https://www.w3.org/Protocols/rfc1341/0_Abstract.html

[156] MIME (wł. typ MIME lub media typ, ang. Multipurpose Internet Mail Extensions) – identyfikator opisujący format plików treści przesyłanych w sieci Internet

[157] W przypadku wykrycia kodowania treści za pomocą języka HTML, konieczne jest pozbycie się wszystkich znaczników HTML, kodowania CSS i pozostawienie jedynie samej treści (bez jakiegokolwiek formatowania).

Tabela 32. Przyjęty podział typów wiadomości email z uwagi na możliwość występowania różnych części składowych.

Rodzaj wykrytych danych/nazw pól	Występowanie				
Część tekstowa	TAK	TAK	NIE	NIE	TAK
Format HTML	NIE	NIE	TAK	TAK	TAK
Wiadomość zawiera wiele części (multipart)*	TAK	NIE	TAK	NIE	TAK
Występowanie pola „boundary[158]”	TAK	NIE	TAK	NIE	TAK
Możliwe występowanie załącznika	TAK	NIE	TAK	NIE	TAK
Przyjęty typ wiadomości	1	2	3	4	5

Bazując na powyższej tabeli można określić:

1. Typ „1” – Wiadomość zawiera część tekstową oraz pole rozgraniczające „boundary”. Wiadomość nie zawiera części formatowanej HTML. Wiadomość może zawierać załączniki lub obrazki graficzne w niej osadzone.
2. Typ „2” – Wiadomość zawiera jedynie część tekstową. Wiadomość nie może zawierać załączników.
3. Typ „3” – Wiadomość zawiera część formatowaną HTML. Wiadomość może zawierać załączniki lub obrazki graficzne w niej osadzone. Wiadomość nie zawiera części tekstowej.
4. Typ „4” – Wiadomość zawiera część formatowaną HTML. Wiadomość nie może zawierać załączników
5. Typ „5” – Wiadomość zawiera wiele części, w tym część tekstową oraz część formatowaną HTML. Wiadomość może zawierać załączniki lub obrazki graficzne w niej osadzone.

Z powyższego opisu można wyszczególnić wiadomości, które zawierają zarówno część tekstową jak i część formatowaną HTML. Usuwając formatowanie HTML, można wówczas porównać treść takich wiadomości.

Tabela 33. Przykład wiadomości zawierającej różne treści.

Lp.	Typ wiadomości	ID próbki
1.	Typ „5”: ▪ Część tekstowa ▪ Część formatowana kodem HTML ▪ Załącznik kodowany base64	P-26-001

[158] Pole „bonduary” – 7 bitowy ciąg ASCII znajdujący się wewnątrz wiadomości email, który definiuje granice poszczególnych fragmentów MIME wiadomości, jeżeli wiadomość zawiera więcej niż jedną część.

III.1.27 Inne załączniki

Złośliwe załączniki, czy też osadzony w dokumencie kod HTML, finalnie prowadzący do pobrania złośliwego elementu, nie są jedynymi elementami wiadomości, na które należy zwrócić uwagę, analizując zjawisko phishingu. Z uwagi na możliwość istnienia w danej wiadomości email, różnych typów plików, istotnym elementem systemu pocztowego staje się rozwiązanie umożliwiające detekcję i analizę różnego rodzaju plików, pod katem możliwości występowania niepożądanego oprogramowania. Omówione w ramach analizy złośliwych załączników skanery antywirusowe skupiają swoje działanie na wykryciu w analizowanym pliku obecności złośliwego oprogramowania. Jednakże z uwagi na niektóre typy plików, nie jest możliwe określenie czy nie zawierają one niepożądanego kodu czy niedozwolonego działania. Do takich typów zaliczyć możemy:

1. archiwa (*.rar, *zip) – archiwa mogą zawierać inne pliki, w tym oprogramowanie złośliwe lub niepożądane[159], których uruchomienie może wywołać niepożądane skutki. Dostawcy darmowych usług pocztowych, zwykle nie poosiadają rozbudowanych możliwości skanowania archiwum (wyjątkiem jest Gmail[160]), co stwarza możliwość dostarczenia złośliwej zawartości bezpośrednio na stację roboczą użytkownika.
2. archiwa zabezpieczone hasłem dostępu – w niektórych rozwiązaniach, zwłaszcza darmowych wersjach), zabezpieczone hasłem archiwum uniemożliwia skanowanie jego zawartości przez silniki antywirusowe (lub inne rozwiązania zapewniające bezpieczeństwo lub zgodność z polityką organizacji).
3. pliki obrazów dysku (ISO/IMG) – pliki te mogą w sobie zawierać inne pliki

Rozszerzenie cech

Zaprezentowane wskaźniki, stanowić mogą podstawowy zbiór wskaźników umożliwiających detekcję ataku phishingowego. Analizując poszczególne składowe wskaźników, przytoczony zbiór, można rozszerzać o inne elementy w obrębie danego wskaźnika. Takie podejście pozwoli na modyfikację wektora cechy (jego rozszerzenie), co może znacznie poprawić jakość wykrywania ataku phishingowego.

[159] Pod pojęciem oprogramowania niepożądanego w danej organizacji, należy rozumieć ogół oprogramowania, które polityka bezpieczeństwa danej organizacji, wyklucza z możliwości korzystania przez użytkowników.
[160] https://support.google.com/a/answer/2364580?hl=pl

III.2 Podobieństwa cech

Określając cechy występujące w wiadomościach phishingowych, należy również rozważyć i zidentyfikować cechy wiadomości określanych jako spam – z uwagi na adaptację w ataku phishingowym, niektórych technik i metod wykorzystywanych w wiadomościach typu spam. Również w normalnej wymianie korespondencji email, można zidentyfikować niektóre z cech przypisywanych wiadomością o charakterze phishingowym.

Tabela 34. Zestawienie podobieństwa opisanych cech.

Cecha	Normalne	Spam	Phishing
Nieprawidłowy odnośnik URL zawarty w wiadomości			TAK
Adres IP w odnośniku URL zawartym w wiadomości			TAK
Wykorzystanie serwisów skracających odnośniku URL		TAK	TAK
Złośliwy załącznik, skrypty wykonywalne (VBS/VBA)			TAK
Groźba w przypadku niepodjęcia przez ofiarę sugerowanych działań (wskaźnik nietechniczny)			TAK
Nieprawidłowy adres email nadawcy			TAK
Niewłaściwy adres nadawcy			TAK
Niespójność nazwy nadawcy			TAK
Wykorzystanie nazwy odbiorcy wiadomości		TAK	TAK
Wykorzystanie nazwy domenowej jako nazwy użytkownika w adresie nadawcy			TAK
Automatyczne generowanie nazwy użytkownika lub domeny w adresie email nadawcy		TAK	TAK
Mechanizm śledzący w wiadomości email		TAK	TAK
Strona wyłudzająca dane (wskaźnik techniczny / nietechniczny)			TAK
Typosquatting / domen udające istniejące			TAK
Złożona nazwa domenowa wraz z subdomenami			TAK
Wiek zarejestrowanej domeny			TAK
Brak zarejestrowanej domeny, wykorzystanie adresów chmurowych	TAK	TAK	TAK
Spoofing instytucji			TAK
Błędy językowe (wskaźnik nietechniczny)	TAK	TAK	TAK
Temat otrzymanej wiadomości			TAK
Niespójna szata graficzna (wskaźnik nietechniczny)			TAK
Nietypowe prośby, niespodziewana treść		TAK	TAK
Niespodziewane załączniki (wskaźnik techniczny/nietechniczny)			TAK
Użycie narzędzi programowych do wysyłki wiadomości email	TAK	TAK	TAK
Wykorzystanie tagowania wiadomości przez serwery pocztowe	TAK	TAK	TAK

Identyczna treść przesłana ze skompromitowanego konta do wszystkich użytkowników kont email w danej organizacji, w krótkim okresie		TAK	TAK
Masowa wysyłka wiadomości email z jednego konta	TAK	TAK	TAK

Rozdział IV – Metoda wykrywania wiadomości phishingowych

Klasyczne[161] metody wykrywania phishingu (do których zaliczyć możemy czarne/białe listy, świadomość użytkowników, kwestie prawne, podobieństwo wizualne, raporty użytkowników – podejście ro opisane np. w pracach: [81], [82]), są oparte na wcześniej zidentyfikowanych nieprawidłowościach, modyfikacji, odstępstw otrzymanej wiadomości w stosunku do normalnej korespondencji. Modyfikacje te mogą być trudno wykrywalne dla człowieka.

Dodatkowym aspektem, zwiększającym szansę powodzenia ataku phishingowego, a utrudniającego analizę otrzymanej wiadomości pod kątem możliwego ataku, jest wykorzystanie inżynierii społecznej. Wykorzystująca silne emocje człowieka inżynieria społeczna z powodzeniem zakłóca proces weryfikację poprawności u odbiorcy wiadomości, co finalnie prowadzi do realizacji zakładanych przez atakujących celów.

W publikowanych dostępnych opracowaniach (np. [48], [83], [63], [64], [84], [60]) jako techniczną metodę[162] wykrywania phishingu, wskazuje się głównie na wykrycie i analizę wartości kilku pól, np.:

1. adres IP występując w odnośniku URL występującym w treści wiadomości.
2. odnośnik URL zawarty w treści wiadomości znajduje się bazach domen phishingowych (np. APWG),
3. wykorzystanie nieszyfrowanej wersji protokołu NP. zamiast HTTPS,
4. ilość kropek w nazwach domenowych,
5. występowanie charakterystycznych dla wiadomości phishingowych frazach znajdujących się zarówno w tytule wiadomości jak i w jej treści (np. „potwierdź aktualizację", „weryfikacja konta", np.),
6. daty zarejestrowania domeny, do której prowadzi odnośnik URL zawarty w wiadomości email,
7. kodowanie treści za pomocą języka HTML,
8. ilość odnośników zawartych w danej wiadomości email.

[161] Wyrażenie „klasyczne", należy rozumieć jako metody, który nie zawierają klasyfikacji, elementów uczenia maszynowego. W tym sensie są to metody i techniki szeroko opisywane i implementowane w rozwiązaniach bezpieczeństwa.

[162] Poprzez techniczną metodę detekcji phishingu należy w tym przypadku rozumieć, możliwość wykrycia i odczytania wartości cechy przez automatyczny mechanizm

Niektóre z wymienionych powyżej wskaźników, z uwagi na zachodzące zmiany technologiczne, rozwój niektórych usług (np. możliwość darmowej rejestracji domeny), nie mogą współcześnie być wykorzystywane jako cecha phishingowa i nie można opierać detekcji bazującej na ich identyfikacji. Do wskaźników tych zaliczyć można:

1. Adres IP występując w odnośniku URL. Wskaźnik ten często wymieniany jest w opracowaniach, jednakże w zgormadzonych do badań materiałach występował w niewielkiej ilości, lub nie występował wcale. Spowodowane jest to znacznie szerszym dostępem do możliwości wykupu i rejestracji domeny przez szersze grono użytkowników. Wpływ na to miało znaczne obniżenie cen oraz udostępnienie przez wielu operatorów możliwość darmowej rejestracji domeny.
2. Wykorzystanie protokołu NP. Z uwagi na upowszechnienie się usługi generowania certyfikatów SSL[163], stosowanie szyfrowanej wersji protokołu NP. systematycznie zaczęło wzrastać. Jedną z przyczyn jest wymuszenie przez twórców przeglądarki Google Chrome[164] od lipca 2018 roku[165] oznaczenie witryn internetowych nie wspierających protokołu HTTPS jako niebezpiecznych. Działanie takie znacznie przyspieszyło proces wdrażania szyfrowanej wersji protokołu HTTPS przez twórców witryn internetowych. Wykorzystanie certyfikatów SSL oraz protokołu HTTPS zostało również zaadoptowane przez atakujących – we wrześniu 2021 roku aż 82%[166] zidentyfikowanych domen phishingowych wykorzystywało protokół HTTPS. Nieszyfrowana wersja protokołu NP. jest więc rzadko spotykana w otrzymywanych odnośnikach URL znajdujących się w wiadomości email – zarówno klasyfikowanej jako normalna korespondencja (z rzeczywistymi, bezpiecznymi odnośnikami), jak i klasyfikowanej jako spam czy phishing (z odnośnikami prowadzącymi do domen oznaczonych jako niebezpieczne). Z tego powodu, znaczenie tego

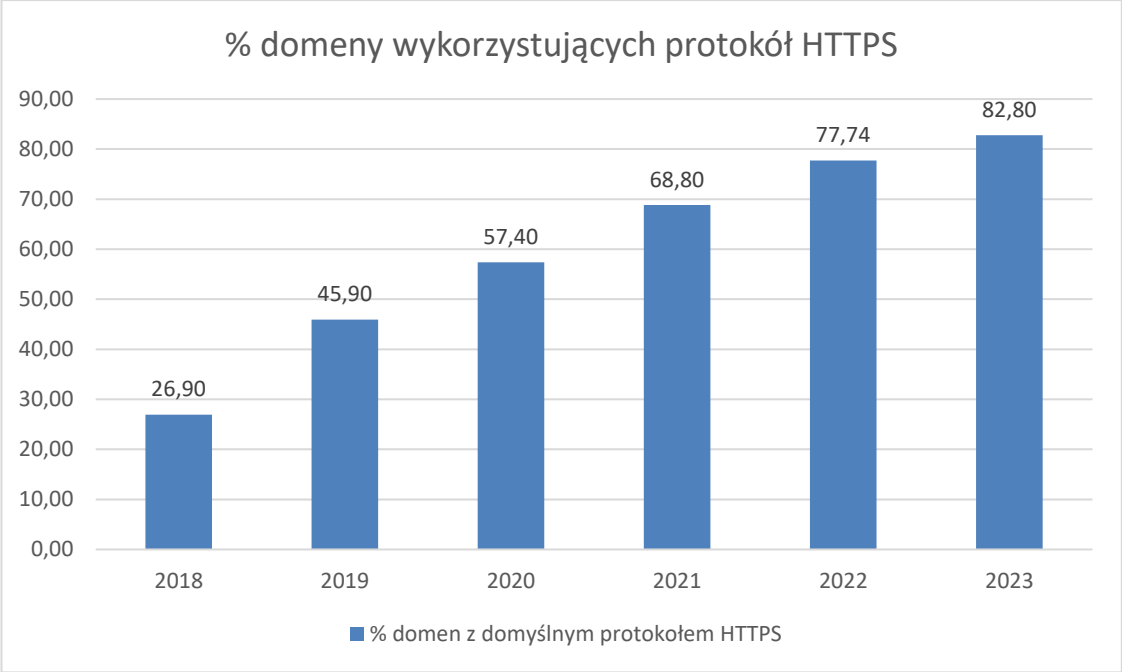
[163] Certyfikat SSL – jest to niewielki plik, generowany zwykle przez niezależny, zaufany podmiot, poświadczający wiarygodność domeny bądź domeny oraz jej właściciela. Potwierdza bezpieczeństwo szyfrowania danych przesyłanych pomiędzy użytkownikiem a serwerem

[164] Wg rankingu opublikowanego przez statcounter, przeglądarka chrome z udziałem w światowym rynku na poziomie 63.51% pozostaje liderem wśród przeglądarek, źródło: https://gs.statcounter.com/browser-market-share [dostęp: 16.05.2023].

[165] Dokładnie od opublikowania przez firmę Google przeglądarki Chrome w wersji 68, źródło: https://developer.chrome.com/blog/new-in-chrome-68/ [dostęp: 16.05.2023].

[166] Raport „Top 10 TLDs Abused" firmy Fortra. Źródło: https://www.phishlabs.com/blog/top-10-tlds-abused/, data publikacji: 14.09.2021 [dostęp: 16.05.2023].

wskaźnika systematycznie (od 2018 roku) maleje i w modelach detekcji może pełnić funkcję wspomagającą inne reguły.



Rysunek 40. Procent domen w sieci Internet wykorzystujących domyślnie protokół HTTPS. Źródło: https://w3techs.com/technologies/history_overview/site_element/all/y [dostęp: 16.05.2023].

3. Kodowanie treści za pomocą języka HTML. Tego typu kodowanie treści może przenosić złośliwy kod (wczytywany ze zdalnego serwera kod CSS, JavaScript, który uruchomiony na komputerze lokalnym może posłużyć do dalszej eskalacji i finalnie zainfekowania złośliwym oprogramowaniem), jednakże współczesne edytory wiadomości email (zarówno cienki[167] jak i gruby[168] klient) oferują swoim użytkownikom jako domyślnie włączoną opcję, tworzenie wiadomości email z wykorzystaniem formatowania za pomocą języka HTML i kaskadowych arkuszy styli CSS. Z tego powodu, większość wysyłanej i odbieranej korespondencji email, zawiera treść formatowaną tym właśnie sposobem – samo

[167] Cienki klient (ang. thin client) – termin oznaczający komputer z zainstalowanym specjalnym oprogramowaniem typu klient lub specjalne oprogramowanie do komunikacji typu klient-serwer, które uruchamiane na komputerze użytkownika, korzystają jednak z zasobów serwerowych (przetwarzanie danych odbywa się po stronie serwera). Zmiana aplikacji serwerowej nie pociąga za sobą konieczności wymiany oprogramowania klienta (np. przeglądarka internetowa wykorzystywana jako klient pocztowy). Jest przeciwieństwem grubego klienta.

[168] Gruby klient (ang. fat client) – termin oznaczający komputer z zainstalowanym systemem operacyjnym i pełnym zestawem aplikacji do komunikacji typu klient-serwer, które uruchamiane są i działają bezpośrednio na komputerze użytkownika, przetwarzając dane z wykorzystaniem jego zasobów lokalnych (dysk twardy, pamięć operacyjna, karta grafiki, itp.) – np. program pocztowy Mozilla Thunderbird. Jest przeciwieństwem cienkiego klienta.

więc wykorzystanie języka HTML do stworzenia wiadomości email nie może być więc wskaźnikiem możliwego ataku phishingowego.

4. Ilość odnośników URL. Kampanie marketingowe wykorzystują mechanizm wczytywania zdalnej zawartości w postaci grafiki reklamowej. Przesłanie i wyświetlenie pojedynczego rozmiaru (często o wysokiej jakości, a więc i odpowiednio dużej wielkości pliku), może wiązać się z opóźnieniami wczytywania, dlatego grafika dzielona jest na kilka (czasem kilkanaście) mniejszych fragmentów i dla każdego z niech generowany jest odnośnik URL do niego prowadzący. Z tego powodu niektóre wiadomości marketingowe (nie będące spamem ani phishingiem) mogą zawierać wiele odnośników URL w kodzie źródłowym wiadomości.

IV.1 Metoda identyfikacji cech wskazujących na potencjalnie phishingowy charakter wiadomości.

Przedstawiona w niniejszej rozprawie autorska metody identyfikacji cech wskazujących na atak phishingowy, wykorzystywać będzie zidentyfikowane i opisane w Rozdziale III wskaźniki techniczne (patrz: Rozdział III – Techniczne i nietechniczne wskaźniki phishingu, strona: 102). Założeniem metody jest, że dana wiadomość zawierać będzie co najmniej jedno pole (lub wiele pól), którego odczytana i przetworzona wartość (wartości) poddana procesowi uczenia maszynowego, pozwoli na określenie do jakiej klasy należy wiadomość. W niektórych przypadkach do przetwarzania wartości odczytanych pól, metoda będzie oprócz zbudowanych reguł wykorzystywała uczenie maszynowe.

Metoda klasyfikacji wiadomości email wykorzystywana do wykrywania ataku phishingowego stosować będzie – jako jedno z kryteriów – badanie podobieństwa wiadomości email na podstawie:

1. Podobieństwo adresów – masowe wiadomości phishingowe rozsyłane są z tego samego adresu email lub posiadają ten sam adres IP serwera nadawcy. Do tej kategorii zaliczać się będą również adresy email różniące się pomiędzy sobą, lecz posiadające wspólne cechy:
    a. Polimorfizm nazwy użytkownika w obrębie tej samej domeny pocztowej

Tabela 35. Polimorfizm nazwy użytkownika w obrębie tej samej domeny pocztowej.

Lp.	Adres email nadawcy
1.	1255.73-**replies**@3067.drivefact.org
2.	1428.06-**replies**@6888.drivefact.org
3.	**arthurcdumas010**+6bsHPl8GJGMPvYGZJ4YY@gmail.com
4.	**arthurcdumas010**+6WGJ9WcW9olkPnQhkfKn@gmail.com

b. Polimorfizm nazwy użytkownika w obrębie różnych domen pocztowych

Tabela 36. Polimorfizm nazwy użytkownika w obrębie różnych domen pocztowych.

Lp.	Adres email nadawcy	Wspólna cecha
1.	**azzfass4**@enviedebienmanger.fr	nazwa użytkownika
2.	**azzfass4**@gmail.com	nazwa użytkownika
3.	**nikola**wersa33@gmail.com	mix fragmentów nazwy
4.	**wersa**73**nikola**@hotmail.com	mix fragmentów nazwy

c. Powtarzalność frazy w nazwie użytkownika i części domenowej

Tabela 37. Powtarzalność frazy w nazwie użytkownika i części domenowej.

Lp.	Adres email nadawcy	Wspólna cecha
1.	**AHMAX323A.AHMAX323A@AHMAX323A**.us	AHMAX323A

d. Tożsamości adresów – adres widoczny w polu nadawcy (pole „Form” / „Od”) może zostać podmieniony (spoofing), wówczas adresy email w polach „Return-Path”, „Reply-To” (o ile występują) będą się różniły pomiędzy sobą. Rzeczywiste wiadomości email mają zachowaną tożsamość adresów email we wszystkich tych polach (o ile wszystkie występują w danej wiadomości – zależy to od indywidualnej konfiguracji danego serwera pocztowego).

Obecnie ocenianie podobieństwa adresów email (z uwagi na podobieństwo występujących w nim fraz, wyrazów, np.) możliwe jest jedynie w przypadku, posiadania obszernej bazy danych zawierających adresy email, które zidentyfikowane zostały jako adresy, z których wysłane zostały wiadomości o charakterze phishingowym. Jest to dość istotne organicznie tej metody i konieczność utrzymywania dużych zasobów danych, ich ciągła aktualizacja – co może generować duże nakłady administracyjne. Analiza podobieństwa adresów email ze względu na wariacje nazwy użytkownika jest obiecującą i perspektywiczną metodą wykrywania możliwego niewłaściwego adresu

nadawcy, z uwagi na brak ograniczenia powodowanego koniecznością posiadania obszernej (i stale aktualizowanej, co powoduje duży koszt administracyjny) bazy adresów email.

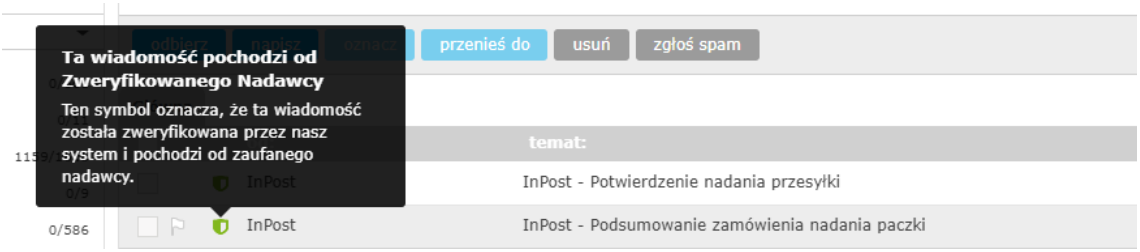
2. Podobieństwo tematu – atak phishingowy kierowany na daną instytucję i przesyłany na konta jej pracowników charakteryzuje się masową wysyłką wiadomości (duże podobieństwo do spamu), w krótkim przedziale czasu (patrz: pkt. 4) o tym samym tytule.
3. Podobieństwo treści – wiadomości email kierowane do różnych odbiorców zawierają tą samą lub bardzo podobną treść, powtarzające się frazy, ten sam układ wiadomości czy identyczną szatę graficzną (podobieństwo do spamu). Podobieństwo treści może bazować również na wykorzystaniu podobieństw znaczeniowych wyrazów (różne wyrazy, lecz o tym samym znaczeniu w kontekście do całości otrzymanej treści).
4. Podobieństwo czasu – wiele wiadomości phishingowych wysyłanych jest masowo w krótkim przedziale/odstępie czasu z tego samego adresu email (adresu IP serwera nadawcy). W odniesieniu do zaprezentowanego modelu faz ataku phishingowego masowa wysyłka wiadomości email następuje w pierwszej fazie ataku (Rysunek 21 – Fazy ataku phishingowego – model uproszczony). Istotny jest odczyt daty otrzymania wiadomości email, porównanie go z datą wystawienia certyfikatu domeny do której się odwołuje czy też daty rejestracji tejże domeny.
5. Podobieństwo cech – wiadomości, w których widoczny nadawca (pole „Form" / „Od") jest różny, różne są również adresy serwerów nadających daną wiadomość, jednakże wykorzystany został podobny schemat układu, konstrukcja.
6. Powtarzalność schematu – wiadomości phishingowe konstruowane wg. Powtarzalnego układu, zawierającego zestawy tych samych cech phishingowych. Do takich zestawów zaliczyć można np.:
    a. Nieznaczna modyfikacja adresu email z pola nadawcy („From" / „Od") i umieszczenie go w polu „Reply-To".
    b. Umieszczenie nazwy użytkownika z adresu email odbiorcy (wartość sprzed znaku „@") w tytule wiadomości lub w treści (np. jako powitanie).

    Wartości tych cech mogą być różne (np. inne adresy widoczne w polu „From" / „Od" dla kolejnych otrzymywanych wiadomości).
7. Reputację domeny (czas jej życia) – jak wykazano [79] krótki czas życia domen o charakterze phishingowym w połączeniu z masowością wysyłki w krótkim

przedziale/odstępie czasu jest cechą wskazującą na możliwość ataku phishingowego.

Ponadto, metoda ta obejmować będzie:

1. Analizę treści wiadomości:
    a. wykrywanie możliwego szantażu wpłaty okupu w postaci kryptowaluty (BitCoin). Atak oparty na inżynierii społecznej, brak jest więc technicznych elementów wskazujących na phishing, analiza treści ujawni natomiast złośliwy charakter wiadomości,
    b. wykrywanie nagromadzonych błędów językowych,
    c. umieszczania dodatkowych treści mających wskazywać na automatycznie dodany tag przez dany serwer pocztowy, mający uwiarygodnić autentyczność wiadomości. Z uwagi że oznaczenie takie nie pojawia się zwykle w treści wiadomości a jedynie w metadanych lub w inny sposób uwidocznione (np. na liście odebranych wiadomości w danym kliencie pocztowym), wykrycie jej w treści wiadomości wskazywać może na próbę ataku phishingowego.



Rysunek 41. Przykład rzeczywistego oznaczenia (tagowanie) wiadomości pochodzącej od zaufanego odbiorcy, wykorzystywane przez operatora poczty Wirtualna Polska. Źródło: opracowanie własne.

    d. Umieszczenie niewidocznych treści (część tekstowa i część formatowana za pomocą języka HTML).
2. Analizę domen oraz odnośników URL – zgodnie z przytoczonymi badaniami [69], [70], [71] oraz własnymi analizami (patrz: IV.3.4 Moduły uczące, str. 183) –wartości ilości kropek, ukośników („/”) czy długości odnośników URL i domen phishingowych odbiegają ilościowo od tych samych wartości domen i odnośników rzeczywistych, bezpiecznych zasobów. Analiza obejmować będzie również:
    a. wykrywanie w danym odnośniku serwisu skracającego odnośniki

3. Wykrywanie stosowanych narzędzi programistycznych i niestandardowych typów oprogramowania klienckiego (USER-AGNET) służących do wysyłki wiadomości.
4. Analizę wykorzystania mechanizmów śledzących. Mechanizm śledzący osadzony jest zwykle w treści wiadomości email formatowanej za pomocą języka HTML (osadzony niewidoczny znacznik <IMG>). Osadzenie w kodzie języka HTML elementów mogących być niebezpiecznymi dla użytkownika lub ujawniające informacje o nim (np. USER-AGENT, adres IP, rodzaj przeglądarki, np.) będą wykrywane za pomocą analizy formatowania HTML.

Przedstawiona powyżej analiza, pozwoli na opracowanie nowej metody, pozbawionej słabości opisanych w Rozdziale II, która wykryje wcześniej nie stosowane schematy ataku oraz używanie nieznanych do tej pory wzorców.

IV.1.1 Założenia wiadomości phishingowej

Konstruując metodę wykrywania wiadomości phishingowych, konieczne jest opracowanie modelu, jaki będzie spełniała wiadomość, która może być uznana za phishingową:

1. opisane i wybrane cechy możliwe są do wykrycia w sposób algorytmiczny,
2. wiadomość phishingowa zawierać będzie co najmniej jedną z opisanych cech,
3. sugestywna treść wiadomości zachęcająca do podjęcia określonego działania, którego wykonanie stanowić będzie szkodę dla odbiorcy tej wiadomości (ofiary ataku),
4. co najmniej jedna z opisywanych cech, znajdzie się w drugiej (lub kolejnej) otrzymanej wiadomości – pod warunkiem realizacji sugerowanych działań w wiadomości inicjalizującej.

IV.1.2 Klasyfikacja wiadomości typu spam.

Klasyfikacja i filtrowanie spamu, może odbywać się z wykorzystaniem różnych metod. Do najpopularniejszych należą:

1. Czarne listy – działanie opiera się na utworzonych listach zawierających adresy email nadawców uznanych za niepożądanych w danej organizacji. Wada takiego rozwiązania opisana została w rozdziale: Rozdział II – Analiza metod detekcji phishingu.

2. Systemy regułowe – filtracja bazująca na utworzonych regułach detekcji cech charakterystycznych dla wiadomości typu spam, np. określone słowa, frazy w nagłówku, odnośniki kierujące do określonych domen (uznawanych za phishingowe), np.
3. Filtrowanie treści – analiza wiadomości pod kątem zawartych w niej pewnych sformowań lub słów kluczowych. Konieczne jest posiadanie obszernej bazy słów kluczowych i fraz niezbędnych do skutecznego działania systemu klasyfikacji (konieczność ciągłego aktualizowania bazy o nowe sformułowania i wyrażenia użyte w zmieniających się szablonach wiadomości).
4. Analiza nagłówka wiadomości – identyfikacja adresów email nadawców. Podobnie jak w przypadku filtrowania treści, konieczne jest posiadanie obszernej bazy danych zwierających adresy email, z których rozsyłane są wiadomości typu spam. Rozwiązanie to nie jest skuteczne, z powodów:
    a. łatwość utworzenia kont email – operatorzy darmowych usług email, nie ograniczają ilości zakładanych kont email, nadawcy wiadomości typu spam, mogą więc utworzyć dowolną ilość adresów email (różniących się pomiędzy sobą nazwą użytkownika) i wykorzystywać je partiami do rozsyłki poszczególnych kampanii wiadomości spam.
    b. łatwości generowania adresów email – tworząc własną infrastrukturę do rozsyłki wiadomości email, operator spamu może wygenerować dozwoloną ilość adresów email, różniących się zarówno nazwą użytkownika jak i adresem domenowym.
    c. łatwości pozyskania (kompromitacji) adresu email – przejęte konta email (w wyniku np. ataku phishingowego) mogą zostać wykorzystane do rozsyłki masowych wiadomości.
5. Naiwny Klasyfikator Bayesa – metoda uczenia trenowana na zbiorze wiadomości uprzednio oznaczanych przez użytkowników jako spam.

IV.2 Problem poprawności cech

Przedstawione w Rozdziale III cechy wiadomości email, mogące świadczyć o jej phishingowym charakterze, mogą okazać się niewystarczające do poprawnej klasyfikacji danej wiadomości jako phishing. Atak phishingowy korzysta z całego szerokiego spectrum technik inżynierii społecznej (manipulacja, oszustwo) oraz może

wykorzystywać przejętą w nielegalny sposób istniejącą infrastrukturę teleinformatyczną, której reputacja nie wskazuje na jej złośliwy charakter. W ten sposób początkowa, otrzymana wiadomość może zawierać minimalną ilość cech wskazujących na możliwych ataku (lub nie zawierać ich w ogóle – patrz „Atak responsywny"). Należy wówczas w procesie klasyfikacji uwzględnić poniższe kryteria:

1. Wiadomość może zostać wysłana z przejętego, rzeczywistego konta email (np. Google, WP, Onet) – często spotykane jest otrzymanie wiadomości z konta założonego u wiodącego na rynku operatora (np. Google dostawca usługi Gmail) i przesłanie wiadomości z odpowiednio spreparowanym odnośnikiem. Zaimplementowane w danej organizacji (odbiory wiadomości) mechanizmy weryfikujące nadawcę (SPF, DKIM), dokonają sprawdzenia organizacji nadawcy (jako dostawcy usług poczty elektronicznej) czy jest uprawniony do wysyłania wiadomości z tej domeny pocztowej do której należy właśnie adres nadawcy. Wykorzystanie oferowanych usług u publicznych operatorów spowoduje poprawne zweryfikowanie wiadomości przez mechanizmy SPF, DKIM (rekord z wiadomości pokrywał będzie się z rekordem TXT zapisanym na serwerze DNS) i wiadomość zostanie więc finalnie dostarczona do odbiorcy.
2. Wiadomość może zostać wysłana z doraźnie na ten cel założonego konta email u operatora świadczącego darmowe usługi poczty elektronicznej (np. Google, WP, Onet). Wiadomość taka zostanie również prawidłowo zweryfikowana zgodnie z opisanym w poprzednim punkcie mechanizmem.
3. Wiadomość może nie zawierać żadnych załączników – mechanizm antywirusowy oznaczy taka wiadomość jako bezpieczną i finalnie zostanie dostarczona do odbiorcy końcowego.
4. Wiadomość może zawierać spreparowany dokument HTML jako załącznik. Przygotowany dokument jest w zasadzie gotową stroną pozwalającą na wyłudzenie danych (formularz do podania danych, np. danych logowania czy danych osobowych). Zaimplementowany w dokumencie kod JavaScript nawiąże połączenie z serwerem C2 i prześle dane z formularza.
5. Wiadomość może zawierać spreparowany dokument HTML jako załącznik. W kodzie źródłowym dokumentu HTML osadzony został kod JavaScript, który dokonuje dekowanie danych (również osadzonych w tym samym kodzie źródłowym dokumentu HTML) do postaci pliku zawierającego oprogramowanie

złośliwe. Mechanizm antywirusowy skanujący załączniki danej wiadomości (po jej odebraniu przez serwer pocztowy) nie wykryje zagrożenia, gdyż dokument HTML traktowany jest tako plik tekstowy.

6. Wiadomość może zawierać kontenery dysków wirtualnych (ISO, IMG). Kontenery dysków wirtualnych mogą zawierać w sobie złośliwe oprogramowanie – wykrywalne dopiero po jego uruchomieniu na komputerze ofiary – jednakże mechanizm skanujący wiadomość po jej otrzymaniu przez serwer pocztowy, nie przetwarza zawartości kontenerów.
7. Wiadomość może zawierać tylko jeden odnośnik.
8. Odnośnik do domeny phishingowej może znajdować się wewnątrz załączonego dokumentu (np. doc, docx czy pdf). Dołączony dokument nie zawiera złośliwego kodu (makra), nie uruchomi się skanowania dokumentu przez system antywirusowy.
9. Domena, do której kieruje odnośnik URL znajdujący się wewnątrz otrzymanej wiadomości, może nie znajdować się na listach domen uznawanych za phishingowe. Wbudowane reguły antyphisingowe (lub bazy adresów domen phishingowych) w urządzeniach bezpieczeństwa sieciowego (firewall), nie rozpoznają zagrożenia i dopuszczą użytkowników do komunikacji z daną witryną.
10. Domena, do której kieruje odnośnik URL znajdujący się wewnątrz otrzymanej wiadomości, może posiadać długą i dobrą reputację, a w wyniku błędu konfiguracji, może być przejęta przez atakujących i wykorzystana do uwiarygodniania ataku.
11. Odnośnik znajdujący się w wiadomości może kierować do uznawanych za wiarygodne zasoby sieciowe i portale, z których potencjalna ofiara kierowana jest dopiero do właściwej strony phishingowej (np. przekierowanie na zasoby sieciowe Google Drive, wyświetlenie dokumentu z odnośnikiem i przekierowanie pod złośliwy adres).
12. Odnośnik z wiadomości email może posiadać ważny, wystawiony dla danej domeny certyfikat.
13. Wiadomość może nie zawierać żadnych odnośników. Wykorzystanie metod inżynierii społecznej ma za zadanie skutecznie przekonać potencjalną ofiarę, do szybkiej odpowiedzi, kontaktu z odpowiednio przygotowaną infrastrukturą (witryna internetowa, adres email, kanał w komunikatorze).

