UDC: 004.056.55 Info M str. 28-37

DETEKCIJA MALICIOZNIH URL-OVA KORIŠTENJEM METODA MAŠINSKOG UČENJA
DETECTING MALICIOUS URLS USING MACHINE LEARNING METHODS

Jelena Jokić, prof. dr Zoran Đurić

REZIME: U ovom radu analizirane su metode detekcije malicioznih URL-ova koristeći algoritme mašinskog učenja sa ciljem otkrivanja pravilnosti u podacima koje nisu mogle biti detektovane tradicionalnim blacklist pristupima. Posebna pažnja posvećena je određivanju skupa atributa koji će se koristiti i implementaciji prikupljanja vrijednosti odabranih atributa. Kao praktični dio implementirano je obučavanje šest predloženih klasifikacionih algoritama na dva skupa podataka. Urađena je evaluacija dobijenih modela kojom se pokazuje da odabrani klasifikatori, sa predloženim skupom atributa, daju tačnost od 96-99%, dakle sa velikom vjerovatnoćom uspijevaju da tačno detektuju maliciozne URL-ove. Takođe, urađena je analiza greške radi boljeg razumijevanja problema i dati su pravci mogućeg daljeg unapređenja.

KLJUČNE REČI: Sigurnost na Internetu, Mašinsko učenje, Algoritmi, Maliciozni URL-ovi

ABSTRACT: In this paper, we describe methods for detecting malicious URLs using machine learning algorithms with a purpose of discovering rules in data which could not be detected by traditional blacklist approach. The particular challenge was to analyze data sets and choose the most appropriate features. In order to show that efficiently predicting malicious URLs can be done using just URL, without page content, we focus on lexical and host-based features. Further, we implement automated services for gathering and generating all proposed feature values. We explore six binary classifiers using two data sets. The experimental results show that the combination of the proposed URL features and classifiers in this paper can achieve accuracy 96-99%. We also discuss issues and indicate some important open problems for further research.

KEY WORDS: Internet security, Machine learning, Algorithms, Malicious URLs

1. UVOD

Razvojem informacionih tehnologija povećavaju se i mogućnosti malicioznih napada na krajnje korisnike. S druge strane, količina podataka dostupnih za analizu i učenje o malicioznim sistemima svakim danom je sve veća.

Jedinstveni identifikator resursa (eng. *Uniform Resource Locator*, URL) predstavlja jedinstvenu adresu resursa na *web*-u (*World Wide Web*)[1]. Primjer takvog resursa je *web* stranica. Maliciozni URL podrazumijeva postojanje *web* stranice koja distribuira maliciozan ili neželjen sadržaj krajnjim korisnicima. Detekcija malicioznih URL-ova jedan je od važnijih problema u oblasti sigurnosti na Internetu. Ovaj problem, koji je i ranije bio predmet izučavanja u računarskim naukama, sada se sve više rješava primjenom metoda mašinskog učenja. Tradicionalno, detekcija se uglavnom vršila pomoću odgovarajućih listi za filtriranje (eng. *blacklist*) [1]. Ove liste se jednostavno kreiraju, obično od strane velikog broja korisnika, a putem rješenja kao što su PhishTank [2] ili APWG [3]. Kada se potvrdi da je URL maliciozan dodaje se u listu i postaje lako dostupan svim rješenjima i alatima koji datu listu koriste. Korištenje ovakvih listi je krajnje jednostavno, a vrijeme odziva je veoma malo. Takođe, zbog načina na koji se kreiraju ove liste, procenat pogrešno klasifikovanih URL-ova je obično veoma mali, mada su zabilježene i drugačije situacije [4]. Bez obzira na dobre osobine listi za filtriranje, one ipak nisu adekvatne današnjem vremenu. Osnovni razlozi za to su načini kreiranja i adresiranja malicioznih resursa, brzina njihovog kreiranja, tehnike maskiranja, kao i životni vijek malicioznih resursa koji je danas značajno kraći nego ranije. Ova dinamika utiče negativno na efikasnost listi za filtriranje, prvenstveno zbog toga što su one u mogućnosti da detektuju samo one maliciozne URL-ove koji su ranije prijavljeni i potvrđeni kao maliciozni. Često je taj proces vremenski duži od životnog vijeka samog malicioznog resursa. Iz tog razloga se posljednjih godina koriste i drugi pristupi u detekciji malicioznih URL-ova koji su prvenstveno bazirani na tehnikama mašinskog učenja.

Ovaj rad se bavi detekcijom malicioznih URL-ova korištenjem klasifikacionih algoritama mašinskog učenja, a na bazi odgovarajućeg skupa identifikovanih atributa koji opisuju maliciozne URL-ove. U ovom radu analizirane su postojeće metodologije i mogućnosti njihovog unapređenja. Glavni doprinos ovog rada jeste identifikovani skup atributa za obučavanje modela. Pored toga, ovaj rad bavi se i metodama prikupljanja vrijednosti identifikovanih atributa (tj. generisanjem skupa podataka), izborom odgovarajućih algoritama i vrijednosti njihovih parametara u cilju dobijanja najboljeg modela za detekciju malicioznih URL-ova. Atributi za obučavanje modela koji su predloženi u ovom radu obuhvataju leksičke atribute izdvojene iz samih URL adresa, te atribute bazirane na informacijama o DNS (*Domain Name System*) i WHOIS zapisima domena kojim pripadaju analizirane *web* stranice i AS broju. DNS i WHOIS zapisi domena prikupljeni su namjenski razvijenim alatom. Obučavanje i testiranje modela izvršeno je na dva skupa podataka. Veći skup sadrži oko 230 miliona zapisa. Sve manipulacije ovim skupom podataka vršene su u distribuiranom okruženjum na Spark[2] *cluster*-u. Na manjem skupu podataka izvršena je opširnija pretraga hiper parametara (eng. *hyper parameters*) testiranih algoritama, tj. izvršeno je određivanje vrijednosti parametara koje algoritam učenja nije u stanju da optimizuje. Evaluacijom i testiranjem obučenih klasifikatora, pokazano je da generisani modeli daju tačnost od 96-99%.

