Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 4.056
DOI: https://doi.org/10.24867/32OI04Sepa

UPOTREBA VEŠTAČKE INTELIGENCIJE ZA DETEKCIJU „PHISHING“ E-POŠTE: PRISTUP ZASNOVAN NA PRIRODNOJ OBRADI JEZIKA

THE USAGE OF ARTIFICIAL INTELLIGENCE FOR “PHISHING” E-MAIL DETECTION: AN APPROACH BASED ON NATURAL LANGUAGE PROCESSING

Nemanja Šepa, *Fakultet tehničkih nauka, Novi Sad*

Oblast – INFORMACIONA BEZBEDNOST

Kratak sadržaj – *Ovaj rad istražuje primenu naprednih algoritama veštačke inteligencije i mašinskog učenja za detekciju „phishing“ napada analizom sadržaja elektronske pošte. Poređenjem performansi modela Naivnog Bajesa, XGBoost-a, RNN-a i GRU-a, analizirane su tačnost i efikasnost, pri čemu se razmatraju ključne prednosti i ograničenja ovih modela u savremenim sajber bezbednosnim sistemima.*

Ključne reči: *Veštačka inteligencija, informaciona bezbednost, prirodna obrada jezika*

Abstract – *This paper explores the application of advanced artificial intelligence and machine learning algorithms to detect phishing attacks by analyzing email content. By comparing the performance of Naive Bayesian, XGBoost, RNN, and GRU models, accuracy and efficiency are analyzed, while considering the key advantages and limitations of these models in modern cyber security systems.*

Keywords: *Artificial intelligence, information security, natural language processing*

1. UVOD

U savremenom digitalnom svetu, elektronska pošta ostaje ključni alat za komunikaciju, kako u privatnom, tako i u poslovnom okruženju. Međutim, uz ovu rasprostranjenost dolazi i do značajne ranjivosti – „phishing“ napada. „Phishing“ predstavlja oblik sajber kriminala u kome napadači pokušavaju da dobiju osetljive informacije, poput korisničkih imena, lozinki, finansijskih podataka ili drugih ličnih informacija, pretvarajući se da su legitimne institucije. Najčešći medij za ove napade su „phishing“ e-poruke, zbog čega su sistemi elektronske pošte postali kritični cilj sajber kriminalaca.

2. KONCEPTUALNI OKVIR „PHISHING“ NAPADA I UPOTREBE VEŠTAČKE INTELIGENCIJE I MAŠINSKOG UČENJA U SAJBER BEZBEDNOSTI

Eskalacija „phishing“ napada ukazuje na ograničenja tradicionalnih pristupa u sajber bezbednosti.
Konvencionalne metode, kao što su filteri zasnovani na pravilima i antivirusni programi, pokazale su smanjenu efikasnost sa razvojem tehnika koje koriste napadači. Savremeni napadi su sofisticiraniji i koriste tehnike socijalnog inženjeringa, kako bi zloupotrebili psihološke slabosti korisnika, čineći teže prepoznavanje opasnosti za tradicionalne bezbednosne sisteme [1]. Kao odgovor na ovaj izazov, primena veštačke inteligencije (AI) i mašinskog učenja (ML) pojavili su se kao rešenje za detekciju „phishing“ e-poruka. Ove tehnologije, posebno u kombinaciji sa obradom prirodnog jezika (NLP), nude napredan pristup identifikaciji napada analizom sadržaja i strukture e-poruka na dubljem nivou.

Za razliku od statičnih filtera, AI sistemi uče iz novih podataka i kontinuirano poboljšavaju svoje detekcijske mogućnosti, što ih čini prilagodljivijim u suočavanju sa dinamičnim sajber pretnjama.

„Phishing“ napadi su pretrpeli značajnu evoluciju od svog nastanka. Rani „phishing“ napadi bili su relativno jednostavni i pokušavali su da prevare korisnike da otkriju lične podatke. Ovi rani napadi su se odlikovali loše napisanim porukama, generičkim sadržajem i očiglednim znakovima prevare, kao što su pravopisne greške i sumnjivi URL-ovi [2]. Iako su tradicionalni sistemi za detekciju e-poruka prvobitno bili efikasni u prepoznavanju ovakvih napada, sofisticiranost tehnika „phishing“ napada značajno je porasla.

