ŽILINSKÁ UNIVERZITA V ŽILINE

Fakulta riadenia a informatiky

AUTOREFERÁT

DIZERTAČNEJ PRÁCE

Žilina, máj 2024 **Ing. Ivan Škula**

ŽILINSKÁ UNIVERZITA V ŽILINE

Fakulta riadenia a informatiky

Ing. Ivan Škula

Autoreferát dizertačnej práce

Detekcia phishingu vo webových stránkach

na získanie akademického titulu **„philosophiae doctor“ (v skratke PhD.)**

v študijnom programe doktorandského štúdia

aplikovaná informatika

v študijnom odbore:

informatika

Žilina, apríl 2024

Dizertačná práca bola vypracovaná v externej forme doktorandského štúdia **na Katedre informatiky, Fakulte riadenia a informatiky Žilinskej univerzity v Žiline.**

Predkladateľ: Ing. Ivan Škula
Katedra informatiky
Fakulta riadenia a informatiky
Žilinská univerzita v Žiline

Školiteľ: doc. Ing. Michal Kvet, PhD.
Katedra informatiky
Fakulta riadenia a informatiky
Žilinská univerzita v Žiline

Oponent: prof. Ing. Marcel Harakaľ, PhD.
Katedra informatiky
Akadémia ozbrojených síl generála M. R. Štefánika
Liptovský Mikuláš

Oponent: doc. Ing. Jarmila Škrinárová, PhD.
Katedra informatiky
Fakulta prírodných vied Univerzity Mateja Bela
Banská Bystrica

Autoreferát bol rozoslaný dňa: 27. júna 2024

Obhajoba dizertačnej práce sa koná dňa **22. augusta 2024** o **12:30** hod. pred komisiou pre obhajobu dizertačnej práce schválenou pracovnou skupinou odborovej komisie v študijnom odbore **informatika** v študijnom programe **aplikovaná informatika**, vymenovanou dekanom Fakulty riadenia a informatiky Žilinskej univerzity v Žiline.

prof. Ing. Ľudmila Jánošíková, PhD.
predsedníčka pracovnej skupiny odborovej komisie
v študijnom odbore **informatika**
v študijnom programe **aplikovaná informatika**

Fakulta riadenia a informatiky
Žilinská univerzita
Univerzitná 8215/1
010 26 Žilina

Obsah

1. Úvod

Phishing je jednou z najstarších a najbežnejších techník využívaných kyber-zločincami. Je tak rozšírený (počet phishingových emailov sa odhaduje na ≈3.4 miliardy [6]), že s určitým zjednodušením môžeme povedať, že každý človek s e-mailom je denne terčom phishingového útoku. Je nesporné, že phishing je vážny problém, ktorý má vplyv na každého z nás individuálne aj celospoločensky.

Hoci phishing existuje už dlho, jeho rozšírenosť neustále a nepretržite rastie. V posledných rokoch - ako je vidieť v Tabuľke 1 - spolu s digitálnou transformáciou mnohých organizácií a vládnych služieb [5] zaznamenal obrovský nárast. V dôsledku týchto zmien je phishing bežnejší ako kedykoľvek predtým. Šíri sa prostredníctvom mnohých nových kanálov, ako sú SMS, hlasové hovory, nahrané hlasové správy, chatové aplikácie v rámci sociálnych sietí a cez qr kódy popri tradičnejšom e-maile a webe.

Tabuľka 1 Počet obetí phishingu v USA podľa IC3 (FBI) [3],[4]

	2018	**2019**	**2020**	**2021**	**2022**	**2023**
Obete	26,379	114,702	241,342	323,972	300,497	298,878
zmena %		435%	210%	134%	93%	99%

Výskum súvisiaci s phishingom je medzi vedcami a výskumníkmi čoraz populárnejší, pričom v posledných rokoch došlo k výraznému nárastu záujmu. Postupne sa zvyšuje aj počet výskumných publikácií zameraných na techniky detekcie phishingu [2]. Žiaľ, napriek všetkému úsiliu a po mnohých rokoch

výskumu sa dá boj proti phishingu nazvať „hrou na mačku a myš“ medzi zločincami, ktorí prispôsobujú a zdokonaľujú svoje techniky a snažia sa dosiahnuť svoje ciele, a technologickými spoločnosťami, výskumníkmi, bezpečnostnými nadšencami a vládnymi subjektmi, ktorí sa snažia zabrániť alebo zmierniť ich dopady.