14. Wiadomość może być sformatowana poprawnie językowo lub procent popełnianych błędów nie przekracza wartości świadczącej o możliwym ataku (patrz: „Błędy językowe”).
15. Nazwa użytkownika, widoczna dla odbiorcy jako nadawca wiadomości (pole ”From” / „Od”) może być tożsama z nazwą użytkownika w adresie email (brak spoofingu nadawcy).
16. Nazwa użytkownika widoczna dla odbiory jako nadawca wiadomości (pole „From” / „Od”) może być tożsama z nazwą znajdującą się w podpisie/stopce wiadomości.

Spełnienie powyższych warunków powinno gwarantować, że otrzymana wiadomość email nie stanowi niebezpieczeństwa dla odbiorcy, jednakże sposób wykorzystania dostępnych w polach nagłówka wiadomości czy sugestywna treść doprowadzą finalnie do niebezpiecznego zdarzenia z punktu widzenia odbiorcy wiadomości. Z tego powodu konieczna jest analiza większej ilości pól nagłówka, których wartości w połączeniu z pozostałymi wskazują na możliwy atak phishingowy. Analiza ta wraz z automatycznym przetwarzaniem treści wiadomości, pozowali na skuteczniejsze niż dotychczas wykrywanie phishingu.

Kolejnym zagadnieniem jest problematyka ilości cech występujących w wiadomości email. Mniejsza ilość wykrytych cech w wiadomości nie przesądza, ani nie stanowi, że wiadomość nie jest phishingiem w stosunku do wiadomości z większą ilością cech. Problematykę tą można ująć jako:

1. Phishingiem może być wiadomość z minimalną ilością cech (f=1).
2. Phishing może wykorzystywać nowatorskie metody, dla których nie rozpoznano cech i nie opracowano jeszcze żadnych mechanizmów detekcji.

IV.3 Wybór metody wykrywania wiadomości phishingowych na podstawie zidentyfikowanego wektora cech

Do kategorii technicznych wskaźników, zaliczyć można te cechy wiadomości, dla których można utworzyć regułę jej wykrywania w sposób zautomatyzowany, niezależny od człowieka. Reguły będące zdefiniowanymi warunkami logicznymi, dokonują analizy wiadomości w poszukiwaniu określonych pól i wartości. Wśród wskaźników możliwych

do automatycznej identyfikacji, a wskazujących na możliwy atak phishingowy zaliczyć można:

1. odnośnik prowadzący do strony zidentyfikowanej jako phishingowa,
2. nieprawidłowy odnośnik URL zawarty w wiadomości,
3. adres IP w odnośniku URL zawartym w wiadomości,
4. wykorzystanie serwisów skracających odnośniku URL,
5. długość odnośnika URL zawartego w wiadomości,
6. długość domeny utworzonej na podstawie odnośnika URL znajdującego się w wiadomości,
7. groźba w przypadku niepodjęcia przez ofiarę sugerowanych działań,
8. nieprawidłowy adres email nadawcy,
9. niewłaściwy adres nadawcy,
10. wykorzystanie nazwy odbiorcy wiadomości,
11. wykorzystanie nazwy domenowej jako nazwy użytkownika w adresie nadawcy,
12. mechanizm śledzący w wiadomości email,
13. złożona nazwa domenowa wraz z subdomenami,
14. wiek zarejestrowanej domeny,
15. błędy językowe,
16. temat otrzymanej wiadomości,
17. użycie narzędzi programowych do wysyłki wiadomości email,
18. wykorzystanie tagowania wiadomości przez serwery pocztowe,
19. nazwa użytkownika w treści wiadomości.

Zaprezentowana powyżej lista wskaźników jest ostateczną listą, dla której opracowano zestaw funkcji algorytmu i przeprowadzono badanie.

Kategoria wskaźników nietechnicznych obejmuje grupę tych cech, których wykrycie w pewnych szczególnych warunkach jest możliwe[169] bez ingerencji człowieka – eksperta wyposażonego w niezbędną wiedzę, lecz z uwagi na ich charakterystykę wymagają szerokiej i często aktualizowanej bazy informacji, nie dając jednocześnie jednoznacznego wyniku klasyfikacji. Do tej kategorii wskaźników zaliczać się będą:

1. niespójność nazwy nadawcy,

[169] Wskaźnikiem takim może być badanie podobieństwa wizualnego stron www. Zasadniczą wadą tego rozwiązania, mocno ograniczającą jego powszechne stosowanie jest konieczność porównywania kaskadowych arkuszy styli (CSS) witryn. Metoda nie analizuje obrazów (które często wykorzystują atakujący) jak również nie uwzględnia innych aspektów podobieństwa wizualnego witryn.

2. strona wyłudzająca dane – różnego rodzaju formularze lub bramki płatności (w tym fałszywe).
3. typosquatting / domen udające istniejące – oznaczanie domen jako phishingowe wykonywane jest przez szeroką, międzynarodową społeczność (np. PhishTank), bazując na własnych analizach i wiedzy. Z tego powodu, wskaźnik (na podstawie wartości adresu domenowego) ten można połączyć ze wskaźnikiem „Odnośnik prowadzący do strony zidentyfikowanej jako phishingowa",
4. spoofing instytucji – możliwa sytuacja podobieństwa nazw, podobieństw adresów domenowych dla dwóch różnych podmiotów, które to jednocześnie prowadzą legalną działalność.
5. brak zarejestrowanej domeny, wykorzystanie adresów chmurowych – weryfikacja techniczna może być niejednoznaczna:
    a. odnośnik znajdujący się w danej wiadomości może zawierać adres domenowy już nieistniejący (otrzymana odpowiedź NXDOMAIN od serwera DNS), ale w przeszłości była to legalnie funkcjonująca witryna bez osadzonej złośliwej zawartości,
    b. mechanizm próby nawiązania połączenia z nieistniejącą domeną wykorzystują niektóre rodzaje złośliwego oprogramowania,
    c. odnośnik prowadził będzie do zasobu sieciowego oferowanego przez operatora udostępniającego nieodpłatnie zasoby sieciowe (np. Google Drive) a zawartość, do której prowadzi odnośnik nie jest złośliwa,
    d. wyrafinowane ataki zawierają odnośnik do dokumentu umieszczonego na zasobie sieciowym (który to dokument sam w sobie nie zawiera złośliwej zawartości), a zawarty w nim kolejny odnośnik uruchamia pętle przekierowań, finalnie prowadząc do złośliwej witryny.
6. niespójna szata graficzna – wiele firm może przechodzić procesy rebrandingu[170] lub dostosowywać logo i szatę graficzną do aktualnie panującego trendu, zgodnie z panującą modą lub by dotrzeć do innego, specyficznego grona odbiorców. Działanie takie podyktowane jest marketingiem. W trakcie tego procesu użytkownik, może otrzymywać wiadomości, której szata może być pozbawiona niektórych elementów graficznych wcześniej obecnych, co może prowadzić do

[170] Rebranding - proces transformacji wszystkich elementów marki, takich jak oferowane produkty i usługi, jakość obsługi oraz sposób komunikacji, w tym wygląd logo, by osiągnąć lepszą pozycję marki na rynku

podejrzeń o nieautentyczność danej wiadomości. Podejście do bazowania na wizualnym podobieństwie witryn jako podstawa do wykrywania phishingu zaprezentowali A.P.E. Rosiello, E. Kirda, C. Kruegel, F. Ferrandi w opracowaniu [85], jednakże generowało ono fałszywe alarmy (*false positive*), co sami autorzy opisali w materiale konferencyjnym [86].

7. nietypowe prośby, niespodziewana treść – zarówno w wiadomościach stanowiących normalną korespondencję, jak w wiadomościach typu spam i phishing.
8. niespodziewane załączniki – mechanizm weryfikacji załączników został opisany powyżej. Z uwagi na obszerność narzędzi wspomagających w identyfikacji charakteru załącznika, działających niezależnie wskaźnik ten nie będzie rozpatrywany w przedstawianej metodzie.

Wskaźniki, te są identyfikowalne przez eksperta, zaznajomionego z aktualnie występującymi trendami w ataku phishingowym, lecz z powodu opisanych powyżej sytuacji, automatyczne ich wykrywanie i analizowanie, może prowadzić do pojawienia się błędnej klasyfikacji, dlatego w przedmiotowej rozprawie, brane będą pod uwagę jedynie te wskaźniki, których identyfikacja odbywa się automatycznie i niezależnie od przejętej przez atakujących techniki. Nie wymagana jest też do ich identyfikacji obszerne zasoby danych, na podstawie których prowadzana będzie identyfikacja.

Z uwagi na konieczność precyzyjnego dobierania cech, by identyfikowały wiadomości phishingowe, przyjęto następujący podział wiadomości email:

1. wiadomość phishingowa – wiadomość zawierająca dowolną złośliwą treść,
2. wiadomość typu spam – niechciana poczta (w tym agresywny marketing),
3. wiadomość poprawna – normalna korespondencja pomiędzy użytkownikami (firmami, instytucjami).

IV.3.1 Pomijane wskaźniki

Przygotowując opis istniejących wskaźników mogących świadczyć o ataku phishingowym, występujących w wiadomości email, brano pod uwagę, te, które wystąpiły w rzeczywistych kampaniach phishingowych. Wymieniane wskaźniki zaimplementowane masowo w oprogramowaniu antywirusowych i urządzeniach bezpieczeństwa w niniejszej pracy, są pomijane, ze względu na dojrzałość rozwiązania (oprogramowanie antywirusowe) czy też sporadyczne występowanie lub szybko następujące zmiany technologiczne. Do takich cech, zaliczyć można:

1. złośliwy załącznik, skrypty wykonywalne (VBS/VBA),
2. kod JavaScript osadzony w wiadomość email (formatowanej jako HTML) lub w dokumencie HTML dołączonym do wiadomości email jako załącznik,
3. inne załączniki (np. pliki ISO/IMG, *.rar, *.zip.).

Tabela 38. Opis pomijalnych wskaźników

Lp.	Wskaźnik główny	Wskaźnik drugorzędny	Uwagi
1.	Złośliwy załącznik		Wysoka skuteczność istniejących rozwiązań.
2.	Niespodziewany załącznik	Kod JavaScript osadzony wewnątrz dokumentu HTML jako załącznik wiadomości.	Osadzony wewnątrz dokumentu HTML kod JavaScript może zawierać złośliwy kod.
3.	Inne załączniki	Możliwość dołączenia złośliwego pliku do wnętrza kontenera zawierającego inne, niezłośliwe pliki	Brak możliwości skanowania zawartości.

Powyższe wskaźniki silnie wskazują na możliwość ataku i są nośnikiem cech phishingowych. W przedstawionej w niniejszej rozprawie autorskiej metodzie detekcji, są one pomijalne z uwagi na obecność na rynku dojrzałych i skutecznych rozwiązań bazujących w głównej mierze na tych właśnie wskaźnikach. W przedstawionym rozwiązaniu wskaźniki te są pomijane również z powodu przeprowadzonej analizy i przedstawienia inny cech, które obecnie nie występują w literaturze a z uwagi na swoje własności (samodzielnie lub w korelacji z innymi) również mogą z powodzeniem wskazywać na atak phishingowy. Włączenie wyników[171] działania komercyjnych rozwiązań bazujących na tych wskaźnikach do zakodowanego wektora cech, zwiększy skuteczność detekcji.

IV.3.1.1 Złośliwy załącznik

Rozwiązanie antywirusowe, skanujące złączniki wiadomości email, są zwykle implementowane na wejściu serwera pocztowego, a więc przed działaniem opisywanych w niniejszej rozprawie metod - wydobycie cech phishingu następuje na wiadomościach skanowanych przez urządzenie bezpieczeństwa (a więc złośliwa zawartość została już

[171] Otrzymane wyniki należy wówczas zakodować w sposób analogiczny jak przedstawione w niniejszej rozprawie kodowanie pozostałych cech.

zablokowana) – tego powodu wskaźnik ten jest pomijany i wykluczony z przygotowanego wektora cech.

IV.3.1.2 Osadzony kod JavaScript

Wskazana powyżej metoda, jest połączeniem wykorzystania inżynierii społecznej, istniejących funkcjonalności systemu operacyjnego oraz pewnego braku wiedzy i świadomości przez administratorów i użytkowników odnośnie stosowanych technik ataku. Z uwagi na wykorzystywanie istniejących funkcjonalności i mechanizmów systemów operacyjnych (w tym również stosowania rozwiązania antywirusowego na plikach zawartych w przesłanym kontenerze ISO, IMG), obecność kodu JavaScript nie będzie podlegała sprawdzeniu, a oferowane możliwości przez oprogramowanie antywirusowe dokona finalnego sprawdzenia działania kodu (już po jego uruchomieniu) i na tym etapie powstrzyma wykonanie się złośliwej funkcji.

IV.3.1.3 Inne załączniki

Skanowanie i analiza załączników – podobnie jak w przypadku rozwiązania antywirusowego – odbywa się na wejściu systemu pocztowego. Filtracja niepożądanych (niedozwolonych) typów pików, następuje również przez dostarczeniem wiadomości do skrzynki odbiorczej użytkownika.

IV.3.2 Odczyt danych nagłówka wiadomości email.

Przedstawione wskaźniki phishingu mogą występować pojedynczo lub być wykrywalne w większej ilości. Widoczna dla użytkownika treść wiadomości, może być nośnikiem niektórych z nich, zwłaszcza w formatowaniu HTML, gdzie program pocztowy (klient) użytkownika dokonuje renderowania kodu HTML do postaci czytelnej dla człowieka. Treść wiadomości może np. zawierać cechy:

1) groźba szantaż,
2) mechanizm śledzący,
3) błędy językowe,
4) widoczne nazwy nadawcy wiadomości (automat generujący nazwy).

Większe możliwości przenoszenia cech phishingowych oferuje, niewidoczny dla użytkownika nagłówek wiadomości email. Nagłówek wiadomości email zawiera

szczegółowe informacje nt. przesłanej wiadomości, np.: adresy rzeczywistego nadawcy, adresy serwerów pośredniczących w przekazywaniu wiadomości, sposoby kodowania, strefy czasowe serwerów, np. Dane w nagłówku zapisane są w odwrotnej chronologii (najnowsze są na górze). Dokonując analizy nagłówka wiadomości email i ekstrahując wartości określonych pól, można na ich podstawie ocenić, czy dana wiadomość spełnia kryteria phishingu czy też można uznać daną wiadomość za poprawną (nie złośliwą).

Projektując algorytm przyjęto założenie, że nie jest znana konstrukcja nagłówka wiadomości email, nie znana jest treść wiadomości, układ czy sposób kodowania. Nie rozpoznano również charakteru zawartych (lub nie) w wiadomości odnośników URL. Na wejście algorytmu podawany więc jest strumień wiadomości email o nieznanych wartościach zdefiniowanych w dokumentach standaryzacyjnych polach.

Zadaniem algorytmu będzie więc przetworzenie nagłówka i treści wiadomości email w poszukiwaniu opisanych w poprzednim rozdziale wskaźników phishingu. Przetwarzając określone pola wiadomości, na podstawie wartości tych pól i stworzonych reguł kodowania cech, algorytm przypisze danej cesze wartość 0 (zero), gdy nie wskazuje ona na możliwość ataku phishingowego, lub wartość 1 (jeden), gdy na takowy atak wskazuje.

```
Return-Path: <sekretariat@u-kp.pl>
Delivered-To: [redacted]
X-WP-SR: M3/D7OP6wU8sGiY2/AeqX/1Goi//rVbGdvRezk+mqlz0ug==
Received: (wp-smtpd mx.wp.pl 23124 invoked from network); 19 Dec 2019 01:22:31 +0100
Received: from cloudserver046320.home.pl ([89.161.230.92])
          (envelope-sender <sekretariat@u-kp.pl>)
          by mx.wp.pl (WP-SMTPD) with ECDHE-RSA-AES256-GCM-SHA384 encrypted SMTP
          for [redacted]; 19 Dec 2019 01:22:31 +0100
Return-Path: <sekretariat@u-kp.pl>
Received: from rdns7.kaminedar.info (37.228.132.249) (HELO 194.36.189.226)
 by uzdrowisko-kamienpomorski1.home.pl (89.161.230.92) with SMTP (IdeaSmtpServer 0.83.320)
 id b755c2a048355582; Thu, 19 Dec 2019 01:19:34 +0100
From: "Robert Karlicki" <sekretariat@u-kp.pl>
Subject: fv 21/12/1813
To: [redacted]
Content-Type: multipart/alternative; boundary="pqyY4VdOgO94fvb6P3agKOf=_noRKb431K"
MIME-Version: 1.0
Reply-To: "=?utf-8?B?T3NpxYRza2kgTWlyb3PFgmF3?=" <sekretariat2@u-kp.pl>
Organization: Profil Sp.j.
Date: Wed, 18 Dec 2019 16:19:34 -0800
X-WP-DKIM-Status: no signature (id: n/a)
X-WP-MailID: e9ba0bcb19f92156b4f8f2e878788dcb
X-WP-AV: skaner antywirusowy Poczty Wirtualnej Polski
X-WP-SPAM: NO (U9) 0M00010 [IQNH]
Message-ID: <e9ba0bcb19f92156b4f8f2e878788dcb@wp.pl>

This is a multi-part message in MIME format

--pqyY4VdOgO94fvb6P3agKOf=_noRKb431K
Content-Type: text/plain; charset="utf-8"
Content-Transfer-Encoding: quoted-printable
Content-Disposition: inline
```

Rysunek 42. Nagłówek wiadomości phishingowej. Na uwagę zasługują różnice w wartości pól „Return-path" oraz „Reply-To", różne nazwy użytkowników oraz różne nazwy organizacji. Dane odbiorcy zostały zanonimizowane.

Z uwagi na założenia standardu wiadomości email (RFC 5322[172]), zakładającego pewną dowolność w konfiguracji serwerów poczty elektronicznej, występowanie pewnych wartości pół jest opcjonalne, ilość pól serwerów pośredniczących może być różna (w zależności od długości drogim jaką dana wiadomość email przeszła od serwera nadawcy do serwera odbiorcy). Adresy serwerów pośredniczących mogą również być zapisane w dowolnym formacie: jako adres domenowy, jako adres IP, lub zawierać obie te wartości jednocześnie.

Problematycznym zagadnieniem jest również określenie daty wysłania / przekazania danej wiadomości – data wysłania / przekazania znajduje się (zgodnie ze standardem RFC 5322) w tej samej linii co adres (adresy) serwera naddającego / pośredniczącego. Również nie wszystkie serwery dodają do wartości linii z adresem wartość daty wysłania / przekazania wiadomości.

[172] https://datatracker.ietf.org/doc/html/rfc5322

Z uwagi na możliwość instalacji serwera pocztowego pracującego pod różnymi systemami operacyjny, z różnymi ustawieniami językowymi – a więc i z różnymi sposobami kodowania, treść otrzymywanej przez odbiorcę wiadomości email może być kodowana za pomocą różnych metod, z wykorzystaniem różnych standardów.

Występujące problemy podczas odczytu i przetwarzania nagłówka wiadomości email zostały przedstawione w poniższej tabeli:

Tabela 39. Wykaz istotnych pól nagłówka wiadomości email.

Lp.	Wartość	Występowanie	Problem
1.	Pole „From” („Od”)	Zawsze, pojedyncze	Nazwa nadawcy , brak adresu email
2.	Nazwa nadawcy	Opcjonalne, nazwa nadawcy może występować w wielu różnych polach	Brak nazwy nadawcy
			Nazwa nadawcy zakodowana Base64
3.	Pole „Return-Path”	Opcjonalnie wielokrotne	Wielokrotne występowanie
		Opcjonalne, pojedyncze	Występująca nazwa odbiorcy przed adresem email
			Nazwa nadawcy zakodowana Base64
4.	Pole „Reply-To”	Opcjonalne, wielokrotne	W zależności od konfiguracji serwerów pocztowych, wartość w polu może nie wystąpić w wiadomości.
5.	Pole „Received”	wielokrotne	Wielokrotne występowanie zależne do ilości serwerów pośredniczących
			Może składać się z jednej lub więcej linii
			Może zawierać adres w postaci adresu domenowego, adresu IP lub zawierać oba te adresy jednocześnie
		opcjonalne	Nazwa pola może być poprzedzona pewną ilością biały znaków, frazą „X-„
6.	Data wysłania	Wielokrotne, opcjonalne	Data wysłania / przekazania wiadomości może znajdować się w tej samej linii co pole „Received” lub nie występować w niej
		opcjonalne	Data można zawierać nazwę strefy czasowej
		opcjonalne	Data może zawierać wartość różnicy strefy czasowej do czasu serwera docelowego odbiorcy, zawierać wartość różnicy strefy czasowej wraz ze międzynarodowym kodem tej strefy lub nie zawierać żądnej z tych wartości
		pojedyncze	Nazwa miesiąca zakodowana jest skrótem nazwy angielskiej (np. styczeń = january -> jan)

7.	Pole „Subject”	pojedyncze	Brak tytułu wiadomości
8.	multipart	Opcjonalne, wielokrotne	Wiadomość zawiera kilka różnych sekcji, do których należą: - tekst niesformatowany - tekst sformatowany (HTML) - załączniki
9.	kodowanie		Nazwa użytkownika widoczna w polu „Od” (wraz z nazwą email), może: - być zakodowana za pomocą Base64 - używać kodowania tekstu: quoted printable - używać kodowania tekstu utf-8 - nie być kodowana Treść wiadomości może być: - w tekstowej formie niezakodowanej - zakodowana za pomocą Base64
10.	Pole „X-Mailer”	Opcjonalne, pojedyncze	Informacja o wykorzystaniu narzędzi programowych (wbudowanych w język programowania), użytych do wysyłki danej wiadomości email.
11.	Inne występujące pola o niestandardowych nazwach lub/i wartościach	Opcjonalne, wielokrotne, pojedyncze	Pola tworzone i konfigurowane przez danego administratora systemu pocztowego, o nieustalonej nazwie, sposobie zapisu.
12.	Zapis treści wiadomości	Opcjonalne, wielokrotne	Różne sposoby kodowania znaków, różne sposoby formatowania treści, występowanie różnych części wiadomości.

Biorąc pod uwagę wymienione powyżej a mogące wystąpić problemy przy odczytywaniu i przetwarzaniu danej wiadomości email, algorytmy odczytujące cechy wskazujące na możliwy atak phishingowy, powinny spełniać następujące założenia:

1. Opisane są techniczne wskaźniki phishingu.
2. Algorytm odczytuje kolejne linie pliku wiadomości email (.eml), aż do odczytania ostatniej linii (koniec pliku).
3. Zadaniem algorytmu jest wyszukanie w nagłówku wiadomości email (odczytanej linii) ciągów znaków o określonej wartości:

   X = <„Reply-To”, „Return-Path”, „Received”, ”From”, …>
4. Ciągami tymi są nazwy pól z przypisanymi im wartościami.
5. W trakcie odczytu wiadomości, wyszukiwana i przetwarzana (w sposób automatyczny) jest również treść wiadomości email. Treść wiadomości nie jest ujawniana, wyszukiwane są w niej schematy świadczące

o możliwym phishingu (np. wykrycie adresu portfela BitCoin, bez przypisania właściciela, poznania salda, itp.).

6. Treść wiadomości poddana jest również weryfikacji poprawności językowej (wyszukiwanie błędów ortograficznych w treści).
7. W przypadku wykrycia w danej linii poszukiwanego ciągu, uruchamiany jest zestaw funkcji przetwarzających, odpowiedniej dla wykrytego ciągu wartości.
8. Niektóre pola występują zawsze, wywołana więc zostanie funkcja przetwarzająca wartości dla danego pola (co najmniej raz).
9. Niektóre pola są opcjonalne, mogą wystąpić raz, mogą wystąpić dwukrotnie lub nie muszą występować wcale. Wywołanie funkcji przetwarzającej wartości dla tego rodzaju pól jest różne.
10. Niektóre pola lub części wiadomości mogą mieć różne kodowanie – algorytm odczytuje zapisany sposób kodowania danego pola (o ile taka informacja występuje – nie we wszystkich wiadomościach email będzie zapisany standard kodowania). W przypadku braku określenia sposobu kodowania, algorytm stosował będzie domyślne kodowanie dla regionu Europy Środkowej (UTF-8).
11. Wynikiem funkcji przetwarzającej wartości danego pola są tworzone listy wartości. Listy mogą zawierać od jednego do n elementów.
12. Algorytm kończy odczyt pliku, gdy nie ma już więcej linii do odczytania, uruchamiany jest wówczas odpowiedni zestaw funkcji.
13. Odczytane listy wartości są przetwarzane na podstawie zbudowanych reguł, w celu utworzenia zakodowanego wektora cech:

$$F = [x_1, x_2, \dots, x_i] \text{ dla i =1,2,…,n} \qquad (4.1)$$

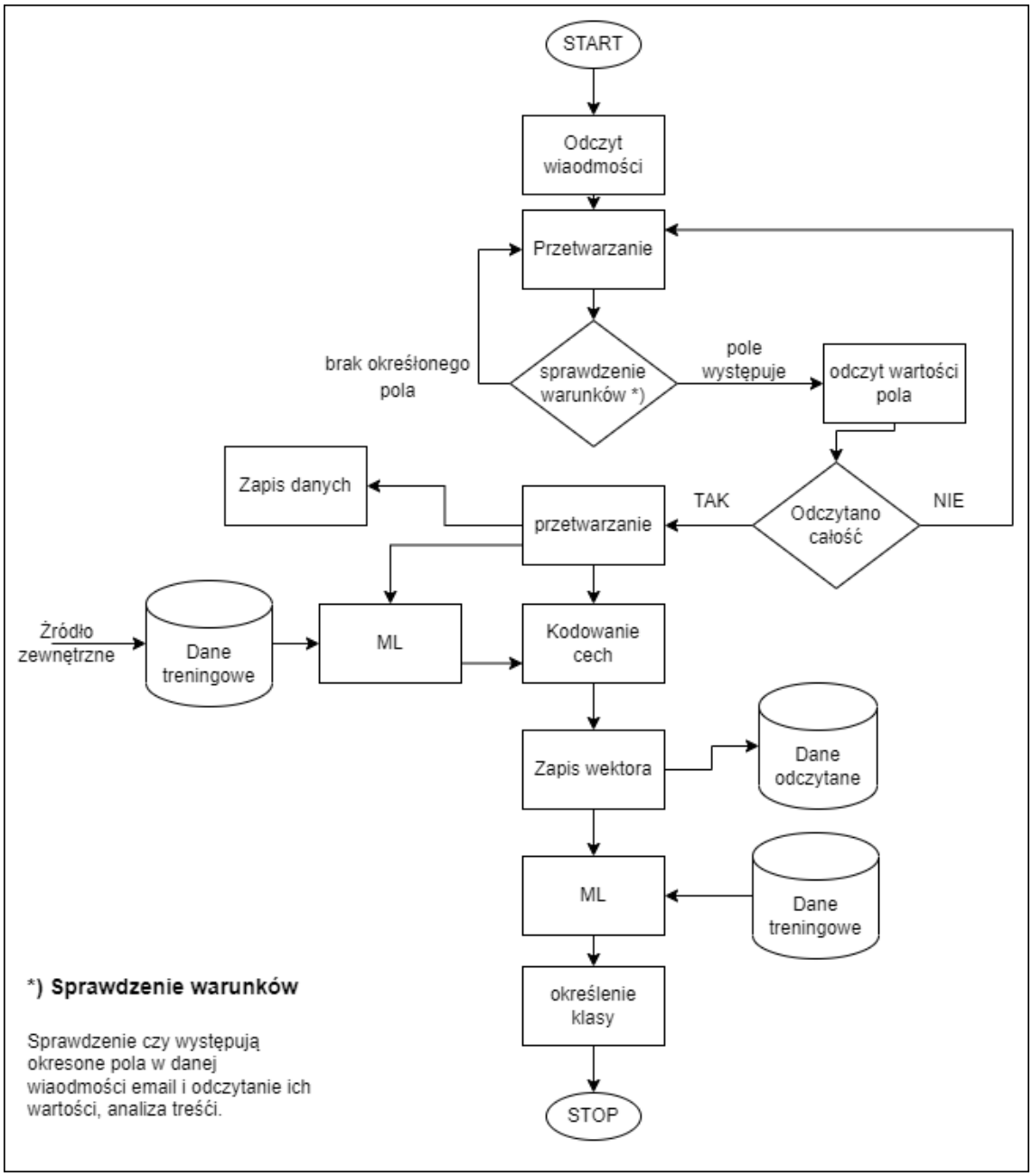
gdzie:
xi –i-ta cecha
n – liczba cech

Liczba cech (n), będąca wynikiem wewnętrznych reguł przetwarzania algorytmu zostanie ustalona na podstawie analizy opisanych w Rozdziale III – Techniczne i nietechniczne wskaźniki phishingu. Ustalona liczba cech będzie stała, niezależna od konstrukcji wiadomości email czy

występowania w niej określonych pól. Ustalając wartość liczby cech (n), wzięto pod uwagę czynniki:

a. wykrywalność wartości danej cechy w sposób automatyczny (bez ingerencji człowieka),
b. dostępność danych uczących dla określonego, właściwego wskaźnika (jeżeli zakodowana wartość cechy jest wynikiem uczenia maszynowego),
c. możliwość przetworzenia wartości danej cechy w sposób automatyczny,
d. dostępność danych (baz) jeżeli wartość cechy ustalana jest w korelacji z innymi danymi.

14. Reguły przetwarzania odczytanych wartości pól nagłówka wiadomości wynikają z opisu technicznych cech phishingu oraz technik jego implementacji.



Rysunek 43. Uproszczony algorytm odczytu wiadomości email.

Projektując algorytm odczytujący dane, należy rozważyć wykonywane operacje pod kątem oszacowania złożoności obliczeniowej. Z uwagi na możliwość przetwarzania dużej ilości danych (jako dane wejście rozumieć należy pełną wiadomość email) i wykonywanie na niej wielu kolejnych operacji, należy dokonać optymalizacji działania algorytmu, z uwagi na ograniczone możliwości czasowe przetwarzania (wiadomość bez zbędnych opóźnień musi finalnie zostać dostarczona do odbiorcy końcowego – pod warunkiem oznaczenia jej jako bezpieczna).

IV.3.2.1 Opis szczegółowy

Danymi wejściowymi są wiadomość email w postaci pliku .eml . Algorytm odczytuje wiadomość email, linia za linią. Przeszukiwana jest zawartości wiadomości w poszukiwaniu atrybutów:

1. Widocznego nadawcy (pole „From" / „Od"), występuje zawsze, wartość pojedyncza.
2. Pola „Delivered-To" określającego do której skrzynki odbiorczej kierowana była dana wiadomość.
3. Pola „To" i „Sender" – pola opcjonalne, występują raz.
4. Tematu wiadomości (pole „Subject"), występuje zawsze, możliwe kodowanie Base64.
5. Ścieżki odpowiedzi (pole „Return-Path"), występuje opcjonalnie.
6. Adresu email służącego do odpowiedzi (pole „Reply-To"), występuje opcjonalnie, maksymalnie jedna wartość.
7. Adresów serwerów pocztowych przekazujących wiadomość (pola „Received"), występuje co najmniej raz.
8. Frazy „for" jako części składowej pola „Received" – dodanie do listy na pozycji ostatniego wystąpienia pola „Received".
9. Frazy „by" jako części składowej pola „Received" – dodanie do listy na pozycji ostatniego wystąpienia pola „Received".
10. Daty wysłania/przekazania wiadomości (pętla po angielskich skrótach nazwy miesiąca) i dodanie do listy na pozycji ostatniego wystąpienia **pola** „Received".
11. Pola „boundary" – informacja o zawartości wielu części w wiadomości, występuje jedynie wówczas, gdy wiadomości zawiera różne części.
12. Pola „Content-Type" oraz „Content-Type-Encoding", świadczących o kodowaniu treści i formatowaniu.
13. Pól świadczących o możliwym wykorzystaniu narzędzi programowych („X-Mailer").
14. Pola „User-Agent".
15. Pól „SPF" i „DKIM" - analiza poprawności i wiarygodności serwera nadawcy, występują opcjonalnie (konieczne skonfigurowanie odpowiednich mechanizmów serwera pocztowego).
16. Przetwarzanie treści wiadomości email w celu wykrycia:

a. szantażu i prób wymuszania okupów,

b. odnośników URL,

c. wykrywania mechanizmu śledzącego,

d. formatowania i wykrycia różnych części.

Odczyt wiadomości email kończy się w momencie, gdy nie ma więcej linii do odczytania.

Algorytm 1: przetwarzania wiadomości email

```
1.  open file
2.  define list lista[], month[] ...
    /*definicje list 179lementów179h179h odczytane wartości pól */
    /*definicje słowników */
3.  read line
    /* odczyt kolejnych linii */
4.    if "NAZWA_POLA" in line
5.      /* wywołanie właściwej funkcji dla określonego pola */
6.      parsefunction(line, NAZWA_POLA)
        /*
7.      if „date" in „Received"
        /* w przypadku pola "Received" uuchomienie wielu funkcji  */
8.        getDate   /* odczyt daty */
9.        getBy     /* odczyt wartości pola „by" */
10.       getFrom   /* odczyt wartości pola „from" */
11.   for i=0 to i=12
12.     if mont[i] in date[]
        /*sprawdzenie występowania nazwy 179lementó w odczytanj
        linii */
13.       read rec_lst[] index
14.       delete „\n"
15.       add to list rec_lst[]
16.   end
17.   if „/n" in line
      /* przetwarzanie wiadomości */
18.     parsemessage()
19. end
20. /*utworzenie wektora wartości cech pierwotnych */
    createValueVector(lst)
21. saveData(lst)
    /*zapis wartości cech pierwotnych */
22. for i=0 to i=21
    /*wektor cech 179lementów179h zawiera 31 wartości służących do
    zakodowania 19 cech */
23.   /*kodowanie wektora cech na podstawie wartości */
      f[i] = parsevalue(data[i])
24. saveFeature(f)
    /*zapis wartości wektora cech */
25. end
```

Dla pola „Received" (zakodowane adresy IP i domenowe serwera wysyłającego, odbierającego, strefa czasowa), konieczne jest wywołanie kilku różnych funkcji. Ilość wywołań różnych funkcji będzie różny dla każdej z wiadomości email oddzielnie. Zależy to od wewnętrznej konstrukcji serwerów (i ich ilości – a więc ilości skoków)

pośredniczących w wymianie wiadomości email. Pole to może być zapisane w wielu liniach, dlatego konieczne jest zbudowanie dodatkowych funkcji przetwarzających odczytaną linię w poszukiwaniu specyficznych wyrażeń – pól „by" i „from", które są integralną częścią składowej wartości pola „Received". Odczytywane kolejne wartości linii z polem „Received" zapisywane są na liście. Pole to jest również istotne z punktu widzenia analizy daty wysłania wiadomości – zawiera datę dla każdego z serwerów przetwarzających wiadomość.

Algorytm 2: Funkcja odczytująca wartość odnalezionego pola

```
1.  /*funkcja odczytu wartości pola – właściwa dla danego typu */
    define parsefunction(line, NAZWA_POLA):
2.      delete „NAZWA_POLA"
        /*usunięcie z odczytanej linii wiadomości nazwy pola */
3.      delete „<"
        /* odnalezienie w pozostałym ciągu znakowym pozycji wystąpienia
        znaku „<" oznaczającego początek wartości pola /*
4.      delete „>"
5.      if „?" == true
        /* Jeżeli znak „?" występuje w linii, wartość zakodowana jest
        za pomocą Base64 /*
6.          explode as q_lst[]
7.          for i=0 to i=count(q_lst)
8.              if q_lst[i] > 5
                /* dekodowanie */
9.                  decode to text
10.                 add to name_lst[]
11.             else
12.                 delete „ „ „
                    /* usunięcie niepotrzebnych danych */
13.                 add to lst[]
14.         end
15.     else
16.         add to lst[]
17.     return lst[]
18. end
```

Wywoływanie funkcji dla poszczególnych pól jest uzależniony od tego czy dane pole występuje w wiadomości (niektóre są opcjonalne) oraz od ilości jego wystąpienia (niektóre pola mogą wystąpić w danej wiadomości wielokrotnie i mają różne wartości).

Algorytm 4: Funkcja kodująca wartości danej cechy

```
1.  /*funkcja kodujaca cechę – właściwa dla danego typu */
    define parsevalue(data):
2.      /*usunięcie z odczytanej linii wiadomości nazwy pola */
3.      if domain == true
        /* konieczność pobrania danych
4.          getIP
5.          getASN /*pobranie numeru AS */
```

```
6.      getWHOIS
7.      compare() /*porównanie uzyskanych wyników */
8.    ml(learningData, data)
      /* uruchomienie modułów ML dla niektórych wartości */
9.    f = {0|1}
10.   return val
11. end
```

Treść wiadomości jest istotna z punktu widzenia zawartości informacji o możliwym mechanizmie śledzących, może zawierać phishingowe odnośniki URL oraz jej format może wskazywać na potencjalny szantaż i wymuszenie okupu. Funkcja przetwarzania treść wiadomości ma również zadanie wykryć, czy wiadomość posiada różne osadzone treści (zgodnie z przyjętym podziałem na typy).