Tulan i Karti ističu da su najuspešnije tehnike za detekciju phishing mejlova u velikoj meri zasnovane na analizi specifičnih segmenata same poruke. Dok se crne liste često koriste kao preventivna mera, njihova efikasnost je ograničena kada je reč o novijim, do tada nepoznatim tehnikama napada. U tom smislu, neophodno je fokusirati se na različite karakteristike poruke kako bi se ostvarila veća tačnost u detekciji. Ovo uključuje analizu spoljašnjih elemenata poruke, kao što su podaci o pošiljaocu, zatim sadržaja samog tela poruke, karakteristika vezanih za URL linkove koji se pojavljuju u poruci, kao i informacija koje se nalaze u zaglavlju poruke. Prema njihovoj analizi, ovi aspekti su od ključnog značaja za izgradnju sistema koji bi mogli da efikasno odgovore na sve složenije pretnje u oblasti fišing napada [3].

3. ETIČKA RAZMATRANJA UPOTREBE VEŠTAČKE INTELIGENCIJE U SAJBER BEZBEDNOSTI

Primena veštačke inteligencije u sajber bezbednosti donosi značajna poboljšanja u otkrivanju sajber pretnji, ali istovremeno i otvara brojna etička pitanja.

NAPOMENA: Ovaj rad proistekao je iz master rada čiji mentor je bio dr Aleksandar Rikalović, red. prof.

Jedan od ključnih izazova je očuvanje privatnosti korisnika, budući da AI sistemi često zahtevaju pristup velikim količinama osetljivih podataka kako bi efikasno otkrivali „phishing“ napade. Ova obrada podataka može narušiti privatnost, jer uključuje pristup informacijama koje nisu uvek relevantne za otkrivanje pretnji.

Dok rešenja poput anonimizacije i enkripcije mogu pomoći u zaštiti privatnosti, potrebno je pažljivo upravljanje podacima da bi se smanjili rizici. Pored toga, transparentnost AI sistema predstavlja još jedan izazov. Mnogi modeli, naročito oni zasnovani na dubokim neuronskim mrežama, funkcionišu kao „crne kutije“ koje donose odluke bez jasnog objašnjenja [4]. Nedostatak transparentnosti ograničava razumevanje korisnika i povećava rizik od nekontrolisanih grešaka, što je posebno važno u slučajevima kada AI donosi odluke koje mogu direktno uticati na poslovne ili bezbednosne procese.

Uvođenje objašnjivih AI (XAI) metoda može povećati razumevanje i prihvatanje ovih sistema, kao i omogućiti bolji nadzor [5]. Pristrasnost je takođe važan aspekt, jer ako modeli nisu obučeni na raznovrsnim i reprezentativnim skupovima podataka, mogu favorizovati određene grupe ili propustiti legitimne obrasce. Ovo može dovesti do nepravednih rezultata, poput lažnih pozitivnih ili negativnih identifikacija, što se može ublažiti redovnom revizijom modela i korišćenjem raznovrsnih podataka. Pored toga, pitanje odgovornosti u slučaju greške predstavlja značajan etički izazov. Ako AI pogrešno identifikuje legitimnu poruku kao zlonamernu ili propusti da označi „phishing“ sadržaj, neophodno je jasno definisati ko snosi odgovornost za nastalu štetu. Uspostavljanje jasnih etičkih okvira i mehanizama nadzora biće od suštinske važnosti za određivanje odgovornosti i efikasno upravljanje rizicima, posebno kako AI sistemi postaju sve više integrisani u kritične infrastrukture.