Táto práca sa zameriava na návrh a implementáciu riešenia na detekciu phishingových webstránok v reálnom čase s využitím algoritmov strojového učenia. K úspešnej formulácii a návrhu riešenia bolo potrebné zanalyzovať a dekomponovať phishing na jeho stavebné časti. Za také môžeme považovať napr. bežné prípady použitia, rôzne techniky používané na uskutočnenie útokov a indikátory identifikujúce škodlivé správy alebo webové stránky. Prehľad a analýza výskumu v tejto oblasti nám umožnili sformulovať teoretické a praktické ciele a nasmerovať výskum k dosiahnutiu efektívneho riešenia s požadovanou mierou detekcie. Naše úlohy, experimenty a samotný výskum následne hľadali odpovede na sformulované hypotézy a definované ciele. Praktická časť práce opisuje proces návrhu a implementácie softvérového riešenia, počnúc zberom údajov, spracovaním súborov údajov, trénovaním a porovnávaním prediktívnych modelov a napokon integráciou natrénovaného modelu do riešenia na detekciu phishingových webstránok, ktoré sme nazvali **PhishCheck**.

2. Súčasný stav

Aktuálne neexistuje jednotné, všeobecné riešenie na elimináciu phishingu v celej svojej variabilite. Avšak aj drobné vylepšenia existujúcich prístupov môžu v súčasnom svete pomôcť znížiť celkové dopady phishingu. Rôzne, v súčasnosti aplikované riešenia, sa snažia riešiť phishing prostredníctvom:

- školení zameraných na osvetu a informovanosť používateľov aj prostredníctvom simulovaných phishingových správ (cieľom tohto procesu je priebežne monitorovať a zlepšovať schopnosť ľudí rozpoznať phishingový útok),
- riešení monitorujúcich počítačové siete, ktoré sledujú a filtrujú podozrivú a škodlivú komunikáciu,
- pokročilých riešení na filtrovanie e-mailov, ktoré hlási alebo priamo odstraňuje pokusy o phishingové útoky (autonómne alebo s podporou používateľa alebo prevádzkovateľa riešenia),
- aplikácií pre koncové body - rozšírenia/pluginy prehliadačov [1] alebo rôzny softvérový agenti - ktorí analyzujú vlastnosti navštívenej webovej stránky a rozhodujú, či sa používateľ pokúša o prístup na bezpečnú alebo phishingovú, prípadne inak nebezpečnú stránku.

3. Ciele práce

V práci sme sa zamerali na niekoľko teoretických (**T**) a praktických (**P**) cieľov v súvislosti so zameraním a názvom našej práce:

T1. Analýza a kategorizácia typov phishingu od historicky najstarších po najnovšie techniky podľa rôznych klasifikačných kritérií (komunikačný kanál, cieľ útoku, aplikované techniky sociálneho inžinierstva, objekt útoku a počet fáz).

T2. Zozbieranie a popis identifikačných znakov phishingových stránok (znakov, vďaka ktorým je možné odlíšiť phishingovú stránku od autentickej).

T3. Popis osvedčených postupov a riešení problémov súvisiacich so získavaním phishingových údajov a úpravami kvality údajov.

T4. Identifikácia, klasifikácia a popis rozlišovacích charakteristík webstránok použiteľných na trénovanie zvolených predikčných algoritmov na detekciu phishingových webstránok.

T5. Vytvorenie úplnej sady krokov vedúcich k vytvoreniu súboru údajov pre použitie algoritmov prediktívnej analytiky v oblasti detekcie phishingových webstránok.

T6. Prehľad charakteristík a metrík presnosti pre zvolené algoritmy strojového učenia.

P1. Implementácia softvérového riešenia na automatizovanú detekciu phishingových webových stránok.

P2. Návrh a implementácia Blacklistu phishingových domén.

Na realizáciu uvedených cieľov – predovšetkým praktických - bolo nevyhnutné získať relevantné phishingové dáta, ktorých všeobecná dostupnosť je problematická. Z tohto dôvodu predstavuje nie malú časť obsahu práce práve popis postupov súvisiacich so zberom údajov a riešenie s tým súvisiacich problémov.