Algorytm 5: Przetwarzanie treści

```
1.  define btc[] = {btc1,….} /*lista wartości prefiksów */
2.  /*funkcja kodujaca cechę – właściwa dla danego typu */
    define parsemessge(data):
3.    /* przetwarzanie treści wiadomości */
4.    if „http” == true
      /* wyszukiwanie odnośników URL w wiadomości */
5.      delete „ „ „
        /* usunięcie niepotrzebnych 181lementów */
6.      add to url_lst[]
7.    if „img” and „width = 1”  == true
      /* mechanizm śledzący */
8.      track = 1
9.    if btc[] in data
10.     */ adres portfela kryptowaluty, uruchomienie modułów uczących
        */
11.     ml(learningData, data)
12.   if „bonduary” == true
13.     /* wiadomość zawiera wiele części */
        findMessageType
14.     extractMessage
15.     getTextEncoding
16.   return val[]
17. end
```

Implementację algorytmu odczytu wartości poszczególnych cech z wiadomości email wykonano z wykorzystaniem języka programowania Python, w wersji 3.9.2[173]. Wykaz dodatkowych modułów niezbędnych do prawidłowego funkcjonowania algorytmu znajduje się w Dodatku B.

[173] Opis modułów i zmian tej wersji języka Python, wydanym przez organizację Python Software Foundation znajduje się pod adresem: https://www.python.org/downloads/release/python-392/

IV.3.3 Przygotowanie wektora cech

Odczytane dane z wiadomości email, zapisane zostają w postaci listy wartości. Dane te, zgodnie z założeniem zostaną następnie zakodowane w postaci wartości <0,1>. Proces kodowania odczytanych wartości wskaźników (opisanych w Rozdziale III) do postaci wektora cech odbywa się zgodnie z opisanymi zasadami – charakterystycznymi dla danego wskaźnika. Implementując rozwiązanie, dla każdego ze wskaźników opracowano jego indywidualną klasę w języku oprogramowania, która na podstawie przekazanych do niej argumentów (wartości wskaźnika), dokonuje ich analizy, przetwarzania oraz zwraca jedną z dwóch wartości:

- 0 – jeżeli, wartość nie wskazuje na phishing,
- 1 – jeżeli wartość może wskazywać na phishing.

Przygotowany i zakodowany w ten sposób wektor cech, poddanie zostanie następnie procesowi uczenia maszynowego.

IV.3.3.1 Funkcje dodatkowe

Przetwarzając poszczególne pola nagłówka wiadomości oraz dokonując automatycznej analizy formatu źródłowego treści wiadomości, można uzyskać dodatkowe informacje, które agregowane i zapisywane w bazie danych, mogą posłużyć do zasilenia baz antyphishingowych (zawierających informacje o używanych we wcześniejszych atakach adresach email, domenach, wykorzystywanych taktykach) lub do opracowania nowych modeli detekcji opartych na uczeniu maszynowym.

Dane zawarte w wiadomości email, mogące zasilić bazy danych antyphisingowe:

1. Adres nadawcy widoczny w polu „Od” („From”),
2. Adresy email odpowiedzi (pola „Reply-To”),
3. Adres email zwrotny (pole „Return-Path”),
4. Adresy IP serwerów pośredniczących w wymianie wiadomości email,
5. Adresy domenowe serwerów pośredniczących w wymianie wiadomości email,
6. Nazwy narzędzi programowych (wartość pola „X-Mailer),
7. Adresy portfeli BitCoin wykorzystywane do otrzymywanie zapłaty w ramach wykonywanego szantażu,

IV.3.4 Moduły uczące

W przypadku kodowania wektora cech, dla niektórych wartości wskaźników, zasadne jest zastosowanie metod uczenia maszynowego do określenia jego wartości (0- nie wskazuje na atak, 1 – możliwy phishing). Cechy te, charakteryzują się pewną zmiennością w czasie i dlatego zastosowanie do określenie ich zakodowanej wartości modułów uczenia maszynowego, pozwoli na uodpornienie się modelu na zmiany, pozwoli z większą precyzją w przewidywaniu klasy oraz pozwoli na właściwą identyfikację nawet w przypadku nieznanego wcześniej modelu ataku. Do takich cech należą:

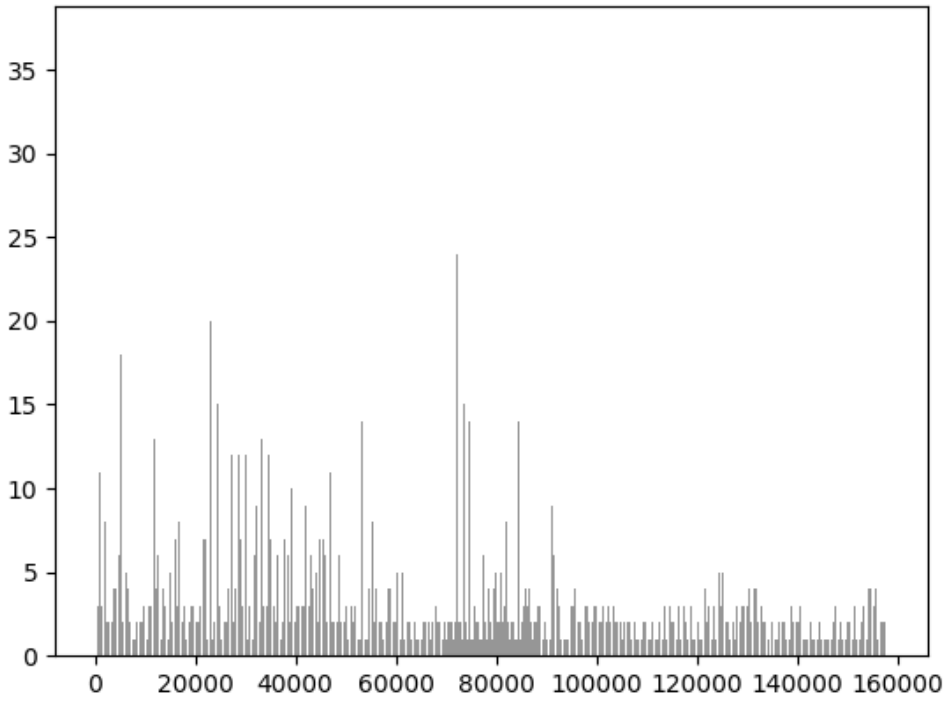
1. charakterystyka odnośnika URL znajdującego się w wiadomości email,
2. charakterystyka adresów domenowych (zarówno w polu nadawcy wiadomości jak i pozostałych polach nagłówka),
3. treść wiadomości (pod kątem wykrycia szantażu i wpłacenia okupu – moduł ten można dodatkowo rozszerzyć pod katem wykrywania innych, charakterystycznych schematów w treści).

Pozyskano bazy odnośników[174] URL zawierającej sumarycznie 551 291 rekordów, które następnie poddano preprocesingowi w celu obliczenia dla każdego wiersza ([69] [70] [71]) :

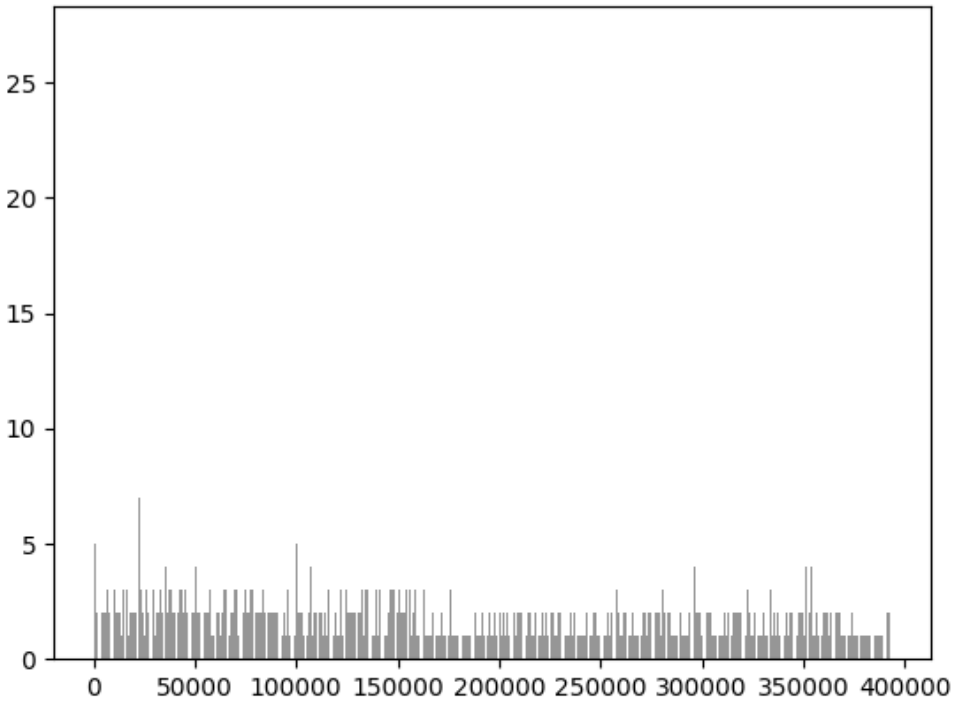
1. ilość kropek występujących w poszczególnych adresach URL,
2. ilość występujących ukośników (slash),
3. długość danego odnośnika / domeny z pominięciem nazwy protokołu („http://” lub „https://”),

Wyliczone w sten sposób parametry odnośnika poddano procesowi uczenia maszynowego z wykorzystaniem Maszyny Wektorów Nośnych (SVM), z uwagi na wysoką skuteczność, brak czułości na możliwe przetrenowanie [87] (pierwotnie pozyskany zbiór danych uczących – odnośników URL – liczył pond 500 000 elementów).

[174] Źródła: https://www.kaggle.com/datasets/taruntiwarihp/phishing-site-urls [dostęp: 01.03.2023], https://openphish.com/phishing_feeds.html [dostęp: 01.03.2023], https://phishstats.info/ [dostęp: 01.03.2023] oraz pozyskiwane w ramach własnych badań.



Rysunek 44. Rozkład ilość kropek w analizowanych domenach phishingowych.



Rysunek 45. Rozkład ilość kropek w analizowanych odnośnikach URL nie będących odnośnikami phishingowymi.

Poniższa tabela przedstawia porównanie wartości średnich dla odnośników oznaczonych jako normalne i phishingowe:

Tabela 40. Wartości średnie odnośników kategorii normal i phishing.

Label	normal	Phishing
Dots average	1.7820777298152524	2.7650499187516204
Slash average	2.3266468429098697	2.7650499187516204
Length average	45.65158687562161	62.1734795171886
Data count	393131	158157

IV.3.4.1 Błędy rozpoznania

Pomimo widocznej różnicy w wartościach średnich, nie można przedstawionego powyżej podejścia uznać za jedyną metodę rozpoznania czy dany URL prowadzi do domeny phishingowej czy też nie. Wśród zgromadzonej bazy danych uczących zawierającej 551290 odnośników URL (zarówno sklasyfikowanych jako phishingowe jak i normalne), liczna domen phishingowych o długości mniej niż 30 znaków wynosiła 50990, co stanowi aż ~32.24%. Dla tych domen, średnia długość wynosiła ~21.94, co jest znacznie poniżej średniej, nawet dla ogółu domen oznaczonych jako „normal" (niebędącymi domenami phishingowymi).

Tabela 41. Statystyka średniej długości domen phishingowej poniżej wartości średniej domen normalnych.

Label	normal	Domeny phishingowe o długości mniejszej niż 30 znaków
Length average	**45.65158687562161**	**21.941262992743674**

Dla tak wygenerowanych domen, mechanizm bazujący na heurystyce[175] odnośnika URL, dysponując odczytanymi wartościami takimi jak: długość odnośnika URL (bez nazwy protokołu – http://, https://), ilością kropek w nazwie domenowej oraz ilością występujących w odnośniku znaków slash („/") okazuje się zawodny i dający błędne wyniki. Dla domen phishingowych stanowiących pozostały zbiór (większa długość domeny, większa ilość kropek lub znaków slash („/"), weryfikacja oparta na heurystyce, okazuje się skuteczna.

Mając na uwadze powyższe wynik uczenia i klasyfikacji odnośników (domen) opary na heurystyce jest obarczony pewnym błędem i może stanowić jedynie tylko jeden ze wskaźników możliwego ataku phishingowego.

[175] Pod pojęciem mechanizmu heurystycznego rozumieć należy mechanizm szacujący parametry odnośnika URL, takie jak: długość odnośnika, długość domeny, liczba występujących znaków „." (kropka), „/" (slash) i na tej podstawie określający przynależność do jednej z klas (np. phishing).

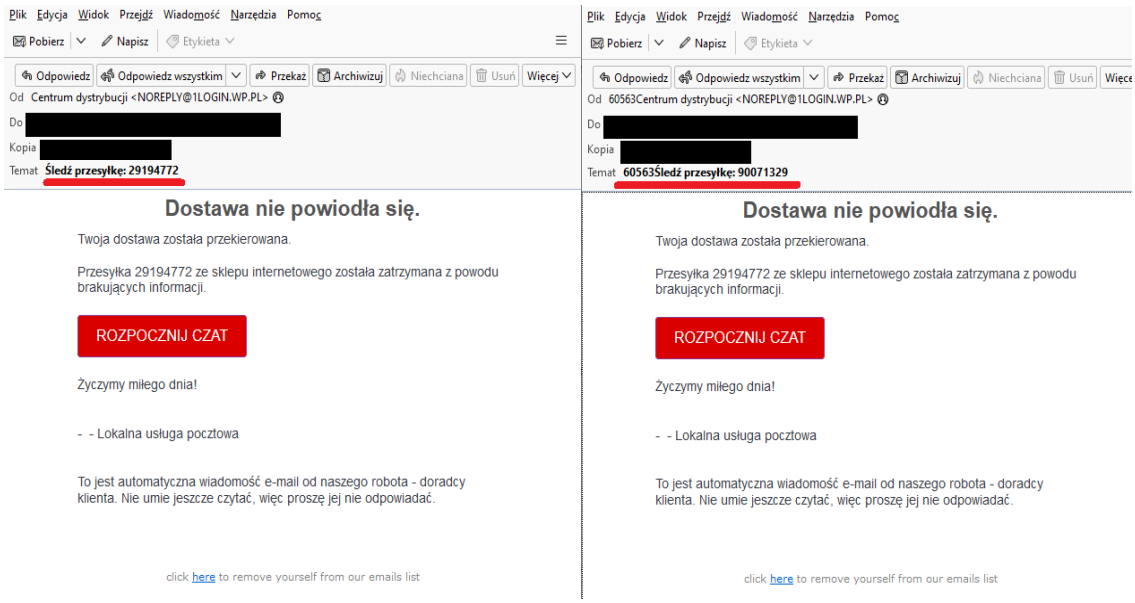
IV.3.4.2 Przygotowanie danych uczących

Opisane cechy mogące wskazywać na możliwy atak phishingowy zapisane są bezpośrednio lub pośrednio w nagłówku lub treści wiadomości email. Do celów testowania mechanizmów detekcji phishingu, pozyskano zbiory danych:

1. 76 wiadomości które nie wykazywały ani cech spamu ani cech phishingu (uznane jako cześć normalnej wymiany korespondencji),
2. 356 wiadomości wstępnie sklasyfikowanych (i oznaczonych) jako spam,
3. 1155 wiadomości, wstępnie rozpoznane i uznane przez zespół CSIRT jako wiadomości o cechach phishingowych.

Tak przygotowany zestaw, poddano przetwarzaniu w celu ekstrakcji występujących w zebranych wiadomościach cechach. Określenia dokładności identyfikacji poszczególnych pól, oraz poprawności kodowania wykrytych cech.

W trakcie przygotowania zbiorów uczących w poszczególnych kategoriach (phishing, spam, normal) dokonano sprawdzania wiadomości pod katem: powtarzalności wiadomości – atakujący, pozbawieni umiejętności technicznych zwykle decydują się ma zakup gotowego rozwiązania do przeprowadzenia ataków (szczegóły opisane w punkcie: „Phishing as a Service"). Identyczna treść wiadomości, odnośniki prowadzące do tego samego zasobu sieciowego, co prawda zwiększają objętość danego zbioru wiadomości, jednakże zawyżają występowanie określonej cechy w zbiorze danych – co jest niepożądane w przypadku konieczności zapewnia różnorodnego zbioru.



Rysunek 46. Przykład wiadomości phishingowych, korzystających z tego samego szablonu. Różnica w wiadomości to data otrzymania oraz modyfikacja tytułu wiadomości (zaznaczona czerwonym podkreśleniem).

Mając na uwadze zapewnienie jak najlepszej jakości zbioru wiadomości, w pierwszym etapie badania usunięto z niego, te wiadomości, które miały identyczną treści, a zawarte w wiadomości email odnośniki, prowadziły użytkownika do tego samego zasobu sieciowego. Dokonano również powiększenia ilości zbioru wiadomości normalnej korespondencji. Po dokonaniu redukcji danych, do kolejnej iteracji badania cech phishingowych uzyskano zbiory:

1. 80 wiadomości normalnej korespondencji,
2. 331 wiadomości typu spam,
3. 967 wiadomości oznaczonych jako phishing.

Źródłem wiadomości email było:

1. wiadomości pochodzące ze skompromitowanego[176] kota email,
2. wiadomości przekazywane przez różnych użytkowników, którzy otrzymywali wiadomości email, co do których posiadali podejrzenia, że mogą być próbą ataku,
3. wiadomości przekazywanych do analizy w ramach prowadzonej w sieci Internet ankiety użytkowników.

Bazując na posiadanej strukturze danych uczących, w przygotowaniu eksperymentu, jako kandydujące metody, wzięto pod uwagę:

1. Maszynę Wektorów Nośnych (SVM),
2. Naiwny klasyfikator Bayesa,
3. Drzewo decyzyjne,
4. Lasy losowe,
5. Regresję logistyczną.

V.2.2.1 Maszyna wektorów nośnych – SVM

SVM (ang. Support Vector Machines – Maszyna Wektorów Nośnych) – klasyfikator wyznaczający hiperpłaszczyznę rozdzielający z maksymalnym marginesem zbiór danych należących do dwóch różnych klas.

[176] Skompromitowane konto email – termin używany przez zespoły cyberbezpieczeństwa w stosunku do konta email, którego dane (np., nazwa użytkownika, hasło dostępu lub hash hasła) znalazły się w wyciekach danych, które zostały opublikowane w sieci Internet, lub konto email co do którego stwierdzono niezautoryzowany dostęp do treści wiadomości. W tym przypadku pod terminem „skompromitowane konto", znajduje się adres email, którego dane (nazwa użytkownika i skrót hasła, znalazł się w wycieku bazy sklepu internetowego morele.pl, wycieku bazy danych użytkowników platformy Adobe oraz w wycieku bazy danych serwisu MyHeritage. Dane o upublicznionych wyciekach dla danego adresu email, uzyskać można za pomocą serwisu „ ';--have i been pwned?" (https://haveibeenpwned.com/).

Posiadając zbiór danych uczących:

$$\{(x_i, y_i)\}_{i=1}^{N} \tag{4.2}$$

gdzie:
xi – wejście danych
yi – reprezentacja klasy dla i-tego wejścia, taka, że $y \in \{-1, 1\}$

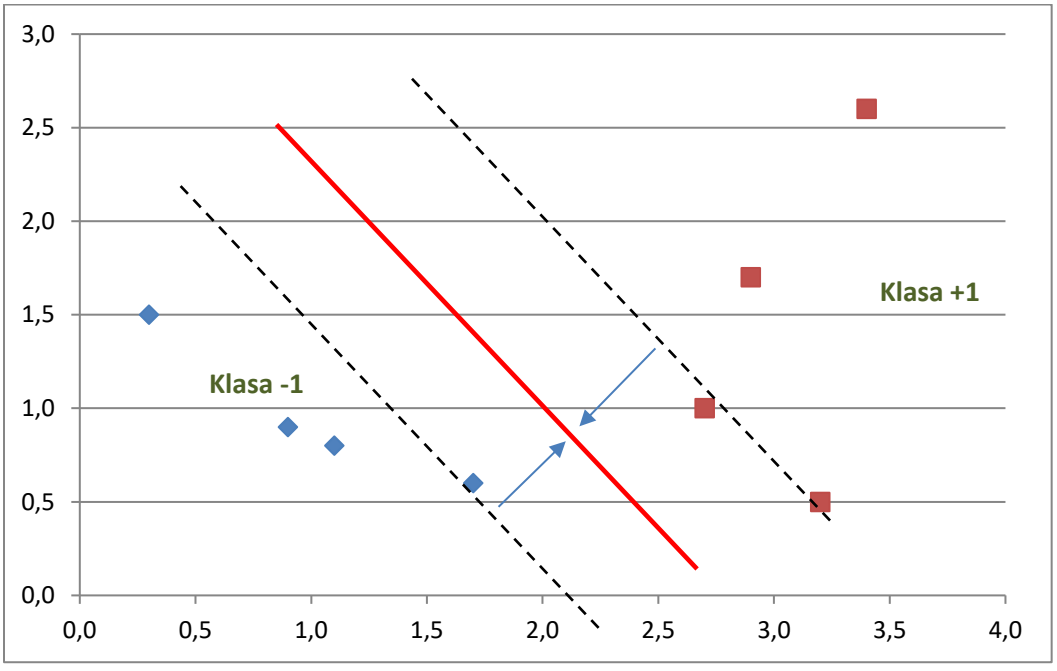
Klasy są separowane linowo, jeżeli istnieje hiperpłaszczyzna H postaci g(x):

$$g(x) = w^T x + b \tag{4.3}$$

gdzie:
w – wektor wag
x – wektor danych wejściowych
b – przesunięcie

Wartości jakie przyjmuje hiperpłaszczyzna rozdzielająca:

$$\begin{cases} g(x_i) > 0 & x_i \in 1 \\ g(x_i) < 0 & x_i \in -1 \end{cases} \tag{4.4}$$



Rysunek 47. Zobrazowanie maszyny SVM.

Zadaniem klasyfikatora SVM jest wyznaczenie możliwie najszerszej granicy dyskryminacji (marginesu klasyfikatora) spośród możliwych, których zwykle istnieje

nieskończona ilość (maksymalizacja marginesu). Hiperpłaszczyznę rozdzielającą dwie klasy można opisać za pomocą zależności:

$$w \cdot x + b = 0 \tag{4.5}$$

Marginesem hiperpłaszczyzny rozdzielającej nazywamy odległość tej hiperpłaszczyzny od najbliższego wektora cech próbki w zbiorze uczącym. By wyznaczyć szerokość marginesu separacji klas należy wyznaczyć wektor wag **w** prostopadły do hiperpłaszczyzny rozdzielającej **y**. Maksymalny margines jest wówczas, gdy element należący do klasy pozytywnej znajduje się na jednym końcu marginesu, element należący do klasy negatywnej znajduje się na drugim końcu marginesu.

Do budowy modelu wykorzystamy bibliotekę ***pandas***[177] oraz ***sklearn***[178] w języku Python. Dane uczące, pozyskane z przetworzenia wiadomości email zapisane zostały jako wartości poszczególnych cech, rozdzielonych przecinkiem (format pliku CSV[179]). Wykorzystując bibliotekę *pandas*, z odczytanych danych utworzony został obiekt z podziałem na zbiór treningowy oraz testowy, w proporcji 80:20 (80% odczytanych danych stanowi zbiór treningowy, 20% danych stanowi zbiór testowy). Zbiory zawierają trzy klasy: phishing, spam i normal – Ta sama technika odczytu i podziału na zbiór treningowy i testowy będzie również wykorzystywana do zbudowania modeli pozostałych klasyfikatorów wykorzystywanych do badania detekcji phishingu.

W wykorzystywanym modelu zdefiniowano maszynę SVM z jądrem liniowym, która uczona była na danych uczących. Model wyposażony został również w funkcję *predict_class*, która przyjmuje nowe dane jako listę cech i zwraca przewidywaną klasę.

Aby uzyskać najlepsze efekty uczenia, można zmieniać parametry modelu SVM, takie jak wartość parametru C, wybór jądra i parametry jądra, oraz inne parametry funkcji train_test_split. Zaprojektowany model maszyny SVM, wykorzystuje również strojenie parametrów z użyciem klasy *GridSearchCV* z biblioteki scikit-learn. W parametrze

[177] Pandas – biblioteka oprogramowania napisana dla języka programowania Python do manipulacji i analizy danych. W szczególności oferuje struktury danych i operacje służące do manipulowania tabelami liczbowymi i szeregami czasowymi. Jest to darmowe oprogramowanie wydane na trzyklauzulowej licencji BSD

[178] Sklearn – bezpłatna biblioteka oprogramowania do uczenia maszynowego dla języka programowania Python.

[179] CSV (z ang. **C**omma-**S**eparated **V**alues, wartości rozdzielone przecinkiem) – format przechowywania danych w plikach tekstowych i odpowiadający mu typ MIME text/csv.

parameters zdefiniowane zostały te wartości, które będą testowane dla każdego parametru modelu. Utworzony obiekt *clf* z klasy *GridSearchCV*, który będzie poszukiwał najlepszych parametrów na podstawie danych uczących. W celu znalezienia najlepszych parametrów, wywoływana była metoda *fit* na obiekcie *clf*.

W celach porównania z innymi modelami, po zakończeniu procesu uczenia, wyświetlone zostają najlepsze parametry za pomocą funkcji *clf.best_params_*. Model uzupełniono o funkcję *predict_class*, która wykorzystuje najlepsze parametry do klasyfikacji nowych danych. W ten sposób, można znaleźć najlepsze parametry modelu SVM dla danych uczących i użyć ich do klasyfikacji nowych danych.

V.2.2.2 Naiwny klasyfikator Bayesa

Klasyfikator bazujący na Twierdzeniu Bayesa o prawdopodobieństwie warunkowym, które jest wrażone wzorem:

$$P(A|B) = \frac{P(B\ |A) * P(A)}{P(B)} \tag{4.6}$$

gdzie:
A, B – zdarzenia
P(B) – prawdopodobieństwo zajścia zdarzenia B, przy czym P(B) > 0.
P(A|B) – prawdopodobieństwo zajścia zdarzenia A, o ile zajdzie zdarzenie B (prawdopodobieństwo warunkowe).
P(B|A) – prawdopodobieństwo zajścia zdarzenia B, o ile zajdzie zdarzenie A (prawdopodobieństwo warunkowe).

Naiwny klasyfikator Bayesa zaliczany jest do metod statystycznych. Algorytm klasyfikatora, wymaga założenia, że zachodzące zdarzenia są niezależne od siebie (stąd słowo „naiwny" w nazwie).

Do budowy modelu wykorzystamy bibliotekę *pandas* oraz *sklearn* w języku Python – analogicznie jak w przypadku maszyny wektorów nośnych SVM. Zdefiniowana została siatka parametrów *param_grid*, która zawiera wartości *alpha* do przetestowania osiągniętego wyniku. Następnie utworzony został obiekt *GridSearchCV,* który służy do przeszukania siatki parametrów w celu znalezienia najlepszego modelu. Użyto pięciokrotnej walidacji krzyżowej (cv=5), aby ocenić jakość modelu na różnych podzbiorach danych treningowych. Po znalezieniu najlepszego modelu, który jest przechowywany w obiekcie *grid_search.best_estimator_,* ponownie jest trenowany klasyfikator clf z użyciem najlepszych wartości parametrów. Dla danych testowych

wykonywane jest przewidywanie tego modelu wraz z oceną jego dokładność. Zdefiniowana funkcja *predict_class* przeznaczona jest do przewidywania klas dla nowych danych, korzystając z najlepszego modelu.

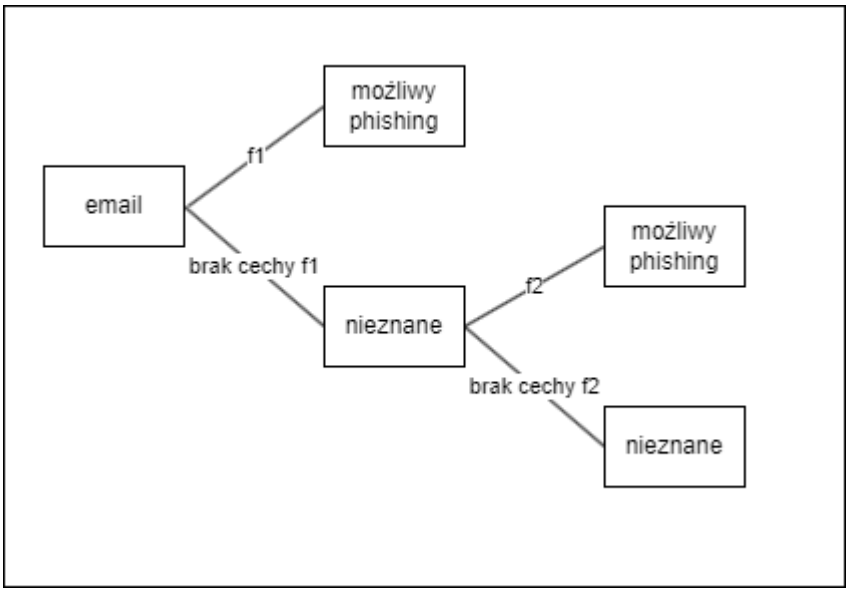
Z uwagi na występowanie nierówności w zbiorach uczących analogicznie jak dla maszyny wektorów nośnych SVM opracowane zostały dwa modele: z uwzględnieniem metod i transformacji dla przekrzywionych danych oraz bez uwzględnienia tych danych. Oba modele (z transformacją i bez) są porównywane ze sobą, w celu znalezienie najlepszego parametru uczenia.

V.2.2.3 Drzewo decyzyjne

Drzewo decyzyjne jest graficzną metodą wspomagania procesu podejmowania decyzji, Algorytm ten, stosowany jest również w uczeniu maszynowym. Drzewa decyzyjne przedstawiają możliwe konsekwencje podjętej decyzji (np. w wyniku klasyfikacji) – stąd w uczeniu maszynowym modele oparte na drzewach decyzyjnych są modelami prognozującymi wynik na podstawie utworzonego modelu.

Zakładając wykorzystanie drzewa decyzyjnego do klasyfikacji phishingu, należy skonstruować drzewo, zawierające odpowiednio zdefiniowane elementy składowe:

a. Korzeń – zbiór wszystkich, odpowiednio oznaczonych próbek treningowych.
b. Węzły – zbiór rozpoznanych i klasyfikowanych cech wiadomości phishingowych. Pojedynczy węzeł zawiera jedną i tylko jedną cechę.
c. Liście – uszeregowane i przekształcone dane pochodzące z odczytanych i wstępnie oznaczonych wiadomości email (zbiór treningowy).



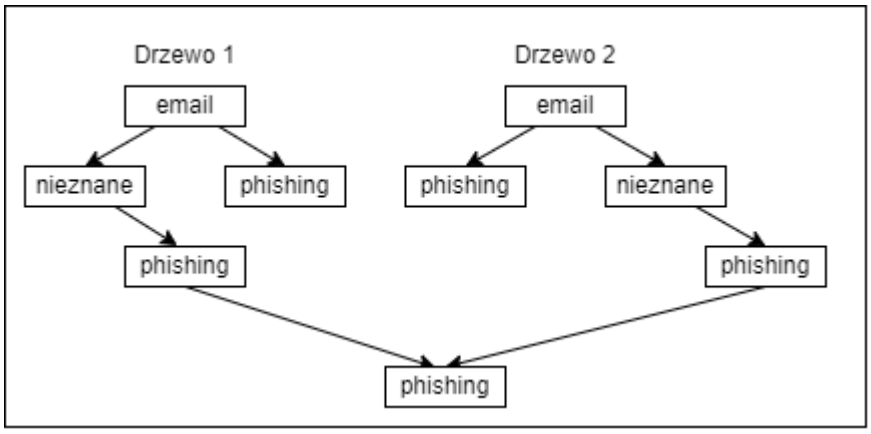
Rysunek 48. Przykład drzewa decyzyjnego do klasyfikacji phishingu.

Metody wykorzystujące drzewa decyzyjne są odporne na nierówność cech w zbiorach uczących.

W tym kodzie również wykorzystano bibliotekę scikit-learn do tworzenia modelu drzewa decyzyjnego oraz podziału danych na zbiór uczący i testowy. Analogicznie jak w poprzednich metodach, wczytano plik CSV z danymi uczącymi, a następnie podzielono je na wektory cech (X) i klasy (y). Dokonano podziału danych na zbiór uczący i testowy w stosunku 80:20 za pomocą funkcji train_test_split z biblioteki scikit-learn. Następnie utworzono obiekt klasyfikatora drzewa decyzyjnego, gdzie dopasowano go do danych uczących za pomocą funkcji *fit.*

V.2.2.4 Lasy losowe

Lasy losowe są metodą statystyczną wykorzystywaną do uczenia maszynowego zarówno dla modeli opartych na klasyfikacji jak i na regresji. Głównym założeniem metody jest generowanie wielu drzew w trakcie procesu uczenia. Generowana jest wówczas klasa, która jest dominantą klas (klasyfikacja) lub przewidywaną średnią (regresja) poszczególnych drzew. Wynik algorytmu modelu lasu losowego na nowych danych (predykcja) to głosowanie większościowe poszczególnych drzew generowanych w trakcie działania algorytmu w procesie uczenia.



Rysunek 49. Przykład lasu losowego do klasyfikacji phishingu.

Przystępując do budowy algorytmu uczenia opartego o lasy losowe, pierwszym krokiem jest ustalenie maksymalnej ilości drzew generowanych w trakcie procesu uczenia. Wykorzystywany jest algorytm Bootstrap Aggregation[180] jako procedurę agregującą poszczególne drzewa ˙w jeden zbiorczy klasyfikator, który wynik klasyfikacji opiera na głosowaniu większościowym. Działanie algorytmu polega na sekwencyjnym losowaniu (ze zwracaniem) nowych ciągów uczących (ze zbioru uczącego), służącego do trenowania nowo wygenerowanego drzewa. Danym uczącym przypisywana jest pewna waga, która wpływa na wylosowanie danego egzemplarza danych w kolejnej iteracji losowania i generowania drzewa.

$$\{(x_1, c_1), (x_2, c_2), \dots, (x_N, c_N)\} \tag{4.7}$$

gdzie:
N – ilość elementów egzemplarzy danych uczących
x – egzemplarz danych uczących (
c – waga przypisana danemu egzemplarzowi danych

Podczas pierwszej iteracji wagi każdego egzemplarza są równe. W kolejnych iteracjach, wagi są zmieniane na zasadzie: większa wartość wagi przypisywana jest temu egzemplarzowi, który został błędnie sklasyfikowany, mniejsza wartość dla poprawnej klasyfikacji. Sprawia to, że egzemplarze, na których drzewo dokonało błędnej klasyfikacji jest częściej losowane, gdyż możliwe jest że dany egzemplarz danych znajduje się blisko granicy decyzyjnej pomiędzy dwoma klasami. Lasy losowe są traktowane jako uogólnienie idei drzew decyzyjnych.

[180] Bootstrap Aggregation jest to metaalgorytm zespołu uczenia maszynowego zaprojektowany w celu poprawy stabilności i dokładności algorytmów uczenia maszynowego używanych w klasyfikacji statystycznej i regresji, mniejsza r wariancję i pomaga uniknąć nadmiernego dopasowania

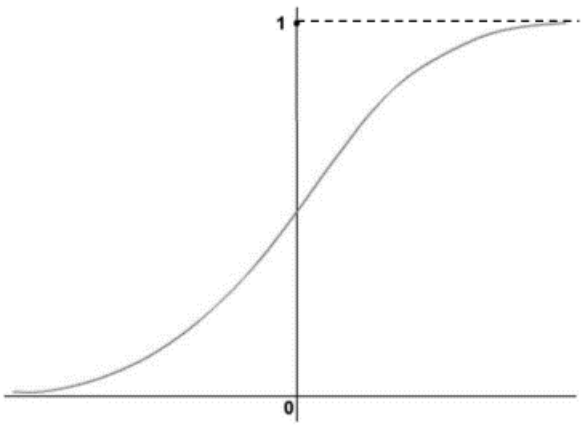
Źródłem danych uczących, podobnie jak w przypadku wcześniej opisywanych modeli był plik formatu *.csv, zawierający zakodowany wektor cech wraz z nazwami klas (phishing, spam, normal). Podobnie również wykorzystano biloteki *pandas* oraz *sklearn* w języku Python. Wykorzystując funkcję *train_test_split* z pakietu sklearn, dane zostały podzielone na zbiór uczący i testowy w stosunku 80 (uczący):20 (testowy). By odnaleźć najlepsze paramenty uczenia modelu lasu losowego, wykorzystana została funkcja GridSearchCV – skąd najlepszy parametr został wykorzystany do uczenia modelu. Model ten podobnie jak wcześniej opisywane, dokonuje rozpoznania nowych danych (również wczytywanych jako zakodowany wektor cech z pliku *.csv).