4. METODOLOGIJA IZRADE MODELA

Metodološki okvir istraživanja zasnovan je na prikupljanju, preprocesiranju i analizi skupa podataka o e-porukama, nakon čega sledi implementacija i evaluacija različitih AI/ML modela. Skup podataka korišćen u ovoj studiji preuzet je sa platforme Kaggle, renomirane u oblasti analize podataka i takmičenja u mašinskom učenju. Dataset sadrži e-poruke koje su klasifikovane kao „phishing“ ili „legitimne“, pružajući raznovrsne primere iz obe kategorije. Ova raznovrsnost je ključna, jer se „phishing“ e-poruke značajno razlikuju po sadržaju, strukturi i nameri, što čini razvoj robusnog sistema za detekciju izazovnim [6].

4.1 Tehnike preprocesiranja podataka

Preprocesiranje teksta i ekstrakcija podataka su ključni koraci u izgradnji pouzdanog modela za otkrivanje phishing imejlova. Phishing poruke često sadrže suptilne znakove koji ih razlikuju od legitimnih imejlova, a identifikovanje tih znakova zahteva primenu različitih tehnika za obradu prirodnog jezika (NLP). U nastavku su opisane neke od najčešće korišćenih tehnika za preprocesiranje i ekstrakciju relevantnih karakteristika iz phishing poruka.

Tokenizacija: Tokenizacija podrazumeva razbijanje teksta e-poruke na pojedinačne reči ili fraze, poznate kao tokeni. Ovaj korak je od suštinskog značaja u NLP-u, jer pretvara sirov tekst u jedinice koje se mogu analizirati pomoću algoritama mašinskog učenja [7].

Pretvaranje u mala slova i uklanjanje stop reči: Svi tekstovi se pretvaraju u mala slova radi doslednosti, kako bi se sprečilo da model različito tretira istu reč u zavisnosti od kapitalizacije (npr. „Password“ u odnosu na „password“). Pored toga, uklonjene su uobičajene reči, poznate kao stop reči („the“, „is“, „and“), kako bi se fokusirali na važnije delove sadržaja e-poruke [8].

Steming i lematizacija: Ove tehnike služe za vraćanje reči na njihov korenski ili osnovni oblik. Steming uključuje uklanjanje sufiksa i prefiksa (npr. „playing“ postaje „play“), dok lematizacija vraća reč u njen ispravan oblik (npr. „better“ postaje „good“), uzimajući u obzir kontekst [9].

Analiza N-grama: N-gram predstavlja sekvencu reči ili karaktera dužine N. Analiza N-gramova omogućava prepoznavanje čestih fraza ili obrazaca koji mogu ukazivati na „phishing“ poruke [10].

Ekstrakcija karakteristika pomoću BoW i TF-IDF: „Bag of Words“ (BoW) metod predstavlja e-poruke kao zbir reči, ignorišući njihov redosled, ali zadržavajući njihovu frekvenciju. TF-IDF dodaje težinu retkim, ali značajnim terminima, smanjujući uticaj uobičajenih, manje informativnih reči [11].

Ekstrakcija metapodataka: Pored tekstualnog sadržaja, analizirani su i metapodaci, kao što su adresa pošiljaoca i struktura domena. Ovi podaci su ključni za otkrivanje sumnjivih elemenata u „phishing“ porukama.

4.2. Odabir modela

Nakon preprocesiranja podataka, primenjeni su različiti AI/ML modeli kako bi se procenila njihova efikasnost u detekciji „phishing“ e-poruka. Svaki od ovih modela ima specifične prednosti i karakteristike koje su doprinele unapređenju rezultata.

Naivni Bajesov klasifikator: Ovaj algoritam, zasnovan na Bajesovoj teoremi, pretpostavlja nezavisnost između karakteristika, što ga čini relativno jednostavnim za primenu. Ipak, i pored ove pojednostavljene pretpostavke, Naivni Bajes je dokazao svoju efikasnost u zadacima klasifikacije teksta, posebno kada se radi o obradi velikih setova podataka.

XGBoost: Ovaj model primenjuje iterativno poboljšavanje predviđanja kroz tehniku bustinga, što ga čini izuzetno moćnim u radu sa složenim i visoko dimenzionalnim podacima. Njegova robusnost i sposobnost da se uspešno nosi sa neuravnoteženim klasama čine ga idealnim za detekciju „phishing“ poruka.