[1] Uniform Resource Identifier (URI): Generic Syntax, RFC 3986

[2] https://spark.apache.org

2. SLIČNA ISTRAŽIVANJA

Posljednjih godina intenzivno se provode istraživanja u kojim se primjenom mašinskog učenja pokušava povećati uspješnost detekcije malicioznih URL-ova. U radu [5] korištena je logistička regresija (eng. *Logistic Regression, LR*) na skupu od osamnaest ručno odabranih atributa za detekciju *phishing* napada. Atribute čine različita rangiranja *web* stranica, pripadnost *white* listama, atributi koji opisuju određeni tip obfuskacije URL-a i atributi koji potvrđuju prisustvo određenih ključnih riječi unutar URL-a. Tačnost testiranog klasifikacionog modela na skupu podataka od 2.500 zapisa je 97,3%. U radu [6] za obučavanje klasifikacionog algoritma korišteni su leksički atributi izdvojeni iz URL adresa i atributi bazirani na informacijama o domenima kojim analizirane *web* stranice pripadaju. Neki od leksičkih atributa korištenih u ovom radu s-u dužina imena hosta, dužina URL-a i broj tačaka u URL adresi. Od važnijih atributa koji su bazirani na informacijama o domenima kojim pripadaju analizirane *web* stranice korištene su informacije o WHOIS registru preko kojeg je domen registrovan, informacije o vlasniku domena, informacije o A, NS i MX DNS zapisima, TTL (*Time to live*) vrijednost, prisustvo domena u određenim *blacklist*-ama, brzina konekcije prema *web* serverima na kojim se nalaze analizirani URL-ovi, te kontinent/država/grad kojoj IP (*Internet Protocol*) adresa pripada. U ovom radu data je komparativna analiza *batch* i *online* algoritama za klasifikaciju. Na skupu podataka od dva miliona zapisa, model obučavan korištenjem *online* algoritma za klasifikaciju dao je tačnost od 99%. U radu [7] se koristi skup strukturalnih atributa i atributa koji se dobijaju statističkom analizom URL-a i predstavljaju distribuciju frekvencije karaktera (slova i specijalnih karaktera). Neki od korištenih strukturalnih atributa su IP adresa, dužina imena domena, dužina URL-a i broj tačaka u URL adresi, ali i prisustvo IP adrese ili specijalnih karaktera unutar URL stringa. U ovom radu testirano je šest algoritama: LR, Bajesov klasifikator (eng. *Naive Bayes, NB*), klasifikaciono stablo odlučivanja (eng. *Decision Tree, DT*), šuma slučajnih stabala (eng. *Random Forest, RF*), metoda potpornih vektora (eng. *Support Vector Machine, SVM*) i *Multi-layer Perceptron* (MLP)**.** Obučavanje modela izvršeno je na dva skupa, skupu podataka od 29.000 zapisa gdje je odnos malicioznih i benignih URL-ova 43:57 i skupu od 62.573 zapisa sa odnosom malicioznih i benignih URL-ova 39,8:61,2. U više provedenih eksperimenata, najbolje rezultate daje RF sa preciznošću i odzivom od 99% na prvom, i preciznošću i odzivom od 93% na drugom skupu. Slična poređenja više klasifikatora opisana su i u radu [8]. Skup podataka koji je korišten u ovom radu sadrži 2,4 miliona zapisa sa 3,2 miliona atributa od kojih su 64 ne-binarni numerički atributi. Skup podataka je podijeljen na tri skupa u zavisnosti od tipa atributa i njihovih vrijednosti. Algoritmi koji su testirani na generisanim skupovima su: NB, MLP, DT, RF i klasifikacija na osnovu k najbližih susjeda (eng. *K-Nearest Neighbor*, kNN). U provedenom eksperimentu najbolje rezultate dao je model obučavan RF algoritmom, sa tačnošću od 97,69%. Poređenje RF algoritma sa drugim klasifikacionim algoritmima opisano je i u radu [9], gdje su korišteni samo leksički atributi. Slično kao i u prethodno navedenim istraživanjima, RF algoritam se pokazao efikasnijim od NB, LR i DT. U radu [10], predloženo je rješenje za detekciju *phishing* URL-ova korištenjem samo deskriptivnih i statističkih karakteristika URL adrese bez korištenja leksičkih, *bag-of-words* ili atributa baziranih na informacijama o hostu. Iako na jednostavnom skupu atributa, nakon evaluacije, RF daje preciznost od 85% i odziv od 87% a SVM preciznost od 90% i odziv od 88%.

3. OPIS PROBLEMA

Maliciozni URL-ovi podrazumijevaju postojanje *web* stranica koje distribuiraju maliciozan ili neželjen sadržaj krajnjim korisnicima. Ovakve aktivnosti provode se u cilju ostvarivanja finansijske dobiti za vlasnika malicioznih *web* stranica, a na štetu krajnjih korisnika koji ovim sadržajima pristupaju, te se često mogu smatrati i kriminalnim aktivnostima [11, 12, 13]. Postojanje takvih domena predstavlja veliku prijetnju za sigurnost krajnjih korisnika na internetu, jer sadržaji koji se putem njih distribuiraju mogu dovesti do gubitka važnih podataka, poput korisničkih kredencijala, narušavanja privatnosti, a veoma često mogu nanijeti i finansijsku štetu krajnjim korisnicima. Neki od primjera malicioznih napada su pokušaji krađe važnih informacija o korisniku navodeći istog da pristupi falsifikovanim *web* stranicama (eng. *phishing*), napadi kojim se krajnji korisnik navodi da uplati novac na račun napadača (eng. *advance-fee scam*), te napadi koji podrazumijevaju preuzimanje datoteka od strane krajnjeg korisnika bez njegove saglasnosti ili bez razumijevanja mogućih posljedica (eng. *drive-by-downloads*).

Jedan od posebno zastupljenih *advance-fee scam* napada je i napad pružanjem tehničke podrške (eng. *Technical Support Scam, TSS)* gdje se krajnji korisnik navodi da uplati novac za uslugu pružanja lažne tehničke podrške. Neki od kanala distribuiranja TSS su: alati za skraćivanje URL-ova (eng. *URL shortening services*), maliciozne reklame servirane putem parkiranih domena, zloupotreba *search engine* algoritama korištenjem *black hat SEO (Search Engine Optimization)* tehnike radi visokog rangiranja pomoćnih domena (domena koji vode ka TSS domenima) u rezultatima pretrage, zloupotreba mreža za reklamiranje itd. Agresivni TSS napadi su napadi koji intruzivnim metodama „zalede“ *web* čitač i time primoraju krajnjeg korisnika da pozove telefonski broj navedene tehničke podrške. Analizirajući DNS i WHOIS karakteristike TSS domena, uočeni su mnogobrojni pasivni TSS napadi gdje maliciozne *web* stranice izgledaju potpuno legitimno, te su samim tim takvi TSS teži za detekciju.

Uspješna detekcija malicioznih URL-ova omogućava unapređenje postojećih antivirusnih rješenja, kao i razvoj novih servisa, poput servisa koji obavještava krajnje korisnike o pokušaju pristupa potencijalno malicioznim *web* stranicama. Da bi se sistem za detekciju malicioznih URL-ova mogao smatrati uspješnim, potrebno je da bude brz i precizan, te da ima mogućnost otkrivanja novokreiranih malicioznih domena.

4. PRIJEDLOG ATRIBUTA

Jedan od najvažnijih koraka u implementaciji rješenja za detekciju malicioznih URL-ova predstavlja odabir atributa nad čijim vrijednostima će biti obučavan klasifikacioni model.

Kao što je već pomenuto, atributi korišteni u ovom istraživanju mogu se podijeliti na leksičke, koji su izdvojeni iz samih URL adresa, te atribute bazirane na informacijama o DNS i WHOIS zapisima domena kojim pripadaju analizirane *web* stranice.

4.1. *Leksički atributi*

Na osnovu vidljive razlike između „izgleda" malicioznih i benignih URL-ova, leksički atributi se nameću kao početni skup atributa za rješavanje problema detekcije malicioznih URL-ova. Leksički atributi odabrani nakon analize prikupljenog skupa podataka su:

− **A1 - Dužina URL-a**

Maliciozni URL-ovi često imaju dužine od nekoliko stotina do nekoliko desetina hiljada karaktera i lako se vizuelno razlikuju od benignih URL-ova.

− **A2 - Dužina imena domena**

Maliciozni URL-ovi često imaju duža imena domena nego benigni URL-ovi. Primjeri takvih malicioznih domena iz skupa podataka korištenih u ovom radu prikazani su na slici 1.

suspicious-bank-login-activity-call-now1.azurewebsites.net
windows-drive-not-responding-07-hacking-attempt-found-0697.s3.amazonaws.com
www.malwar-detect-safety-alert.website
system-security-alert-012-hacking-attempt-found-48.s3-us-west-1.amazonaws.com
securenetworkalert24x7official.monster

Slika 1. Primjeri malicioznih domena sa dužim imenima

− **A3 - Broj tačaka sadržanih u imenu domena**

Maliciozni domeni često sadrže veći broj tačaka u imenu domena dok, nasuprot njima, benigni domeni teže jednostavnosti radi lakšeg pamćenja imena domena od strane krajnjih korisnika. Primjeri malicioznih domena sa većim brojem tačaka sadržanih u imenu domena dati su na slici 2.

newmix-env.jqzmtkfdgc.us-east-2.elasticbeanstalk.com
all-env.873n7wnhgq.us-east-2.elasticbeanstalk.com
mixall-env-1.z4a8vj3kxq.us-east-2.elasticbeanstalk.com
windowsphishingalert187.s3.amazonaws.com

Slika 2. Primjeri malicioznih domena sa većim brojem tačaka u imenu domena

− **A4 - Broj tokena unutar imena hosta**

Token predstavlja skup karaktera razdvojenih tačkama i srednjom crtom. Broj tokena unutar imena hosta takođe je često veći kod malicioznih nego kod benignih domena. Primjeri takvih malicioznih domena iz korištenog skupa podataka dati su na slici 2.