V.2.2.5 Regresja logistyczna

Regresja logistyczna jest jedną z metod statystycznych pozwalającą na opisanie współzmienności zmiennych w przypadku gdy zmienna zależna (zmienna przewidywana) przyjmuje tylko dwie wartości (np. sukces lub porażka). Model ten jest szczególnym przypadkiem uogólnionego modelu liniowego [88]. Model regresji logistycznej oparty jest o funkcję:

$$f(z) = \frac{e^z}{1 + e^z} = \frac{1}{1 + e^{-z}} \tag{4.8}$$

Funkcja logistyczna przyjmuje wartości od 0 (gdy x dąży do minus nieskończoności) do 1 (gdy x zmierza do plus nieskończoności):



Rysunek 50. Przykład wykresu funkcji logistycznej.

Regresja logistyczna opiera się sposobie wyrażenia prawdopodobieństwa zajścia danego zdarzenia zwanego szansą (z ang. *Odds*) – regresja logistyczna pozwala ocenić wpływ wielu różnych cech na szanse zajścia jakiegoś zdarzenia. Jeżeli jako Y przyjmiemy zmienną dychotomiczną (zmienną przyjmującą tylko dwie wartości), to wówczas szansę można zapisać jako:

$$P(Y = 1|x_i) = P(X) = \frac{e^{a_i + \sum_{i=1}^{k} a_i x_i}}{1 + e^{a^0 + \sum_{i=1}^{k} a_i x_i}} \qquad (4.9)$$

gdzie:
a_i – współczynniki regresji dla i-tej zmiennej (i=0,…,k)
x_i – zmienne niezależne (i=0,…,k)

Nierówność zbioru danych uczących może mieć niekorzystny wpływ na proces uczenia się z wykorzystaniem modelu regresji logistycznej - co może zaburzać właściwą interpretację danej cechy.

Do zaprojektowania modelu wykorzystano funkcję *train_test_split* z biblioteki s*cikit-learn* do podziału danych na zbiór uczący i testowy w proporcjach 80 (uczący):20(testowy). Przeprowadzone zostało strojenie hiperparametrów z wykorzystaniem pętli for. W każdej iteracji uczone są modele regresji logistycznej z różnymi wartościami parametrów C i max_iter, a następnie wykorzystano je do predykcji etykiet na zbiorze testowym. Zapisane zostały najlepsze wartości parametrów, które prowadzą do najlepszej dokładności klasyfikacji. Za pomocą klasy *LogisticRegression* dokonano uczenia klasyfikatora danych uczących za pomocą metody *fit*, Analogicznie do pozostałych klasyfikatorów model jest wyposażony w funkcję która na podstawie nowego nieznanego zestawu danych (również wczytywanego z pliku .CSV) dokonuje przewidywania klasy (*predic*t).

Rozdział V – Weryfikacja jakości metody wykrywania wiadomości phishingowych

Weryfikacja jakości metody wykrywania wiadomości phishingowych, obejmowała będzie dwa etapy:

1. Ocenę jakości identyfikacji opisanych cech w wiadomości email przez algorytm wykrywania. Etap ten zawierał będzie również dostosowywanie funkcji odczytujących odpowiednie pola wiadomości email, dostosowanie długości wektora cech.
2. Ocenę jakości klasyfikacji wiadomości email na wiadomości phishingowe, spam oraz stanowiące normalną korespondencję. Etap ten zawierał będzie również badanie zachowania się klasyfikatorów na niezbalansowanym i zbalansowanym różnymi metodami zbiorze danych uczących.

V.1 Weryfikacja jakości identyfikacji cech wskazujących na potencjalnie phishingowy charakter wiadomości

W celu badania poprawności działania algorytmu i prawdziwości założeń wiadomości phishingowej (opisanych w punkcie IV.1.1 Założenia wiadomości phishingowej, na stronie 161), przeprowadzono badanie poprawności działania algorytmu na próbce 50 wiadomości email. Wśród wiadomości testujących poprawności działania algorytmu znajdowały się wiadomości ocenione jako wiadomości phishingowe, wiadomości spełniające warunki uznanie je za wiadomości charakterze spamu, wiadomości marketingowe (w tym wiadomości z zakresu tzw. „agresywnego marketingu").

Testowany zbiór wiadomości, wybrany został w sposób losowy ze zgromadzonego zasobu na potrzeby niniejszej rozprawy. Zbiór testowy liczył 50 wiadomości email, w tym:

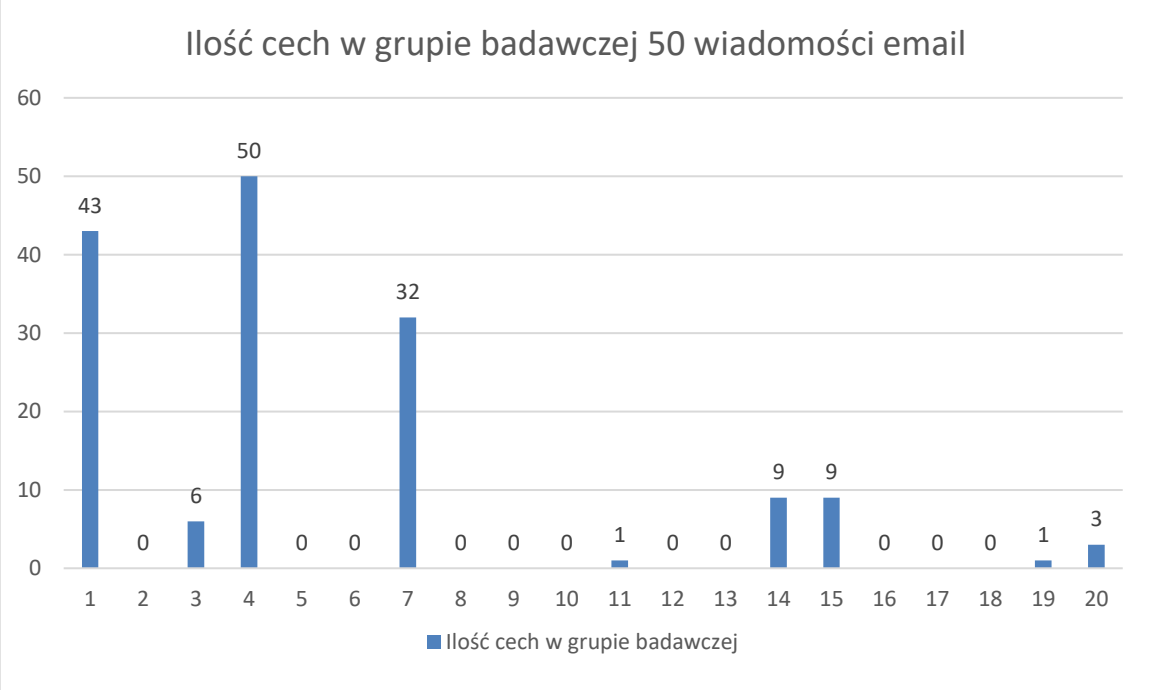
a) 43 wiadomości phishingowe,
b) 7 wiadomości o charakterze marketingowym (normalne).

Głównym założeniem przeprowadzenie próby kontrolnej – zadaniem działania algorytmu na obecnym etapie – było wykrycie i poprawne zaklasyfikowanie wykrytej

przez algorytm cechy mogącej świadczyć o potencjalnym ataku phishingowym a nie poprawne przypisanie poszczególnych wiadomości do grup (phishing, spam, normalne). Celem więc było przetestowanie poprawnego rozpoznawania poszczególnych cech w bazie różnych wiadomości.

V.1.1 Zestawienie danych testowych

Uzyskane wyniki ujawniły, że nie każdy opisany, wykryty i poddany analizie wskaźnik występuje w wiadomościach uznawanych za wiadomości phishingowe:



Rysunek 51.Wynik działania algorytmu wykrywania cech phishingu na losowej grupie 50 wiadomości email.

1. Błędnie sklasyfikowano cechę nr 4 – różnice w dacie rejestracji domeny wynikającej z części domenowej adresu email a daty otrzymania wiadomości email pochodzącej z tego adresu. Błędy wykrywania tej cechy spowodowane są konstrukcją zewnętrznej usługi (*whois*) i importowanej biblioteki języka Python (*python-whois*), za pomocą których, otrzymuje się informację podawane przez właściciela w trakcie rejestracji danej domeny oraz inne techniczne informacje. Z podawanych danych pobrano datę rejestracji domeny (lub daty odnowienia certyfikatu). Dokonano konwersji tej daty do standardowego modelu (RRRR-MM-DD), wykorzystywanego np. w systemach bazodanowych. Z nagłówka

wiadomości również pobrano datę otrzymania wiadomości i również dokonano jej konwersji do formatu RRRR-MM-DD. Daty te zostały porównane ze sobą. Dla całego zbioru testowego (dla każdej indywidualnej wiadomości email – zbiór testujący poprawność algorytmu zawierał 50 wiadomości) funkcja zwróciła wartość: TRUE – oznaczająca, że różnice obu tych dat (data rejestracji domeny lub odnowienia certyfikatu oraz data otrzymania wiadomości) jest niewielka, na tyle że można daną wiadomość, dla tej cechy oznaczyć jako możliwy phishing. Zwracana wartość zawsze wynosiła True, bez względu jak duża wynosiła ta różnica. W próbce znajdowały się wiadomości, które posiadały odnośniki do domen zarejestrowanych w 1999 roku. Różnicą pomiędzy tą datą, a datą otrzymania tej wiadomości, która zawierała wskazany odnośnik wynosiła 22 lata (wiadomość otrzymano w 2021 roku) – jednakże wynik działania połączenia serwisu „*whois*", biblioteki „python-whois"" oraz konwersji dat, dał wynik TRUE – oznaczający, że czas pomiędzy rejestracją domeny a datą otrzymania wiadomości jest mniejszy niż 7 dni. Wartość tą ustawiono z uwagi na wyniki badania opublikowanego przez zespół Farsight Security[181], które wskazuje, że domeny phishingowe przestają być aktywne w ciągu 7 dni od daty rejestracji, a większość (84%) przestaje być aktywna w ciągu 24h.

2. Nie wykrycie cech nr 5 – niewłaściwy adres email nadawcy wiadomości. Próbka testowa nie zawierała tych danych, co potwierdziło prawidłowe działanie funkcji algorytmu.
3. Nie wykryto żadnej cechy nr 8 – żadne odnośniki wewnątrz analizowanych wiadomości email nie prowadziły do uznanych za phishingowe domen. Na potrzeby wykrywania domen phishingowych korzystano z publikowanej i zweryfikowanej przez zespół CERT POLSKA (NASK) bazy domen phishingowych[182]. Próba testowa nie zawiera odnośników do domen, które przez zespoły NASK zostały uznane za phishingowe i wpisane na listę.
4. Nie wykryto żadnej cechy nr 10 – odnośniku URL wewnątrz analizowanych wiadomości email nie zawierały złożonych nazw domenowych
5. W jednej wiadomości phishingowej wykryto cechę, która tam nie występowała (adres portfela BTC) – działanie niepożądane. Wykryty został ciąg

[181] https://www.farsightsecurity.com/assets/media/download/VB2018-study.pdf
[182] https://hole.cert.pl/domains/domains.txt [dostęp: 18.06.2022].

znaków, którego początek sugerował wartość portfela BitCoin (BTC). Adresy portfeli BitCoin zaczynają się na określony prefix:

['1', '2', '3', '5', '9', 'bc1', 'tb1', 'xpub', 'xprv', 'm', 'n', 'c', 'tpub', 'tprv']

W celu wyeliminowania tego błędu zmodyfikowano mechanizm sprawdzający całkowitą długość adresu portfela BitCoin.

6. W dwóch wiadomościach phishingowych nie wykryto mechanizmu śledzącego (mechanizm śledzący tam występował) – działanie niepożądane
7. W wiadomościach legalnych nie wykryto cech phishingowych, co jest działaniem pożądanym – algorytm nie generuje cechy „False Positive" (FP).

V.1.2 Metoda weryfikacji

Wynikiem działania algorytmu wykrywania cech phishingowych w wiadomości email jest zakodowany binarnie wektor cech. W wyniku przetwarzania wiadomości email, wynikowy wektor cech na pozycjach przypisywanym poszczególnym cechom, może zawierać następujące wartości:

a. wartość „1" – oznacza, że dana cecha występuje w wiadomości email a jej wartość lub władności wskazują na możliwy atak phishingowy,
b. wartość „0" – przetwarzana wiadomość nie zawiera danej cechy lub zawiera tą cechę jednakże jej wartość lub własności nie wskazują na atak phishingowy.

Algorytm dokonuje więc binarnej klasyfikacji wartości cech. W trakcie przetwarzania danych (odczytu wiadomości), możliwa jest sytuacja, że na podstawie odczytanych wartości, dana cecha pozytywna zostanie błędnie oznaczona jako negatywna (lub cecha negatywna zostanie oznaczona jako pozytywna). Wszystkie takie sytuacje przedstawia poniższa tablica pomyłek (macierz błędów).

TP Klasa prawdziwie pozytywna (ang. True Positive). Cecha phishingowa poprawnie zidentyfikowana jako phishingowa	**FP** Klasa fałszywie pozytywna (ang. False Positive). Zidentyfikowano cechę phishingową w wiadomości, w której nie występuje (błędy pierwszego rodzaju[183]).
FN Klasa fałszywie negatywna (ang. False Negative) Występująca cecha phishingowa nie została wykryta (błędy drugiego rodzaju[184]).	**TN** Klasa prawdziwie negatywna (ang. True Negative) Nie wykryto cech phishingowych w wiadomościach nieposiadających cech phishingowych

By ocenić skuteczność algorytmu w wykrywaniu i przypisywaniu rzeczywistych klas, należy przeprowadzić obliczenia stosując poniże miary:

1. Czułość – odsetek cech prawdziwie pozytywnych, stosunek wyników prawdziwie dodatnich do sumy prawdziwie dodatnich i fałszywie ujemnych. Czułość na poziomie 100% oznacza, że wszystkie cechy mogące świadczyć phishingu zostaną rozpoznane i we właściwy sposób oznaczone (f_i=1 dla i=1,…,19). Parametr ten obrazuje zdolność algorytmu do prawidłowego wykrycia cech i przypisania ich do właściwej klasy.

$$TPR = \frac{\sum TP}{\sum TP + \sum FN} \tag{5.1}$$

gdzie:
∑TP – suma cech które otrzymały klasę True Positive
∑FN – suma cech które otrzymały klasę False Negative

2. Swoistość – odsetek cech prawdziwie negatywnych, stosunek wyników prawdziwie ujemnych do sumy prawdziwie ujemnych i fałszywie dodatnich. Swoistość na poziomie 100% oznacza, że nie wykryto żądnej cechy świadczącej o phishingu w wiadomościach, faktycznie pozbawionych tych cech.

[183] Błąd pierwszego rodzaju - błąd polegający na odrzuceniu hipotezy zerowej, która w rzeczywistości nie jest fałszywa. Oszacowanie prawdopodobieństwa popełnienia błędu pierwszego rodzaju oznacza się symbolem α i nazywa poziomem istotności testu.
[184] Błąd drugiego rodzaju - błąd polegający na nieodrzuceniu hipotezy zerowej, która jest w rzeczywistości fałszywa.

$$TNR = \frac{\sum TN}{\sum FP + \sum TN} \quad (5.2)$$

gdzie:
∑TN – suma cech które otrzymały klasę True Negative
∑FP – suma cech które otrzymały klasę False Positive
∑TN – suma cech które otrzymały klasę True Negative

3. Precyzja – określana jako stopień zgodności między wynikami (liczbą cech określonych jako phishingowe) z wielokrotnych pomiarów tej samej wielkości. Precyzja określa jaka część wyników określona przez algorytm jako dodatnie jest faktycznie dodatnia.

$$PPV = \frac{\sum TP}{\sum TP + \sum FP} \quad (5.3)$$

gdzie:
∑TP – suma cech, które otrzymały klasę True Positive
∑FP – suma cech, które otrzymały klasę False Positive

4. Dokładność – można określić jako stopień zgodności wartości rzeczywistej (ilość wykrytych cech) ze średnią arytmetyczną wyników uzyskanych dla wielości całej populacji cech phishingowych. Im dokładniejsza metoda pomiaru, tym uzyskiwane wyniki są bliższe wartości prawdziwej.

$$ACC = \frac{\sum TP + \sum TN}{P} \quad (5.4)$$

gdzie:
∑TP – suma cech, które otrzymały klasę True Positive
∑TN – suma cech, które otrzymały klasę True Negative
P – wielkość populacji (ilość wszystkich cech wskazujących na phishing)

$$\frac{TN + TP}{TN + TP + FN + FP} \quad (5.5)$$

V.1.3 Wyniki weryfikacji

Przeprowadzoną próbę można ocenić w kategoriach:

a. Całość populacji liczy 105 cech wskazujących na możliwy atak phishingowy. W cechach tych nie uwzględniono wielkości populacji klasy cechy różnicy dat, pomiędzy rejestracją domeny, do której prowadzą odnośniki URL wewnątrz wiadomości a datą otrzymania wiadomości, z powodu wadliwego działania

połączenia usługi udostępniającej informacje o domenie (*whois*), a biblioteką języka Python (python-whois).

b. W 2% przypadków (jedna wiadomość) wykryto cechę, która nie występowała („False Positive" – FP).
c. W 4% przypadków (dwie wiadomości) nie wykryto cechy, która występował („False Negative" – FN).
d. W 100% wiadomości legalnych nie wykryto cech phishingowych (nie występowały, „True Negative" – TN).

Powyższym wskaźnikom możemy przypisać wartości liczbowe:

Tabela 42. Wykaz wartości liczbowych przypisanych do poszczególnych wskaźników.

Lp.	klasa	Wielkość	Opis klasy
1.	∑TP	**104**	Suma cech, które otrzymały klasę True Positive
2.	∑FN	**2**	Suma cech, które otrzymały klasę False Negative
3.	∑TN	**0**	Suma cech, które otrzymały klasę True Negative
4.	∑FP	**1**	Suma cech, które otrzymały klasę False Positive
5.	P	**105**	Wielkość populacji (ilość wszystkich cech wskazujących na phishing)

Podstawiając wartości liczbowe dla poszczególnych parametrów, możemy wyliczyć:

a. Czułość

$$TPR = \frac{104}{104 + 2} \cong 0.98 \tag{5.6}$$

b. Precyzję

$$PPV = \frac{104}{104 + 1} \cong 0.99 \tag{5.7}$$

c. Dokładność

$$ACC = \frac{104 + 0}{105} \cong 0.99 \tag{5.8}$$

Przedstawiony powyżej test działania algorytmu wykonany został na zbiorze 50 wiadomości email, pochodzących z różnych serwisów, wysłanych i przetwarzanych przez różne serwery pocztowe, wysłane w różnych przedziałach czasowych (data otrzymywania to lata 2019 – 2021). Badany zbiór posiadł znaczną różnorodność, by

wyczerpać wszelkie możliwości konstrukcji pól nagłówka i ich wartości. Z uwagi na zakładaną pewną dowolność w konstrukcji nagłówka wiadomości email (standardy RFC), nie można wykluczyć, że istnieje pewna grupa wiadomości email, dla której wykrywanie i klasyfikacja cech nie została opisana regułami przedstawionymi w Rozdziale III. Dla tej grupy wiadomości email, z powodu różnic pomiędzy wartościami spodziewanymi a rzeczywistymi, algorytm może błędnie rozpoznawać cechy, nie rozpoznawać ich w ogóle, błędnie klasyfikować rozpoznane cechy, a więc i uzyskane wartości czułości, precyzji i dokładności mogą odbiegać od obecnych wyników. Konieczne będzie wówczas zmodyfikowanie jedynie wąskiego wycinka algorytmu – gdyż opisany proces wykrywania i kodowania cech będzie taki sam.

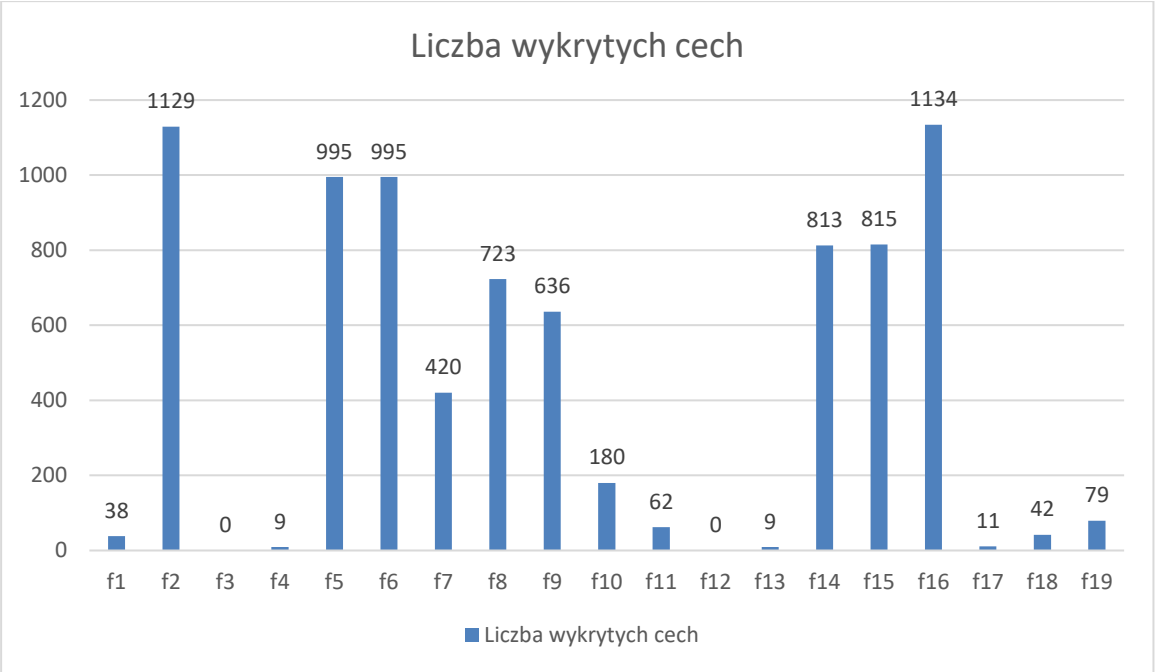
V.2 Weryfikacja metody wykrywania wiadomości phishingowych na podstawie zidentyfikowanego wektora cech

V.2.1 Zestaw danych testowych

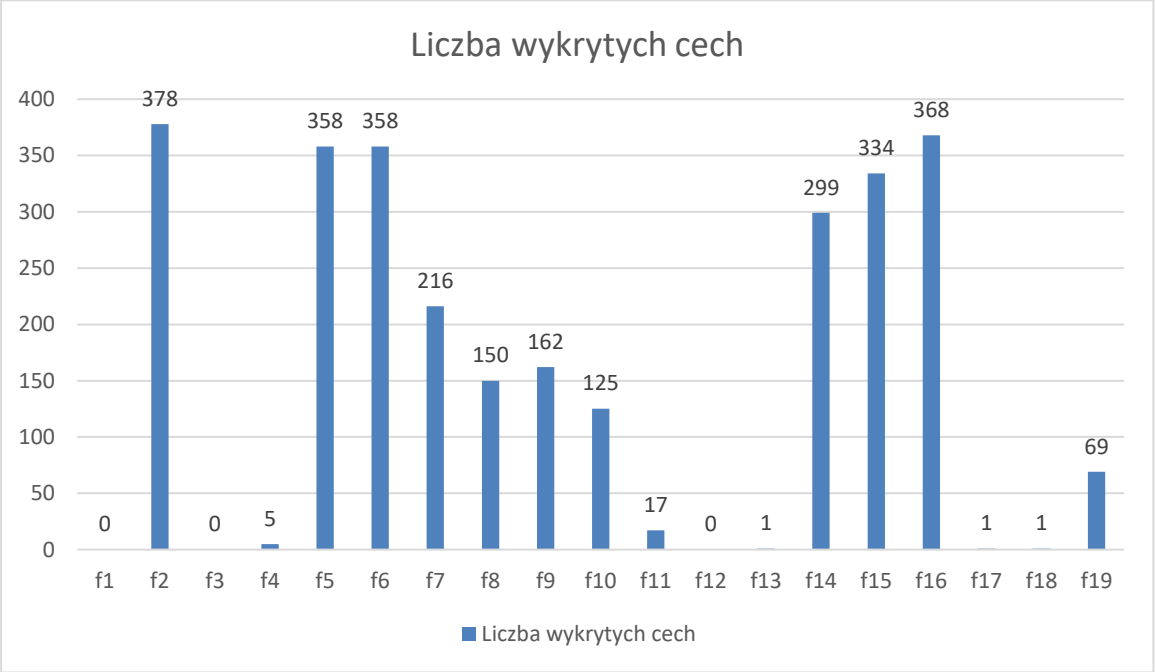
Przygotowany zbiór wiadomości email podzielony był na trzy oddzielne kategorie:

1. wiadomości normalne – 80 wiadomości, co stanowiło 4,92% zbioru uczącego,
2. wiadomości typu spam – 378 wiadomości, co stanowiło 23,25% zbioru uczącego,
3. wiadomości phishingowe – 1168 wiadomości co stanowiło 71,83% zbioru uczącego.

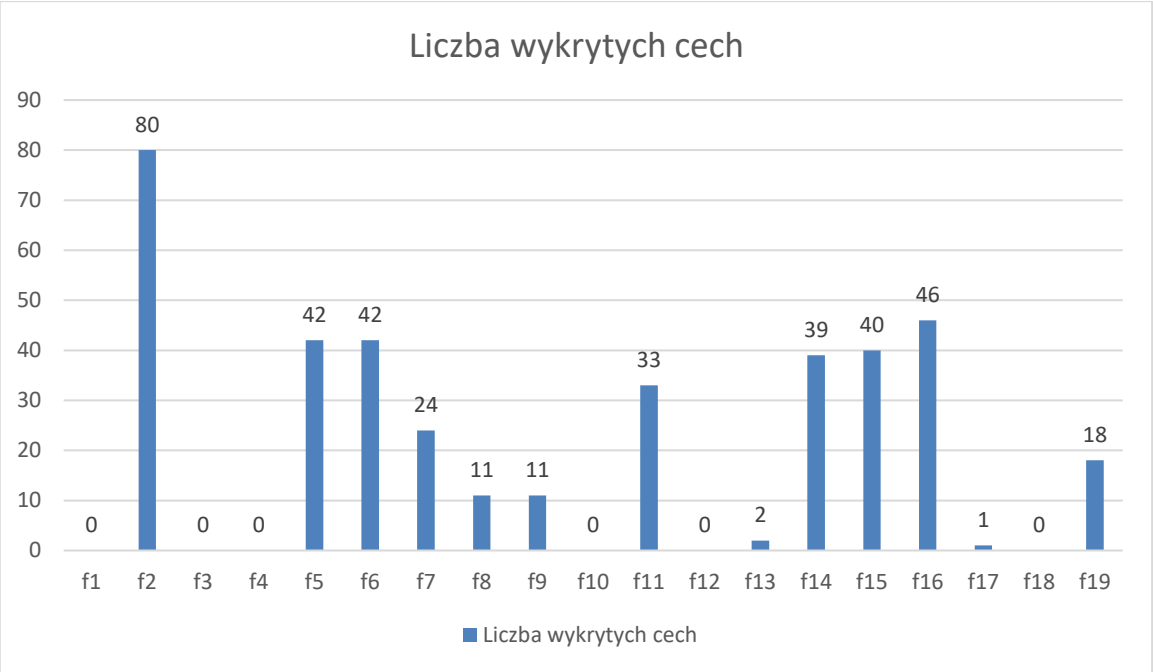
Każdy ze zbiorów wiadomości (phishing, spam, normalne) był odczytywany oddzielenie. Dla zbioru wiadomości phishingowych uzyskano listę cech, ilościowo przedstawioną na poniższym wykresie:



Dla wiadomości typu spam:



Dla wiadomości typu „normal”:



V.2.1.1 Błędy odczytu

W trakcie pierwszej iteracji działania algorytmu przetwarzającego wiadomości email (odczyt pliku *.eml wiadomości), funkcja odczytu nie mogła odczytać 14 plików. Wiadomości te zostały usunięte ze zbioru wiadomości przetwarzanych przez algorytm.

Za główną przyczynę niemożności odczytu danego pliku wiadomości wskazuje się:

1. Możliwe błędy podczas zapisu treści wiadomości (wraz z nagłówkiem i załącznikami) jako plik.
2. Błędnie rozpoznane kodowanie lub błędy zapisu kodowania. Domyślnym kodowaniem podczas zapisu wiadomości do pliku formatu .eml było wybrane kodowanie UTF-8[185]. Wadą kodowania UTF-8 jest wykorzystanie 3 bajtów do zapisuj znaków CJK[186] oraz 2 bajtów dla liter alfabetu niełacińskiego. Z tego powodu niektóre symbole specjalne mogą być zapisywane na pełnych 4 bitach informacji lub też mogą wystąpić nieoczekiwane bity kontynuacji podczas zapisu w danym formacie plików. Wadą kodowania UTF-8 są również możliwe błędy powstające z powodu prostego obcięcia łańcucha znaków – zakończenie ciągu

[185] UTF-8 - system kodowania Unicode, wykorzystujący od 1 do 4 bajtów do zakodowania pojedynczego znaku.
[186] CJK - używane w informatyce określenie systemów pisma wywodzących się z pisma chińskiego.

przez końcem znaku, co uniemożliwia prawidłowy odczyt przez wiele dekoderów (w tym również wykorzystywany w języku Python, w którym opracowano PoC).

Analizując pliki, które nie zostały odczytane przez algorytm przetwarzający wiadomości, przyczyną braku odczytu wiadomości był błędny sposób zapisu wiadomości podczas jej przetwarzania w systemach o innym kodowaniu niż UTF-8. Wiadomości eksportowane z serwera pocztowego, gdzie bazowym systemem operacyjnym był Microsoft Serwer, mogą mieć ustawioną stronę kodową jako Windows-1250, CP-1250, co prowadzić może do wspomnianych błędów odczytu. W celu wyeliminowania błędu odczytu wspomnianych plików wiadomości wykonano:

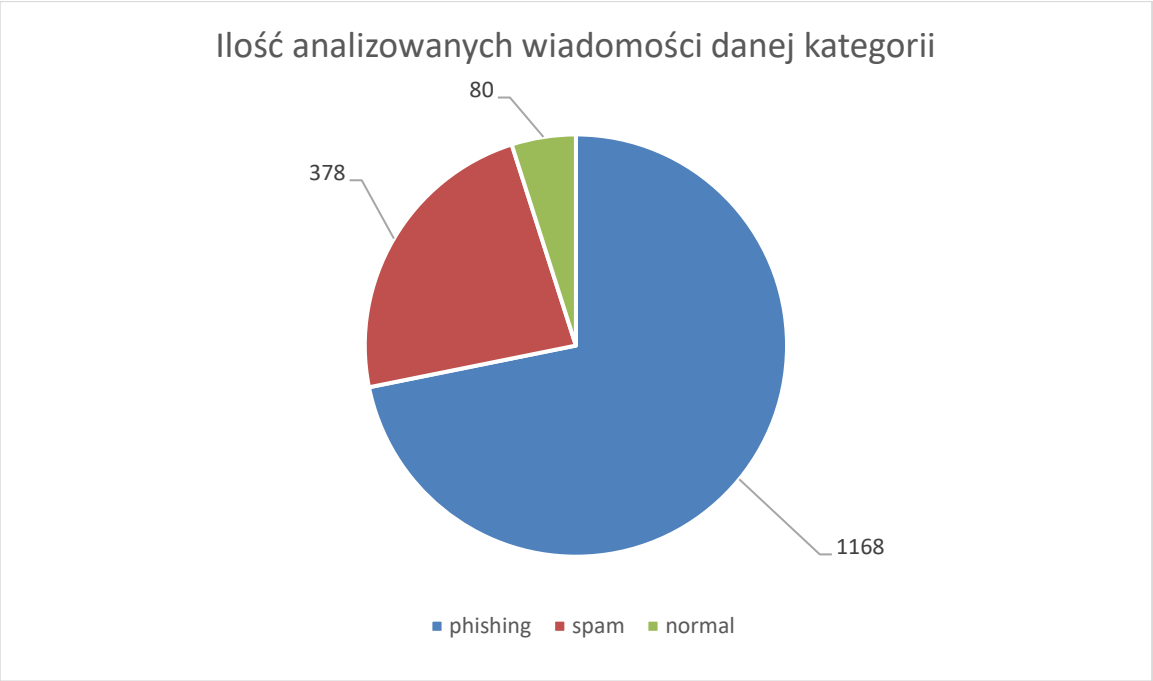
1. wczytano treść źródłową wiadomości do edytora tekstu (Notepad++ v. 8.4.8),
2. ustawiono kodowanie tekstu jako utf8,
3. dokonano ponownego zapisu (bez modyfikacji treści) wiadomości pod tą samą nazwą, lecz ze zmienioną stroną kodową,
4. Podano na wejście algorytmu – nie odnotowano błędu odczytu, plik wiadomości został prawidłowo przetworzony.

Powtórny zapis wiadomości nie modyfikuje wartości żadnych z potencjalnie występujących cech mogących wskazywać na atak phishingowy, nie wpływa więc na wynik odczytu i wykrywania cech.

V.2.1.2 Nierówność zbioru

Poniższy wykres przedstawia nieodzwierciedlającą w rzeczywistość, nierówność[187] zbioru wiadomości email. Według statystyk [89] dziennie przesyłanych jest około 347,3 miliarda wiadomości, wobec 3,4 miliarda[188] wiadomości dziennie identyfikowanych jako phishing, co stanowi zaledwie około 1% danych. W zbiorze uczącym oraz testowym, klasa opisująca wiadomości phishingowe jest reprezentowana przez znacznie więcej przykładów niż pozostałe dwie klasy (spam i normalne).

[187] W literaturze nierówność zbioru określana jest pojęciem „Skew Data" (z ang. dane przekrzywione), co oznacza że dane posiadają odchylenie w prawo lub w lewo od standardowego rozkładu normalnego.
[188] https://aag-it.com/the-latest-phishing-statistics/



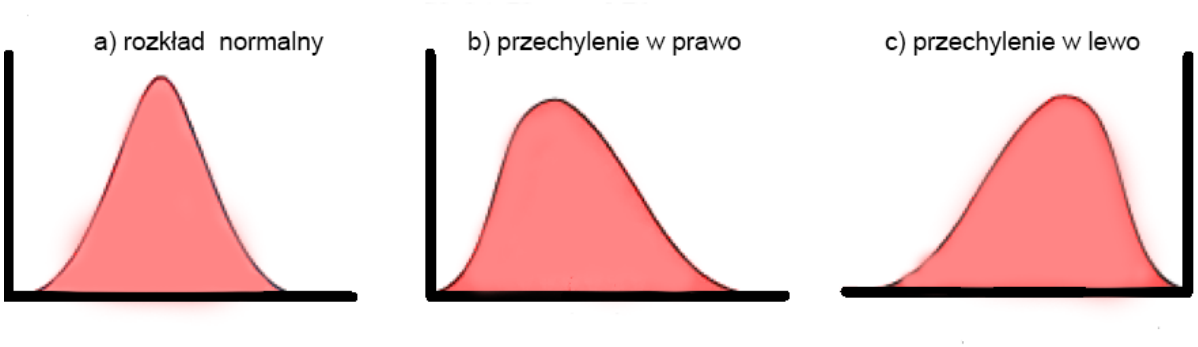
Rysunek 52. Ilość analizowanych wiadomości danej kategorii

Problem nierówności danych (dane skośne – skewed data) jest często występującym scenariuszem w zadaniach klasyfikacji, gdzie jedna klasa jest dominująca, podczas gdy inne klasy są słabo reprezentowane.

Trudnością w pozyskaniu zbioru wiadomości stanowiącej normalną korespondencję, była niechęć odbiorców do ujawniania swojej prywatnej korespondencji, zawierających prywatne treści, zasłanianie się rozporządzeniem o ochronie danych osobowych (RODO). Trudności ten powodujące brak wystarczającej ilości danych do konstrukcji właściwego zbioru uczącego, wymuszaj, transformację nierównego zbioru danych.