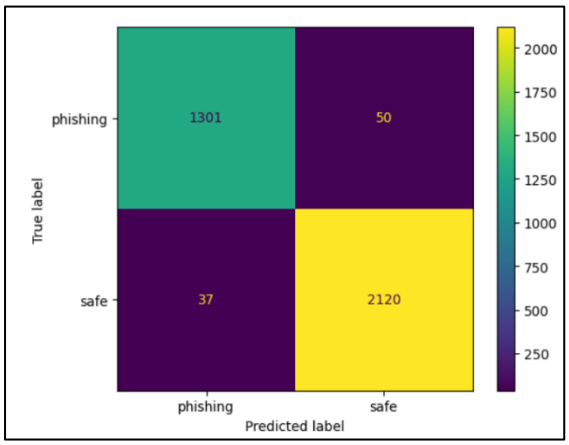
Jednostavne rekurentne neuronske mreže (RNN): RNN modeli su posebno pogodni za obradu sekvencijalnih podataka, jer uzimaju u obzir vremenski kontekst i redosled reči. Ova sposobnost je od velike važnosti u detekciji „phishing“ e-poruka, gde su sekvence reči ključne za razumevanje poruke.

Mreže sa gejt-rekurentnim jedinicama (GRU): GRU modeli predstavljaju naprednu verziju RNN-a i omogućavaju zadržavanje važnih informacija kroz duže vremenske sekvence. Zahvaljujući ovome, GRU mreže su veoma efikasne u analizi dugačkih i kompleksnih tekstualnih poruka, što ih čini veoma primenljivim za detekciju „phishing“ e-poruka.

4.3 Evaluacija modela

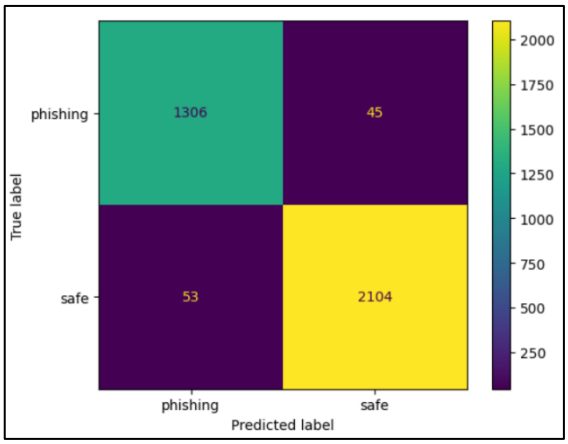
Evaluacija svih modela izvršena je korišćenjem metrika kao što su tačnost, preciznost, odziv i F1-skor. Pored toga, konfuziona matrica je korišćena za vizualizaciju rezultata klasifikacije, omogućavajući jasnu sliku o performansama svakog modela u kontekstu detekcije „phishing“ e-poruka.

Model zasnovan na Naivnom Bajesu je pokazao izuzetne performanse sa tačnošću od 97,52%, preciznošću od 98%, odzivom od 0.98 i F1-skorom od 0.98, što se može pripisati njegovoj jednostavnoj arhitekturi zasnovanoj na verovatnoći i pretpostavci o nezavisnosti karakteristika. U praksi, ovaj model je izvršio 1301 tačnu identifikaciju „phishing“ e-poruka (True Positive) i 2120 tačnih identifikacija legitimnih e-poruka (True Negative), sa relativno niskim brojem False Positive (50) i False Negative (37) (Slika 1). Ovi rezultati ukazuju na to da je model veoma pouzdan u konzistentnoj klasifikaciji „phishing“ poruka. Iako je jednostavan, njegova sposobnost da brzo obrađuje podatke čini ga praktičnim izborom za velike tekstualne skupove podataka, posebno kada su obrasci teksta konzistentni.