− **A5 - Broj tokena unutar URL putanje**

Token unutar URL putanje predstavlja skup karaktera razdvojenih kosom (eng. *slash*), srednjom i donjom crtom. Vlasnici malicioznih domena često automatizovano generišu URL putanje sa većim brojem tokena različite (često veće nego prosječno) dužine nego što je to slučaj sa benignim domenima. Na osnovu analize podataka uočene su sličnosti u izgledu URL putanja na različitim domenima što potencijalno upućuje na istog malicioznog aktera. Takav primjer predstavljen je na slici 3 gdje šest različitih domena distribuiraju URL-ove sa istim brojem tokena unutar URL putanje i (skoro) istom dužinom tokena. Ovi domeni ne bi mogli biti otkriveni na osnovu atributa baziranih na DNS i WHOIS zapisima zbog raznolikosti konfiguracije DNS zapisa i podešene WHOIS zaštite.

https://techno-site925.ml/Win10001010_100101.kme/MCH01010101010X0M/21
https://query-tech292.ml/Win01011010_101010.kpr/MCH01010101010X0M/37
https://helpissue-server482.ml/Wi01010110_0101010.poi/MCH01010101010X0M/70
https://zipped-limited532.ml/Wi0101001_hep0100101.kmw/MCH01010101010X0M/32
https://network-operate244.ml/Wi100110_101001.kpe/MCH01010101010X0M/60
https://implement-server824.ml/Winhelp101010110_1199119.lmt/MCH01010101010X0M/87

Slika 3. Primjeri malicioznih URL-ova sa različitih domena sa sličnim leksičkim karakteristikama URL adrese

− **A6 - Dužina najdužeg tokena unutar URL putanje**

Dužina najdužeg tokena unutar URL putanje, takođe, često je signal da se radi o malicioznom URL-u. Analizom malicioznih URL adresa, pokazalo se da mnogi URL-ovi sadrže tokene unutar URL putanje dužine i do više desetina hiljada slučajno generisanih karaktera. Primjeri malicioznih URL-ova sa izrazito dugim tokenima unutar URL putanje dati su na slici 4.

https://www.cozamarket.com/CH01010101010X/012345678910111
213141516171819202122232425262728293031323334353637383
940414243444546474849505152535455565758596 06….28581285
912860128611286212863128641286512866128671286812869128
701287112872128731287412875128761287712878128791288012
881128821288312884128851288612887128881 2889 (dužina tokena je 53.360 karaktera)

https://pelospes.best/1981/Windows/012345678910111213141516171819202122232425262728293031323334353637383940414243445464748495051525354555657585960616263646 56….41484149415041514152415341544155415641574158415941604161416241634164416541664167416841694170417141724173417441754176417741784179418041814182418341844185418641874188418941904191419241934194 (dužina tokena URL-a 15.683 karaktera)

Slika 4. Primjeri malicioznih URL-ova sa tokenima velike dužine

− **A7 - Dužina *query* stringa.** Na sličan način kao i URL putanja, i *query* dio malicioznih URL-ova sadrži slučajno generisane tokene dužine preko više desetina hiljada karaktera. Primjeri takvih malicioznih URL-ova dati su na slici 5.

https://03-secure.com/us/kw/index-og_c7.php?et=1581093584&diet=1&skin=1&teeth=0&hair=0&mskin=0&mhair=0&brain=0&ecig=0&muscle=0&watch=1&beard=0&goggles=0&phone=0&bb=0&fl=0&fbit=0&edt=0&bpack=0&purse=0&cbd=1&enhance=1&subtitle=2&jewel=0&cbdc=1&sxid=2ojkszu959sr&tvid=33bfb92479317035000001702081 2d99d&...OTJ4OVh4TzcvZnkKUVBqZE9KMD08L2RzOlg1MDlDZXJ0aWZpY2F0ZT48L2RzOlg1MDlEYXRhPjwvZHM6S2V5SW5mbz48L2RzOlNpZ25hdHVyZT48L3NhbWwycDpBdXRoblJlcXVlc3Q%25252B%2526RelayState&root=https%3A%2F%2Fnsuok.edu%2F&is_ts=1 (dužina *query* dijela URL-a je 4.404)

https://www.promotionsonlineusa.com/reredirect.aspx?https%3A//www.surveyandshop.com/default.aspx%3FFlow%3DA815225423ED8E076275BA24A4108BD6892C355F%261%3D1%261%3D1%26email%3Drchfuenning@gmail.com%26firstname%3Drichard%26lastname%3Dfuenning%26gender%3DTrue%26dobmonth%3D04%26dobday%3D15%26dobyear%3D1961%26phonecode%3D480%26phoneprefix%3D580%26phonesuffix%3D3307%26addre...k%20of%20America%20Visa%20Gift%20Card%26SubAff%3D22101-698011-238196_192331_85866_boavisa1000_BM%26AffSecID%3D%26EntranceVID%3D%257CUJokFwIflh7cxaY7KhsaA2 (dužina *query* dijela URL-a je 3.803)

Slika 5. Primjeri malicioznih URL-ova sa *query* dijelom URL-a velike dužine

4.2. *Atributi bayirani na informacijama o DNS, WHOIS zapisima domena i AS broju*

Atributi bazirani na informacijama o DNS, WHOIS zapisima domena i AS broju su korisni jer maliciozni domeni često koriste *cloud* servise, te servise za *web* hosting, DNS hosting i/ili WHOIS registre sa lošom reputacijom. Takođe je jasno da se maliciozne aktivnosti često skrivaju generisanjem novih domena koji budu aktivni kratak vremenski period. Tako je u nekim istraživanjima pokazano da je prosječan životni vijek *phishing* domena 62h, dok medijana iznosi 20h [14]. Medijana životnog vijeka TSS domena pasivnog tipa je 100 dana, dok je to kod domena agresivnog tipa svega devet dana [15]. Primjeri domena iz analiziranog skupa podataka koji su kratko trajali su *s5ry8jh.info* i *w4hfg7k.info*. Navedeni domeni su kreirani 7. februara a NS serveri *ns19.domaincontrol.com* i *ns19.domaincontrol.com* su uklonjeni 10. februara.

Odabrani atributi bazirani na informacijama o DNS, WHOIS zapisima domena i AS broju koji se koriste u ovom radu su:

− **A8 - *Top-level* domen**

Top-level domeni koji se generalno smatraju sigurnijim su državni *top-level* domeni, kao i istorijski generički *top-level* domeni, poput *.com*, *.org*, *.net*, *.gov*, *.edu* i *.mil*. Na osnovu analize korištenog skupa podataka, najveći broj malicioznih *web* stranica se distribuira sa jeftinih domena sa *top-level* domenima kao što su: *.xyz*, *.club*, *.info, .pro*, *.best*, *.support*, *.online*, *.site*, *.top*, *.life*, *.download*, *.work*, *.live, .website* i *.zone*. Takođe, i pojedini državni *top-level* domeni se zloupotrebljavaju za izvršavanje malicioznih aktivnosti. Primjeri takvih *top-level* domena u analiziranom skupu podataka su: *.gq* (Ekvatorijalna Gvineja)*, .cf* (Centralnoafrička Republika), *.ml* (Mali), *.ga* (Gabon), *.ch* (Švajcarska) i *.to* (Tonga).

− **A9 - DNS A zapis**

A zapis predstavlja vezu između imena hosta i njegove IP adrese. Drugim riječima, A zapis se može koristiti za otkrivanje *web* hosting kompanije na čijim serverima se nalazi maliciozna *web* stranica, te adresnog prostora koji data kompanija koristi, a koji se koristi u maliciozne svrhe.