4. Výsledky práce a diskusia

Zatiaľ čo ľudia sa pri identifikácii phishingových webových stránok zameriavajú na vizuálne znaky, počítačové algoritmy môžu využívať rôzne techniky. Tie jednoduchšie, ako sú Blacklist, alebo Whitelist, sa spoliehajú na rýchle a spoľahlivé uchovanie údajov (napr. domény alebo IP adresy), ktoré boli už v minulosti klasifikované ako phishingové alebo naopak ako bezpečné. Iné - ako napríklad expertne formulované pravidlá - využívajú podmienenú logiku na posúdenie charakteristík webovej stránky a snažia sa identifikovať spoločný identifikačný prvok pre legitímne alebo phishingové webové stránky s cieľom odvodiť konečnú klasifikáciu. Väčšina v súčasnosti-skúmaných techník využíva algoritmy strojového učenia na odhalenie hlbších alebo nie ľahko pozorovateľných korelácií v rámci rôznych pozorovateľných charakteristík s cieľom rozhodnúť o konečnom výsledku klasifikácie.

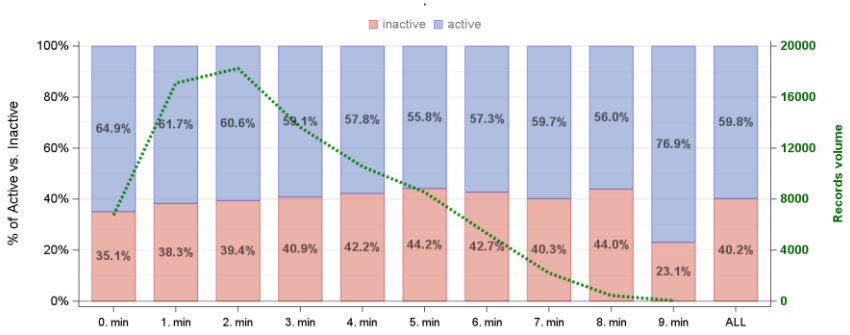
V práci sme sa venovali všetkým vyššie uvedeným technikám. Prvá analýza sa zamerala na relevantnosť používania Blacklistu a Greylistu (zoznam domén, ktoré nie je možné jednoznačne klasifikovať ako phishingové alebo legitímne – napr. webhosting domény, ktoré sú dostupné zdarma ako sub-domény jednej hlavnej domény), najmä v posledných rokoch. Tento výskum potvrdil náš pôvodný predpoklad o klesajúcom účinku Blacklistu - v roku 2022 bolo iba 6,1 % domén takých, ktoré boli už v minulosti hlásené ako podozrivé. Ide teda o maximálny potenciálny efekt Blacklistu k odhaľovaniu phishingu, keďže sa dá aplikovať iba na domény, ktoré sme už v minulosti klasifikovali ako phishingové. Napriek tomuto nízkemu číslu a samotnému klesajúcemu trendu

opakovane sa vyskytujúcich domén sa domnievame, že toto nízke číslo je v skutočnosti aj dôsledkom stáleho používania Blacklistov a dôvodom, prečo musia útočníci registrovať stále nové a nové domény.

Ďalší výskum, ktorý sme uskutočnili a publikovali, sa zameral na expertne definované pravidlá. Na konkrétnom príklade - rozšírenosti bežných techník ktoré majú za cieľ zmiasť používateľa maskovaním skutočnej URL adresy (angl. obfuscation). Tieto techniky sa stále používajú aj v novo publikovaných výskumných článkoch ako vynikajúci indikátor phishingu. Na základe našej analýzy nás prekvapilo to, ako zriedkavo sa tieto techniky vyskytujú medzi phishingovými URL adresami. V našom výskume sme analyzovali a zdokumentovali výskyt sedem takýchto techník na phishingových dátach v rozmedzí rokov 2009-2023 a konečné číslo bolo, že menej ako 3% všetkých phishingových URL využíva aspoň jednu takúto techniku. Detekčné riešenie spoliehajúce sa iba na týchto sedem techník by preto malo minimálnu úspešnosť a schopnosť detegovať phishing v celej svojej variabilite.

Ďalšia podrobná analýza sa zamerala na životnosť phishingových webových stránok. Náš výskum potvrdil veľmi krátku životnosť phishingových webových stránok, 35,1 % nahlásených phishingových webových stránok bolo nedostupných ihneď po ich sprístupnení v najčastejších zdrojoch phishingových URL adries - PhishTank a OpenPhish.

Po prvých piatich minútach – viď Obrázok 1 - bolo nedostupných ďalších ≈9% nahlásených URL adries; po 24 hodinách je dostupných už len ≈41 % a po 48 hodinách len 36% nahlásených URL adries.