Nierówność zbioru danych uczących może przybrać formę „pochylenia” (skosu) w dwóch rodzajach przypadków:

1. ujemne pochylenie – wykres reprezentujący zebrane dane pochylony jest w lewo w stosunku do wykresu rozkładu normalnego,
2. dodatnie pochylenie – wykres reprezentujący zebrane dane pochylony jest w prawo w stosunku do wykresu rozkładu normalnego.



Rysunek 53. Przykład zbiorów danych: a) rozkład normalny, b) rozkład danych przechylonych w prawo, c) rozkład danych przechylonych w lewo. Źródło: opracowanie własne

Występowanie nierówności danych w zbiorach jest naturalnym odzwierciedleniem wielu zachodzących rzeczywistych procesów i sytuacji (wspomniane powyżej nierówność rzeczywistego zbioru wszystkich wiadomości email przesyłanych jednego dnia w porównaniu do zbioru wiadomości phishingowych). Jednakże, występowanie przekrzywienia danych w modelach uczenia maszynowego może prowadzić do niedokładności klasyfikacji, nieproporcjonalnego ważenia klas, niestabilności modelu a przez to do zniekształcenia przewidywanych wyników.

Nieproporcjonalność zestawu danych uczących może wpływać na:

1. Model regresji liniowej: niewłaściwa interpretacja wyników – dane pochylone w lewo posiadają, mniejszą ilość obserwacji o wartościach mniejszych od średniej. Przy modelowaniu regresji, gdzie zakładamy liniową zależność między zmiennymi, taki rozkład może prowadzić do błędnej interpretacji wyników. Model może błędnie szacować wpływ zmiennych niezależnych na zmienną zależną, ponieważ bardziej skoncentruje się na ekstremalnych wartościach.
2. Model regresji liniowej: wrażliwość danych odstających – dane te w modelu regresji mogą wpływać na wytrenowaniu modelu i jego tendencji do skrajnego przewidywania danych w skrajnych obszarach.
3. Model drzewa decyzyjnego: nieproporcjonalne ważenie poszczególnych klas, szczególnie w przypadku klasyfikacji binarnej. Jeśli jedna z klas jest bardziej liczna, model drzewa decyzyjnego może być bardziej skłonny przewidywać tę klasę jako prawidłowy wynik. Może to prowadzić do niedokładności klasyfikacji i zniekształconego odwzorowania rzeczywistości.

4. Model KNN[189]: zakłócenie obliczania odległości między poszczególnymi obserwacjami. Jeśli jedna z klas jest bardziej liczna, to modele k-najbliższych sąsiadów mogą skłaniać się ku przypisywaniu większej wagi dla tej klasy podczas klasyfikacji na nowym zestawie danych. Może to prowadzić do niedokładności klasyfikacji dla mniejszościowej klasy
5. Model SVM: niewłaściwą optymalizację hiperpłaszczyzny separacji. Jeśli jedna z klas jest bardziej liczna, optymalizacja może skupić się na separacji dominującej klasy, pomijając mniejszościową klasę. W rezultacie, model oparty na maszynie wektorów nośnych może mieć trudności w skutecznym rozróżnianiu między klasami i może generować niedokładne wyniki klasyfikacji.
6. System oparty na regułach: wygenerowanie reguł, które są skupione na dominującej, większościowej klasie, pomijając klasę zawierającą mniejszą ilość danych (elementów). Klasyfikatory oparte na regułach, takie jak algorytmy asocjacyjne czy algorytmy korelacyjne, mogą mieć trudności w wykrywaniu wzorców i zależności występujących w mniejszościowej klasie.
7. Sieci neuronowe: możliwe wystąpienie błędów aktualizowania poszczególnych wag, poprzez takie ustalenie wartości wagi, dzięki której dany neuron będzie preferował większościową klasę. W takim przypadku sieć neuronowa będzie miała trudności w nauczeniu się poprawnej reprezentacji mniejszościowej klasy i może skłaniać się do przewidywania klasy z większą ilością danych.

Nierówność zbioru danych uczących można zniwelować stosując np.: próbkowanie równoważące klasy, metody równoważenia zbiorów, zastosowanie wag klas lub inne techniki przetwarzania danych mogą pomóc w radzeniu sobie z tymi wyzwaniami.

Wykorzystując metody równoważenia zbioru uczącego z przewagą liczebną jednej z klas (klasa większościowa) lub w zbiorze danych z występującymi brakami w pokryciu unikalnych cech, może czasem wystąpić zjawisko pogorszenia wyników rozpoznawania nowych wzorców, po przeprowadzonym treningu nowym, zrównoważonym zbiorem. Problem ten może dotyczyć przewidywania klas mniejszościowych. Przyczynami takiego stanu rzeczy mogą być:

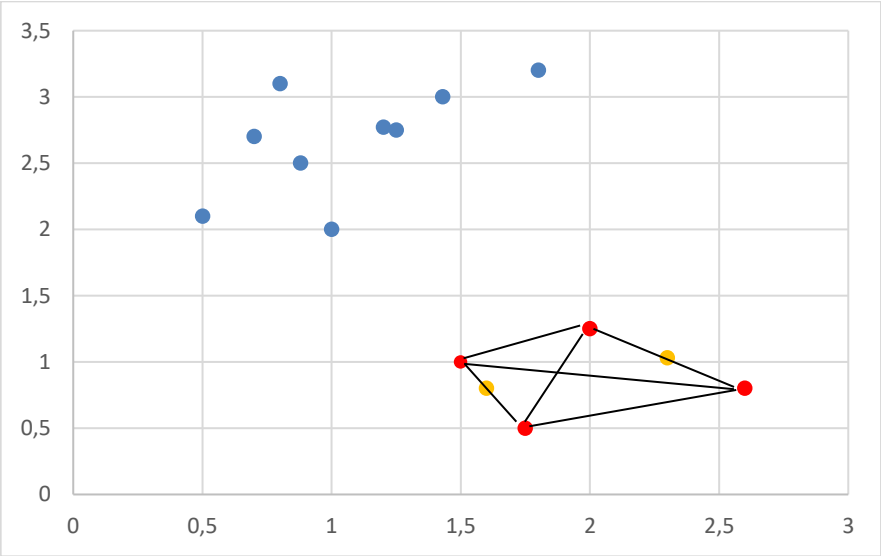
[189] KNN – (k-najbliższych sąsiadów) - jeden z algorytmów regresji nieparametrycznej używanych w statystyce do prognozowania wartości pewnej zmiennej losowej, zwłaszcza wtedy, gdy zależność między zmiennymi objaśniającymi a objaśnianymi jest złożona lub nietypowa (np. niemonotoniczna), czyli trudna do modelowania w klasyczny sposób.

1. Strata informacji – niektóre metody generując nowe egzemplarze danych dla poszczególnych klas mniejszościowych, mogą wprowadzić duplikaty informacji do modelu, co może prowadzić do przeszkodzenia procesowi uczenia. W skrajnych przypadkach, gdy liczba duplikatów jest duża, model może "pamiętać" konkretne przypadki zamiast uogólniać.
2. Przeuczenie (overfitting) – dodanie dużych ilości syntetycznych danych do klas mniejszościowych może zwiększyć ryzyko przeuczenia, zwłaszcza w przypadkach powielania egzemplarzy danych na bazie już istniejących. Model może zbyt mocno dopasować się do nielicznych przypadków, co powoduje trudności w rozpoznaniu nowych, nieznanych danych.
3. Niewłaściwe dobieranie parametrów – nieprawidłowe ustawienie parametrów może spowodować wygenerowanie dużej ilości identycznych egzemplarzy (zbyt jednolity zbiór). Podczas powielania danych, może również dojść do istotnej utraty informacji co wpływa negatywnie na możliwości rozpoznawania nowych wzorców, czy też niewłaściwe ustawienie wag poszczególnych klas mogą zaburzać proces uczenia.
4. Charakterystyka danych – dla dobrze zdefiniowanych zbiorów (nawet w przypadku klas mniejszościowych), charakteryzujących się unikalnymi dla danej klasy cechami, wygenerowanie nowych egzemplarzy danych (powielenie ich), może prowadzić do utraty unikalności tych cech, co przełoży się na pogorszenie procesu rozpoznawania nowych wzorców.
5. Odpowiedni algorytm – właściwy wybór modelu uczenia maszynowego jest niezwykle istotny w przypadku posiadania danych niezrównoważonych. W przypadku Naiwnego Klasyfikator Bayesa wartości odstające mają wpływ na obliczenie ogólnego prawdopodobieństwa [90], co przekłada się na rozpoznawanie nowych wzorców. Pozytywny wpływ równoważenia danych niezbalansowanych jest w przypadku użycia algorytmów poszukujących dobrego podziału danych (np. Maszyna Wektorów Nośnych – SVM lub Drzewo Decyzyjne) [91].

V.2.1.3 Metoda SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) – technika generująca syntetyczne dane dla klasy mniejszości zaproponowana przez Chawla, Bowyer, Hall,

Kegelmeyer [92]. Nowe dane klasy mniejszościowej (syntetyczne) powstają poprzez łączenie w sposób losowy „linią” w pary k-punktów z klasy mniejszościowej (k- neighbor), a następnie wygenerowanie nowego obiektu znajdującego się dokładnie na tej linii. Wybór punktów z klasy mniejszościowej następuje w sposób losowy, w sposób losowy następuje również wybór otoczenia k-sąsiadów. Powstające w ten sposób egzemplarze danych, posiadają jednak inne cechy niż oryginale dane na podstawie których wytworzony został nowy egzemplarz.



gdzie:

- klasa mniejszościowa
- klasa większościowa
- wygenerowany egzemplarz danych klasy mniejszościowej za pomocą SMOTE

Metoda SMOTE zasadniczo różni się od innych metod generowania egzemplarzy danych klasy mniejszościowej, generowaniem obiektów w przestrzeni cech a nie w przestrzeni danych, wzmacniając tym samym klasę mniejszościową Zaproponowany przez Chawla, Bowyer, Hall, Kegelmeyer [92] algorytm obejmował grupę do 5 najbliższych sąsiadów wylosowanego egzemplarza danych, z której to grupy losowane było dwóch sąsiadów i na ich podstawie generowany był nowy egzemplarz. Metodę tą można również wykorzystać do danych binarnych:

a. losowany jest egzemplarz danych z klasy mniejszościowej,
b. losowo wybieranych jest jego k-sąsiadów,

c. generowany jest nowy egzemplarz danych poprzez losowe kombinowanie cech wylosowanego egzemplarza i jego sąsiada.

V.2.1.4 SMOTE-ENN

Algorytm SMOTE-ENN (SMOTE with Edited nearset neighbor) działa w dwóch krokach:

1. Wygenerowanie egzemplarzy klasy mniejszościowej za pomocą metody SMOTE,
2. Wykorzystanie algorytmu ENN dla próbek większościowych w celu usunięcia otoczenia wstępnie sklasyfikowanych najbliższych sąsiadów [93].

Zasadniczym działaniem algorytmu ENN jest wstępna klasyfikacja na podstawie próbek większościowych a następnie badanie (z wykorzystaniem algorytmu KNN) otoczenia egzemplarza danych x_i ($i \in N$, gdzie $N -$ zbiór egzemplarzy danej klasy). Egzemplarz x_i zostanie usunięty, jeżeli w zbadanym sąsiedztwie dominuje liczba egzemplarzy innej klasy [94].

V.2.1.5 ADASYN

Zasadniczą ideą metody ADASYN (Adaptive Synthetic Sampling) jest wykorzystanie ważonego rozkładu dla różnych przykładów klas mniejszościowych [95]. W metodzie tej w pierwszym kroku oblicza się stopień nierównowagi dla klasy mniejszościowej (stosunek egzemplarzy danych klasy mniejszościowej do egzemplarzy klasy większościowej), a następnie obliczona zostaje ilość koniecznych danych do wygenerowania. Im większa nierówność pomiędzy klasa mniejszościową a większościową, tym więcej egzemplarzy syntetycznych zostanie wygenerowanych przez ten algorytm. Stopień nierównowagi klasy mniejszościowej możemy obliczyć zgodnie ze wzorem

$$d = \frac{m_s}{m_l} \tag{5.9}$$

gdzie:
ms – liczba egzemplarzy danych klasy mniejszościowej

ml – liczba egzemplarzy danych klasy większościowej

W drugim kroku dla każdego egzemplarza klasy mniejszościowej, za pomocą algorytmu K-NN wybierani są jego sąsiedzi. Generowanie nowego syntetycznego egzemplarza danych polega na obliczeniu różnic pomiędzy wybranym, oryginalnym egzemplarzem danym a wylosowanym jego sąsiadem – analogiczny proces jest dla danych binarnych.

V.2.1.6 Random Over Sampler

Algorytm Random Over Sampler (ROS) działa na zasadzie losowego powielania egzemplarzy klasy mniejszościowej [91]. Tak wygenerowane nowe egzemplarze dodawane są do zbioru uczącego. Egzemplarze z klasy mniejszościowej mogą być wybierane wielokrotnie aż do powstania zrównoważonego zbioru z jednakową ilością egzemplarzy w poszczególnych klasach.

Metodę ROS można bez modyfikacji zastosować do danych binarnych.

V.2.1.7 Random Under Sampler

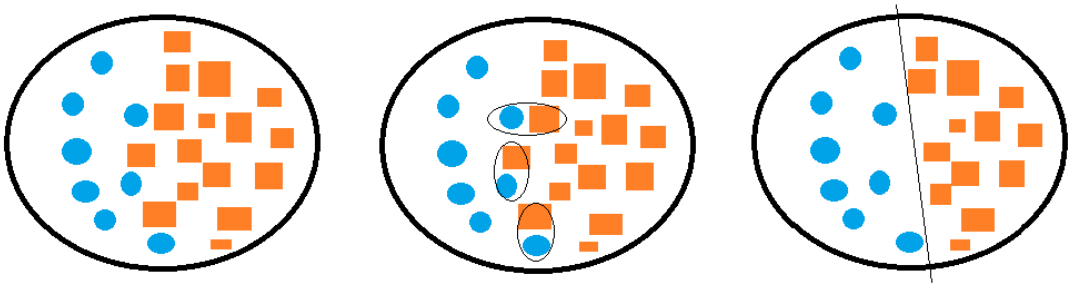
Działanie algorytmu Random Under Sampler (RUS) jest zasadniczo podobne do działania algorytmu ROS, z tym, że zasadniczym celem jest zmniejszenie ilości klasy większościowej poprzez losowe usunięcie egzemplarzy danych [91], tak aby zbiór uczących był zrównoważony (posiadał jednakową ilość egzemplarzy danych we wszystkich istniejących klasach).

Metodę RUS można bez modyfikacji zastosować do danych binarnych

V.2.1.8 Tomek Links

Technika undersamplingu (redukcji zbioru uczącego), jedna z modyfikacji algorytmu CNN (Condensed Nearest Neighbors) [96]. Technika ta łączy ze sobą punkty znajdujące się ze sobą w najbliższym sąsiedztwie, lecz należących do różnych klas, tak by punkt z klasy sąsiedniej znajdował się tuż za granicą klas. Sąsiedztwo określne jest poprzez najkrótszą, euklidesową odległość. Tak wyselekcjonowane punkty z klasy

większościowej usuwane są ze zbioru oryginalnego. W przypadku danych binarnych, egzemplarze dane są uprzednio sortowane – każdej obserwacji binarnej, przypisywana jest jej najbliższa obserwacja przeciwna.



Rysunek 54. Przykład działania algorytmu Tomek Links, źródło: opracowanie własne

V.2.2 Metoda weryfikacji

Metoda weryfikacji działania mechanizmu detekcji wiadomości phishingowych podzielona zostanie na kilka następujących po sobie etapów:

1. Wytworzenie zbioru danych na podstawie zgromadzonych wiadomości email. W trakcie tworzenia zakodowanego wektora cech, każdemu egzemplarzowi danych zostanie nadana odpowiednia etykieta:
    - phishing – dla wiadomości, które wstępnie zostały zakwalifikowane jako wiadomości o charakterze phishingowym (w tym wiadomości z szantażem, wymuszeniem okupu, itp.),
    - spam – wiadomości zakwalifikowane jako spam, niechciana korespondencja,
    - normal – wiadomości stanowiące normalną korespondencję email.
2. Analizę uzyskanych cech pod kątem poprawności ich zakodowania. Analiza ta obejmowała będzie również wprowadzenie ewentualnych poprawek do modułów odpowiedzialnych za odczytywanie wiadomości email, ekstrakcję cech i ich kodowanie.
3. Trenowanie klasyfikatorów pozyskanym zbiorem zakodowanych cech i obliczenie jakości ich przewidywań (accuracy). Wybranych zostało 5 modeli, na których prowadzone będą badania:
    - Model Regresji Logistycznej,

- Model Lasów Losowych,
- Model Drzewa Decyzyjnego,
- Model Maszyny Wektorów Nośnych,
- Model Naiwnego Klasyfikatora Bayesa.

4. Ocenę jakości przewidywań (patrz: tabele jakości przewidywań w opisie etapów) poszczególnych klasyfikatorów. Ocena ta obejmował będzie również wprowadzanie zmian i modyfikacji mających na celu poprawę jakości przewidywań danego modelu.
5. Z uwagi na występowanie niezrównoważonych klas mniejszościowych w posiadanym zbiorze uczącym, zostanie dokonane porównanie pomiędzy jakością klasyfikacji na oryginalnym, niezrównoważonym zbiorze oraz zostaną przeprowadzone badania jakości z wykorzystaniem algorytmów dokonujących równoważenia klas.
6. Porównanie zostanie dokonane dla dwóch podejść równoważnia zbioru:
    a. Wczytanie oryginalnego, niezrównoważonego zbioru a następnie dokonywanie „w locie" jego równoważenie. Tak zbalansowany zbiór podany zostanie na wejście modelu klasyfikacji.
    b. Dokonanie równoważenia zbioru za pomocą metody SMOTE i zapisanie go jako nowy zbiór. Tak powstały nowy zbirów po wczytanie podany zostanie na wejście modelu klasyfikacji. Wybór tej metody podyktowany został dużą różniącą ilości egzemplarzy danych pomiędzy klasą mniejszościową „normal" a klasą większościową „phishing", a metoda zapewnia generowanie syntetycznych modeli egzemplarzy danych o unikalnych cechach klasy mniejszościowej.
    c. Porównanie wyników poszczególnych modeli dla obu podejść (patrz: Rysunek 62).
7. Równoważenie zbioru uczącego przeprowadzone zostanie po ewentualnych etapach związanych z wprowadzaniem poprawek funkcji kodujących cechy czy modeli klasyfikatorów.
8. Opracowanych zostanie kilka metoda równoważenia danych. Te same metody zastosowane zostaną do równoważnia oryginalnego zbioru uczącego dla każdego z wybranych modeli klasyfikatorów. Wybrane metody równoważenia to:
    - SMOTE,
    - SMOTE-ENN,

- ADASYN,
- Random Over Sampler (ROS),
- Random Under Sampler (RUS),
- Tomek Links.

9. Opisy czynności podejmowanych w poszczególnych etapach będą opisane w dalszej części pracy i oznaczone jako „etap 1”, etap 2”, itd. Ilość etapów uzależniona będzie od uzyskiwanych wyników, konieczności wprowadzania poprawek

Posiadany zbiór zawierający dane uczące w trakcie treningu poddano następującemu procesowi:

a. Wczytanie zbioru danych uczących (zbiór oryginalny) – ten sam zbiór będzie wczytywany na wejście wszystkich metod równoważnie zbiorów oraz po zrównoważeniu za pomocą danej metody będzie służył jako dane trenowania klasyfikatorów.
b. Równoważenie zbioru oryginalnego za pomocą wymienionych powyżej metod. Jeden z obiektów oryginalnych danych (kopania wczytanych egzemplarzy) nie został poddany równoważeniu przez żadną z metod. Trenowanie klasyfikatorów na tak niezrównoważonym zbiorze (oryginalnym) jest celem porównania wyników zwracanych przez dany model w stosunku do wyników tego samego modelu, ale na danych poddanych procesowi równoważnia zbiorów.
c. Podział zrównoważonego zbioru na zbiór treningowy i zbiór testowy w proporcjach **80** (treningowy) : **20** (testowy).
d. Rozpoczęcie procesu uczenia się klasyfikatorów na zbiorze uczącym.
e. Testowanie stanu rozpoznawania nowych danych po zakończonym procesie uczenia się (wykorzystanie zbioru testowego).
f. Obliczenie parametrów dopasowania nowych danych do nauczonego wzorca i wyświetlenie wartości.
g. Obliczenie i wyświetlenie macierzy pomyłek dla wszystkich klasyfikatorów, podstawie których obliczony zostanie wskaźnik accuracy. Wskaźnik ten prezentowany będzie w tabelach opisujących progres modeli klasyfikacji (dla poszczególnych etapów). Wartości macierzy pomyłek zaprezentowane zostaną w Dodatku D.

Wskaźnik accuracy określa poprawność przewidywania danego modelu. Wskaźnik ten zdefiniować można jako liczba poprawnych przewidywań do liczby całkowitych przewidywań modelu (dla wszystkich występujących klas w danych).

$$accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{5.9}$$

Ekstrakcja cech, na podstawie których można zidentyfikować wiadomości phishingowe zaimplementowana została z wykorzystaniem języka programowania Python. Z tego powodu, implementacja modelu klasyfikacji zostanie również wykonana z wykorzystaniem tego samego języka z powodu łatwości integracji wewnątrz jednolitego środowiska poszczególnych modułów uczących. Język ten oferuje również szereg zestawów bibliotek *machine learning*, które uchodzą za jedne z najlepszych do uczenia maszynowego[190].

V.2.3 Wynik weryfikacji

W wyniku przeprowadzonych badań na podstawie opisanych czynności przyjętych jako metoda weryfikacji mechanizmu detekcji kampanii phishingowych, zrealizowanych zostało sześć etapów, które obejmowały:

1. etap 1 – początkowe testy wybranych modeli, bez strojenia parametrów uczenia się oraz na zestawie cech (długość wektora na tym etapie wynosiła 16 cech), które początkowo nie zostały zoptymalizowane i posiadały pewne drobne błędy logiczne systemu regułowego:
   1.1. błędnie interpretowano długi czas oczekiwania odpowiedzi dla zapytania o czas założenia domeny – wykorzystano bibliotekę *whois* języka Python, która dla niektórych zapytań o dane identyfikacyjne domeny zwracała w przypadku niepowodzenia wartość nieustaloną, interpretowaną jako TRUE (prawda),
   1.2. błędnie interpretowano generowane wyjątki przez bibliotekę *ipinfo* dla puli adresów prywatnych oraz zarezerwowanych,[191]

[190] https://www.kdnuggets.com/2023/01/7-best-libraries-machine-learning-explained.html
[191] Zarezerwowane adresy IP – adresacja IP zarezerwowana do określonych usług, która nie może być przydzielona indywidualnym komputerem do komunikacji sieciowej (np. adres localhost – 127.0.0.1, loopback -127.x.x.x, adres domyślny – 0.0.0.0 czy adresacja usług systemowych – 169.x.x.x.).

1.3. występowały błędy w metodzie wykrywania adresu portfela BitCoin i tym samym niewłaściwe kodowanie cechy,

1.4. etap ten nie uwzględniał również równoważenia zbioru.

Tabela 43. Wyniki klasyfikatorów etapu 1

Nazwa klasyfikatora	Accuracy
Maszyna Wektorów Nośnych (SVM),	**83.125**
Naiwny klasyfikator Bayesa - Gaussian	**27.5**
Wielomianowy Naiwny klasyfikator Bayesa	**77.5**
Drzewo decyzyjne,	**83.125**
Lasy losowe	**83.125**
Regresja logistyczna	**0.0**

2. etap 2 – wprowadzenie do poszczególnych modeli funkcji pozwalającej na wybór optymalnego parametru uczenia się oraz poprawiono wszystkie wykryte w etapie pierwszym błędy kodowania cech. W etapie tym, całkowicie zmieniono również sposób kodowania poszczególnych cech. Wprowadzone zmiany obejmują:

   2.1 dodanie obsługi wyjątków generowanych przez bibliotekę *whois,*

   2.2 dodanie obsługi wyjątków generowanych przez bibliotekę *ipinfo* odpowiedzialną za odczyt numeru ASN, na podstawie którego dokonywana jest identyfikacja czy adresy nadawcy pochodzą od tego samego operatora – wykluczono pulę adresów prywatnych oraz zarezerwowanych, dla których biblioteka zgłaszała wyjątki,

   2.3 zmiana elementów wektora cech – długość wektora cech na tym etapie: 19 cech, zmiany obejmowały:

      a. wykrywanie czy wiadomość email została wysłana za pomocą narzędzi programowych,
      b. wykrywanie czy w treści wiadomości umieszczono elementy imitujące tagowanie wiadomości email (fałszywe tagowanie),
      c. wykrywanie czy nazwa użytkownika z pola nadawcy (pole „From”/”Od”) została używa w treści wiadomości (np. jako element podpisu wiadomości w stopce).

   2.4 Dodano oraz zmieniono sposób kodowania cech: długości odnośnika URL, długości nazwy domeny (wraz ze subdomenami) oraz ilości kropek – „.” oraz ukośników (slash) – „/” w adresach domenowych i odnośnikach URL. Wartość danej cechy jest wynikiem działania modułów uczenia

maszynowego, określającego wartość cechy (0 lub 1) na podstawie pozyskanego zbioru uczącego zawierającego odnośniki URL oraz domeny wraz ze subdomenami z przypisanymi im etykietami (phishing lub normal). Jako algorytm uczenia wykorzystany został Algorytm Maszyny Wektorów Nośnych (SVM).

2.5 Z uwagi na charakter danych (wektor cech kodowanych wartościami dyskretnymi) jako Naiwny Klasyfikator Bayesa wybrano Multinomial Naive Bayes, który w pierwszym etapie osiągnął lepszy wynik uczenia niż Gaussian Naive Bayes (77.5 do 27.5). Należy natomiast pamiętać, że w tym etapie wartości poszczególnych cech są bardziej precyzyjnie i dokładniej wyliczane, stąd też i stopnień nauczenia się poszczególnych wersji Naiwnego Klasyfikatora Bayesa może się różnić od wyników zaprezentowanych w etapie nr 1.

2.6 Zmieniono funkcję odpowiedzialną za wykrywanie w treści wiadomości email adresu portfela BitCoin – dodano szereg warunków sprawdzających czy wykryty ciąg jest rzeczywiście możliwym adresem portfela BitCoin.

2.7 Zmieniono sposób kodowania cechy informującej o możliwym szantażu – po wykryciu w treści wiadomości adresu portfela BitCoin, treść poddawana klasyfikacji z wykorzystaniem Naiwnego Klasyfikatora Bayesa. W celu poprawnej klasyfikacji, zgromadzono zbiór treści wiadomości zawierających groźbę ujawnienia rzekomo kompromitujących materiałów, a w zamian za nie ujawnienie tych materiałów żądanie wpłaty określonej sumy w postaci kryptowaluty BitCoin. Z uwagi na konieczność wyselekcjonowania z treści wiadomości jedynie słów, na podstawie których oceniane będzie czy dana wiadomość zawiera groźbę/szantaż czy też nie, rozbudowano algorytm przetwarzający wiadomość o metody które normalizują treść:

   a. usunięcie formatowania tekstu (kodowanie treści za pomocą języka HTML, usunięcie formatowania CSS),
   b. usunięcie wszelkich znaków interpunkcyjnych,
   c. usunięcie znaków specjalnych,
   d. usunięcie cyfr (z wyjątkiem adresu portfela BitCoin),
   e. usunięcie wyrazów o liczbie liter mniejszej lub równiej liczbie 3 (np. że, z, ale, itp.) – wyraz te często pełnią rolę łączników i nie

niosą żadnej wartości informacyjnej, której istnienie by się przysłużyła w procesie klasyfikacji do określenia typu wiadomości.

2.8 Zmieniono funkcję odczytującą temat wiadomości i wprowadzono mechanizm odczytujący temat zapisany w kilku niezależnych liniach (dotychczasowe założenie nie weryfikowało zapisania tematu w więcej niż jednej linii).

2.9 Uzupełniono listę adresów serwisów oferujących usługę skracania odnośnika URL.

2.10 Poprawiono metodę odczytywania różnych części wiadomości email, zwłaszcza posiadających treści zapisanej za pomocą różnego kodowania: części tekstowej i części formatowej z wykorzystaniem języka HTML. Z uwagi, że obie te części wiadomości mogą być nośnikiem różnych treści, wprowadzono metodę całkowicie usuwającą formatowanie HTML oraz dodano metodę porównującą treść z części tekstowej oraz treść z części formatowanej HTML (po uprzednim jej usunięciu). Oprócz różnic pod względem innej zawartości treści, każda z części wiadomości email może posiadać inną stronę kodową tekstu, dlatego konieczne okazało się dodanie sposobu odczytu strony kodowej dla każdej części wiadomości oddzielenie.

2.11 Poprawiono funkcję odpowiedzialną za rozpoznanie kodowania treści formatu Base64. Zapis sposobu kodowania może zawierać się w więcej niż jednej linii, dlatego koniczne okazało się wykrywanie wszelkich form zapisu kodowania z wykorzystaniem metody Base64

2.12 Dodano funkcję dokonującą dekodowania treści tematu wiadomości email, zakodowanego za pomocą Base64. Poprawne odczytanie treści wiadomości konieczne jest do zakodowania cechy wykrywającej czy w temacie nie występuje nazwa użytkownika z adresu email odbiorcy wiadomości (pole "Delivered-To").

2.13 Dokonano usunięcia duplikatów wiadomości (pod względem treści). Powtarzająca się identyczna treść wiadomości email (z lekkimi modyfikacjami początku lub końca tytułu wiadomości, inną datą wysyłki oraz

innym adresem IP serwera[192] nadawcy) nie zwiększa różnorodności wykrywanych cech, a jednocześnie zwiększa ilość (powiela) cechy już zidentyfikowane.

2.14 Etap nie uwzględniał równoważenia zbioru.

Tabela 44. Wyniki klasyfikatorów etapu 2 z wprowadzonymi poprawkami.

Accuracy / Klasyfikator	Bez poprawek reguł kodowania cech	Z wprowadzonymi poprawkami	zmiana
Maszyna Wektorów Nośnych (SVM),	77,60736196319019	79,34782608695652	+ 1,7404641237663
Wielomianowy Naiwny klasyfikator Bayesa	76,07361963190185	72,82608695652174	- 3,2475326753801
Drzewo decyzyjne	78,22085889570552	80,07246376811594	+ 1,8516048724104
Lasy losowe	79,44785276073619	80,43478260869565	+ 0,9869298479595
Regresja logistyczna	74,84662576687117	78,2608695652174	+ 3,4142437983462

3. etap 3 – wprowadzenie testowych funkcji wykonujących równoważnie klas (dla obu zbiorów: uczących i testowych) dla różnych klasyfikatorów. Z uwagi na występujący problem nierówności zbioru uczącego (z dominacją jednej klasy nad innymi), moduł uczenia maszynowego rozbudowano o funkcję dokonujące równoważenia danych przed rozpoczęciem procesu uczenia się.

Tabela 45. Wyniki klasyfikatorów z wprowadzonymi metodami transformacji danych.

Nazwa klasyfikatora	Metoda równoważenia	Accuracy	zmiana
Maszyna Wektorów Nośnych (SVM),	Random Over Sampler	66.37931034482759	negatywna
Maszyna Wektorów Nośnych (SVM),	Standard Scaler	79.34782608695652	neutralna
Wielomianowy Naiwny klasyfikator Bayesa		58.96551724137931	negatywna
Drzewo decyzyjne,		80.43478260869566	pozytywna
Lasy losowe		64.4927536231884	negatywna
Regresja logistyczna		8.333333333333333	negatywna

[192] Wiadomości email wysłane z tego samego serwera pocztowego lub z tej samej infrastruktury operatora oferującego usługi poczty email, mogą posiadać różne adresy IP. Dzieje się tak, dlatego, ze operatorzy często w ramach balansowania ruchu sieciowego korzystają z wielu różnych fizycznych urządzeń, które posiadają różne adresy IP widoczne jako jednio urządzenie wirtualne. Z tego powodu zaproponowane rozwiązanie weryfikuje różnice nie w adresacji IP a w przydzielonych numerach ASN (ten sam operator na wszystkich swoich urządzeniach i zarządzanych przez niego adresach IP będzie miał przypisany ten sam numer ASN).

4. etap 4 – porównanie wykrywania cech dla poszczególnych klas. Analizując wykryte i zakodowane cechy dla poszczególnych typów wiadomości oraz weryfikując poszczególne zbiory, dokonano niewielkiej zmiany liczebność zbiorów:
    a. zbiór oznaczony jako „phishing" (wiadomości wstępnie uznane za phishingowe) – zwiększono liczebność zbioru z 966 wiadomości do 981. Jedna wiadomość nie została odczytana – przyczyną nieodczytania był błąd wewnętrzy podczas komunikacji biblioteki *whjois* z zewnętrznym serwisem internetowym. Na podstawie ponownej weryfikacji poszczególnych zbiorów, 6 wiadomości wstępnie oznaczonych jako „normal" przeniesiono do zbioru phishing z powodu występowania cechy wskazującej na możliwy phishing.
    b. zbiór oznaczony jako „normal" (wiadomości stanowiące normalną korespondencję) zmniejszono liczebność zbioru z 90 wiadomości do 84 wiadomości. Cztery wiadomości nie zostały odczytane. Przyczyną nieodczytania było wykrycie niewłaściwego kodowania pliku.

Tabela 46. Wyniki klasyfikatorów etapu 4

Accuracy / **Klasyfikator**	**Bez poprawek reguł kodowania cech**	**Z wprowadzonymi poprawkami**	**zmiana**
Maszyna Wektorów Nośnych (SVM),	79.34782608695652	81.36200716845879	+ 2,01418108150227
Maszyna Wektorów Nośnych (SVM), z wykorzystaniem transformacji	79,34782608695652	80.64516129032258	+ 1,29733520336606
Wielomianowy Naiwny klasyfikator Bayesa	72,82608695652174	77.06093189964157	+ 4,23484494311984
Drzewo decyzyjne,	80,07246376811594	80.3987395467216	+ 0,326275778605663
Lasy losowe	80,43478260869565	80.28673835125448	- 0,148044257441171
Regresja logistyczna	78,2608695652174	79.56989247311828	+ 1,30902290790088

5. etap 5 – wprowadzone poprawki uwzględniające przeprowadzoną analizę wykrywalności poszczególnych cech etapu 4. Analizując źródłowy wektor zakodowanych cech, będący zbiorem danych uczących, pod względem prawidłowości ich wykrycia i zakodowania, nie stwierdzono zakodowania cechy nr 4 – odnośnik URL zapisany w formie skróconej (skracacz linku), pomimo jej faktycznego występowania w źródłowym zbiorze pierwotnym. Analiza

przyczyny braku cechy w wektorze wykazała, że specyfika funkcji wykorzystywanej w języku Python do wczytywania zbioru domen serwisów skracających nie ignorowała znaku końca linii. Dodawany biały znak – „\r" – do odczytanej domeny, powodował, że funkcja nie rozpoznawała występowanie skracacza linku w przetwarzanej wiadomości email. Usprawnienie funkcji spowodowało wykrycie występowania tej cechy w zbiorze wiadomości, a tym samym w ponownym procesie uczenia się poszczególnych klasyfikatorów, wpłynęło na ich wynik.

Tabela 47. Wyniki klasyfikatorów etapu 5.

Accuracy / Klasyfikator	Bez poprawek reguł kodowania cech	Z wprowadzonymi poprawkami	zmiana
Maszyna Wektorów Nośnych (SVM),	81.36200716845879	80.6451612903225	- 0,716845878
Maszyna Wektorów Nośnych (SVM), z wykorzystaniem transformacji	80.64516129032258	82.07885304659499	+ 1,433691756
Wielomianowy Naiwny klasyfikator Bayesa	77.06093189964157	78.13620071684588	+ 1,075268817
Drzewo decyzyjne,	80.3987395467216	80.30864945663151	- 0,09009009
Lasy losowe	80.28673835125448	80.64516129032258	+ 0,358422939
Regresja logistyczna	79.56989247311828	79.56989247311828	Bez zmian

6. etap etap 6 – poszerzono zbiory posiadanych wiadomości email we wszystkich klasach („phishing", „spam" i „normal") oraz zastosowano selekcję by poprawić jakość danych stanowiących wartość wektora cech (unikalność cech). Zastosowano metody równoważenia zbioru wejściowego by zniwelować problem nierówności zbioru uczącego (przewaga większościowej klasy „phishing"). Badaniu podlegały dwa podejścia równoważenie zbioru:
    a. wykorzystanie różnych technik równoważenia zbioru (po jego wczytaniu),
    b. utworzenie nowego, zrównoważonego zbioru na bazie zbioru oryginalnego.