− **A10 - DNS NS zapis**

NS zapisom definišu se *name* serveri koji su odgovorni za datu DNS zonu. Vrijednost NS zapisa daje informaciju o DNS hosting servisu vezanom za domen. U korištenom skupu podataka, najčešći DNS hosting servisi za maliciozne domene su: Cloudflare, NameCheap, Enom, te Microsoft Azure DNS.

− **A11 - DNS SOA zapis**

SOA (*Start of Authority*) zapis sadrži administrativne informacije o domenu. SOA *Serial number* predstavlja trenutnu verziju DNS baze za posmatrani domen. Domeni kod kojih se radi često ažuriranje DNS zapisa spadaju u grupu potencijalno malicioznih domena. Veoma često se SOA serijski brojevi generišu po obrascu YYYYmmddss, gdje prve četiri cifre označavaju godinu, naredne četiri mjesec i dan u mjesecu, a posljednje dvije redni broj promjene DNS zapisa u tekućem danu. Na ovaj način moguće je detektovati česta ažuriranja DNS zapisa.

− **A12 - Datum kreiranja domena**

Najizraženiji problem detekcije malicioznih domena je frekventnost njihovog nastajanja. Mnogi napadači koriste automatizovane algoritme za generisanje velikog broja naziva domena (eng. *Domain Generation Algorithms*, DAG) da bi zaobišli tradicionalne pristupe detekcije i dodavanja u liste za filtriranje. Ovo istraživanje pokazalo je da je datum kreiranja domena jedan od najvažnjih atributa. Neki od malicioznih domena koji su kreirani istog dana kada je generisan i skup podataka korišten u ovom radu dati su na slici 6.

competition3175.takeprize9.life
xxx.bbffee.com
trustedsystemhelp.com
game8042.takeprize100.life
errorcode-2702.xyz
app4114.takeprize53.life
onlineprizesurvey.com
yebpettlkcqm.com
www.ar98.ryatov.info

Slika 6. Primjeri malicioznih domena kreiranih istog dana kada je i generisan skup podataka

− **A13 - WHOIS zaštita/privatnost domena**

Veliki broj domena, uglavnom državnih, ima podrazumijevano definisanu politiku privatnosti na nivou *top-level* domena, te se informacije o vlasniku domena i datumu kreiranja domena ne objavljuju putem WHOIS servisa. Primjeri takvih *top-level* domena prisutnih u analiziranom skupu podataka su: *.gov, .au, .de, .im, .ir, .lv, .nz, .pe, .tm, .nl, .eu, .es, .to, .at, .bg, .lu, .ml* i *.ae.* Takođe, mnogi registri domena nude usluge zaštite privatnosti domena maskirajući informacije o vlasniku domena generičkim informacijama o samom registru. Primjer WHOIS odgovora sa skrivenim informacijama o domenu dat je na slici **7**.

```
Domain name:
   UBITECH-NETWORK249.ML

Organisation:
   Freedom Registry, Inc.
   2225 East Bayshore Road #290
   Palo Alto CA 94303
   United States
   Phone: +1 650-681-4172
   Fax: +1 650-681-4173

Domain Nameservers:
   NS01.FREENOM.COM
   NS02.FREENOM.COM
   NS03.FREENOM.COM
   NS04.FREENOM.COM

Your selected domain name is a domain name that has been
cancelled, suspended, refused or reserved at the Point ML Registry

It may be available for re-registration at http://www.point.ml

In the interim, the rights for this domain have been automatically
transferred to Freedom Registry, Inc.

Please be advised that the Point ML Registry, Freenom and
Freedom Registry, Inc. cannot be held responsible for any content
that was previously available at this domain name.

Due to restrictions in Point ML 's Privacy Statement personal information
about the previous registrants of the domain name cannot be released
to the general public.
```

Slika 7. Primjer WHOIS odgovora sa skrivenim informacijama o domenu

− **A14 - ASN**

ASN (*Autonomous System Number*) predstavlja oznaku autonomnog sistema, tj. oznaku administrativnog tijela (najčešće ISP - *Internet Service Provider*) u čijoj je nadležnosti

određeni skup IP adresa. Analizom skupa podataka korištenog u ovom radu utvrđeno je da se kod određenog broja autonomnih sistema znatno češće pojavljuju maliciozni URL-ovi. U tabeli 1. prikazani su AS brojevi gdje prva tri sadrže od 2 do 6 puta više malicoznih domena u odnosu na ostale AS brojeve sa prosječnim brojem malicioznih domena.

ASN	Vlasnik	Ukupno domena	Maliciozni domeni	Benigni domeni	Maliciozni domeni [%]
2914	NTT-COMMUNICATIONS-2914, US	12,007	43	11,964	0.36%
14618	AMAZON-AES, US	71,609	139	71,470	0.19%
13335	CLOUDFLARENET, US	123,296	235	123,061	0.19%
8075	MICROSOFT-CORP-MSN-AS-BLOCK, US	23,459	21	23.438	0.09%
16625	AKAMAI-AS, US	16,470	12	16,458	0.07%
19551	INCAPSULA, US	8,241	5	8,241	0.06%

Tabela 1. ASN primjeri sa brojem malicioznih i benignih domena

5. ANALIZIRANI ALGORITMI

Za potrebe detekcije malicioznih URL-va korišćenjem predloženog skupa atributa, analizirano je šest algoritama. Prilikom implementacije obučavanja algoritama izvršena je *grid-search* optimizacija hiper parametara i unakrsna validacija (eng. *cross-validation*) dijeleći skup za obučavanje na pet dijelova. Za obučavanje algoritama i testiranje modela na većem skupu podataka korišteno je distribuirano okruženje i Apache Spark MLlib[3] dok je za manji skup podataka korištena Scikit-learn [4] biblioteka. U nastavku je prikazan kratak opis analiziranih algoritama i razlozi njihovog odabira.

Logistička regresija

LR je izabrana kao jedan od algoritama za testiranje jer generalno daje dobre performanse i jedan je od najpopularnijih klasifikacionih algoritama. LR za novu vrijednost ulazne promjenljive (vektor svih vrijednosti atributa) vraća predikciju koristeći sigmoid funkciju. Obučavanje modela se svodi na izbor optimalnih parametara modela tako da funkcija greške nad primjerima iz skupa za obučavanje bude minimalna. Pogodna funkcija za ocjenu greške na pojedinačnom primjeru za obučavanje je funkcija logističkog gubitka (eng. *log loss*). Za ubrzavanje traženja minimuma funkcije greške, SparkMLlib i Scikit-learn implementacije LR podrazumijevano koriste L-BFGS (*Limited memory Broyden–Fletcher–Goldfarb–Shanno*). Takođe, obje implementacije podrazumijevano koriste L2 regularizaciju.

Metoda potpornih vektora

SVM podrazumijeva definisanje hiperravni koja klasifikuje sve ulazne vektore u dvije klase (u slučaju binarne klasifikacije). Ukoliko postoji više takvih hiperravni potrebno je odabrati onu sa maksimalnim rastojanjem do najbliže tačke u svakoj od klasa, tj. kriterijum za razdvajanje je maksimalna margina između klasa. Metoda za klasifikaciju potpornim vektorima je jednostavna i intuitivna ako je granica između klasa linearna. U slučaju klasifikacije sa nelinearnim granicama, korištenjem kernel funkcija, podaci se preslikavaju u prostor sa većim brojem dimenzija u kome ih je moguće linearno razdvojiti.

SVM je izabran kao jedan od algoritama za testiranje zbog generalno dobrih performansi na širokom spektru problema. Spark MLlib SVM implementacija podržava samo linerni kernel [16]. Scikit-learn implementacija tokom pretrage hiper parametara, koristeći *grid-search* pristup, validira i polinomijalni i kernel sa radijalnom osnovom (eng. *radial basis function*, RBF). Nedostatak SVM-a u odnosu na ostale testirane algoritme je primjetno duže vrijeme obučavanja modela.