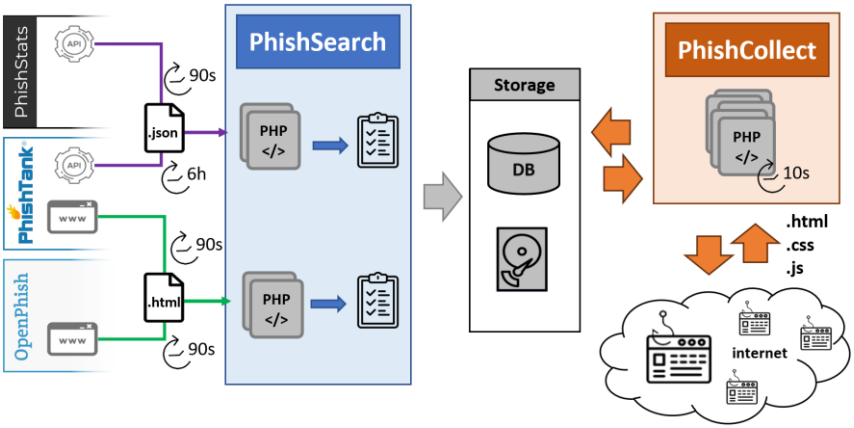


Obrázok 1 Dostupnosť phishingových webstránok v prvých 10 minútach

Každý, kto plánuje zbierať phishingové údaje, musí navrhnúť riešenie zberu tak, aby zachytil všetky zbierané údaje (HTML, whois, ping apod.) ihneď po nahlásení a sprístupnení URL adries.

V ďalšej časti práce sme sa venovali procesu zberu údajov. Tri webové aplikácie (viď Obrázok 2) sú zodpovedné za zber údajov o phishingu:

1. **PhishSearch** - zodpovedá za zber novo nahlásených adries URL z phishingových zdrojov (PhishTank, PhishStats a OpenPhish), ich čítanie, spracovanie a ukladanie do databázy v reálnom čase.
2. **PhishCollect** - monitoruje novo zachytené adresy URL z PhishSearch, zbiera a archivuje všetky relevantné podrobnosti vťahujúce sa k URL adrese na ďalšie spracovanie (adresa URL, HTML stránka, favicon, whois, ping, atď.).
3. **PhishLongevity** - zodpovedá za monitorovanie a zachytávanie stavu dostupnosti novo zaznamenaných adries URL v databáze vo vopred definovaných časových obdobiach, až kým sa obsah umiestnený na adrese URL stane nedostupným.



Obrázok 2 Proces zberu phishigových údajov cez PhishSearch a PhishCollect

Keďže zber údajov prebieha kontinuálne, z praktického hľadiska bolo potrebné zapracovať aj rýchlo dostupné reporty, ktoré zobrazujú aktuálny priebeh zberu údajov. Súčasne, v prípade výpadku, ho umožnia okamžite spozorovať. Takéto riešenie sme nazvali **PhishReport** (Obrázok 3), ide o grafický dashboard, ktorý je implementovaný s použitím open-source riešenia Grafana, ktoré číta údaje z našej databázy.

Obdobným spôsobom, akým zbierame URL adresy pre phishingové údaje, sme implementovali samostatnú webovú aplikáciu s názvom **URLCollect** na zhromažďovanie charakteristík pre legitímne stránky. Z hľadiska návrhu poskytuje **URLCollect** analogickú funkcionalitu ako **PhishCollect**.

Ďalším dôležitým krokom po zbere údajov je príprava súboru dát pre strojové učenie. V práci sme detailne rozpísali celý postup takejto prípravy od zdrojov dát, krokov súvisiacich s dátovou kvalitou ako de-duplikácia údajov,