Wykorzystanie technik równoważenia zbiorów

Poszerzony o nowe wiadomości (we wszystkich trzech klasach: „phishing", „normal", „spam"), w dalszym ciągu zawierał przewagę ilościową jednej z klas

(klasa „phishing”), co może być pewnym utrudnieniem w procesie uczenia. W tym celu dla każdego wykorzystywanego klasyfikatora, opracowano metody transformacji danych by zniwelować problem nierówności klas. Do transformacji wykorzystano algorytmy:

1. SMOTE,
2. SMOTE-ENN,
3. ADASYN,
4. Random Over Sampler,
5. Random Under Sampler,
6. Tomek Links.

Tabela 48. Wyniki klasyfikatorów etapu 6 - balansowanie oryginalnego zbioru

	Regresja logistyczna (accuracy)	**Lasy losowe (accuracy)**	**Drzewo Decyzyjne (accuracy)**	**SVM (accuracy)**	**Naiwny Klasyfikator Bayesa (accuracy)**
Wartość bez balansowania	79,71530	80,42705	80,42705	80,78292	79,35943
SMOTE	82,82313	82,82313	82,82313	83,16327	81,97279
SMOTE-ENN	100	100	99,73545	100	64,55026
ADASYN	81,43813	81,27090	81,27090	81,10368	68,89632
Random Over Sampler	82,31293	82,82313	82,48299	82,99320	82,48299
Random Under Sampler	82,35294	82,35294	82,35294	82,35294	66,66667
Tomek Links	79,71530	80,42705	80,42705	80,78292	79,35943

gdzie:

	pogorszenie
	poprawa wyniku
	bez zmian

Efektem działań algorytmów równoważących klasy zbioru uczącego jest poprawa wykrywania wiadomości phishingowych na podstawie zakodowanego wektora zidentyfikowanych cech. Najbardziej efektywnym algorytmem równoważenie okazała się metoda SMOTE – która przyniosła poprawę dla wszystkich klasyfikatorów

(zauważalne wyższe wartości algorytmu SMOTE-ENN dotyczyły 4 z 5 wykorzystywanych w procesie uczenia klasyfikatorów, dla Naiwnego Klasyfikatora Bayesa, odnotowano zauważalnie niższe wartości niż dla procesu uczenia na oryginalnym, niezrównoważonym zbiorze). Generowanie nowych egzemplarzy danych na bazie sąsiedztwa istniejących z wykorzystaniem metody SMOTE, spowodowało, że do zbioru uczącego dodane zostały egzemplarze o unikalnych cechach, co przełożyło się na jakość klasyfikacji.

Utworzenie nowego zbioru.

Problem niezbalansowania zbioru danych uczących można również zniwelować stosując zbalansowanie zbioru z wykorzystaniem metody SMOTE. Posiadany zbiór uczący, powstały na bazie odczytu wiadomości email i zakodowaniu cech mogących wskazywać na atak phishingowy poddano procesowi równoważnia ilościowego klas. Nowe, wygenerowane syntetyczne obiekty zostały wraz z danymi oryginalnymi zapisane jako nowy zbiór danych uczących.

Tabela 49. Wyniki klasyfikatorów etapu 6 - balansowanie zbioru przez wczytaniem danych.

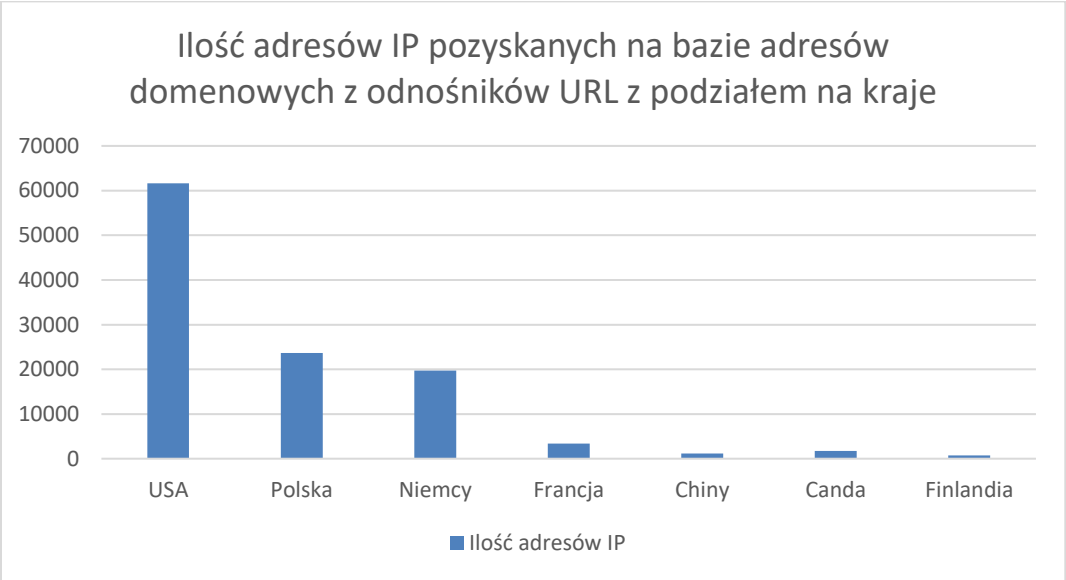
	Regresja logistyczna	Lasy losowe	Drzewo Decyzyjne	SVM	Naiwny Klasyfikator Bayesa
Wartość bez balansowania	79,71530	80,42705	80,42705	80,78292	79,35943
Wartości z balansowaniem	85,37415	84,86395	85,37415	85,54422	85,37415
% różnicy	**+6,63**	**+5,23**	**+5,79**	**+5,57**	**+7,05**

Równoważenie zbioru – zarówno algorytmami na oryginalnym zbiorze uczącym, jak i zbalansowanie samego zbioru - przynosi poprawę uczenia się poszczególnych modeli – generowane są nowe egzemplarze danych, które nie występują w oryginalnym zbiorze danych

W trakcie przetwarzania treści wiadomości email, pozyskano 138579 adresów IP (z czego unikalnych adresów jest 8617). Adres IP pozyskiwany jest na podstawie:

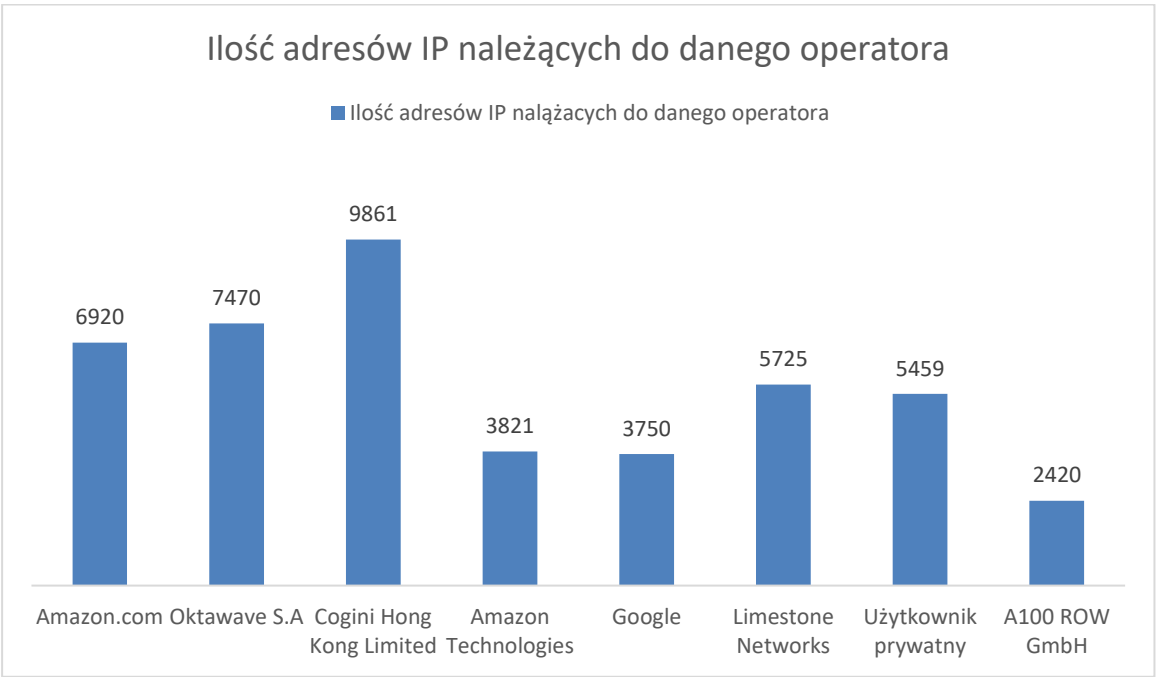
a. odnośników URL znajdujących się w treści danej wiadomości email,
b. adresów serwerów pośredniczących w wymianie korespondencji email i odpowiedzialnych za jej dostarczenie do odbiorcy końcowego,

c. pozyskanych na podstawie domeny adresów email znajdujących się w polach „From”, „Reply-To” (o ile występuje) i „Return-Path” (o ile występuje).



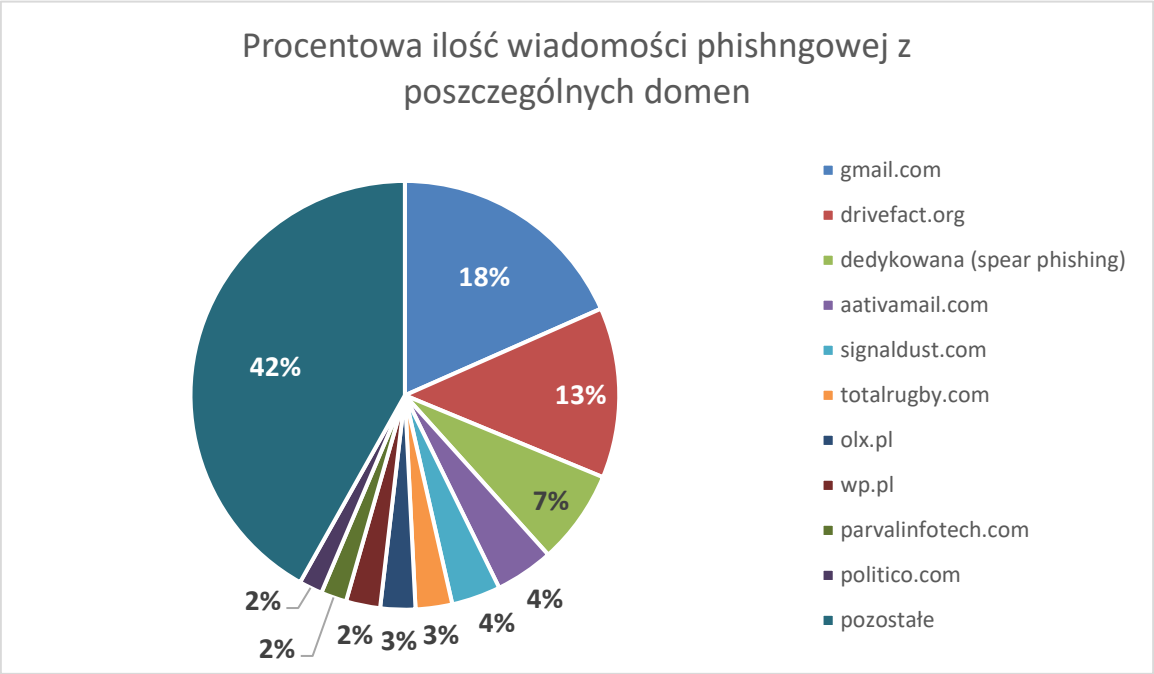
Rysunek 55. Ilość adresów IP pozyskanych na bazie adresów domenowych z odnośników URL z podziałem na kraje

Analizując dostawców usług (hostingu, operatorzy pocztowi, itp.), na podstawie uzyskanych adresów IP można zidentyfikować:



Rysunek 56. Ilość adresów IP należących do danego operatora.,

W trakcie działania algorytmu wyselekcjonowano również 1665 adresy email pochodzące ze zbioru uczącego oznaczonego jako wiadomości phishingowe.



Rysunek 57. Procentowa ilość wiadomości phishingowej z poszczególnych domen.

Z powyższego zestawienia wynika, że największy udział w dystrybucji wiadomości identyfikowanych jako phishingowe pochodzi z adresów email oferowanych w ramach usługi oferowanej przez firmę Google (gmail.com).[193] Na popularność tego operatora (i co za tym idzie wielkość przeprowadzanych ataków z wykorzystaniem adresacji do niego należącej) składa się kilka elementów:

1. Usługa założenia i prowadzenia konta email jest darmowa.
2. Wysoka dostępność usług – serwis oferowany przez największego operatora zapewnia nieprzerwalne działanie, co w przypadku kampanii phishingowych o krótkim cyklu życia jest istotne.
3. Anonimowość – wykorzystanie adresacji operatora do dystrybucji wiadomości w ramach danej kampanii phishingowej nie ujawnia wykorzystywanej adresacji atakującego. Dostęp do poczty może również odbywać się z wykorzystaniem tunelu VPN.[194]

[193] Potwierdzają to również raporty firm zajmujących się cyberbezpieczeństwem, m.in. Baracuda (https://blog.barracuda.com/2021/11/10/threat-spotlight-bait-attacks/) [dostęp: 20.05.2023].
[194] Na rynku istnieją serwisy oferujące anonimowy dostęp do usługi VPN, nie jest więc możliwe określenie lokalizacji użytkownika rzeczywistego (np. usługa CyberGhost, https://www.cyberghostvpn.com/en_US/).

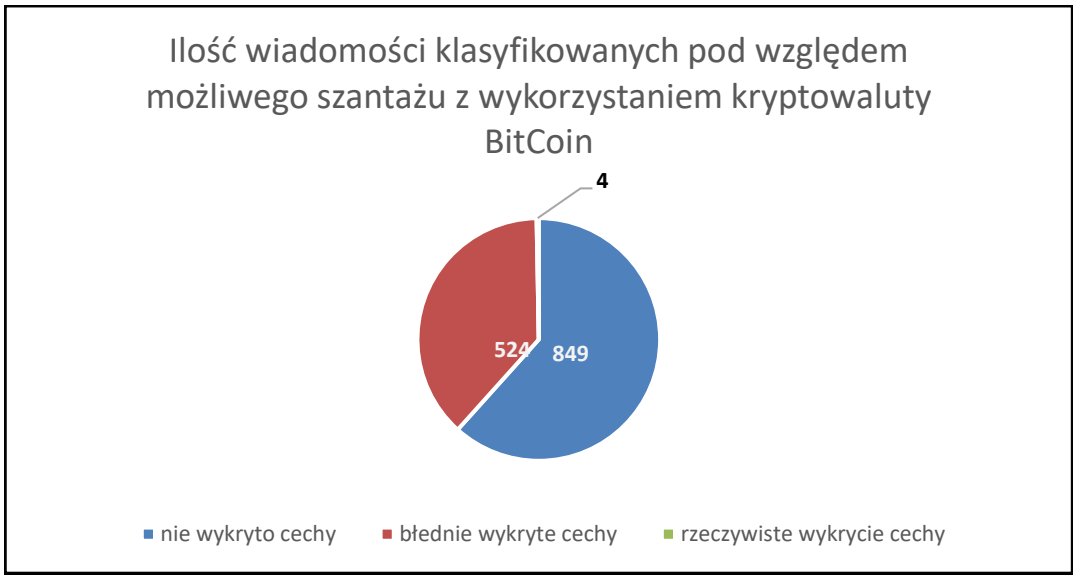
4. Możliwość kategoryzowania wysyłanych i otrzymywanych wiadomości – co ułatwia zarządzanie kampaniami phishingowymi.
5. Usypia czujność potencjalnej ofiary – otrzymanie wiadomości od innego użytkownika posługującego się popularnym adresem (gmail.com), traktowane jest jako bezpieczna forma komunikacji (poczta Google uchodzi na jedną z najbezpieczniejszych usług poczty elektronicznej).
6. Dostęp do szeregu innych usług powiązanych (m.in. Dysk sieciowy Google Drive umożliwiający hosting plików, skryptów, prezentacji, dokumentów, itp.) – co w połączeniu z inżynierią społeczną (zaufanie użytkowników do operatora) umożliwia skuteczniejsze przeprowadzenie kampanii.
7. „Email-przynęta" - schemat opisany jako atak responsywny – wysłanie wiadomości, które nie zawierają żadnej złośliwej treści co ma w pierwszej kolejności za zadanie sprawdzenie czy konto jest aktywne oraz sprawdzić czy użytkownik podejmie współpracę – odbiorcy zazwyczaj odpisują ma wiadomości zachęcające go kontynuowania korespondencji, gdy otrzymują taką wiadomość z konta email należącego do z konta znanego i popularnego operatora[195] [196], którego jakość usług ma wysoką renomę[197] i uchodzi za bezpieczną pocztę.

V.3 Wnioski

Wyniki poszczególnych etapów należy taktować rozłącznie (zwłaszcza nr 1, 2, 3) z uwagi na zachodzące zmiany w sposobie kodowania poszczególnych cech (dotyczy różnic w kodowaniu cech pomiędzy etapem nr 1 a etapem nr 2). Zmiany sposobu kodowanie nie dokonano pomiędzy etapem nr 2 a etapem nr 3 – wprowadzone zmiany to równoległe dodanie funkcji dokonujących transformację danych.

Zbyt mały zbiór uczący zawierający wiadomości o charakterze szantażu i wymuszających przesłanie okupu w formie przelewu kryptowaluty BitCoin, powodował generowanie błędnego kodowania cechy nr 7.

[195] https://www.wisestamp.com/blog/free-email-providers/
[196] https://mailchimp.com/resources/most-used-email-service-providers/
[197] https://www.softwaretestinghelp.com/email-service-providers/



Rysunek 58. Klasyfikacja wiadomości pod względem możliwego szantażu z wykorzystaniem kryptowaluty BitCoin

Błędna klasyfikacja i przypisanie wartości 1 cechy świadczącej o możliwym szantażu z wykorzystaniem kryptowaluty BitCoin na podstawie analizy treści, może być spowodowane kilkoma czynnikami:

a. zbyt mały zbiór uczący – do celów uczenia pozyskano 11 wiadomości, których treść jasno wskazuje na szantaż z wykorzystaniem kryptowaluty BitCoin,
b. zbyt jednorodny zbiór uczący – w posiadanym do wykorzystanie zbiorze wiadomości email, przeznaczonych do procesu uczenia, które zawierały treści szantażujące zaobserwowano jedynie niewielkie różnice w treści.

```
Podobieństwo między tekstami 1 i 3: 49.70
Podobieństwo między tekstami 1 i 5: 11.86
Podobieństwo między tekstami 1 i 7: 10.59
Podobieństwo między tekstami 1 i 8: 11.46
Podobieństwo między tekstami 1 i 9: 50.00
Podobieństwo między tekstami 1 i 11: 11.46
Podobieństwo między tekstami 2 i 4: 10.67
Podobieństwo między tekstami 2 i 6: 49.61
Podobieństwo między tekstami 2 i 10: 49.69
Podobieństwo między tekstami 2 i 11: 8.63
Podobieństwo między tekstami 3 i 5: 11.54
Podobieństwo między tekstami 3 i 7: 10.59
Podobieństwo między tekstami 3 i 8: 11.46
Podobieństwo między tekstami 3 i 9: 49.70
Podobieństwo między tekstami 3 i 11: 11.46
Podobieństwo między tekstami 4 i 5: 12.54
Podobieństwo między tekstami 4 i 6: 10.57
Podobieństwo między tekstami 6 i 10: 49.92
Podobieństwo między tekstami 7 i 9: 10.59
Podobieństwo między tekstami 8 i 9: 11.46
Podobieństwo między tekstami 8 i 11: 49.68
Podobieństwo między tekstami 9 i 11: 11.46
```

Rysunek 59. Podobieństwa treści wybranych wiadomości zbioru uczącego dla rozpoznawania wiadomości zawierającej szantaż.

Poszczególne wiadomości zawierają te same wyrazy używane w takiej samej frazie, podobnie brzmiącej lub mającej takie samo znacznie, co istotnie wpływa na wynik.

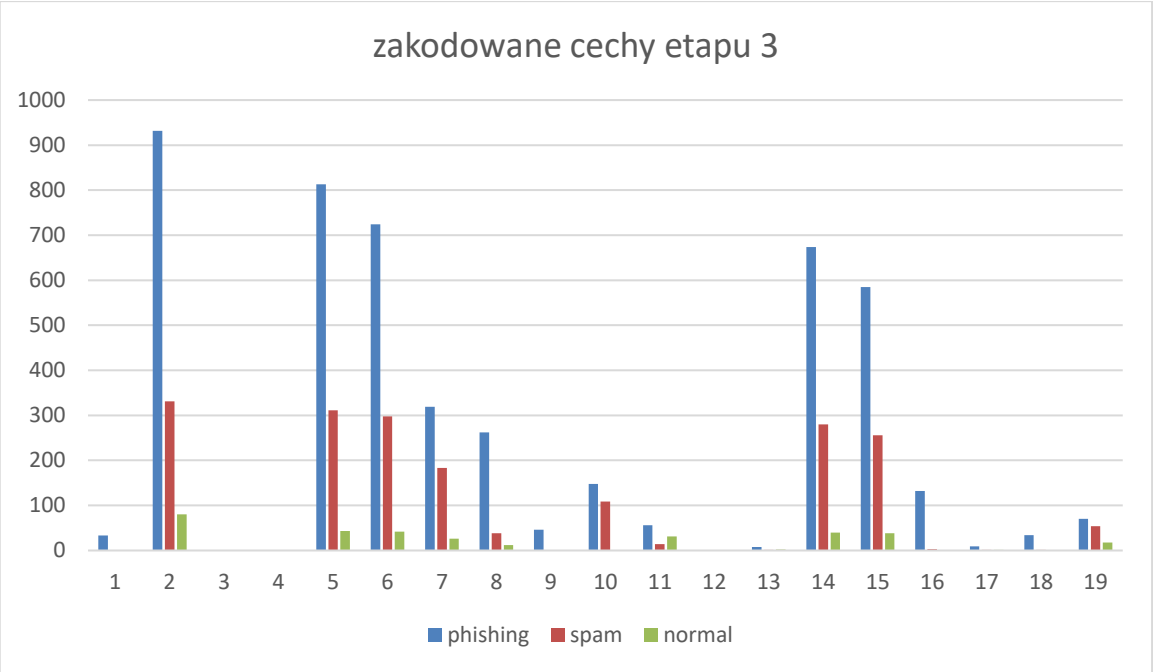
Tabela 50. Wyborne podobieństwa wiadomości stanowiących zbiór uczący.

Tekst nr 1	**Tekst nr 2**
Pozdrawiam to twoje ostatnie ostrzeżenie	Pozdrawiam to twoje ostatnie ostrzeżenie
Twojej listy kontaktów telefonicznych	Twojej listy kontaktów telefonicznych
wszystkie zdjęcia media społecznościowe czaty kontakty	zdjęcia dostęp wszystkich komunikatorów także poczty elektronicznej

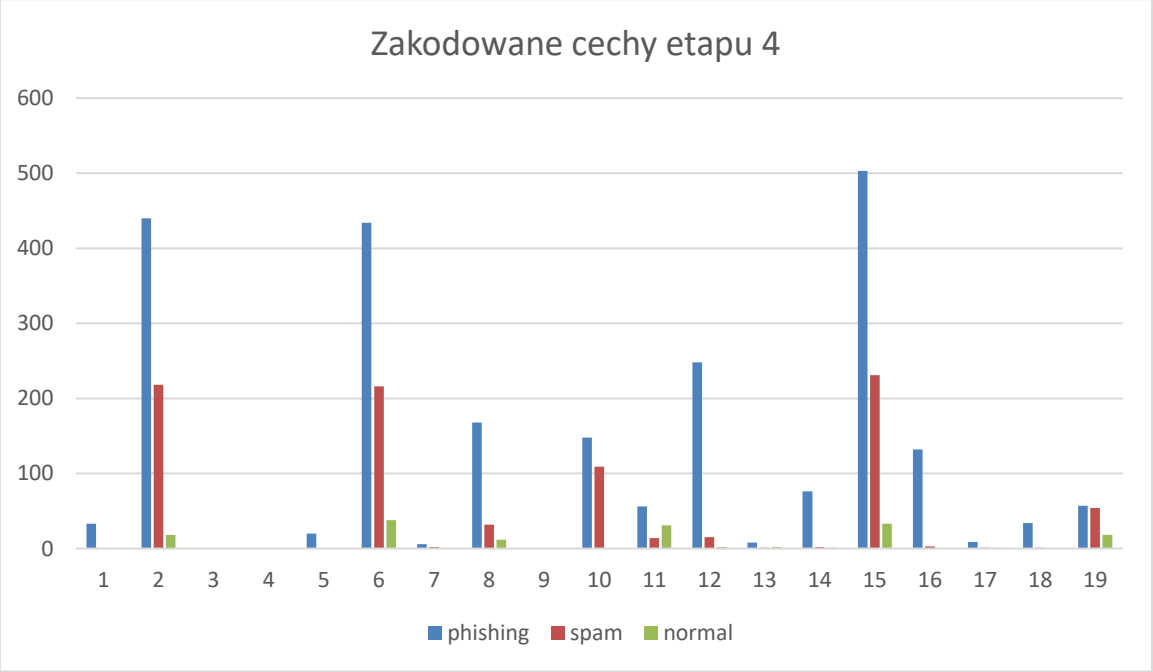
c. Częste występowanie wyrazów znajdujących się w wiadomościach stanowiących zbiór uczący w innych wiadomościach, które zarówno mogą być wiadomościami phishingowymi jak i stanowić normalną korespondencję email.

Problem pogorszenia procesu nauki [97] większości wybranych klasyfikatorów, widoczny w etapie nr 3 był dostrzeżony i opisywany. Dane uczące – wektor cech phishingowych zakodowany danymi binarnymi (0,1) co wg autorów [97], nie jest optymalną techniką - „techniki normalizacji danych nie są pomocne w problemach klasyfikacją danych jednorodnych”.

Polepszenie jakości klasyfikacji etapu nr 4, możliwe było dzięki zastosowaniu lepszych warunków logicznych podczas wykrywania i kodowania poszczególnych cech. Pomimo nieznacznej różnicy w wielkości zbiorów (odpowiednio +15 dla phishingu oraz -6 dla klasy „normal”), wykresy przedstawione na Rysunek 60 oraz Rysunek 61 pokazują większe różnicę pomiędzy wykrytymi i zakodowanymi cechami odpowiednio dla etapu nr 3 i etapu nr 4, niż to wynika ze zmiany liczebności poszczególnych zbiorów.



Rysunek 60. Rozkład kodowanych cech dla etapu 3.



Rysunek 61. Rozkład kodowanych cech dla etapu 4.

Wysoka ilość wykrywalności cechy nr 15 (nagromadzenie błędów językowych) w zbiorze wiadomości. Duży odsetek zidentyfikowanych błędów znajdujących się we wszystkich typach wiadomości, na który miały wpływ:

a. Duża ilość anglicyzmów[198] – obserwowana w mowie potocznej ekspansja anglicyzmów wynika z globalnego wpływu angielszczyzny na inne języki (w tym język polski). Nowe zapożyczenia, nie posiadają jeszcze ustalonej pisowni[199], nie występują też w Słowniku Języka Polskiego, stad też brak jego występowania w bazie, taktowany jest przez algorytm kodujący cechy jako błąd językowy i wpływa na ogólną statystykę błędów danej, przetwarzanej wiadomości, która po przekroczeniu pewnego progu (ilości nagromadzonych błędów w pojedynczej wiadomości), interpretowana jest jako możliwy wskaźnik ataku phishingowego.
b. Przenikanie się terminologii technicznej do języka codziennego – fachowe słownictwo pochodzące z różnych języków (w tym z języka angielskiego, który odpowiada za większość występujących w mediach słów – anglicyzmów). Multikulturalne społeczeństwa Europy (w tym i Polski), posługujące się więcej niż jednym językiem, chętnie zapożyczają terminologię z jednego języka do drugiego.
c. Niedoskonałość tłumaczenia – firmy i osoby prywatne nie znające języka natywnego swojego kontrahenta, korzystają z serwisów oferujących automatyczne tłumaczenie tekstu (np. translate.google.com). Pomimo dużego zawansowania, w pewnych specyficznych wyrażeniach, fraz i zwrotów charakterystycznych wyłączenie dla danej nacji, system taki może nie podać właściwej interpretacji słowa - co dla odbiorcy posługującego się natywnie językiem, na który zostało przeprowadzenie tłumaczenie, okazuje się błędnym sformułowaniem. Dla docelowego odbiorcy, kontekst może być w pełni zrozumiały, jednakże funkcja sprawdzająca, w takich przypadkach zwracała będzie błąd.
d. Stosowanie slangu[200], skróty myślowe, itp. – popularne zwłaszcza wśród młodzieży czy specyficznych grup dzielących wspólne zainteresowania, mogą pojawiać się jako forma komunikacji w wiadomościach email i nie będą wówczas

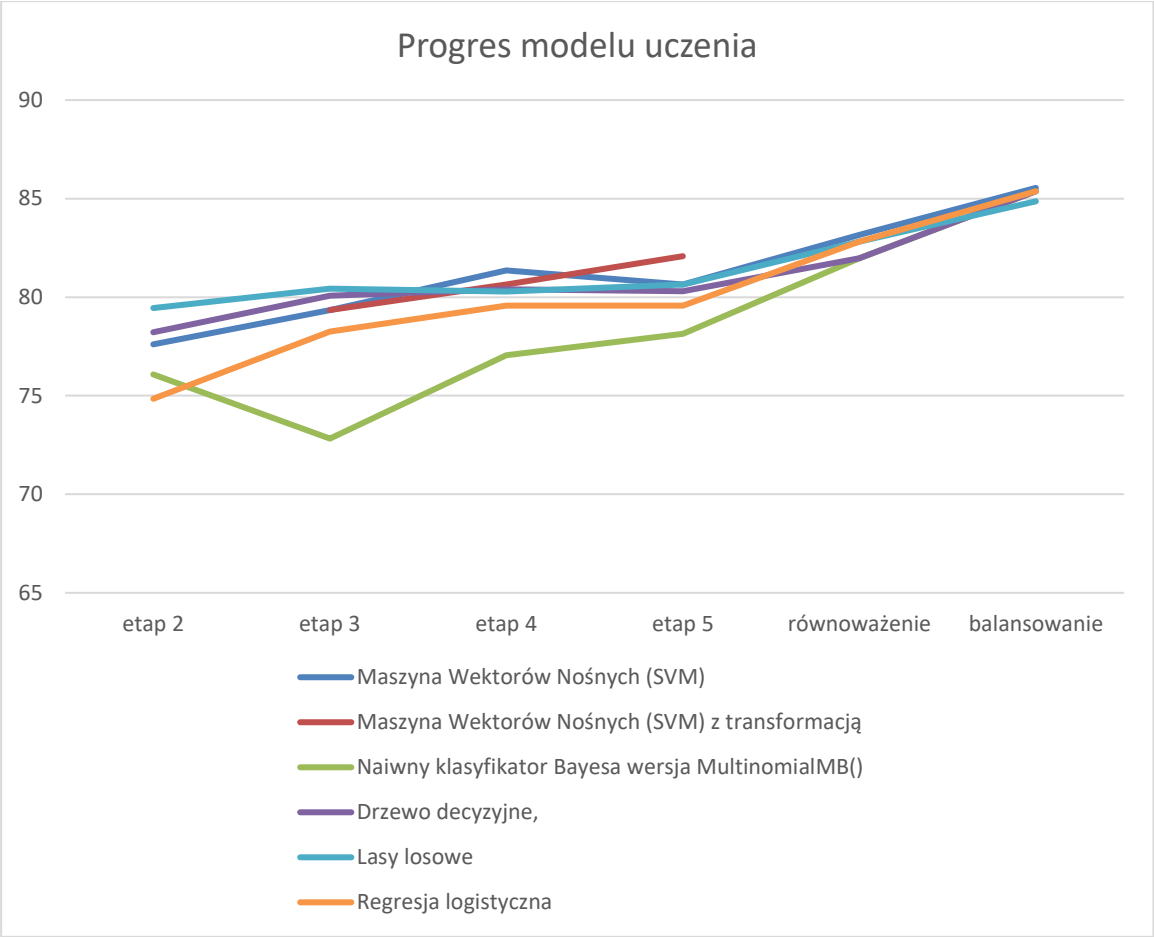
[198] Anglicyzm - element językowy występujący w języku polskim, zaczerpnięty z języka angielskiego.
[199] Wiele neologizmów (anglicyzm jest jednym z neologizmów – nowy element języka polskiego, wykształcony na podstawie języka angielskiego) stopniowo zadomawia się w mowie codziennej, co wiąże się z adaptacją również ich pisowni, dostosowanej do charakterystyki języka polskiego. Jednym z p[przykładów kształtowania się pisowni jest anglicyzm – wyraz „mecz", jeszcze w latach 30. XX wieku zapisywano je jako „match".
[200] W tym kontekście oznacza wyraz, zespół wyrazów, które są zrozumiałe jedynie dla pewniej wyodrębnionej grupy a stanowiące wynik pewnej spontanicznej twórczości słownej.

rozpoznane przez funkcję opartą na słowniku – pomimo potencjalnego pełnego jej zrozumienia przez docelowego odbiorcę.

e. Specyficzne ustawienia językowe – popularne w kręgach specjalistów IT stosowanie ustawień językowych wskazujących na język angielski wraz z typowym dla niego układem klawiatury. Tworząc wiadomość email w takim systemie, w którym domyślnym (i często jedynym) językiem jest język angielski, nie jest możliwe poprawne stosowanie polskich znaków diakrytycznych, więc finalnie wiadomość, zrozumiała będzie dla odbiorca końcowego. Funkcja przetwarzająca potraktuje tak wytworzona treść jako błędną.

Poprawa logiki systemu regułowego (szczegółowo przedstawiona w opisach dokonywanych zmian w poszczególnych etapach prowadzonego doświadczenia), przełożyła się również na poprawienie się procesu uczenia się poszczególnych klasyfikatorów (Rysunek 62).

Na przedstawionych wykresach (Rysunek 60 oraz Rysunek 61) uwidocznione są cechy, występujące jedynie w wiadomościach phishingowych (cechy nr: 1, 5, 7, 13, 14, 17, 18) – nie występuję one ani w zbiorze wiadomości typu spam ani w zbiorze normalnej korespondencji. Cechy te nie występują w dostępnej literaturze traktującej o problematyce phishingu. Z tego powodu istotnym autorskim wkładem w metodę detekcji phishingu jest ich zidentyfikowanie, opisanie i wykorzystywanie do kodowania wektora cech.



Rysunek 62. Progres modelu uczenia w miarę doskonalenia systemu regułowego.

Wszystkie powyższe obserwacje, uzależniają uzyskanie dobrego wyniku (czy to na etapie klasyfikacji, czy na wcześniejszym – wykrywaniu wskaźników i kodowaniu wektora cech), wymagają zestawu odpowiednio przygotowanych danych:

a. Źródeł wskaźników – wiadomości email, które zapewnią w danych uczących obecność takich egzemplarzy danych, które zawierać będą poszczególne wymienione i analizowane wskaźniki. Ilość egzemplarzy musi zapewniać natomiast obecność co najmniej jednego wskaźnika w całym zbiorze danych. Zebrany zbiór wiadomości, służących jako dane uczące procesu klasyfikacji, nie pokrywały całości opisanych i możliwych do identyfikacji wskaźników ataku phishingowego.

Tabela 51. Wykaz niezidentyfikowanych cech w zbiorze uczącym.

Lp.	Numer niezidentyfikowanej cechy	Opis cechy
1.	**3**	adres IP w odnośniku URL zawartym w wiadomości

2.	**9**	niewłaściwy adres nadawcy

Brak występowania adresu IP w odnośnikach URL, obecnie przesyłanych wiadomości email (w tym również o charakterze phishingowym) może być spowodowane głównie:

1. Łatwością rejestracji domeny – istnieje wiele firm hostingowych, dostawców usług, które w swoje ofercie posiadają możliwość wykupienia i rejestracji dowolnej[201] nazwy domenowej. Cena rejestracji nowej domeny jest stosunkowo niska – ceny zaczynają się już od 5.98$[202]
2. Dostępność subdomeny w ramach wykupionej (lub oferowanej za darmo) usługi hostingu. Usługodawca umożliwia (często automatycznie, na podstawie nazwy użytkownika do panelu logowania) utworzenie subdomeny dla klienta np.:

   `mojanazwa.nazwahostingu.tld`

   gdzie:
   tld – z ang. Top Level Domanin, domena najwyższego rzędu, np.: .pl, .com

3. Szeroką dostępnością adresów darmowych usług umożlwiających rejestrację nazwy domenowej (np. *.tk, *.ga, *.ml, *.cf). Wielu usługodawców[203] oferują darmową rejestrację domeny.
4. Łatwością wykrywania adresu IP w odnośniku URL [98] [99].
5. Wywarciem większego efektu – działania z wykorzystaniem inżynierii społecznej. Nazwa zakładanej domeny (subdomeny), jest podobna (lub zawiera frazę) innej domeny i w przypadku złożonych nazw jest to pierwszy człon, skupiający na sobie uwagę użytkownika. W ten sposób wykorzystująca nazwy domenowa atakujący skuteczniej wykorzystują nieuwagę / niewiedzę użytkownika – obecność adresu IP nie zadziała tak samo przekonywująco, dlatego atakujący odchodzą od tej techniki.