Klasifikaciono stablo odlučivanja

Jedan od algoritama koji je prilično jednostavan, a često se koristi u istraživanjima sličnim ovom je DT. Postupak formiranje stabla odlučivanja je rekurzivan postupak. Korak rekurzije podrazumijeva biranje atributa od kog se kreira čvor i vrši dalje grananje u zavisnosti od vrijednosti datog atributa. Proces izbora atributa najpogodnijeg za grananje se vrši tako što se bira atribut koji pruža najveću informacionu dobit (eng. *information gain*), mada su mogući i drugi načini izbora atributa najpogodnijeg za grananje. Pritom, tokom svakog grananja, posmatra se samo trenutno stanje, ne uzima se u obzir kako će se vršiti dalje grananje i koje bi dovelo do najboljih rezultata. Postupak za jednu granu se zaustavlja ukoliko su iskorišteni svi atributi ili ukoliko su svi listovi čisti, što znači da su u svakom listu primjeri samo jedne klase. Nedostatak DT je što su takvi modeli često pretrenirani (eng. *overfitting*).

Šuma slučajnih stabala

RF predstavlja jednu od metoda ansambl učenja (eng. *ensemble learning*) gdje skup više stabala zajedno donosi krajnju odluku. Za svako stablo se slučajno bira skup podataka koji će se koristiti za obučavanje. Ovaj koncept ansambl učenja naziva se *bagging*. Pored podataka, slučajno se bira i skup atributa koje će koristiti pojedinačna stabla. Slučajnim odabirom podataka i atributa obezbjeđuje se nepristrasnost odlučivanja.

Kao što je ranije pomenuto, RF se u dosadašnjim istraživanjima pokazao kao algoritam sa često najboljim rezultatima među testiranim algoritmima. [7, 8, 9]

LightGBM

Pošto se RF pokazao kao jedan od najefikasnijih algoritama za rješavanje problema detekcije malicioznih domena, logično je slijedio odabir jednog od algoritama sa gradijentnim pojačavanjem (eng. *gradient boosting*), koji je takođe zasnovan na skupovima stabala za odlučivanje, kao kandidata za rješavanje ovog problema.

[3] https://spark.apache.org/mllib/

[4] https://scikit-learn.org

LightGBM model se kreira iterativno tako što se u svakoj iteraciji na postojeći model, na osnovu greške prethodne iteracije, dodaje novi „slabi klasifikator“, u ovom slučaju stablo odlučivanja, i time se „pojačava“ model. Što se tiče formiranja stabla, LightGBM nema ograničenja po pitanju balansiranosti čvorova na nivou i može se opisati kao *leaf-wise*.

Za distribuiranu *grid-search* optimizaciju hiper parametara implementirana je aplikacija za paralelno obučavanje LightGBM modela sa različitim kombinacijama vrijednosti hiper parametara. Pritom, za validaciju LightGBM modela odvojeno je 15% primjera iz skupa za obučavanje.

Multi-layer perceptron

MLP je neuronska mreža bez povratnih veza (eng. *feed-forward*) sa više slojeva koji su organizovani u ulazne, izlazne i skrivene slojeve. Tok podataka od ulaznog do izlaznog sloja je isključivo unaprijed. Svaki sloj je u potpunosti povezan sa sljedećim slojem i svaki neuron (osim ulaznih) ima nelinearnu aktivacionu funkciju. Svaki neuron ažurira svoje vrijednosti uzimajući u obzir vrijednosti povezanih neurona i težine ovih veza. Spark MLlib impementacija kao aktivacionu funkciju podrazumijevano koristi sigmoid, dok je pri obučavanju modela koristeći Scikit-learn testirana sigmoid i hiperbolička tangens funkcija. MLP za obučavanje modela koristi *back-propagation* metod.

6. TESTIRANJE I REZULTATI

Kao praktični dio ovog rada implementirano je prikupljanje vrijednosti odabranih atributa, generisanje skupova podataka, obučavanje odabranih klasifikatora te njihova evaluacija. Takođe, urađena je i detaljnija analiza dobijenih rezultata.

6.1. PRIKUPLJANJE PODATAKA

Za prikupljanje podataka o URL-ovima korištena je ekstenzija ugrađena u *web* čitač. Na ovaj način, od strane većeg broja korisnika koji su se saglasili sa upotrebom ove ekstenzije, prikupljeni su podaci koji sadrže realan omjer malicioznih i benignih URL-ova. Labelisanje podataka izvršeno je kreiranjem skupa pravila na osnovu empirijske analize i korištenjem nekoliko alata koji koriste liste za filtriranje malicioznih domena. Potom, urađena je i detaljna manuelna validacija malicioznih i potencijalno malicioznih URL-ova.

Prvi skup podataka, skup A, sadrži oko 230 miliona zapisa. Drugi skup, skup B, sadrži dva miliona zapisa. Balansiranost skupova podataka prikazana je u tabeli 2.

Skup podatka	Broj benignih URL-ova	Broj malicioznih URL-ova	Procenat malicioznih URL-ova
Skup A	199,452,505	32,158,238	13.88%
Skup B	1,437,260	562,740	28.13%

Tabela 2. Balansiranost skupova podataka

Podaci u skupu A generisani su aktivnošću većeg broja korisnika *web* čitača sa ugrađenom ekstenzijom za prikupljanje podataka o posjećenim URL-ovima, u periodu od 48 sati. Analizom je utvrđeno da se tokom jednog dana generiše veliki broj zahtjeva ka domenima koji su kreirani u tom istom danu. Kako bi testiranje modela odgovaralo realnom scenariju, gdje se predviđanje u tekućem danu vrši pomoću modela koji je obučavan na osnovu podataka iz prethodnog (prethodnih) dana, odlučeno je da se za obučavanje modela izdvoji 50% primjera iz skupa podataka, a preostalih 50% je korišteno za testiranje modela.

Kao što je ranije pomenuto, vrijednosti atributa baziranih na informacijama o DNS i WHOIS zapisima domena kojim pripadaju analizirane *web* stranice prikupljene su namjenski razvijenim alatom.

Realizovan je automatizovan pristup DNS zapisima o domenima korištenjem *dig alata*. Za svaki domen iz skupa podataka generisani su dig upiti za A, NS i SOA zapise. Primjer dig upita i odgovora za maliciozni domen iz skupa podataka A dat je na slici 8. Nakon toga, izvršeno je parsiranje odgovora i generisanje vrijednosti odgovarajućih atributa. Atribut A9 koji predstavlja A zapis opisan je IP adresom navedenom u A zapisu (jednom ili više njih). Atribut A10 koji predstavlja NS zapis opisan je sa prva tri okteta IP adrese *name* servera (jedne ili više njih). Ovaj pristup izabran je iz razloga što je tokom ovog istraživanja utvrđeno da postoje blokovi IP adresa na kojim se nalaze DNS serveri koji su autoritativni za maliciozne domene. Ovo se posebno odnosi na DNS hosting servise, poput Cloudflare-a. Ako postoji definisan SOA zapis za dati domen i ako SOA serijski broj sadrži datum u formatu yyyyMMdd, onda je vrijednost atributa A11 koji predstavlja SOA zapis data kao broj dana koji predstavlja razliku između datuma navedenog u serijskom broju i datuma generisanja skupa podataka.

```
jelena@jjokic-pc ~ $ dig onlineuserit.com any +noall +answer

; <<>> DiG 9.11.3-1ubuntu1.11-Ubuntu <<>> onlineuserit.com any +noall +answer
;; global options: +cmd
onlineuserit.com.       60      IN      A       1.1.1.1
onlineuserit.com.       1800    IN      TXT     "v=spf1 include:spf.efwd.registrar-servers
.com ~all"
onlineuserit.com.       1800    IN      MX      20 eforward5.registrar-servers.com.
onlineuserit.com.       1800    IN      MX      10 eforward3.registrar-servers.com.
onlineuserit.com.       1800    IN      MX      10 eforward2.registrar-servers.com.
onlineuserit.com.       1800    IN      MX      15 eforward4.registrar-servers.com.
onlineuserit.com.       1800    IN      MX      10 eforward1.registrar-servers.com.
onlineuserit.com.       3601    IN      SOA     dns1.registrar-servers.com. hostmaster.reg
istrar-servers.com. 1581070343 43200 3600 604800 3601
onlineuserit.com.       1800    IN      NS      dns2.registrar-servers.com.
onlineuserit.com.       1800    IN      NS      dns1.registrar-servers.com.
```