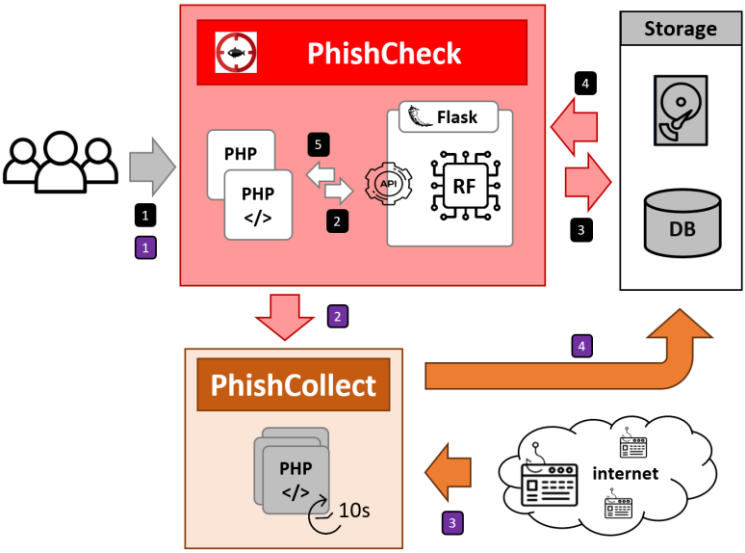
mazanie nerelevantných záznamov, až po analýzu vplyvu rôznych charakteristík na modeli zvolených algoritmov strojového učenia - ako veľkosť dátovej sady, početnosť charakteristík, štruktúra súboru dát (pomer legitímnych a phishingových stránok, zastúpenie rôznych typov priemyslu, jazykov a iné).

Obrázok 3 Monitorovací report vo PhishReport

Záverečná časť práce popisuje kroky návrhu a implementácie softvérového riešenia PhishCheck. **PhishCheck** je webová aplikácia vytvorená na vyhodnotenie, či je zadaná adresa URL phishingová alebo legitímna v reálnom čase (viď Obrázok 4 a Obrázok 5).

Používateľské rozhranie je vytvorené pomocou v jazyku PHP. Flask napísaný v jazyku Python je nositeľom prediktívneho modelu - v našom prípade modelu Random Forest s využitím 30-tich zvolených charakteristík s testovanou presnosťou ≈97.7%, ktorý klasifikuje URL adresu ako phishingovú alebo

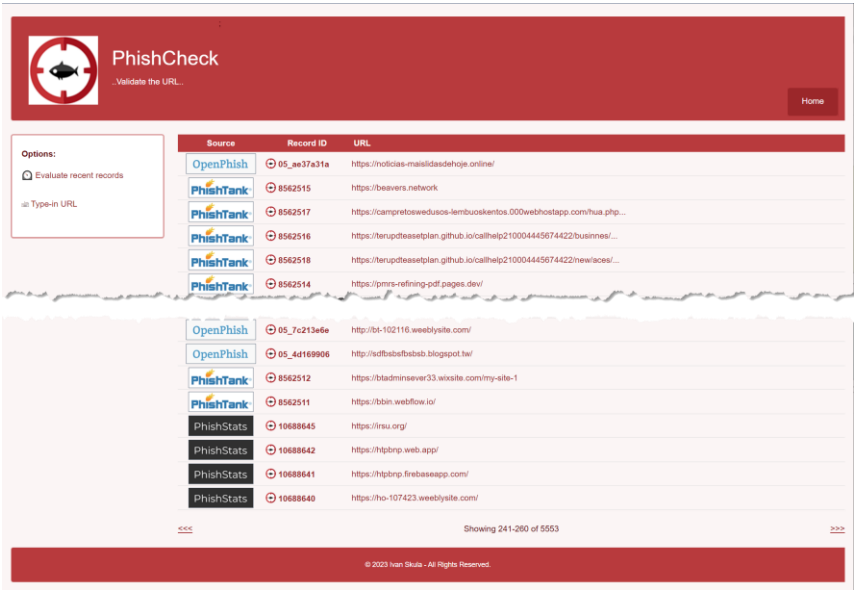
legitímnu a je interaguje s PHP kódom prostredníctvom vrstvy REST API vytvorenej vo Flask-u.



Obrázok 4 Kroky spracovania nahlásenej URL vo PhishCheck

PhishCheck má vo aktuálne implementované dva pracovné postupy:

- Klasifikácia aktuálne spracovaných nahlásených adries URL - používateľ si môže vybrať z nedávno nahlásených adries URL, ktoré spracoval program **PhishCollect**. Interakcia sa začína výpisom nedávno nahlásených adries URL z nášho zdroja údajov (PhishTank, PhishStats a OpenPhish), ako je znázornené na Obrázok 4 – kroky sú reprezentované číslami uvedenými v čiernom štvorci. Výber z aktuálne spracovaných záznamov je vidieť na Obrázok 5.