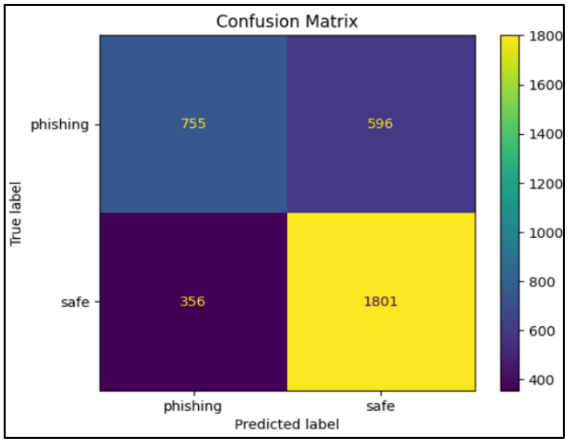
Slika 1: *Konfuziona matrica modela zasnovanog na Bajesovom klasifikatoru*

XGBoost algoritam, iako komplikovaniji, pokazao je solidne rezultate sa tačnošću od 97,21%, preciznošću od 97% i F1-skorom od 0.97. Ovaj model je identifikovao 1306 „phishing“ e-poruka (True Positive) i 2104 legitimne e-poruke (True Negative), uz niže False Positive (45) vrednosti od Naivnog Bajesa, ali nešto veći broj False Negative (53) (Slika 2). XGBoost se oslanja na složeniju arhitekturu zasnovanu na gradijentnom pojačavanju, koja često poboljšava prediktivne performanse putem ponovljenog podešavanja, ali zahteva više vremena i računarskih resursa. U ovom slučaju, model nije uspeo da nadmaši jednostavniji Naivni Bajes, što ukazuje da njegova kompleksnost možda nije uvek potrebna za binarnu klasifikaciju tekstualnih podataka.



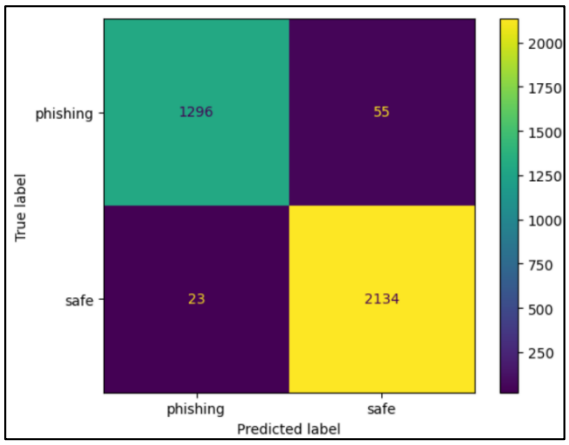
Slika 2: *Konfuziona matrica modela zasnovanog na XGBoost algoritmu*

Jednostavna rekurentna neuronska mreža (RNN) dala je najslabije rezultate, sa tačnošću od 73%, preciznošću od 72,86% i F1-skorom od 0.72. Ovaj model je detektovao 755 „phishing“ e-poruka (True Positive) i 1801 legitimnu e-poruku (True Negative), ali je postigao visoke vrednosti False Positive (596) i False Negative (356) (Slika 3), što ukazuje na probleme u pravilnom klasifikovanju kako „phishing“, tako i legitimnih e-poruka. RNN modeli su podložni problemu „izumiranja gradijenata“, što uzrokuje gubitak informacija tokom dužih sekvenci. To utiče na njihovu sposobnost da se nose sa tekstualnim podacima koji zahtevaju dugoročno zadržavanje informacija.



Slika 3: *Konfuziona matrica modela zasnovanog na jednostavnoj rekurentnoj neuronskoj mreži*

Model zasnovan na GRU neuronskoj mreži je pokazao najbolje performanse među rekurentnim mrežama, sa tačnošću od 97,78%, preciznošću od 98% i F1-skorom od 0.98. Ovaj model je ostvario 1296 True Positive i 2134 True Negative, uz veoma nizak broj False Positive (55) i False Negative (23) (Slika 4). GRU, zahvaljujući svojim „gejt“ mehanizmima, bolje zadržava relevantne informacije tokom dužih sekvenci, što mu omogućava da nadmaši osnovni RNN u tekstualnim zadacima kao što je klasifikacija „phishing“ poruka. Zbog svoje poboljšane arhitekture, GRU je sposoban da preciznije identifikuje i „phishing“ i legitimne poruke, što ga čini efikasnim izborom za zadatke koji zahtevaju dugoročni kontekst.