Slika 8. Primjer dig upita i odgovora za maliciozni domen iz skupa A

Prikupljanje datuma kreiranja domena realizovano je automatizovanim pristupom podacima o internet domenima putem WHOIS servisa. Za sve domene iz skupa podataka, za koje prethodno nije utvrđen datim kreiranja, šalju se upiti WHOIS serverima. Nazivi WHOIS servera zaduženih za *top-level* domene preuzeti su iz baze o *root* zonama (eng. *Root Zone Database*) [17]. Dobijeni odgovori se parsiraju i radi se ekstrakcija datuma kreiranja domena korištenjem različitih formata datuma koje koriste WHOIS serveri, a koji su identifikovani tokom ovog istraživanja. U slučaju pronalaska datuma sa drugom vremenskom zonom, radi se konverzija u PST (*Pacific Standard Time*) vremensku zonu. Ova vremenska zona je izabrana iz razloga što su i vremena generisanja za-

htjeva od strane krajnjih korisnika koji se nalaze na različitim geografskim lokacijama (u različitim vremenskim zonama), takođe konvertovana u PST vremensku zonu. Na ovaj način spriječena je mogućnost nekonzistentnosti u podacima koja se odnosi na vrijeme registracije novog domena i upućivanja zahtjeva ka istom od strane krajnjeg korisnika. Primjer WHOIS upita i odgovora za maliciozni domen iz skupa podataka A dat je na slici 9. Pritom, različite infrastrukture WHOIS servera i veličina odgovora (najčešće u *plaintext* formatu) znatno utiču na brzinu prikupljanja podataka. Mnogi WHOIS serveri definišu ograničenje broja upita (eng. *rate limit*) na različite vremenske intervale i time onemogućavaju efikasno prikupljanje vrijednosti ovog atributa za velike skupove podataka. Iz tog razloga, postupak prikupljanja se ponavlja uzastopno više puta u toku jednog dana. Na ovaj način omogućena je pokrivenost od 85,5% ovog atributa u analiziranim skupovima podataka. Potrebno je napomenuti da razlog nepostojanja datuma kreiranja domena u generisanom skupu podatka može biti i ranije pomenuta privatnost domena. Privatnost domena (WHOIS zaštita) opisana je atributom A13 koji sadrži binarnu vrijednost koja označava da li su za dati domen informacije o domenu sakrivene ili ne. Pri tome, iz skupa podataka su izbačeni URL-ovi sa domenima za koje su WHOIS serveri vratili odgovor da ne postoji WHOIS zapis za takav domen. Takvi URL-ovi se smatraju nevalidnim i nisu korisni za obučavanje klasifikatora. U generisanim skupovima podataka, datum kreiranja opisan je atributom A12 i predstavlja broj dana od datuma kreiranja domena do datuma generisanja skupa podataka.

```
jelena@jjokic-pc ~ $ whois -h whois.nic.best pelospes.best
Domain Name: PELOSPES.BEST
Registry Domain ID: D169874686-CNIC
Registrar WHOIS Server: whois.namecheap.com
Registrar URL: https://namecheap.com
Updated Date: 2020-03-17T12:19:20.0Z
Creation Date: 2020-02-07T08:38:47.0Z
Registry Expiry Date: 2021-02-07T23:59:59.0Z
Registrar: Namecheap
Registrar IANA ID: 1068
Domain Status: ok https://icann.org/epp#ok
Registrant Organization: WhoisGuard, Inc.
Registrant State/Province: Panama
Registrant Country: PA
Registrant Email: Please query the RDDS service of the Registrar of Record iden
tified in this output for information on how to contact the Registrant, Admin,
or Tech contact of the queried domain name.
Admin Email: Please query the RDDS service of the Registrar of Record identifie
d in this output for information on how to contact the Registrant, Admin, or Te
ch contact of the queried domain name.
Tech Email: Please query the RDDS service of the Registrar of Record identified
 in this output for information on how to contact the Registrant, Admin, or Tec
h contact of the queried domain name.
Name Server: DNS1.REGISTRAR-SERVERS.COM
Name Server: DNS2.REGISTRAR-SERVERS.COM
DNSSEC: unsigned
Billing Email: Please query the RDDS service of the Registrar of Record identif
ied in this output for information on how to contact the Registrant, Admin, or
Tech contact of the queried domain name.
Registrar Abuse Contact Email: abuse@namecheap.com
Registrar Abuse Contact Phone: +1.6613102107
URL of the ICANN Whois Inaccuracy Complaint Form: https://www.icann.org/wicf/
>>> Last update of WHOIS database: 2020-05-15T13:21:35.0Z <<<

For more information on Whois status codes, please visit https://icann.org/epp

>>> IMPORTANT INFORMATION ABOUT THE DEPLOYMENT OF RDAP: please visit
https://www.centralnic.com/support/rdap <<<
```

Slika 9. Primjer WHOIS upita i odgovora

6.2. *Evaluacija*

Za evaluaciju performansi dobijenih modela posmatrane su metrike: tačnost (eng. *accuracy*) i F1-mjera (eng. *F1-score, F1*).

Tačnost je odabrana jer je to najčešće primjenjivana metrika u radovima koji rješavaju problem detekcije malicioznih URL-ova, a kako bi se rezultati dobijeni u ovom istraživanju mogli porediti sa rezultatima drugih istraživanja (koliko je to moguće s obzirom na različite skupove podataka). Tačnost se definiše kao odnos ispravnih predikcija i ukupnog broja primjera, odnosno,

$$A = \frac{TP + TN}{TP + FP + TN + FN}$$

pri čemu je TP (*True Positives*) broj ispravno klasifikovanih primjera pozitivne klase, TN (*True Negatives*) broj ispravno klasifikovanih primjera negativne klase, FP (*False Positives*) primjeri negativne klase koji su pogrešno klasifikovani i FN (False Negatives) primjeri pozitivne klase koji su pogrešno klasifikovani.

Pri radu sa nebalansiranim podacima, najčešće korištena metrika je F1-mjera. F1-mjera predstavlja harmonijsku sredinu preciznosti (eng. *precision, P*) i odziva (eng. *recall, R*) i definiše se na sljedeći način

$$P = \frac{TP}{TP + FP},\ R = \frac{TP}{TP + FN},\ F = \frac{2 * P * R}{P + R}$$

6.3. *Rezultati testiranja i analiza greške*

Sa obzirom na nebalansiranost skupa podataka, za sve probabilističke klasifikatore je vršeno pretraživanje vrijednosti praga klasifikacije (eng. *threshold*) te određivanje praga klasifikacije koji daje najbolje vrijednosti preciznosti i odziva. Jedini neprobabilistički klasifikator među predloženima je SVM gdje je izlaz modela samo klasifikaciona odluka.

Rezultati evaluacije šest predloženih klasifikatora sa odabranim atributima na testnom dijelu skupa A dati su u tabeli 3.

Algoritam	Preciznost	Odziv	F1-mjera	Tačnost
LR	97.03	77.53	86.19	96.11
SVM	93.46	76.81	84.32	95.53
DT	99.27	96.07	97.64	99.27
RF	99.29	99.44	99.36	99.80
MLP	89.05	86.67	87.84	96.24
LGBM	98.66	97.24	97.95	99.36

Tabela 3. Rezultati evaluacije predloženih klasifikatora nad testnim dijelom skupa A.