Obrázok 5 Používateľské rozhranie PhishCheck – výpis hlásených URL adries

- Posúdenie ručne zadanej URL - tento proces je zobrazený na Obrázok 4 a začína na kroku 1 a končí na kroku 4 vo fialovom štvorci, kedy je zadaná URL najskôr spracovaná pomocou **PhishCollect** (pre zozbieranie charakteristík z HTML a whois) a následne je používateľ navigovaný na krok 2 v čiernom štvorci a proces pokračuje ako v predošlom postupe - podľa krokov uvedených v čiernych štvorcoch.

Počas nášho výskumu sme sa snažili venovať veľkú pozornosť možným obmedzeniam, ktoré by mohli skresľovať výsledky. Aby sme eliminovali potenciálnu závislosť od zdroju dát v našej štúdii, zapracovali sme zber z viacerých zdrojov a vytvorili súbory údajov s použitím týchto zdrojov. Údaje,

ktoré sme zozbierali a použili v našom výskume, by mali obsahovať dostatočnú rozmanitosť phishingových variantov, keďže pochádzajú z viacerých zdrojov. To isté platí aj pre legitímne údaje, ktoré pochádzajú z najlepšej možnej reprezentácie webu - Common Crawl. Takisto použité údaje (od roku 2022) boli zozbieraná okamžite po ich nahlásení, takže máme k dispozícii najviac rozmanitú sadu nahlásených phishingových webových stránok, ktorú bolo možné zaznamenať. Napriek tomu môžu existovať určité typy phishingových útokov, ktoré sú zriedkavé alebo majú extrémne krátku životnosť, ktoré nemusia byť v zozbieraných údajoch dostatočne zastúpené (napr. môžu spadať do množiny 35 % nahlásených phishingových URL adries, ktoré sa nedali načítať už v momente ich zverejnenia). V prípade rizika "pretrénovania" modelu sme využili samostatný testovací súbor údajov s dostatočnou veľkosťou a bez prieniku s trénovacími dátami a jasne sme uviedli vybrané ukazovateľ presnosti modelu pre tréningový a testovací súbor údajov oddelene.

S ohľadom na mnohé publikované práce a algoritmy, ktoré sa používajú na detekciu phishingu, možno nami zvolený súbor algoritmov strojového učenia (logistická regresia, rozhodovací strom, SVM, Random Forest, K-najbližší sused, Naive Bayes, Gradient Boost, Adaptive Boost) považovať za limitujúci. Išlo však o vedomé rozhodnutie, pri ktorom náš súbor algoritmov považujeme za východiskový a v budúcnosti plánujeme vyhodnotiť aj iné algoritmy s cieľom ďalej vylepšiť presnosť modelu. Pokiaľ ide o charakteristiky webstránok, snažili sme sa zapracovať široké spektrum, odvodených z rôznych prvkov súvisiacich s webstránkou. Celkovo sme použili 253 charakteristík.

5. Využitie poznatkov v praxi

Všetky ciele práce, spolu s publikovaným výskumom, majú praktický a hmatateľný dopad nielen na ciele našej práce, ale aj na ďalšie možné smery výskumu v tejto oblasti.

Jedným z pozorovaných problémov pri súčasnom skúmaní phishingu je obmedzená možnosť porovnávania výsledkov od rôznych výskumníkov. Aj pri použití rovnakého súboru údajov, tieto často upravujú takým spôsobom, ktorý znemožňuje alebo limituje porovnateľnosť získaných výstupov. Mnohí vynechávajú relevantné detaily o zdroji, alebo o období, počas ktorého boli údaje zbierané. Úpravy zozbieraných údajov, ako je pridávanie legitímnych stránok k zozbieraným phishingovým údajom, rôzne techniky čistenia údajov od duplicitných údajov, odstraňovanie extrémnych hodnôt alebo dokonca dopĺňanie údajov, sú len niektoré z úprav, ktoré negatívne ovplyvňujú porovnateľnosť výsledkov výskumu, ak nie sú zachytené a transparentne zadokumentované. Štandardizačný alebo unifikačný rámec pre kroky prípravy súboru údajov, ktorý sme vytvorili a publikovali, má poskytnúť presne definované kroky, ktoré pomôžu výskumníkom s touto úlohou.

Ďalší praktický príspevok k výskumu v tejto oblasti je zachytený v kapitolách týkajúcich sa údajov - zhrnutie problémov, ktoré sa vyskytli počas nášho procesu zberu údajov. Jedným z pozorovaných problémov je krátka dostupnosť phishingových webových stránok. Zhrnutie ďalších, ako filtrovanie obsahu webu (v prípade niektorých krajín), vplyv antivírusových programov

alebo technológií proti phishingu - spolu s navrhovanými riešeniami poskytuje praktickú pomoc ďalším výskumníkom pri zbere údajov.