Slika 4: *Konfuziona matrica modela zasnovanog na „gejt“ neuronskoj mreži*

5. ZAKLJUČAK

Istraživanje o primeni veštačke inteligencije i mašinskog učenja u detekciji „phishing“ napada ukazuje na značajan potencijal ovih tehnologija za unapređenje sajber bezbednosti. Rezultati pokazuju da su modeli Naivnog Bajesa i GRU postigli najvišu tačnost i pouzdanost, dok je XGBoost pokazao solidne performanse uz povećane zahteve za računarskim resursima. U suprotnosti s tim, RNN model je imao ograničenja zbog problema sa izumiranjem gradijenata, što je rezultovalo slabijom identifikacijom „phishing“ poruka.

Ovi rezultati ističu važnost odabira odgovarajućih modela na osnovu specifičnih zahteva aplikacija u oblasti sajber bezbednosti, s naglaskom na efikasnost, brzinu i preciznost. Pored toga, etička razmatranja, kao što su očuvanje privatnosti korisnika i transparentnost algoritama, predstavljaju ključne aspekte koji moraju biti integrisani u razvoj i implementaciju AI sistema.

U svetlu brzog razvoja sajber pretnji, kontinuirano unapređenje i prilagođavanje modela postaje imperativ. Aktivno uključivanje zaposlenih u procese detekcije i prevencije „phishing“ napada od suštinske je važnosti za postizanje uspešnih rezultata u borbi protiv sajber kriminala. Takođe, važno je promovisati kulturu kontinuiranog učenja i prilagođavanja unutar organizacija, kako bi se osigurala otpornost na sve složenije pretnje u digitalnom okruženju.

4. LITERATURA

[1] Jansson, K.; von Solms, R. (2011-11-09). "Phishing for phishing awareness". Behaviour & Information Technology. 32 (6): 584–593. doi:10.1080/0144929X.2011.632650. ISSN 0144-929X. S2CID 5472217

[2] Toolan F, Carthy J (2009). Phishing detection using classifier ensembles. In: eCrime researchers summit, IEEE conference Tacoma, WA, USA, 2009, pp 1–9

[3] Toolan F, Carthy J (2010). Feature selection for spam and phishing detection. E-Crime Researchers Summit, Dallas, pp 1–12

[4] Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. Department of Technology and Society, Lund University

[5] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82-115

[6] https://www.kaggle.com/datasets/subhajournal/phishingemails

[7] Kadhim, Ammar. (2018). An Evaluation of Preprocessing Techniques for Text Classification. International Journal of Computer Science and Information Security,. 16. 22-32

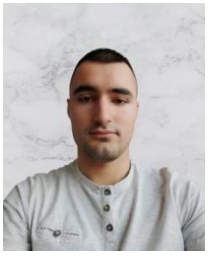
[8] Hassan Saif, Miriam Fernandez, Yulan He, and Harith Alani. 2014. On Stopwords, Filtering and Data Sparsity for Sentiment Analysis of Twitter. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pages 810–817, Reykjavik, Iceland. European Language Resources Association (ELRA)

[9] Khyani, Divya & B S, Siddhartha. (2021). An Interpretation of Lemmatization and Stemming in Natural Language Processing. Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology. 22. 350-357

[10] Cavnar, William & Trenkle, John. (2001). N-Gram-Based Text Categorization. Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval

[11] https://tita.lecturer.pens.ac.id/TextMining_SDT/03.%20Structured%20Data/NLP_%20Bag%20of%20words%20and%20TF-IDF%20explained!%20_%20by%20Koushik%20kumar%20_%20Medium.pdf

Kratka biografija:



Nemanja Šepa rođen je u Novom Sadu 2000. god. Osnovne akademske studije završio je 2022. godine na Vojnoj akademiji Univerziteta odbrane u Beogradu – studijski program Menadžment u odbrani. Trenutno zaposlen kao oficir Vojske Srbije.
kontakt: nemanjasepa@yahoo.com