Kao što je prikazano, obučeni klasifikatori koristeći odabrane atribute postižu tačnost od 95,53-99,80%. Klasifikator koji postiže najbolje rezultate je RF koji uz tačnost od 99,80% postiže preciznost od 99,29% i odziv od 99,44%. Drugi algoritam sa najboljim performansama je LightGBM koji daje tačnost od 99,36%, preciznost od 98,66% i odziv od 97,24%.

Veliki broj primjera u skupu A za posljedicu ima sporije izvršavanje skaliranja, normalizacije, te obučavanja modela pri optimizaciji hiper parametara. Da bi se prethodno navedeni rezultati potvrdili, generisan je i manji skup podataka, skup B, koji je lakši za manipulaciju i detaljniju optimizaciju hiper parametara.

Rezultati evaluacije šest analiziranih klasifikacionih algoritama sa odabranim atributima na testnom dijelu skupa B dati su u tabeli 4.

Algoritam	Preciznost	Odziv	F1-mjera	Tačnost
LR	98.70	91.87	95.16	96.69
SVM	98.15	91.70	94.82	96.45
DT	99.61	71.66	83.36	89.86
RF	99.79	99.87	99.83	99.88
MLP	98.22	97.42	97.82	98.46
LGBM	99.76	99.86	99.81	99.87

Tabela 4. Rezultati evaluacije predloženih klasifikatora nad testnim dijelom skupa B.

Na osnovu rezultata prikazanih u tabeli 4 primjetno je da odabrane metode klasifikacije i sa skupom podataka B daju slične rezultate kao i u prethodnom testiranju nad skupom A. Najbolje rezultate daje RF sa tačnošću od 99,88%, preciznošću od 99,79% i odzivom od 99,87% te LightGBM sa skoro istim rezultatima.

Na osnovu testiranja predloženih klasifikatora jasno je da najbolje rezultate daju algoritmi bazirani na skupovima stabala za odlučivanje kao što su RF i LightGBM. Posmatrajući tačnost, predloženo rješenje daje bolje ili slične rezultate u poređenju sa rezultatima ranije analziranih istraživanja. Za razliku od drugih istraživanja gdje su skupovi podataka sastavljeni iz različitih izvora benignih i malicioznih URL-ova, u ovom istraživanju korišteni su skupovi podataka nastali aktivnošću većeg broja korisnika *web* čitača sa ugrađenim ekstenzijama za prikupljanje podataka o posjećenim URL adresama. Skupovi podataka nastali u ovom istraživanju sadrže realan udio malicioznih URL-ova u ukupnom skupu podataka.

6.4. *Analiza greške*

Iako se evaluacijom dobijenih modela postiže visoka tačnost, urađena je analiza pogrešno klasifikovanih URL-ova radi boljeg razumijevanja problema, uvida u potencijalne slabosti predloženog rješenja, te mogućnosti unapređenja. Prilikom analize, korišten je LightGBM model, obučen i testiran na skupu podataka A.

Analizom greške, može se pretpostaviti da su najčešći razlozi za pogrešno maliciozno klasifikovanje benignih URL-ova hostovanje benignih domena na adresama na kojim su u velikom broju hostovani maliciozni domeni, isti ASN, noviji datum kreiranja domena i/ili leksičke karakteristike slične karakteristikama malicioznih URL-ova. Primjeri benignih domena koji distribuiraju *false positive* URL-ove dati su na slici 10.

www.swagelok.com
ecompliance.training
thebagmarket.store
sportsport.ba
financebooks.online
hopeacademy.online
dietketo.website

Slika 10. Primjeri benignih domena koji distribuiraju FP URL-ove

Takođe, neke kategorije benignih domena koje su primjetne među *false positive* URL-ovima su:

- *browser hijacker* koji mijenja *search engine* i izgled početne strane *web* čitača. Primjeri takvih domena dati su na slici 11.

search.myway.com
int.search.myway.com
search.hdownloadmyemailhub.com
web-start-page.com
free.formfetcherpro.com

Slika 11. Primjeri *browser hijacker* FP domena

- domeni koji nude različite alate/servise korisnicima. Primjeri takvih domena dati su na slici 12.

search.yourweatherinfonow.com
free.internetspeedutility.net
download.pdfconverttools.com
free.propdfconverter.com
free.onlineformfinder.com

Slika 12. Primjeri FP domena koji nude različite servise

- domeni za redirekciju. Primjeri su dati su na slici 13.

ay9244.com
hapzop.com
dising-optors.icu
passeura.com
gwudu.com
vauoy.voluumtrk.com

Slika 13. Primjeri FP domena koji predstavljaju domene za redirekciju

- različite (potencijalne *scam)* ponude. Primjeri takvih domena dati su na slici 14.

www.amarktflow.com
eatsleepburn.com
paidamerican.com
pro.mytraffic.biz
unlimdate.com

Slika 14. Primjeri FP domena koji predstavljaju različite (potencijalne *scam)* ponude

Ono što se može primijetiti, za navedene kategorije domena nije jednostavno ni manuelno odrediti da li su distribuirani URL-ovi maliciozni ili ne. Naime, prethodno navedeni alati/servisi poput *free.internetspeedutility.net* ili *search.yourweatherinfo.now* mogu biti svjesno instalirani i korisni krajnjem korisniku, a istovremeno predstavljati neželjen sadržaj ili *spam* nekom drugom korisniku. Takvo ponašanje znatno otežava i labelisanje i detekciju malicioznih URL adresa.

Analizirajući *false negative* URL-ove, primjetno je da je najviše pogrešno klasifikovanih malicioznih domena hostovano na Microsoft Azure Web Sites ili Amazon Web Services hosting platformama. Razlog za pogrešnu detekciju ovakvih URL-ova je „skrivanje“ vrijednosti atributa baziranih na informacijama o DNS, WHOIS zapisima domena i AS broja iza hosting domena. Samim tim, dešava se da su URL-ovi sa pomenutih domena klasifikovani kao *true positive* i/ili *false negative* na osnovu vrijednosti leksičkih atributa. Primjeri FN domena prikazani su na slici 15.

microopera.azurewebsites.net
hokwsdons.azurewebsites.net
windowsphishingalert186.s3.amazonaws.com
net-server-otp.azurewebsites.net

Slika 15. Primjeri FN domena

Sve prethodno opisane sporne URL adrese je jako teško ispravno klasifikovati korištenjem samo skupa atributa generisanih na osnovu URL-a. Analizom sadržaja *web* stranice i ove URL adrese bi vjerovatno bilo moguće u velikom broju slučajeva ispravno klasifikovati.

Nakon svih navedenih primjera, potrebno je napomenuti da su maliciozni domeni koji su navedeni u ovom radu, u trenutku prikupljanja podataka bili maliciozni. To ne znači da ovi domeni trenutno (ili trajno) nisu promijenili svoj status iz malicioznog u benigni promjenom DNS zapisa ili mijenjanjem malicioznog URL-a. Ovakvo ponašenje je u toku istraživanja primijećeno kod velikog broja analiziranih domena. Na primjer, uočen je veći broj domena kod kojih su tokom generisanja skupa podataka postojali A i NS zapisi, a naknadno, tokom analize, primijećeno je da je NS zapis obrisan, a A zapis promijenjen tako da sadrži IP adresu nekog benignog servisa (poput *www.google.com* pretraživača). Neki od tih domena su dati na slici 16.

xblogger-global988.gq
monitoring-site249.ga
ubitech-network249.ml
helpissue-server245.gq
filing-operate322.gq

Slika 16. Primjeri malicioznih domena kojima je naknadno, tokom analize, primjećena promjena DNS zapisa

Dakle, i maliciozni akteri su upoznati sa atributima koji se koriste pri detekciji malicioznih URL adresa i njihovim izmjenama pokušavaju da maskiraju svoje maliciozne aktivnosti. Opisani problem promjene vrijednosti labele kroz vrijeme za iste primjere iz skupa podataka (eng. *Concept Drift*) [18, 19], pri čemu u našem slučaju maliciozan URL postaje benigni, riješen je obučavanjem modela na dnevnom nivou.