Nami navrhnuté a implementované softvérové nástroje na zber údajov je možné použiť pre akýkoľvek projekt zachytávajúci dáta z internetu.

Prínosom je aj detailný popis návrhu a implementácie riešenia na detekciu phishingu v reálnom čase spolu so zadefinovaním implementačných nástrojov. Nápomocné by mali byť aj popisy a výsledky experimentov, ktoré pomohli navrhnúť parametre finálneho algoritmu a modelu strojového učenia – Random Forest s 30 vybranými charakteristikami.

Tieto experimenty analyzovali:

a) Výber množiny najrelevantnejších ukazovateľov na rozlíšenie legitímnej URL/webovej stránky od phishingu z celkovej sady 253 charakteristík.
b) Algoritmus strojového učenia s najvyššou presnosťou pri rôznych veľkostiach súboru dát a s rôznymi charakteristikami zo súboru 8 algoritmov.
c) Optimálnu veľkosť trénovacej sady údajov s cieľom dosiahnutia čo najpresnejšieho modelu.

6. Ďalší rozvoj

Existuje viacero potenciálnych možností, ako rozšíriť, zlepšiť alebo doplniť našu prácu. Na strane zberu údajov navrhujeme prehodnotiť a zlepšiť proces zberu charakteristík webstránky – predovšetkým údajov o polohe IP adresy, údajov z whois a ping-u (všetky odvodené od registrovanej domény). Zlepšeniu v tejto oblasti by pomohlo zapracovanie viacúrovňovej Top-Level Domain (TLD, napr. **co.uk** namiesto iba **.uk,** alebo **co.au** namiesto iba **.au**).

Okrem toho navrhujeme optimalizovať proces zachytávania charakteristík, keďže súčasný proces ukladá väčšinu údajov súvisiacich s HTML na disk vo forme súborov, ktoré sú na periodickej báze archivované. Extrakcia charakteristík z týchto údajov si vyžaduje manuálne kopírovanie veľkého množstva údajov z archívu na server, ich extrakciu a prípravu na spracovanie vrátane nezanedbateľných požiadaviek na voľný diskový priestor ako aj dodatočné požiadavky na udržiavanie samostatného procesu na extrakciu príslušných charakteristík z týchto archivovaných súborov.

Tiež navrhujeme vyhodnotenie a zapracovanie ďalších charakteristík - favicon, ako aj analýzu snímky webovej stránky. Ďalšou oblasťou je vyhodnotenie presnosti iných typov algoritmov a techník - najmä neurónových sietí a potenciálnych multi-modálnych modelov využívajúcich textové údaje spolu s obrázkami na rozpoznanie phishingu.

Zaujímavou, ale málo preskúmanou oblasťou je stabilita detekčného modelu počas dlhšieho obdobia. Otázky typu - ako dlho funguje model v rámci vopred definovaného rozsahu hraničných hodnôt? Ktoré charakteristiky sú

menej stále a spôsobujú rýchle zhoršenie modelu a ktoré sú stabilnejšie a stále dobre predikujú phishingovú stránku? To sú príklady výskumných otázok súvisiacich s touto témou.

Jedna z prioritných vetiev výskumu by sa mohla zamerať na identifikáciu imitovanej značky, keďže väčšina phishingových webových stránok spadá do tejto kategórie phishingu – imitácia webu dobre známej značky. Ak by sme dokázali presne identifikovať napodobňované značky, mohli by sme s vysokou presnosťou identifikovať značnú časť phishingu jednoduchých porovnaním adresy URL s domovskou stránkou týchto značiek. Detekcia značiek by nám umožnila vytvárať upozornenia a monitorovať pokusy o phishing konkrétnych značiek.

7. Záver

Rozpoznať phishingovú stránku nie je jednoduchá úloha ani pre ľudí, ani pre počítače. Útočníci, ktorí vytvárajú phishingové webové stránky, sú neuveriteľne prispôsobiví a agilní. Phishing je nástrojom širokej škály útočníkov - od oportunistických a technicky málo zdatných až po tých, ktorí kybernetické útoky vykonávajú ako súčasť svojej práce alebo ako hlavný zdroj príjmu. Niektoré útoky sú pripravené veľmi podrobne a vyladené na starostlivo vybraného príjemcu, zatiaľ čo iné sú masovo rozposielané podľa bežných vzorov, ktoré sa už osvedčili. Je dôležité vedieť rozpoznať a uvedomiť si mnohé podoby, ktoré môže phishingový útok mať, pretože bez týchto znalostí by bolo vytvorenie praktického riešenia na detekciu phishingu ťažké, ak nie nemožné.