Niewłaściwy adres nadawcy, należy rozumieć jako adres email, który jest niewłaściwy w kontekście nazwy użytkownika, odpowiednio zawartej w wiadomości treści. Podczas kodowania tej cechy, przyjęto założenie, ze nazwa

[201] Pod pojęciem „dowolnej" należy rozumieć taką nazwę co do której nie występują roszczenia osób trzecich, nie jest obecnie wykupioną i zarejestrowaną, jak również jej nazwa nieobraźliwa.
[202] https://www.namecheap.com/promos/new-com-promo/ [dostęp: 29.07.2023].
[203] np. https://www.freenom.com/

użytkownika (pole „From”/”Od” zgodnie z dokumentami standaryzacyjnymi powinno zawierać konstrukcję „Nazwa użytkownika <adres email>” – lecz dopuszczane są również inne kombinacje tych wartości) może wystąpić w nazwie domenowej (1), lub też nazwa użytkownika jest pełnoprawnie zarejestrowaną domeną (2) z odmienią końcówką tld (top level domain).

Sprawdzenie wykonywano dla wydzielonej nazwy użytkownika, o ile występowała – nie wszystkie analizowane wiadomości email w polu „From”/”od” posiadały nazwę użytkownika, co też przekładało się na wykrywanie tej cechy w analizowanych zbiorach. Nazwa ta łączona była z domeną tld i dla tak utworzonej wartości dokonywano sprawdzenia w zewnętrznym serwisie (*whois* lub *ipinfo*). Prawidłowa odpowiedź od jednego z tych serwisów wskazała na istnieje domeny o wskazanej wartości. Samo istnienie domeny nie przesądzało natomiast o zakodowaniu jej jako możliwy atak phishingowy, gdyż warunkiem była różnica w numeracji ASN.

Przyczyną braku występowania tej cechy w analizowanych zbiorach jest:

a. Korzystanie z usług hostujących u tego samego dostawy - popularne rozwiązanie obniżające koszty utrzymywania własnej domeny. Usługodawca oferując swoje usługi hostingu, posiada wykupioną pewną pulę adresacji IP, będącej w jego zarzadzaniu. Cała ta pula, w polu „org name[204]” posiada identyczny wpis związany z numerem ASN. Ten sam usługodawca oferuje również hosting lub serwery VPN z którego korzystają cyberprzestępcy. Działanie takie zwiększa możliwości pozostania niewykrywalnym z powodu dobrej reputacji danego adresu.
b. Innym sposobem kodowania liter – wykorzystanie złudzenia optycznego człowieka, które bazuje na wizualnym podobieństwie pewnych znaków (lub ciągów znaków). Podobieństwo liter jest elementem inżynierii społecznej, lecz dla przetwarzającego go algorytmu są dwa różne wyrażenia – stąd brak wartości otrzymywanych od serwisów *whois* i *ipinfo*.

[204] Jedno z pól właściwości usługi *whois* lub *ipinfo* opisujących nazwę domenową lub adres IP, pozwalający na określenie właściciela (zarządcy) danej domeny lub dostawy którego przydzielono dany adres IP.

b. Danych uczących w procesie rozpoznawania wiadomości o charakterze szantażu z wykorzystaniem kryptowaluty BitCoin.
c. Danych uczących w procesie rozpoznawania heurystyk odnośników URL zawartych w analizowanej wiadomości.
d. Danych uczących uzyskanych w procesie pozyskiwania wskaźników z wiadomości email

Prawdopodobieństwo wystąpienia cech.

Bazując na pozyskanych zbiorze wiadomości phishingowych, obliczono prawdopodobieństwo wystąpienia danej cechy w wiadomości email:

Tabela 52. Prawdopodobieństwo wystąpienia poszczególnych cech w zbiorze uczącym.

L.p.	Cecha	Prawdopodobieństwo wystąpienia
1.	Odnośnik prowadzący do strony zidentyfikowanej jako phishingowa	0,03
2.	nieprawidłowy odnośnik URL zawarty w wiadomości	0,45
3.	adres IP w odnośniku URL zawartym w wiadomości	0,00
4.	wykorzystanie serwisów skracających odnośniku URL	0,00
5.	długość odnośnika URL zawartego w wiadomości	0,02
6.	długość domeny utworzonej na podstawie odnośnika URL znajdującego się w wiadomości	0,44
7.	groźba w przypadku niepodjęcia przez ofiarę sugerowanych działań	0,01
8.	nieprawidłowy adres email nadawcy	0,17
9.	niewłaściwy adres nadawcy	0,00
10.	wykorzystanie nazwy odbiorcy wiadomości	0,15
11.	wykorzystanie nazwy domenowej jako nazwy użytkownika w adresie nadawcy	0,06
12.	mechanizm śledzący w wiadomości email	0,25
13.	złożona nazwa domenowa wraz z subdomenami	0,01
14.	wiek zarejestrowanej domeny	0,08
15.	błędy językowe	0,51
16.	temat otrzymanej wiadomości	0,13
17.	użycie narzędzi programowych do wysyłki wiadomości email	0,01
18.	wykorzystanie tagowania wiadomości przez serwery pocztowe	0,03
19.	nazwa użytkownika w treści wiadomości.	0,06

Należy jednak zaznaczyć, że w danej wiadomości phishingowej może wystąpić jednocześnie wiele cech. Przedstawione wartości wskazują, że najczęściej identyfikowalnym wskaźnikiem phishingu były:

1. Błędy językowe (0,51) – najczęściej występująca cecha wskazująca na możliwy atak. Wartość ta może być jednak zaburzona z uwagi możliwe

błędy na omówione w rozdziale „III.2.19 Błędy językowe”. Duży odsetek wiadomości zawierających tą cechę, wskazuje, że wiadomości phishingowe, które adresowane są do polskiego odbiorcy, mogą być przygotowywane przez międzynarodowe zespoły, że słabą znajomością języka polskiego – co wskazuje na globalizację zjawiska. Wysoki odsetek wiadomości zawierających ten wskaźnik pokrywa się z badaniami [27].

2. Nieprawidłowy odnośnik URL zawarty w wiadomości (0,45) – wykorzystywanie. Cecha ta występowała niemal w połowie analizowanych wiadomości, wskazuję na podejście polegające na kierowaniu potencjalnych ofiar ataku phishingowego do odpowiednio spreparowanych witryn – gdzie w łatwy sposób można pozyskać interesujące atakujących dane.
3. Długość domeny utworzonej na podstawie odnośnika URL znajdującego się w wiadomości (0,44) – charakterystyka długość nazwy domenowej (ilości kropek i znaków „/” występujących w odnośniku URL) jest obiecującą metodą na weryfikację prawidłowości domeny. Wykorzystane w niniejszej rozprawie moduł uczenia maszynowego do rozpoznania charakterystyki będzie odporny na zmiany i zapewni prawidłowość rozpoznawania tej cechy.
4. Mechanizm śledzący w wiadomości email (0,25) – częste występowanie mechanizmu śledzącego powodowane jest chęcią identyfikacji aktywnego użytkownika, rozpoznania czasu odczytania, itp. Mechanizm zaadoptowany z wiadomości typu „spam”, identyfikacja tej cechy wskazuje na potencjalnie niebezpieczną wiadomość (w przypadku spamu, również wykonanie czynności przez użytkownika może przynieść mu szkodę) – jest więc to mocna przesłanka wskazująca na niebezpieczną zawartość.

Tabela 53. Prawdopodobieństwo wystąpienia w wiadomości phishingowej par najczęstszych cech.

cecha	1 (15)	2 (2)	3 (6)	4 (12)
1 (15)	-----	0,655102041	0,806122449	0,631632653
2 (2)	0,655102041	-----	0,719387755	0,584693878
3 (6)	0,806122449	0,719387755	-----	0,503061224
4 (12)	0,631632653	0,584693878	0,503061224	-----

Połączenie wskazanych powyżej, identyfikowanych cech w pary (zgodnie z przedstawieniem w powyższej tabeli), może znacznie zwiększyć szybkość metody detekcji (konieczność analizy jedynie 4 wartości pól z wielu).

Podsumowanie

W trakcie analizowania zjawiska phishingu, realizując postawione w niniejszej rozprawie zadanie naukowe:

1. Dokonano analizy porównawczej wiadomości email pod kątem podobieństwa czasowego, tematu, treści, adresów domenowych i odnośników URL. Analiza ta zidentyfikowała niektóre cechy występujące w wiadomościach email będących phishingiem.
2. Przeprowadzono analizę cech wiadomości email mogących wskazywać na atak phishingowy. W tym celu opracowany został autorski algorytm odczytujący wartości odpowiednich pól (przedstawionych i opisanych w niniejszej rozprawie, będących również autorską propozycją) wiadomości, analizujący ich wartości.
3. Opracowano metodę automatycznej analizy treści wiadomości, pozwalającej na wykrywanie błędów pisowni (z wykorzystaniem Słownika Języka Polskiego).
4. Opracowano metodę identyfikację szantażu i wymuszenia okupu zawartej w treści otrzymywanej wiadomości email. Metoda oparta na uczeniu maszynowym.
5. Dokonano analizy reputacji domen uzyskanych z odnośników URL, adresów email nadawcy oraz zwrotnego. Wykorzystano publicznie dostępne bazy danych odnośników URL uznanych za phishingowe, jako dane uczące metody określającej heurystyki domen i odnośników, na podstawie których można określić czy analizowana wartość może być atakiem czy też nie.
6. Wykorzystano metody data mining do wykrycia nieznanych zależności i schematów występujących w phishingowych wiadomościach email. Wykorzystano uczenie maszynowe do rozwiązywania problemów data mining podczas analizowania zbiorów danych. W trakcie prowadzonych badań dokonywano strojenia parametrów uczenia maszynowego, by osiągnąć najlepsze rezultaty.
7. Przeprowadzono eksperymenty związane z równoważeniem zbioru, celem usunięcia problematyki nierówności klas. Do równoważenia wykorzystano metody oversampingu (powielanie danych, generowanie nowych próbek), undersampingu (redukcji liczebności klasy dominującej). Badanie przeprowadzono z wykorzystaniem różnych technik i różnych algorytmów – dla

każdego zbudowano macierz pomyłek dla dwóch rodzajów zbiorów: orginalnego (niezrównoważonego) oraz zbalansowanego metodą SMOTE.

Prowadzone eksperymenty oparte na uczeniu maszynowym cech powstałych z ekstrakcji zidentyfikowanych wskaźników phishingu, pozwoliły na weryfikację postawionej w niniejszej rozprawie tezie: połączenie analizy treści wiadomości email, detekcji nie opisywanych wcześniej w literaturze wskaźników (cechy nr: 1, 5, 7, 13, 14, 17, 18) pozwoliło na identyfikację wiadomości o phishingowym charakterze. Zaproponowana metoda pozbawiona jest wad klasycznych metod, do prawidłowego działania nie jest wymagane posiadanie bazy regułowej, budowania i aktualizacji list dostępowych. Metoda również nie wymaga ingerencji człowieka. Postawione zadanie naukowe, zostało więc zrealizowane.

Przeprowadzone eksperymenty z wykorzystaniem opisanych w pracy wskaźników i identyfikacji cech ataku, z udziałem metod uczenia maszynowego na pozyskanych danych (phishingowe wiadomości email) jasno ukazały, że phishing, pomimo opisywanej w literaturze jego pozornej prostoty (np. [100]), jest złożonym zjawiskiem, trudnym do identyfikacji zarówno przez zespoły ludzkie wyposażone w niezbędną wiedzę i doświadczenie, jak i przez automatyczne algorytmy, oparte na uczeniu maszynowym. Pomimo możliwości identyfikacji poszczególnych cech ataku phishingowego (zarówno przez automatyczny algorytm jak i przez doświadczonego eksperta), zważywszy na ciągle zmieniające się metody stosowane przez atakujących i dzielnie niektórych cech z normalną korespondencją i wiadomościami typu „spam", część otrzymanych wiadomości, nie będzie można jednoznacznie przypisać do odpowiedniej klasy (phishing, spam, normalna korespondencją).

Prowadzone eksperymenty wykazały, jednakże, że stosując metody uczenia maszynowego do zidentyfikowanych i opisanych wskaźników mogących wskazywać na atak phishingowy, można z dużym prawdopodobieństwem zidentyfikować taką wiadomość (progres uczenia się modelu - Rysunek 62), bazując na analizie wartości poszczególnych pół nagłówka, korelacji tych wartości z zidentyfikowanymi uprzednio innymi wiadomościami o phishingowych charakterze.

Do istotnych osiągnięć autorskich należy:

1. Dokonanie przeglądu istniejących metod detekcji phisingu, wykazanie ich wad, które nie są odporne na zmieniające się metody ataku.

Zaprezentowana metoda będąca autorską propozycją, jest wolna od tych wad – algorytmy uczenia maszynowego wykryją nowy, nieznany wcześniej wzorzec ataku.

2. Zaprojektowanie wektora cech wskazujących na możliwy atak phishingowy. Poszczególne zakodowane cechy odpowiadają zidentyfikowanym przez autora opisanym wskaźnikom phishingu. Wektora zawiera cechy, które nie były dotychczas brane pod uwagę, jego długość jest również większa od dotychczas występujących w literaturze przedmiotu.
3. Pozyskanie zbiorów uczących dla poszczególnych modułów uczenia maszynowego: zbiór odnośników URL (zarówno w klasie phishing, jak i klasie normal).
4. Opracowanie założeń i zaprogramowanie algorytmu odczytującego poszczególne pola nagłówka wiadomości email – pole te nie występują w dostępnej literaturze traktującej o problematyce phishingu, a co wykazano w niniejszej rozprawie, wartości tych pól wskazują na możliwość wystąpienia takiego ataku. Zasadniczą różnica pomiędzy dostępnymi na rynku rozwiązaniami, potrafiącymi odczytywać wartości nagłówka wiadomości email, a prezentowanym autorskim rozwiązaniem, jest przetwarzanie znacznie większej ilości pól nagłówka, wykorzystanie badania poprawności językowej treści wiadomości, analiza treści z wykorzystaniem uczenia maszynowego pod katem wykrywania szantażu. Dotychczasowe modele automatycznie analizujące treść wiadomości nie brały pod uwagę możliwości wystąpienia w niej szantażu (w połączeniu z żądaniem okupu).
5. Zaprojektowanie modułów uczenia maszynowego bazującego na metodach heurystycznych analizy odnośnika URL. Opisywane w literaturze przedmiotu modele badające charakterystykę odnośnika, nie uwzględniały osobnego charakteru domeny, a jedynie traktowały całościowo odnośnik. Przedstawiona metoda bazując na odczytanej wartości odnośnika URL dokonuje również niezależnie badania domeny zawartej w danym odnośniku – co zwiększa możliwości detekcji.
6. Utworzenie na bazie zbioru uczącego wiadomości o charakterze phishingowym, zbioru wartości cech pierwotnych (adresy email, lista

narzędzi programowych, adresy portfeli BitCoin, itp.) służących jako rozszerzenie danych uczących poszczególnych modułów metody.

Przedstawiona koncepcja identyfikacji cech wskazujących na możliwość ataku phishingowego oraz bazująca na nich metoda wykrywania tego ataku, posiada możliwości dalszego jej rozwijania, m.in. poprzez:

1. Zwiększenie ilości cech (długość wektora) poprzez:
    a. Rozbicie cechy wskazującej na odnośnik phishingowy (badanie jego heurystyki) na poszczególne elementy składowe: 1 – długość odnośnika, 2 – długość domeny, 3 – ilość kropek w odnośniku, 4 – ilość znaków „/” („slash”) w odnośniku. Wartość poszczególnej nowej cechy (0 lub1), wskazującej na możliwość ataku phishingowego, podobnie jak w przypadku pierwotnej cechy może być wynikiem uczenia maszynowego. Zastosowanie tego podejścia może polepszyć proces uczenia się poszczególnych klasyfikatorów – poprawa jakość wykrywania cechy pomiędzy poszczególnymi etapami przyniosła poprawę procesu uczenia się i wykrywania ataku na danych testowych.
    b. Dodanie cechy wskazującej na obecność złośliwego załącznika – cecha ta silnie wskazuje na phishingowy charakter wiadomości.
2. Wprowadzenie większej ilości modułów uczenia maszynowego na etapie kodowania wektora cech dla nowododanych cech, co zwiększy odporność modelu na nowe techniki czy metody stosowane przez atakujących.
3. Pozyskanie zbioru wiadomości phishingowych reprezentujący wszystkie opisywane cechy, celem uzyskania egzemplarzy danych pokrywających całe spektrum cech (pojedynczy egzemplarz może zawierać od jednej do kilku cech). Pozyskanie w takiej samej ilości wiadomości, które zostają przypisane do klas „normal” oraz „spam”. Różnorodność egzemplarzy danych zapewni unikalność cech.

Bibliografia

[1] M. Golka, „Czym jest społeczeństwo informacyjne?," Wydział Prawa i Administracji UAM, Poznań, 2005.

[2] J. Kos-Łabędowicz i S. Talar, „Rola Internetu w procesie konwergencji rozwojowej współczesnej gospodarki światowej," Wydawnictwo Uniwersytetu Ekonomicznego w Katowicach, Katowice, 2013.

[3] D. Dean, S. Digrande, D. Field, A. Lundmark, J. O'Day, J. Pineda i P. Zwillenberg, „The Internet Economy in the G-20. The $4.2 Trillion Growth Opportunity," The Boston Consulting Group, Boston, 2012.

[4] Organizacja Narodów Zjednoczonych, „Human Development Indices and Indicators 2018 Statistical Update," United Nations Development Programme, 2019.

[5] Ministerstwo Infrastruktury i Rozwoju, „Strategia komunikacji Programu Operacyjnego Polska Cyfrowana lata 2014-2020," Ministerstwo Infrastruktury i Rozwoju, Warszawa, 2015.

[6] Główny Urząd Statystyczny, „Społeczeństwo informacyjne w Polsce. Wyniki badań statystycznych z lat 2014-2018," Szczecin, 2018.

[7] IBIMS, IBRIS, „Skąd Polacy czerpią informacje?," IBIMS, Częstochowa, 2021.

[8] AAG, „The Latest 2023 Phishing Statistics (updated April 2023)," AAG, 06 04 2023. [Online]. Available: https://aag-it.com/the-latest-phishing-statistics/#:~:text=Yes%2C%20phishing%20is%20the%20most,emails%20are%20sent%20every%20day. [Data uzyskania dostępu: 21 04 2023].

[9] pwc, „Cyber-ruletka po polsku," pwc, Warszawa, 2018.

[10] CrowdStrike, „10 Most common types of cyber attack," 13 02 2023. [Online]. Available: https://www.crowdstrike.com/cybersecurity-101/cyberattacks/most-common-types-of-cyberattacks/. [Data uzyskania dostępu: 25 09 2023].

[11] P. Robinson, „15 Common Types of Cyber Attacks and Threats," 14 02 2023. [Online]. Available: https://www.lepide.com/blog/the-15-most-common-types-of-cyber-attacks/. [Data uzyskania dostępu: 25 09 25].

[12] Fortinet, „Types of Cyber Attacks,” 2022. [Online]. Available: https://www.fortinet.com/resources/cyberglossary/types-of-cyber-attacks. [Data uzyskania dostępu: 25 09 2023].

[13] Gov.pl, „Czym jest PHISHING i jak nie dać się nabrać na podejrzane wiadomości e-mail oraz SMS-y?,” 2023. [Online]. Available: https://www.gov.pl/web/baza-wiedzy/czym-jest-phishing-i-jak-nie-dac-sie-nabrac-na-podejrzane-widomosci-e-mail-oraz-sms-y. [Data uzyskania dostępu: 25 09 2023].

[14] CERT Polska, „Krajobraz bezpieczeństwa polskiego internet. Raport roczny za 2018 z działalności CERT Polska,” NASK, Warszawa, 2019.

[15] E. Hutchins, M. Cloppert i R. Amin, „Intelligence-Driven Computer Network Defense Informed by Analysis of Adversary Campaigns and Intrusion Kill Chains,” Lockheed Martin Corporation, Bethesda, 2011.

[16] H. Penney, „Winning the Kill Chain Competition,” 28 07 2023. [Online]. Available: https://www.airandspaceforces.com/article/winning-the-kill-chain-competition/. [Data uzyskania dostępu: 25 09 2023].

[17] A. Hahna, R. K. Thomas, I. Lozano i A. Cardenasc, „A multi-layered and kill-chain based security analysis framework for cyber-physical systems,” *International Journal of Critical Infrastructure Protection,* tom 11, pp. 39-50, 12 2015.

[18] M. Chris, „The Phishing Kill Chain,” 05 08 2014. [Online]. Available: https://www.agari.com/email-security-blog/phishing-kill-chain/. [Data uzyskania dostępu: 01 07 2020].

[19] A.-P. W. Group, „Phishing Activity Trends Reports 1H 2017,” APWG, Washington, 2017.

[20] FBI, „2020 Internet Crime Report,” Federal Bureau of Investigation, Waszyngton, 2020.

[21] Trend Micro Incorporated;, „Spear-Phishing Email: Most Favored APT Attack Bait,” Trend Micro, Cupertino, 2012.

[22] I. Karambelas, „Spear Phishing: The Secret Weapon Behind the Worst Cyber Attacks,” 16 01 2016. [Online]. Available: https://www.cloudmark.com/en/blog/spear-phishing-secret-weapon-behind-worst-cyber-attacks. [Data uzyskania dostępu: 22 09 2022].

[23] M. Laudon, „Security Brief: Mobile Phishing Increases More Than 300% as 2020 Chaos Continues,” 02 11 2020. [Online]. Available: https://www.proofpoint.com/us/blog/threat-protection/mobile-phishing-increases-more-300-2020-chaos-continues. [Data uzyskania dostępu: 20 06 2021].

[24] Next Caller, „Covid 19 Fraud Report,” 01 05 2020. [Online]. Available: https://nextcaller.com/blog/next-caller-covid-19-fraud-report. [Data uzyskania dostępu: 20 06 2021].

[25] M. Szutiak, „CERT Polska: Przyszedł SMS od PGE z informacją o zaległości? To może być oszustwo,” 21 06 2021. [Online]. Available: https://www.telepolis.pl/wiadomosci/bezpieczenstwo/cert-polska-pge-falszywy-sms-o-zaleglosci-oszustwo. [Data uzyskania dostępu: 21 06 2021].

[26] Niebezpiecznik, „Jak przeprowadzono atak na KNF i polskie banki oraz kto jeszcze był na celowniku przestępców?,” 01 04 2017. [Online]. Available: https://niebezpiecznik.pl/post/jak-przeprowadzono-atak-na-knf-i-polskie-banki-oraz-kto-jeszcze-byl-na-celowniku-przestepcow/. [Data uzyskania dostępu: 26 06 2021].

[27] Darktrace, „Generative AI: Impact on Email Cyber-Attacks,” Darktrace, London, 2023.

[28] F. Sharevski, A. Devine, E. Pieroni i P. Jachim, „Gone Quishing: A Field Study of Phishing with Malicious QR Codes.,” w *Conference acronym 'XX*, Woodstock, 2022.

[29] A. Alnajim i M. Munro, „An Approach to the Implementation of the Anti-Phishing Tool for Phishing Websites Detection,” w *2009 International Conference on Intelligent Networking and Collaborative Systems*, 2009.

[30] T. Hurley, J. E. Perdomo i A. Perez-Pons, „HMM-Based Intrusion Detection System for Software Defined Networking,” w *15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Anaheim, 2016.

[31] statista, „Average daily spam volume worldwide from October 2020 to September 2021,” statista, 2023.

[32] Proofpoint, „2020 State of the phish. Annual Report,” proofpoint.com, Sunnyvale, 2020.

[33] R. Dahamija, J. D. Tygar i . M. Hearst, „Why phishing works,” w *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 2006.

[34] S. E. Schechter, R. Dhamija, A. Ozment i I. Fischer, „The Emperor's New Security Indicators," w *2007 IEEE Symposium on Security and Privacy (SP '07)*, Berkeley, 2007.

[35] Tessian, „Understand the mistakes that compromise your company's security," Tessian Limited, Londyn, 2021.

[36] Krajowy Rejestr Długów, „Młodzi Polacy świadomi zagrożeń i przygotowani do ochrony danych osobowych w czasie pandemii," ChronPESEL.pl, Warszawa, 2021.

[37] S. H. Apandi, J. Sallim i R. M. Sidek, „Types of anti-phishing solutions for phishing attack," w *The 6th International Conference on Software Engineering & Computer Systems*, Pahang, 2019.

[38] Y. Wang, R. Agrawal i C. Baek-Young, „Light Weight Anti-Phishing with User Whitelisting in a Web Browser," w *2008 IEEE Region 5 Conference*, Kansas City, 2008.

[39] A. K. Jain i B. B. Gupta, „A novel approach to protect against phishing attacks at client side using auto-updated white-list," *EURASIP Journal on Information Security*, 06 05 2016.

[40] W. Han, Y. Cao, E. Bertino i J. Yong, „Using automated individual white-list to protect web digital identities," *Expert Systems with Applications*, tom 39, nr 15, pp. 11861-11869, 2012.

[41] M. K. R. R. K. M. G. Pawan Prakash, „PhishNet: Predictive Blacklisting to Detect Phishing Attacks," *IEEE Conference Publications*, pp. 1-5, 02 2010.

[42] M. Felegyhazi, C. Kreibich i V. Paxson, „On the Potential of Proactive Domain Blacklisting," w *Proceedings of the 3rd USENIX conference on Large-scale exploits and emergent threats: botnets, spyware, worms, and more*, San Jose, 2010.

[43] S. Sheng, B. Wardman, G. Warner i L. Cranor, „An Empirical Analysis of Phishing Blacklists," w *6th Conference on Email and Anti-Spam*, Mountain View, 2010.

[44] G. Xiang, J. I. A. Hong, C. P. Rosé i L. F. Cranor, „CANTINA+: A Feature-Rich Machine Learning Framework for Detecting Phishing Web Sites," *ACM Transactions on Information and System Security*, tom 14, nr 2, pp. 1-28, 2011.

[45] D. L. Cook, V. K. Gurbani i M. Daniluk, „Phishwish: A Stateless Phishing Filter Using," w *Financial Cryptography and Data Security (12th International Conference)*, Cozumel, 2008.

[46] N. Sanglerdsinlapachai i A. Rungsawang, „Using Domain Top-page Similarity Feature in Machine Learning-Based Web Phishing Detection," w *2010 Third International Conference on Knowledge Discovery and Data Mining*, Phuket, 2010.

[47] N. M. Shekokar, C. Shah, M. Mahajan i S. Rachh, „An ideal approach for detection and prevention of phishing attacks," *82-91*, tom 49, pp. 82-91, 06 17 2015.

[48] V.Shreeram, M.Suban, P.Shanthi i K.Manjuha, „Anti-phishing detection of phishing attacks using genetic algorithm," *ICCCCT*, pp. 447-450, 2010.

[49] A. Aljofey, Q. Jiang, A. Rasool, H. Chen, W. Liu, Q. Qu i Y. Wang , „An effective detection approach for phishing websites using URL and HTML features," *Scientific Reports*, nr 12, 25 05 2022.

[50] A. K. Jain i B. B. Gupta, „A machine learning based approach for phishing detection using hyperlinks information," Journal of Ambient Intelligence and Humanized Computing, 2018.

[51] D. Sánchez, M. A. Vila, L. Cerda i J. M. Serrano, „Association rules applied to credit card fraud detection," *Expert Systems with Applications*, tom 36, nr 2, pp. 3630-3640, 2009.

[52] P. Kumaraguru, S. Sheng, A. Acquisti, L. F. Cranor i J. Hong, „Teaching Johnny not to fall for phish," *ACM Transactions on Internet Technology*, tom 10, nr 2, pp. 1-31, 2010.

[53] Y. Mei , „Anti-phishing system," Växjö University , VÄXJÖ, 2008.

[54] E. Harris, „The Next Step in the Spam Control War: Greylisting," 21 08 2003. [Online]. Available: http://projects.puremagic.com/greylisting/whitepaper.html. [Data uzyskania dostępu: 2022 12 28].

[55] M. Khonji, Y. Iraqi i A. Jones, „Phishing Detection: A Literature Survey," *IEEE Communications Surveys & Tutorials*, tom 15, nr 4, pp. 2091 - 2121, 15 4 2013.

[56] M. Jain, „5 Key Limitations of Doing Threat Detection with Rules," 18 1 2017. [Online]. Available: https://www.logichub.com/blog/5-key-limitations-of-doing-threat-detection-with-rules. [Data uzyskania dostępu: 2022 12 28].

[57] D. E. Goldberg, Algorytmy genetyczne, Warszawa: Wydawnictwo Naukowo-Techniczne, 2003.

[58] C. Sinclair, L. Pierce i S. Matzner, „An Application of Machine Learning to Network Intrusion Detection," w *ACSAC'99*, Phoenix, 1999.

[59] A. A. F. T. Neda Abdelhamid, „Phishing detection based Associative Classifications data mining," *Expert system with Applications*, tom 41, nr 13, pp. 5948-5959, 2014.

[60] W. Gansterer i D. Pölz, „E-mail Classification for Phishing Defense," University of Vienna, Vienna, 2014.

[61] Y. Pan i X. Ding, „Anomaly Based Web Phishing Page Detection," w *22nd annual computer security applications conference*, Miami Beach, 2006.

[62] E. Villar-Rodriguez, J. Del Ser i S. Salcedo-Sanz, „On a Machine Learning Approach for the Detection of Impersonation Attacks in Social Networks," *Intelligent Distributed Computing*, tom VIII, pp. 259-268, 2015.

[63] B. Issac, R. Chiong i S. M. Jacob, „Analysis of Phishing Attacks and Countermeasures," Swinburne University of Technology, Kuching, 2006.

[64] A. A. Akinyelu i A. O. Adewumi, „Classification of Phishing Email Using Random Forest Machine Learning Technique," *Journal of Applied Mathematics*, 2014.

[65] M. Sparshot, „The psychology of phishing," Help NET Security, 23 lipiec 2014. [Online]. Available: https://www.helpnetsecurity.com/2014/07/23/the-psychology-of-phishing/. [Data uzyskania dostępu: 01 lipiec 2020].

[66] F. Carroll, J. A. Adejobi i R. Montasari , „How Good Are We at Detecting a Phishing Attack? Investigating the Evolving Phishing Attack Email and Why It Continues to Successfully Deceive Society," *SN Computer Science*, 23 02 2022.

[67] S. Lewis, „Why Is It So Difficult To Detect Phishing Emails?," Southeastern Technical, 23 07 2020. [Online]. Available: https://www.setechnical.net/data-security/why-is-it-so-difficult-to-detect-phishing-emails/. [Data uzyskania dostępu: 2023 08 16].

[68] K. Singh, P. Aggarwal, P. Rajivan i C. Gonzalez, „What makes phishing emails hard for humans to detect?," *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, tom 64, nr 1, pp. 431 - 435, 2021.

[69] S. C. Jevva i E. B. Rajsingh, „Intelligent phishing url detection using association rule mining," *Human-centric Computing and Information Sciences,* tom 6, nr 10, 10 06 2016.

[70] E. Bübera, „Phishing URL Detection with ML," 18 02 2018. [Online]. Available: https://towardsdatascience.com/phishing-domain-detection-with-ml-5be9c99293e5. [Data uzyskania dostępu: 10 01 2023].

[71] R. B. Basnet, A. H. Sung i Q. Liu, „Learning to detect phishing URLs," *International Journal of Research in Engineering and Technology,* tom 03, nr 06, pp. 11-24, 2014.

[72] D. K. McGrath i M. Gupta, „Behind Phishing: An Examination of Phisher Modi Operandi," Indiana University, 04 04 2008. [Online]. Available: https://www.usenix.org/legacy/events/leet08/tech/full_papers/mcgrath/mcgrath_html/mcgrath_gupta.html. [Data uzyskania dostępu: 15 06 2022].

[73] Flexera, „Software Vulnerability Manager (Cloud Edition) Help Library," Flexera, 08 2023. [Online]. Available: https://docs.flexera.com/csi/Content/helplibrary/Criteria_for_the_Threat_Score_Calculation.htm#appendthreatintel_2398103115_1101416. [Data uzyskania dostępu: 25 09 2023].

[74] J. Postel, „RFC 5321 (Simple Mail Transfer Protocol)," Network Working Group, 08 1982. [Online]. Available: https://datatracker.ietf.org/doc/html/rfc5321. [Data uzyskania dostępu: 07 06 2021].

[75] D. H. Crocker, „RFC 5322 (Internet Message Format)," Network Working Group, 13 08 1982. [Online]. Available: https://datatracker.ietf.org/doc/html/rfc5322. [Data uzyskania dostępu: 08 06 2021].

[76] I. Vojinovic , „Save Your Data with These Empowering Password Statistics," DataProt, 07 02 2021. [Online]. Available: https://dataprot.net/statistics/password-statistics/. [Data uzyskania dostępu: 10 08 2021].

[77] J. Johnson, „How many of your accounts use the same password for online logins?," statista.com, 25 01 2021. [Online]. Available: https://www.statista.com/statistics/763091/us-use-of-same-online-passwords/. [Data uzyskania dostępu: 10 08 2021].

[78] J. Topf, „Characters in the local part of a mail address,” 16 05 2018. [Online]. Available: https://www.jochentopf.com/email/characters-in-email-addresses.pdf. [Data uzyskania dostępu: 10 05 2022].

[79] P. Foremski i P. Vixie, „The Modality of Mortality in Domain Names,” Farsight Security, San Mateo, 2018.

[80] M. Ogrodniczuk i M. Kopeć, „Lexical Correction of Polish Twitter Political Data,” w *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, Vancouver, 2017.

[81] M. Alanezi, „Phishing Detection Methods: A Review,” *Technium,* tom 3, nr 9, pp. 19-35, 2021.

[82] B. Wei, R. A. Hamad, L. Yang, X. He, H. Wang, B. Gao i W. L. Woo, „A Deep-Learning-Driven Light-Weight Phishing Detection Sensor,” *Sensors,* 30 09 2019.

[83] Y. Zhang, S. Egelman, L. Cranor i J. Hong, „Phinding Phish: Evaluating Anti-Phishing Tools,” *Carnegie Mellon University*, 01 styczeń 2006.

[84] N. S. A. T. Ian Fette, „Learning to Detect Phishing Emails,” w *Security, Privacy, Realiability and Ethics,* Pittsburgh, 2007.

[85] A. . P. E. Rosiello, E. Kirda, C. Kruegel i F. Ferrandi, „A Layout-Similarity-Based Approach for Detecting Phishing Pages,” [Online]. Available: https://sites.cs.ucsb.edu/~chris/research/doc/securecomm07_antiphishdom.pdf . [Data uzyskania dostępu: 01 09 2023].

[86] A. P. E. Rosiello, E. Kirda, C. Kruegel i F. Ferrandi, „A layout-similarity-based approach for detecting phishing pages,” w *Third International Conference on Security and Privacy in Communications Networks and the Workshops - SecureComm 2007,* Nice, 2007.

[87] D. Majerek, Eksploracja danych, Lublin: Politechnika Lubelska, 2020.

[88] M. Płaszczyca, „Regresja Logistyczna,” 01 01 2013. [Online]. Available: https://www.statystyka.az.pl/regresja-logistyczna.php. [Data uzyskania dostępu: 2023 05 30].

[89] L. Ceci, „Number of sent and received e-mails per day worldwide from 2017 to 2026,” statista, New York, 2022.

[90] G. Ch, „How I was using Naive Bayes (Incorrectly) till now — Part-2,” 20 01 2021. [Online]. Available: https://towardsdatascience.com/how-i-was-using-

naive-bayes-incorrectly-till-now-part-2-d31feff72483. [Data uzyskania dostępu: 06 12 2023].

[91] J. Brownlee, „Random Oversampling and Undersampling for Imbalanced Classification,” 5 01 2021. [Online]. Available: https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/. [Data uzyskania dostępu: 06 12 2023].