7. PRAVCI DALJEG ISTRAŽIVANJA

Mogući pravci daljeg istraživanja obuhvataju dodavanje novih atributa poput porta koji koristi HTTP server na kojem se nalazi maliciozna *web* stranica, broja specijalnih karaktera sadržanih u URL adresi (kao što su '%', '#', '@' i '$'), ključnih riječi u imenu domena (kao što su *secure, alert, help* i *support*), NLP (eng. *Natural Language Processing*) kategorije na osnovu URL adrese, *bag-of-words* i *n-gram* pristup [20]. *Bag-of-words* pristup podrazumijeva predstavljanje skupa riječi, u ovom slučaju tokena u imenu domena (ili URL adresi), i frekvencija njihovog pojavljivanja. *N-gram* predstavlja sekvencu susjednih elemenata, u ovom slučaju karaktera ili riječi sadržanih u imenu domena (ili URL adresi) koji bi se mogli koristiti kao atributi. Pored toga, jedan od pravaca daljeg istraživanja jeste i poređenje *online* i *batch* algoritama sa odabranim atributima.

8. ZAKLJUČAK

U ovom radu opisan je prijedlog rješenja za detekciju malicioznih URL-ova. Ovim radom pokazano je da je za uspješnu detekciju moguće koristiti informacije koje su sadržane u URL adresi, informacije o DNS i WHOIS zapisima domena i AS broju, bez korištenja tekstualnog sadržaja *web* stranice. Atributi koji su predloženi obuhvataju leksičke i atribute bazirane na informacijama o DNS i WHOIS zapisima domena, te AS broju. Posebna pažnja je posvećena implementaciji automatizovanih servisa za prikupljanje i generisanje vrijednosti izabranih atributa. Uspješnim odabirom atributa i korištenjem skupa podataka od oko 230 miliona zapisa sa realnim omjerom malicioznih i nemalicioznih URL adresa, omogućeno je da predloženi klasifikatori daju tačnost od 96-99%, tj. da predloženi klasifikatori sa velikom vjerovatnoćom tačno detektuju maliciozne URL-ove. Takođe, pokazano je da algoritmi bazirani na skupovima stabala za odlučivanje, poput šume slučajnih stabala i LightGBM algoritma, daju veću preciznost i odziv nego ostali algoritmi. Na kraju, dati su primjeri sa analiziranim razlozima pogrešne klasifikacije, te su navedeni najveći problemi pri detekciji i mogući pravci daljeg istraživanja.

LITERATURA

[1] J. Zhang, P. Porras i J. Ullrich, „Highly predictive blacklisting," u *Proceedings of the 17th conference on Security symposium, SS'08*, Berkeley, CA, USA, 2008.

[2] „PhishTank - anti-phishing site," [Na mreži]. Available: https://www.phishtank.com. [Poslednji pristup jun 2020].

[3] „APWG: Anti-Phishing Working Group," [Na mreži]. Available: https://apwg.org. [Poslednji pristup jun 2020].

[4] S. Sinha, M. Bailey i F. Jahanian, „Shades of Grey: On the effectiveness of reputation-based "blacklists"," *InMALWARE 2008. IEEE,* 2008.

[5] S. Garera, N. Provos, M. Chew i A. D. Rubin, „A framework for detection and measurement of phishing attacks," *WORM '07: Proceedings of the 2007 ACM workshop on Recurring malcode, Alexandria, VA,* 2007.

[6] J. Ma, L. K. Saul, S. Savage i G. M. Voelker, „Learning to detect malicious URLs," *ACM Transactions on Intelligent Systems and Technology,* 2011.

[7] C. Liu, L. Wang, B. Lang i Y. Zhou, „Finding effective classifier for malicious URL detection," *Proceedings of the 2018 2nd International Conference on Management Engineering, Software Engineering and Service Sciences, ACM,* 2018.

[8] F. Vanhoenshoven, G. Napoles, R. Falcon, K. Vanhoof i M. Koppen, „Detecting malicious URLs using machine learning techniques," *2016 IEEE Symposium Series on Computational Intelligence (SSCI), Athens,* 2016.

[9] M. Weedon, D. Tsaptsinos i J. Denholm-Price, „Random forest explorations for URL classification," *2017 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA), London,* 2017.

[10] O. Christou, N. Pitropakis, P. Papadopoulos, S. McKeown i W. J. Buchanan, „Phishing URL Detection Through Top-level Domain Analysis:A Descriptive Approach," *6th International Conference on Information Systems Security and Privacy, Valletta, Malta,* 2020.

[11] M. Lin, C. Chiu, Y. Lee i H. Pao, „Malicious URL filtering — A big data application," *2013 IEEE International Conference on Big Data, Silicon Valley, CA,* pp. 589-596, 2013.

[12] B. Alghamdi, J. Watson i Y. Xu, „Toward Detecting Malicious

Links in Online Social Networks through User Behavior," *2016 IEEE/WIC/ACM International Conference on Web Intelligence Workshops (WIW), Omaha, NE,* pp. 5-8, 2016.
[13] S. Albakry, K. Vaniea i M. Wolters, „What is this URL's Destination? Empirical Evaluation of Users' URL Reading.," *In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20). Association for Computing Machinery, New York, USA,* pp. 1-12, 2020.
[14] T. Moore i R. Clayton, „Examining the Impact of Website Takedown on Phishing," *Proceedings of the Anti-Phishing Working Groups 2nd Annual ECrime Researchers Summit,* 2007.
[15] B. Srinivasan, A. Kountouras, N. Miramirkhani, N. Miramirkhani, N. Nikiforakis, M. Antonakakis i M. Ahamad, „Exposing Search and Advertisement Abuse Tactics and Infrastructure of Technical Support Scammers," *Proceedings of the 2018 World Wide Web Conference,* 2018.
[16] „Classification and regression," [Na mreži]. Available: https://spark.apache.org/docs/latest/ml-classification-regression.html. [Poslednji pristup maj 2020].
[17] „Root Zone Database," IANA, [Na mreži]. Available: https://www.iana.org/domains/root/db. [Poslednji pristup maj 2020].
[18] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama i G. Zhang, „Learning under Concept Drift: A Review," *IEEE Transactions on Knowledge and Data Engineering,* 2019.
[19] G. Tan, P. Zhang, Q. Liu, X. Liu, C. Zhu i F. Dou, „Adaptive Malicious URL Detection: Learning in the Presence of Concept Drifts," *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE),* 2018.
[20] H. Zhao, G. Bao, Z. Chang i X. Zeng, „Malicious Domain Names Detection Algorithm Based on N -Gram," *Journal of Computer Networks and Communications ,* 2019.
[21] „Supervised learning," [Na mreži]. Available: https://scikit-learn.org/stable/supervised_learning.html. [Poslednji pristup maj 2020].

Jelena Jokić, dipl. ing, Bravo Systems d.o.o. Banja Luka, RS, BiH / Elektrotehnički fakultet, Univerzitet u Banjoj Luci, RS, BiH
E-mail: jelena.jokic@bravosystems.com
Oblasti interesovanja: mašinsko učenje, data science, sigurnost, data engineering, objektno-orijentisano programiranje i modelovanje, Internet programiranje

prof. dr Zoran Đurić, Elektrotehnički fakultet, Univerzitet u Banjoj Luci, RS, BiH
E-mail: zoran.djuric@etf.unibl.org
Oblasti interesovanja: sigurnost, kriptografija, PKI, platni sistemi i protokoli, formalna verifikacija, mašinsko učenje, data science, objektno-orijentisano programiranje i modelovanje, Internet programiranje, razvoj mobilnih aplikacija, XML-bazirana međuoperativnost, web servisi, računarske mreže, penetration testing, sistem integracija