Ak sa detekcia phishingových webových stránok vykoná presne a rýchlo, významne ovplyvní vnímanie používateľov pri interakcii s online svetom a ich pocit bezpečnosti. Takéto riešenie zmierňuje priame straty spôsobené phishingom. Okrem toho pomáha znižovať vplyv mnohých ďalších, komplexnejších kybernetických útokov, ktoré využívajú phishing na konkrétny účel v rámci oveľa komplexnejšej schémy útoku.

Cieľom práce bolo navrhnúť a implementovať softvérové riešenia na detekciu phishingových webových stránok s využitím strojového učenia a s cieľom maximalizovať presnosť detekcie pre širokú škálu phishingových útokov získaných z aktuálnych phishingových dát.

Použitá literatúra

[1] M.A. Adebowale, K.T. Lwin, E. Sánchez and M.A. Hossain, "Intelligent web-phishing detection and protection scheme using integrated features of Images, frames and text," in *Expert Systems with Applications*, 2019, vol. 115, pp.300-313

[2] A. Das, S. Baki, A.E. Aassal, R.M. Verma and A. Dunbar, "SoK: A Comprehensive Reexamination of Phishing Research From the Security Perspective," in IEEE Communications Surveys & Tutorials, 2020, vol. 22, pp. 671-708.

[3] FBI Internet Crime Complaint Center, 2021 Internet Crime report, Accessed on: Mar. 29, 2024. [Online]. Available: https://www.ic3.gov/Media/PDF/AnnualReport/2021_IC3Report.pdf

[4] FBI Internet Crime Complaint Center, 2023 Internet Crime report, Accessed on: Mar. 29, 2024. [Online]. Available: https://www.ic3.gov/Media/PDF/AnnualReport/2023_IC3Report.pdf

[5] J. Reis, M. Amorim, N. Melão and P. Matos, "Digital Transformation: A Literature Review and Guidelines for Future Research," in Trends and Advances in Information Systems and Technologies, 2018, pp. 411–421.

[6] StationX, "Top Phishing Statistics for 2024: Latest Figures and Trends," Accessed on: Mar. 30, 2024. [Online]. Available: https://www.stationx.net/phishing-statistics/#:~:text=1.,trillion%20phishing%20emails%20per%20year.

Zoznam publikovaných prác

[IS1] J. Bohacik, I. Skula and M. Zabovsky, "**Data Mining-Based Phishing Detection**," in *15th Conference on Computer Science and Information Systems (FedCSIS)*, 2020, pp. 27-30, doi: 10.15439/2020F140.

[IS2] I. Skula, J. Bohacik and M. Zabovsky, "**Use of Different Channels for User Awareness and Education Related to Fraud and Phishing in a Banking Institution**," in *18th International Conference on Emerging eLearning Technologies and Applications (ICETA)*, 2020, pp. 606-612, doi: 10.1109/ICETA51985.2020.9379220.

[IS3] I. Skula, "**Automated detection techniques of phishing**," in *Mathematics in Science and Technologies (MIST)*, 2021, pp. 60-67.

[IS4] I. Skula, M. Kvet, "**Domain Blacklist Efficacy for Phishing Web-page Detection Over an Extended Time Period**," in *33rd Conference of Open Innovations Association (FRUCT)*, 2023, pp. 257-263, doi:10.23919/FRUCT58615.2023.10142999.

[IS5] I. Skula, M. Kvet, "**URL and Domain Obfuscation Techniques - Prevalence and Trends Observed on Phishing Data**," in *IEEE 22nd World Symposium on Applied Machine Intelligence and Informatics*, 2024, pp. 283-290, doi:10.1109/SAMI60510.2024.10432841.

[IS6] I. Skula, M. Kvet, "**Phishing Webpage Longevity**," in *WorldCist2024*, 2024, doi:10.1007/978-3-031-60328-0_21

[IS7] I. Skula and M. Kvet, "**A Framework for Preparing a Balanced and Comprehensive Phishing Dataset**," in IEEE Access, doi:10.1109/ACCESS.2024.3387437.