[92] N. V. Chawla, K. W. Bowyer , L. O. Hall i W. P. Kegelmeyer, „SMOTE: Synthetic Minority Over-sampling Technique,” *Journal of Artificial Intelligence Research,* nr 16, p. 321–357, 2002.

[93] D. L. Wilson, „Asymptotic Properties of Nearest Neighbor Rules Using Edited Data,” *IEEE Transactions on Systems, Man, and Cybernetics ,* Tomy %1 z %2SMC-2, nr 3, pp. 408 - 421, 1972.

[94] M. M. Nishat, F. Faisal, I. J. Ratul, A. Al-Monsur, A. M. Ar-Rafi, S. M. Nasrullah, T. Reza i R. H. Khan, „A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset,” *Scientific Programming,* nr 2022, 2022.

[95] H. He, Y. Bai, E. A. Garcia i S. Li, „ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning,” w *International Joint Conference on Neural Networks*, Hong Kong, 2008.

[96] I. Tomek, „Two Modifications of CNN,” *IEEE Transactions on Systems, Man, and Cybernetics,* Tomy %1 z %2SMC-6, nr 11, pp. 769 - 772, 1976.

[97] I. M. Pires, F. Hussain, N. M. Garcia, P. Lameski i E. Zdravevski, „Homogeneous Data Normalization and Deep Learning: A Case Study in Human Activity Classification,” *Basel,* tom 12, nr 11, 2020.

[98] C. D. Xuan, H. D. Nguyen i T. V. Nikolaevich, „Malicious URL Detection based on Machine Learning,” *International Journal of Advanced Computer Science and Applications,* tom 11, nr 1, pp. 149-153, 2020.

[99] B. Banik i A. Sarma, „Phishing URL detection system based on URL features using SVM,” *International Journal of Electronics and Applied Research,* tom 5, nr 2, pp. 40-55, 2018.

[100] S. Gupta, A. Singhal i A. Kapoor, „A literature survey on social engineering attacks: Phishing attack," w *IEEE*, Greater Noida, India, 2016.

Spis rysunków

Spis tabel

Dodatek A – analiza próbek złośliwego oprogramowania

Wyniki analizy próbek złośliwego oprogramowania stanowiącego załączniki do próbek badawczych.

Lp.	ID próbki	Odnośnik do analizy
1.	P-02-002	https://www.hybrid-analysis.com/sample/e4e38b8688d4908e6daf0e21e305ad4d0c6c9373f20c952b43369e61583339c9

Dodatek B - Komponenty wymagane do zainstalowania.

Lp.	Komponent	Środowisko	Opis	Uwagi
1.	pip	python	Domyślny system zarządzania pakietami dla środowiska języka Python, korzystający z dedykowanego repozytorium pakietów o nazwie Python Package Index lub z innych zdalnych oraz lokalnych repozytoriów	Instalowany domyślnie w pakiecie instalacyjnym środowiska Microsoft Windows.
2.	ipinfo	python	Pobieranie danych odnośnie adresu IP (Operator, ASN, geolokalizacja, itp.).	pip install ipinfo
3.	IPWhois	python	Pobieranie danych odnośnie adresu IP (Operator, ASN, geolokalizacja, itp.).	pip install ipwhois
4.	whois	python	Pobieranie informacji o zarejestrowanej domenie W trakcie prowadzonego eksperymentu, oryginalny skrypt został zmodyfikowany.	pip install python-whois
5.	bs4	python	Oczyszczenie ciągu znaków (typu string) z encji HTML (wydobycie treści)	pip install bs4
6.	imblearn	python	Zestaw bibliotek wspierających transformację danych (w przypadku występowania nierówności zbiorów)	pip install imblearn

Dodatek C - Opis ataku phishingowego na użytkownika z wykorzystaniem wiadomości SMS (smishing)

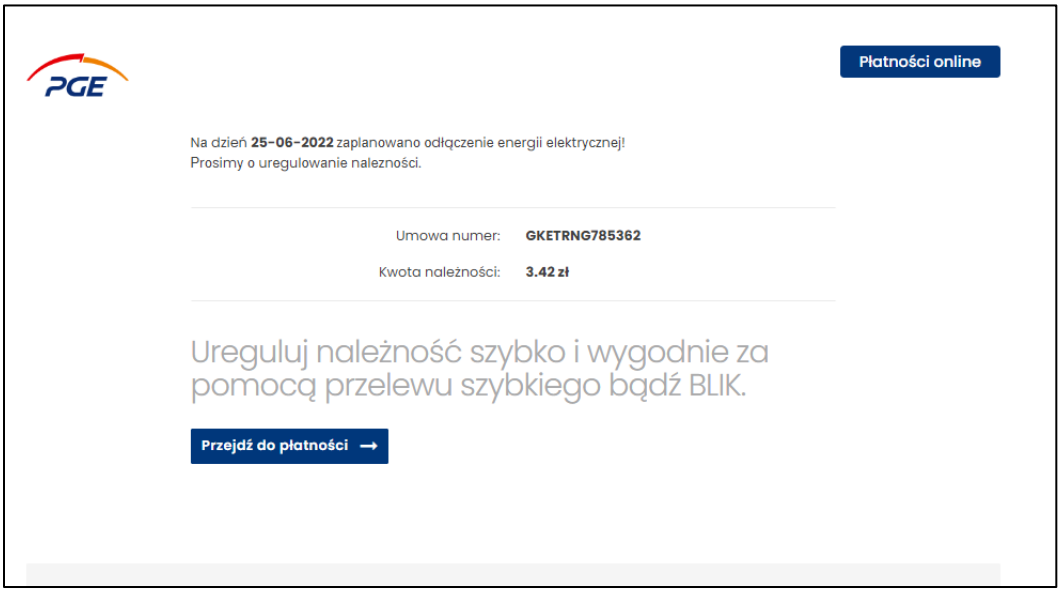
Użytkownik w dniu 24.06.2022r. otrzymał wiadomość SMS o treści:

> PGE: Na dzień 25.06 zaplanowano odłączenie energii elektrycznej!
> Prosimy o uregulowanie należności: https://t2m.io/ba7kUdn

Po wpisaniu do paska przeglądarki przesłanego adresu w wiadomości SMS, użytkownik przenoszony jest pod adres:

```
https://feanoys.com/pge/87315823/1895292/
```

Użytkownikowi wysyłane jest ciasteczko o nazwie: PHPSESSID, oraz wyświetlona zostaje strona:

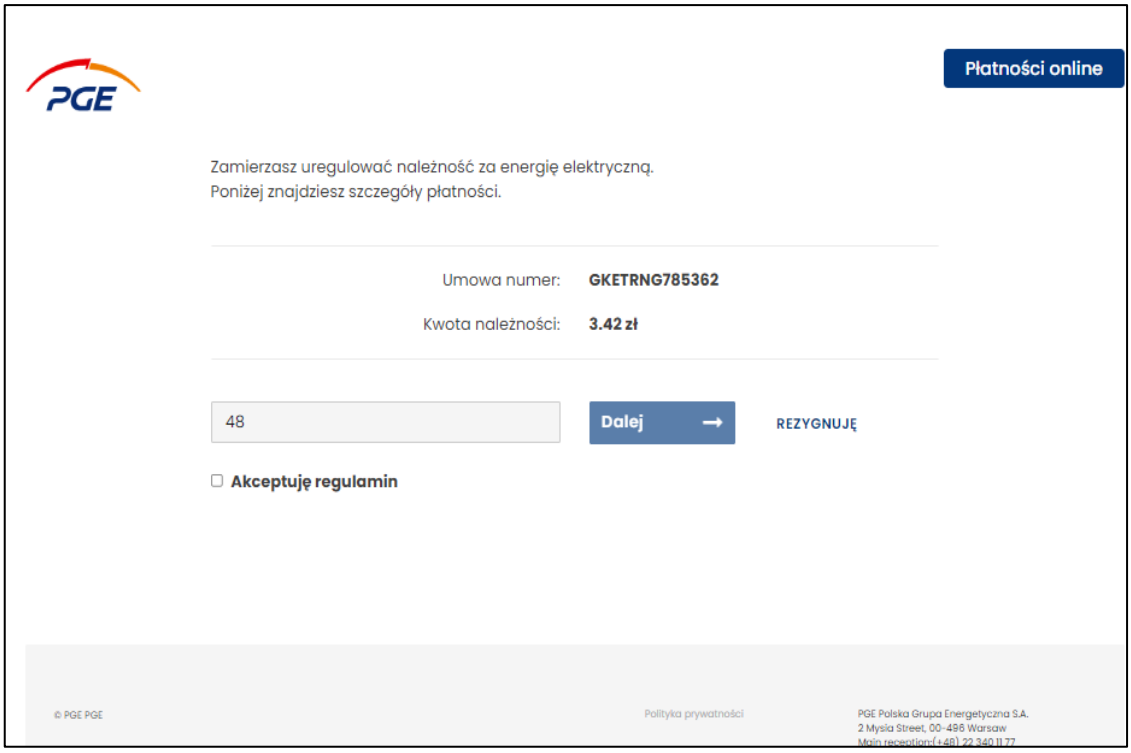


Rysunek 63. Strona wyłudzająca dane, wektor ataku phishingowego "dopłata PGE".

Kliknięcie przycisku „**Przejdź do płatności->**” powoduje przeniesienie użytkownika pod adres

```
https://feanoys.com/pge/start-transaction/index.php
```

Pod powyższym adresem wyświetlony zostanie monit o konieczności uzupełnienia numeru telefonu komórkowego użytkownika, który otrzymał wiadomość SMS. Wymaganie podania numeru komórkowego sugeruje, że dane atak phishingowy nie jest dedykowany i kierowany przeciwko konkretnej osobie, a raczej przeciwko szerokiemu gronu odbiorców, konieczne jest więc poznanie numeru komórkowego ofiary ataku.

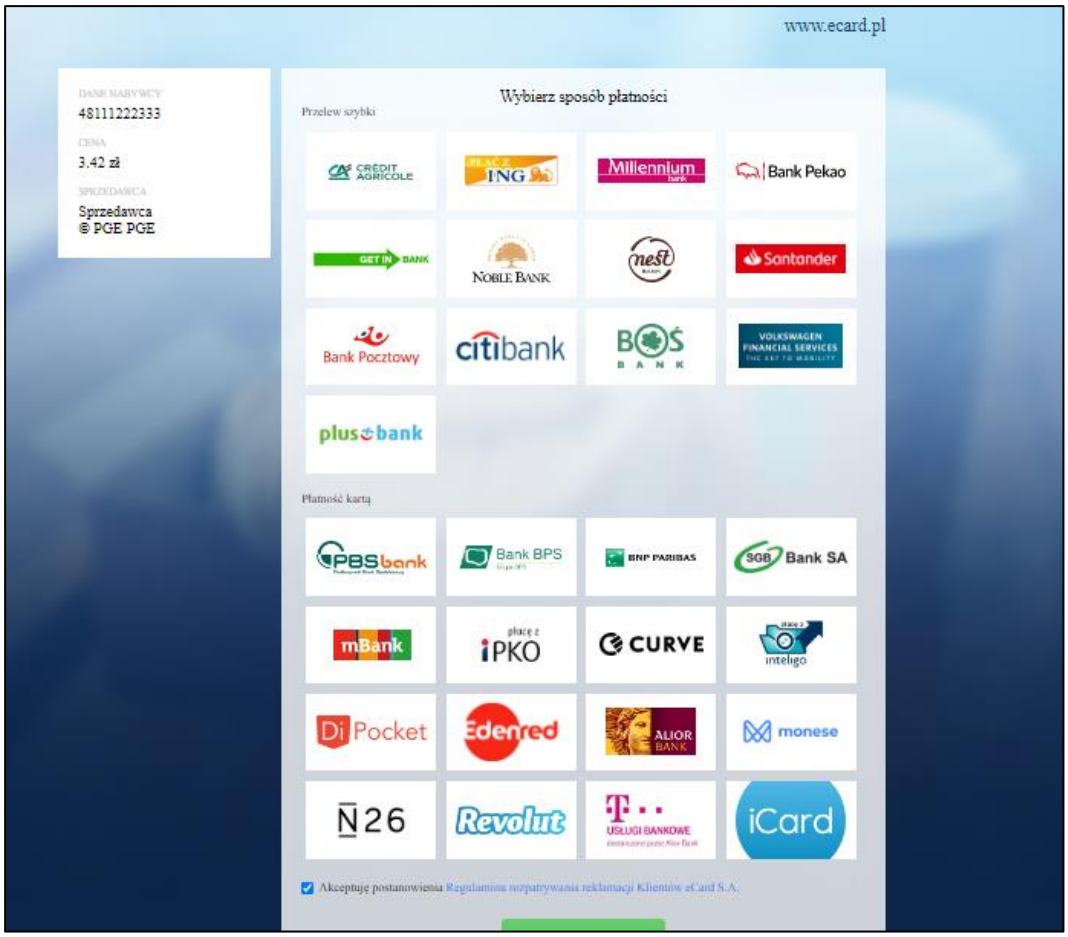


Rysunek 64. Etap 2 ataku - formularz podania numeru komórkowego odbiorcy wiadomości.

By przejść dalej wpisana została wartość: **111222333** (jako numer telefonu komórkowego) i po kliknięciu przycisku „**Dalej ->**” użytkownik przenoszony jest pod adres:

```
https://feanoys.com/pge/pay.php
```

Wyświetlona zostanie strona:

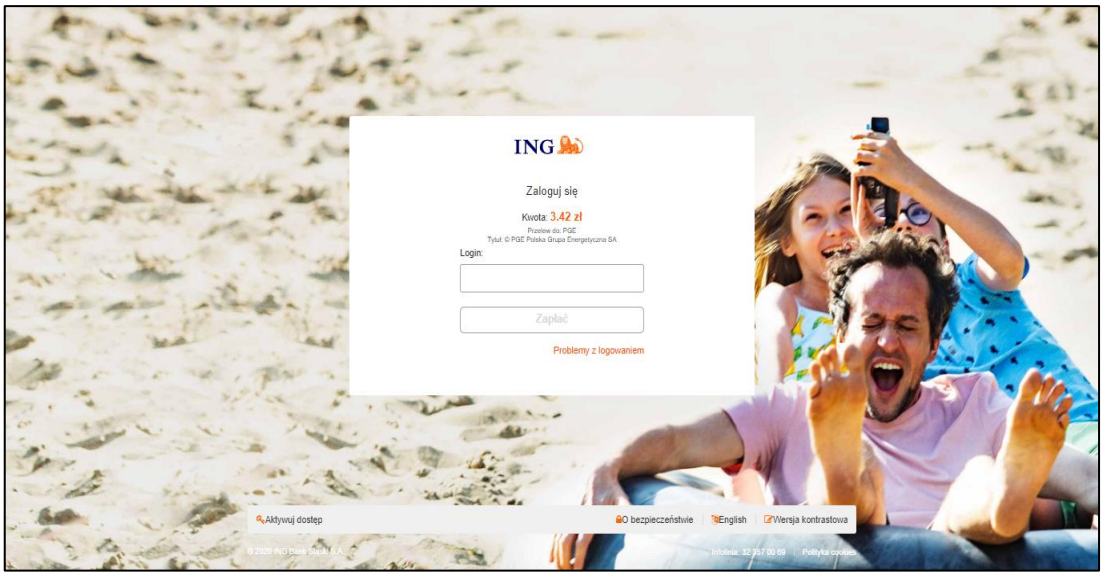


Rysunek 65. Etap 3 - rzekomy pośrednik płatności z dużą ilością banków.

Przedstawiona strona udaje pośrednika płatności ecard.pl. Duża ilość polskich banków sugeruje, że atakujący przygotowali infrastrukturę pod różnorodne i szerokie grono polskich odbiorców. W celach testowych wybrano bank: ING. Po kliknięciu na ikonkę użytkownik przenoszony jest pod adres:

```
https://feanoys.com/ing/mojeing/paybylink/login/ctxid/9GoyqJEdCT
          DZPMHGR12sv6FN9sfab4cr/index.php?pay&b=ing
```

Użytkownikowi podstawiona jest strona imitująca panel logowania się do wybranego przez niego banku.



Rysunek 66. Etap 4 - panel logowania się użytkownika wybranego przez niego banku.

Zaimplementowany mechanizm sprawdzania, waliduje wprowadzoną nazwę użytkownika (zgodnie z wymaganiami banku wybranego w kroku wcześniejszym przez użytkownika – co świadczy o dobrym rozpoznaniu standardów nazewniczych polskich banków) i dopiero po wpisaniu nazwy pasującej do przyjętego przez bank ING standardu nazewniczego, przycisk „**zapłać**” zostaje podświetlony i staje się aktywny. Jako nazwę użytkownika wpisano: **alamk11233** (losowa wartość zgodna ze standardem nazewniczym).

Rysunek 67. Etap 4 - walidacja nazwy użytkownika banku ING.

Po kliknięciu przyciski użytkownikowi wyświetlony jest kolejny panel udający system weryfikacji banku ING. Użytkownik proszony jest o podanie numeru PESEL.

Rysunek 68. Etap 5 - panel żądający podanie numeru PESEL użytkownika.

Za pomocą dostępnego on-line generatora numerów PESEL [205] wygenerowany fałszywy numer PESEL: **51093065186**. Numer wpisano w formularz. Użytkownik następnie jest proszony o podanie kodu PIN.

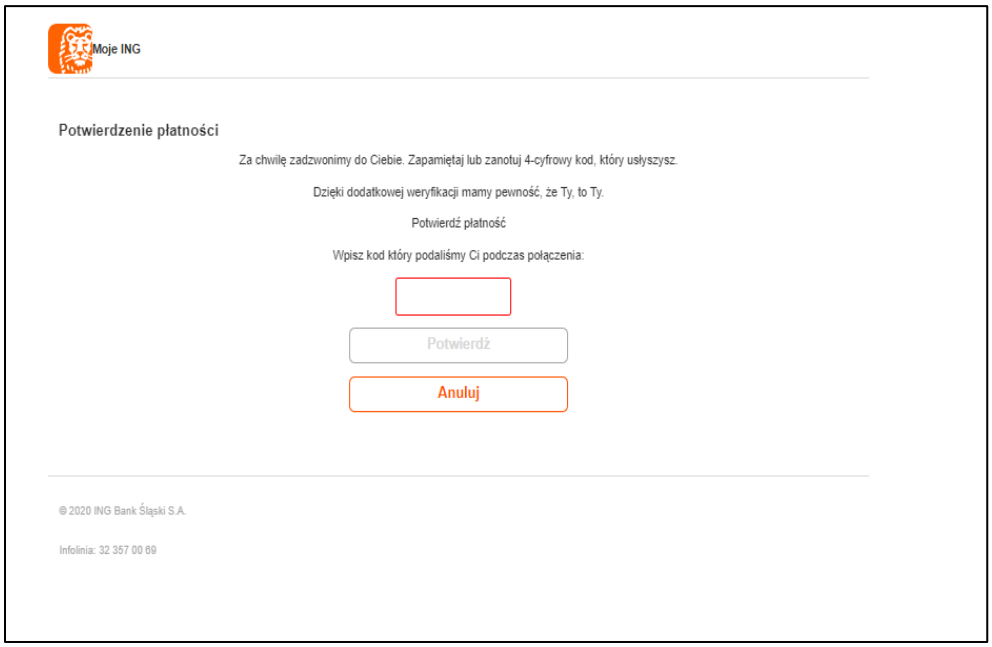
Rysunek 69. Etap 6 - wymaganie podania kodu PIN.

[205] https://pesel.cstudios.pl/o-generatorze/generator-on-line

Na potrzeby testu podano PIN: **0000**. Kod ten jest zwykle odrzucany przez mechanizmy bezpieczeństwa banków.

Rysunek 70. Etap 7 – żądanie podania kodu SMS, celem uwierzytelnienia transakcji.

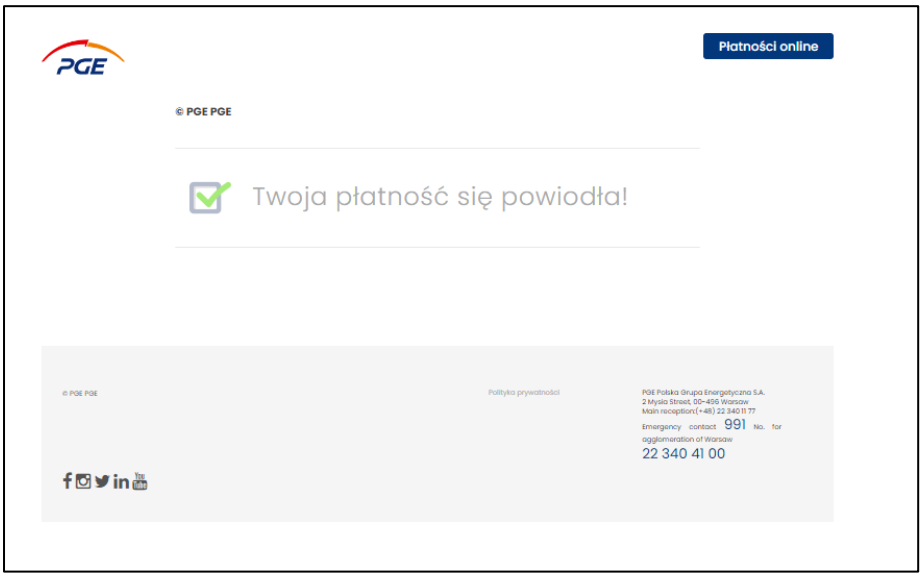
Atak ma symulować przeprowadzenie standardowej transakcji, która potwierdzana jest otrzymanym w wiadomości SMS kodem (stad też prośba we wcześniejszych etapach o podanie numeru telefonu komorowego ofiary ataku). Użytkownik proszony jest o podanie kodu z wiadomości SMS. Z uwagi że we wcześniejszych etapach, podano nieistniejący numer telefonu komórkowego (111222333), nie było możliwości weryfikacji, czy rzeczywiście taki kod jest wysyłany. Celem dalszych testów wpisano kod: **000000**. Uruchomiony na stronie mechanizm dokonał walidacji długości kodu. Po wpisaniu kodu użytkownikowi wyświetlona została kolejna podstrona, prosząca o podanie kodu jaki ma otrzymać podczas wykonanej do niego rozmowy głosowej.



Rysunek 71. Etap 8 - żądanie wpisania kodu otrzymanego podczas połączenia przychodzącego na telefon użytkownika.

Z uwagi na podanie we wcześniejszych etapach nieistniejącego numeru telefonu (111222333), nie otrzymano żadnego połączenia przychodzącego. Jako kod wpisano: 1234. Po wpisaniu kodu, użytkownik przenoszony jest na stronę:

```
https://feanoys.com/payment/successful/
```

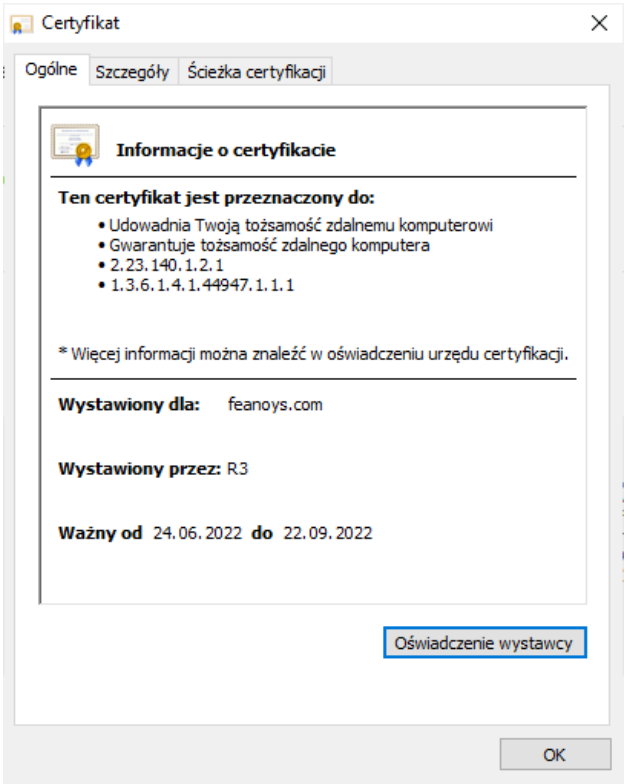


Rysunek 72. Etap 9 - finalizacja

Przeniesienie użytkownika na tą stronę świadczy o braku walidacji tego etapu – pomimo braku otrzymania kodu w połączeniu przychodzącym (nie otrzymano takiego połączenia – podanie błędnego numeru telefonu). Użytkownikowi wyświetlany jest

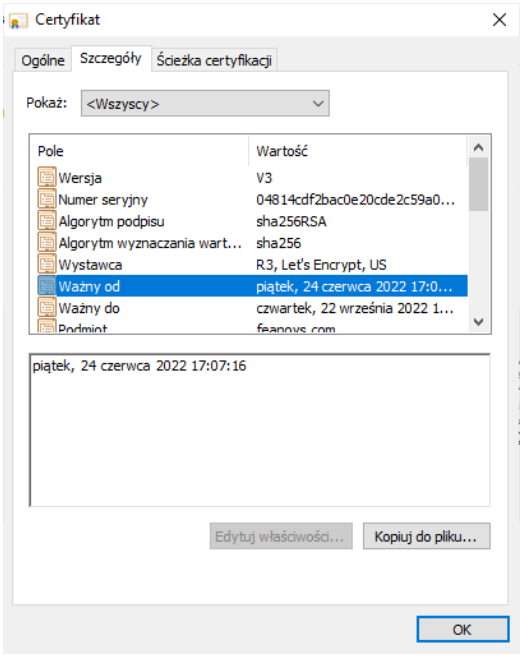
komunikat o powodzeniu procesu weryfikacji i wykonaniu płatności. Jest to ostatni element procesu.

System posiada wdrożony mechanizm weryfikacji stacji roboczej użytkownika, gdyż po zmianie przeglądarki czy też włączenia trybu incognito i próbie ponownego przejścia procesu weryfikacji, użytkownikowi cały czas serwowana jest witryna z komunikatem o powodzeniu wykonania transakcji. Strona *feanoys.com* posiada wystawiony certyfikat:

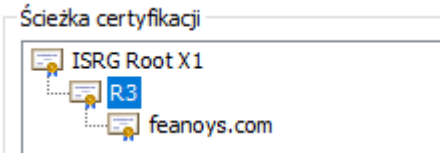


Rysunek 73. Dane certyfikatu witryny

Certyfikat został wystawiony w dniu **24.06.2022** – co jest zgodne z datą otrzymania wiadomości SMS z informacją o rzekomej dopłacie.



Rysunek 74. Data wystawienia certyfikatu.



Rysunek 75. Ścieżka certyfikacji.

Tabela 54. Dane identyfikacyjne domeny phishingowej.

Nazwa domenowa	feanoys.com
Adres IP	149.62.37.153
ASN	AS47583
Hosting	Hostinger International Limited
Geo	São Paulo, Brazylia

```
Domain Name: FEANOYS.COM
Registry Domain ID: 2706198330_DOMAIN_COM-VRSN
Registrar WHOIS Server: whois.hostinger.com
Registrar URL: https://www.hostinger.com
Updated Date: 2022-06-24T15:55:44Z
Creation Date: 2022-06-24T15:55:43Z
Registrar Registration Expiration Date: 2023-06-24T15:55:43Z
Registrar: Hostinger, UAB
Registrar IANA ID: 1636
Domain Status: clientTransferProhibited
https://icann.org/epp#clientTransferProhibited
Registry Registrant ID: Not Available From Registry
Registrant Name: Olena Kircenko
Registrant Organization:
Registrant Street: Reanovasti 31
Registrant City: Kiev
Registrant State/Province: Kiev
Registrant Postal Code: 10029
Registrant Country: UA
Registrant Phone: +380.665445553
Registrant Email: defenamjakla@gmail.com
Registry Admin ID: Not Available From Registry
Admin Name: Olena Kircenko
Admin Organization:
Admin Street: Reanovasti 31
Admin City: Kiev
Admin State/Province: Kiev
Admin Postal Code: 10029
Admin Country: UA
Admin Phone: +380.665445553
Admin Email: defenamjakla@gmail.com
Registry Tech ID: Not Available From Registry
Tech Name: Olena Kircenko
Tech Organization:
Tech Street: Reanovasti 31
Tech City: Kiev
Tech State/Province: Kiev
Tech Postal Code: 10029
Tech Country: UA
Tech Phone: +380.665445553
Tech Email: defenamjakla@gmail.com
Name Server: ns1.dns-parking.com
Name Server: ns2.dns-parking.com
DNSSEC: Unsigned
Domain Name: FEANOYS.COM
Registry Domain ID: 2706198330_DOMAIN_COM-VRSN
Registrar WHOIS Server: whois.hostinger.com
Registrar URL: https://www.hostinger.com
Updated Date: 2022-06-24T15:55:44Z
Creation Date: 2022-06-24T15:55:43Z
Registrar Registration Expiration Date: 2023-06-24T15:55:43Z
Registrar: Hostinger, UAB
Registrar IANA ID: 1636
Domain Status: clientTransferProhibited
https://icann.org/epp#clientTransferProhibited
Registry Registrant ID: Not Available From Registry
Registrant Name: Olena Kircenko
Registrant Organization:
Registrant Street: Reanovasti 31
Registrant City: Kiev
Registrant State/Province: Kiev
Registrant Postal Code: 10029
Registrant Country: UA
Registrant Phone: +380.665445553
Registrant Email: defenamjakla@gmail.com
Registry Admin ID: Not Available From Registry
Tech Email: defenamjakla@gmail.com
Name Server: ns1.dns-parking.com
Name Server: ns2.dns-parking.com
```

Rysunek 76. Dane właściciela phishingowej domeny, widoczne w publicznym rejestrze whois.

Dodatek D – macierze pomyłek dla zbiorów testowych

Macierz pomyłek dla niezrównoważonego zbioru danych:

		Regresja logistyczna				Lasy losowe				Drzewo Decyzyjne				Maszyna Wektorów Nośnych (SVM)				Naiwny Klasyfikatorr Bayesa			
		predict			Σ	predict			Σ	predict			Σ	predict			Σ	predict			Σ
Brak	real	7	9	0	16	7	9	0	16	7	9	0	16	7	9	0	16	0	16	0	16
		5	171	16	192	5	171	16	192	5	171	16	192	5	172	15	192	0	183	9	192
		5	22	46	73	5	20	48	73	5	20	48	73	5	20	48	73	0	33	40	73
		17	202	62	**281**	17	200	64	**281**	17	200	64	**281**	17	201	63	**281**	0	232	49	**281**
SMOTE	real	181	0	0	181	181	0	0	181	181	0	0	181	181	0	0	181	181	0	0	181
		25	147	29	201	25	149	27	201	25	149	27	201	25	152	24	201	25	155	21	201
		42	5	159	206	42	7	157	206	42	7	157	206	42	8	156	206	42	18	146	206
		248	152	188	**588**	248	156	184	**588**	248	156	184	**588**	248	160	180	**588**	248	173	167	**588**
SMOTEENN	real	119	0	0	119	119	0	0	119	119	0	0	119	119	0	0	119	0	119	0	119
		0	132	0	132	0	132	0	132	0	131	1	132	0	132	0	132	12	120	0	132
		0	0	127	127	0	0	127	127	0	0	127	127	0	0	127	127	3	0	124	127
		119	132	127	**378**	119	132	127	**378**	119	131	128	**378**	119	132	127	**378**	15	239	124	**378**
ADASYN	real	187	0	0	187	187	0	0	187	187	0	0	187	187	0	0	187	100	0	87	187
		27	152	40	219	27	154	38	219	27	156	36	219	27	154	38	219	15	153	51	219
		43	1	148	192	43	4	145	192	43	6	143	192	43	5	144	192	30	3	159	192
		257	153	188	**598**	257	158	183	**598**	257	162	179	**598**	257	159	182	**598**	145	156	297	**598**
ROS[206]	real	181	0	0	181	181	0	0	181	181	0	0	181	181	0	0	181	181	0	0	181
		25	145	31	201	25	155	21	201	25	146	30	201	25	149	27	201	25	153	23	201
		44	4	158	206	44	11	151	206	44	4	158	206	44	4	158	206	44	11	151	206

[206] ROS – Random Over Sampler

		250	149	189	**588**	250	166	172	**588**	250	150	188	**588**	250	153	185	**588**	250	164	174	**588**
RUS[207]	real	17	0	0	17	17	0	0	17	17	0	0	17	17	0	0	17	7	10	0	17
		2	10	0	12	2	10	0	12	2	10	0	12	2	10	0	12	0	12	0	12
		4	3	15	22	4	3	15	22	4	3	15	22	4	3	15	22	3	4	15	22
		23	13	15	**51**	23	13	15	**51**	23	13	15	**51**	23	13	15	**51**	10	26	15	**51**
Tomek Links	real	7	9	0	16	7	9	0	16	7	9	0	16	7	9	0	16	0	16	0	16
		5	171	16	192	5	171	16	192	5	171	16	192	5	172	15	192	0	183	9	192
		5	22	46	73	5	20	48	73	5	20	48	73	5	20	48	73	0	33	40	73
		17	202	62	**281**	17	200	64	**281**	17	200	64	**281**	17	201	63	**281**	0	232	49	**281**

Macierz pomyłek dla zrównoważonego zbioru danych:

		Regresja logistyczna				Lasy losowe				Drzewo Decyzyjne				Maszyna Wektorów Nośnych (SVM)				Naiwny Klasyfikatorr Bayesa			
		predict			∑	predict			∑	predict			∑	predict			∑	predict			∑
Brak	real	185	0	0	185	185	0	0	185	185	0	0	185	185	0	0	185	185	0	0	185
		22	154	24	200	22	152	26	200	22	155	23	200	22	157	21	200	22	163	15	200
		34	6	163	203	34	7	162	203	34	7	162	203	34	8	161	203	34	15	154	203
		241	160	187	**588**	241	159	188	**588**	241	162	185	**588**	241	165	182	**588**	241	178	169	**588**
SMOTE	real	185	0	0	185	185	0	0	185	185	0	0	185	185	0	0	185	185	0	0	185
		22	154	24	200	22	152	26	200	22	155	23	200	22	157	21	200	22	163	15	200
		34	6	163	203	34	7	162	203	34	7	162	203	34	8	161	203	34	15	154	203
		241	160	187	**588**	241	159	188	**588**	241	162	185	**588**	241	165	182	**588**	241	178	169	**588**
SMOTEENN	re	198	0	0	198	198	0	0	198	198	0	0	198	198	0	0	198	198	0	0	198
		0	138	0	138	0	138	0	138	0	138	0	138	0	138	0	138	0	138	0	138

[207] RUS – Random Under Sampler

		0	0	66	66	0	0	66	66	0	0	66	66	0	0	66	66	0	15	51	66
		198	138	66	**402**	198	138	66	**402**	198	138	66	**402**	198	138	66	**402**	198	153	51	**402**
ADASYN	real	185	0	0	185	185	0	0	185	185	0	0	185	185	0	0	185	185	0	0	185
		22	154	24	200	22	152	26	200	22	157	21	200	22	157	21	200	22	163	15	200
		34	6	163	203	34	7	162	203	34	8	161	203	34	8	161	203	34	15	154	203
		241	160	187	**588**	241	159	188	**588**	241	165	182	**588**					241	178	169	**588**
ROS[208]	real	185	0	0	185	185	0	0	185	185	0	0	185	185	0	0	185	185	0	0	185
		22	154	24	200	22	152	26	200	22	155	23	200	22	157	21	200	22	163	15	200
		34	6	163	203	34	7	162	203	34	7	162	203	34	8	161	203	34	15	154	203
		241	160	187	**588**	241	159	188	**588**	241	162	185	**588**	241	165	182	**588**	241	178	169	**588**
RUS[209]	real	200	0	0	200	200	0	0	200	200	0	0	200	200	0	0	200	89	0	111	200
		21	160	22	203	21	163	19	203	22	161	20	203	22	162	19	203	16	165	22	203
		38	4	143	185	38	6	141	185	38	6	141	185	38	6	141	185	30	13	142	185
		259	164	165	**588**	259	169	160	**588**	260	167	161	**588**	260	168	160	**588**	135	178	275	**588**
Tomek Links	real	185	0	0	185	185	0	0	185	185	0	0	185	185	0	0	185	185	0	0	185
		22	154	24	200	22	152	26	200	22	155	23	200	22	157	21	200	22	163	15	200
		34	6	163	203	34	7	162	203	34	7	162	203	34	8	161	203	34	15	154	203
		241	160	187	**588**	241	159	188	**588**	241	162	185	**588**	241	165	182	**588**	241	178	169	**588**

[208] ROS – Random Over Sampler
[209] RUS – Random Under Sampler