URL STRUKTURAVIY XUSUSIYATLARI ASOSIDA FISHING HAVOLALARNI MASHINALI O‘QITISH USULLARI YORDAMIDA ANIQLASH

Qo’yliyeva Feruzaxon Alisher qizi
Toshkent davlat agrar universiteti assisenti
Email: qoyliyevaferuza@gmail.com
Mirzapolatova Nozima Abdunabi qizi
Toshkent davlat agrar universiteti Stajor o’qituvchisi
Email: mirzapolatovanozima@gmail.com

Annotatsiya. *Ijtimoiy tarmoqlar va internet xizmatlaridan foydalanishning kengayishi fishing havolalar orqali amalga oshiriladigan kiberhujumlar sonining ortishiga olib kelmoqda. Fishing havolalarni aniqlashda URL manzilining strukturaviy xususiyatlaridan foydalanish veb-sahifa tarkibini yuklamasdan turib shubhali havolalarni tezkor baholash imkonini beradi. Ushbu maqolada URL manzilining uzunligi, domen va yo‘l uzunligi, subdomenlar soni, raqamli belgilar, nuqta, defis, qiya chiziq va boshqa maxsus belgilar soni, HTTPS protokolidan foydalanish hamda shubhali tokenlarning mavjudligi kabi strukturaviy xususiyatlar asosida fishing havolalarni aniqlash usuli tadqiq etilgan. Klassifikatsiya uchun Random Forest va XGBoost algoritmlaridan foydalanilgan. Tajriba 20 000 ta URLdan tashkil topgan ma’lumotlar to‘plamida o‘tkazilib, ma’lumotlarning 80 foizi o‘qitish va 20 foizi testlash uchun ajratilgan. Modellarning samaradorligi Accuracy, Precision, Recall va F1-score mezonlari asosida baholangan. Tajriba natijalariga ko‘ra, Random Forest modeli 89,4% aniqlikni, XGBoost modeli esa 92,1% aniqlikni ko‘rsatgan. Olingan natijalar URLning strukturaviy xususiyatlari fishing havolalarni avtomatik aniqlashda samarali belgilar sifatida qo‘llanilishi mumkinligini ko‘rsatadi.*

Kalit so‘zlar: *fishing URL, zararli havola, URL strukturaviy xususiyatlari, mashinali o‘qitish, Random Forest, XGBoost, klassifikatsiya, kiberxavfsizlik.*

Kirish

Hozirgi kunda internet xizmatlari va ijtimoiy tarmoqlardan foydalanish ko‘lamining ortishi axborot almashish jarayonini sezilarli darajada tezlashtirdi. Shu bilan birga, internet foydalanuvchilariga qaratilgan turli kiberxavflar ham ko‘payib bormoqda. Shunday xavflardan biri fishing hujumlari bo‘lib, bunda foydalanuvchini soxta yoki zararli veb-sahifaga yo‘naltirish orqali uning login va parollari, bank ma’lumotlari yoki boshqa shaxsiy ma’lumotlarini qo‘lga kiritishga harakat qilinadi. Fishing hujumlarini amalga oshirishda zararli URL havolalardan keng foydalaniladi.

Fishing havolalarni aniqlashning muhim yo‘nalishlaridan biri URL manzilining o‘ziga xos strukturaviy xususiyatlarini tahlil qilish hisoblanadi. URL tarkibida uning umumiy uzunligi, domen nomi va yo‘l uzunligi, subdomenlar soni, raqamlar, nuqtalar, defislar, qiya chiziqlar va boshqa maxsus belgilar soni kabi ko‘rsatkichlar mavjud. Bundan tashqari, HTTPS protokolidan foydalanilishi va URL tarkibida shubhali so‘z yoki tokenlarning uchrashi ham havolaning ishonchliligini baholashda muhim belgi bo‘lishi mumkin.

URLning strukturaviy xususiyatlaridan foydalanishning asosiy afzalligi shundaki, bunda havola yo‘naltirayotgan veb-sahifaning to‘liq tarkibini yuklash talab etilmaydi. Kerakli belgilar URL satrining o‘zidan ajratib olinadi va klassifikatsiya modeli uchun kirish ma’lumotlari sifatida shakllantiriladi. Bu esa ko‘p sonli havolalarni avtomatik qayta ishlash va potensial xavfli URLlarni dastlabki bosqichda aniqlash imkonini beradi.

Fishing URLlarni strukturaviy belgilar asosida aniqlashda turli mashinali o‘qitish algoritmlaridan foydalanish mumkin. Bunday algoritmlar URL tarkibidan ajratib olingan belgilar va havolaning tegishli sinfi o‘rtasidagi bog‘lanishlarni o‘rganadi. Ayniqsa, Random Forest va gradient boosting oilasiga mansub algoritmlar ko‘p sonli belgilar o‘rtasidagi murakkab bog‘lanishlarni hisobga olish imkoniyatiga ega. Random Forest bir nechta qaror daraxtlari natijalarini umumlashtirish asosida klassifikatsiyani amalga oshirsa, XGBoost ketma-ket quriladigan modellar yordamida klassifikatsiya xatolarini kamaytirishga yo‘naltirilgan.

Shu bilan birga, fishing URLlar doimiy ravishda o‘zgarib borishi ularni aniqlash masalasini murakkablashtiradi. Zararli havolalar oddiy foydalanuvchiga ishonchli ko‘rinishi uchun haqiqiy domenlarga o‘xshash nomlar, qo‘shimcha subdomenlar, raqamlar yoki turli maxsus belgilar yordamida shakllantirilishi mumkin. Shu sababli URLning faqat bitta xususiyatiga asoslanish yetarli bo‘lmasligi mumkin. Bir nechta strukturaviy belgilarni birgalikda tahlil qilish va ularni mashinali o‘qitish algoritmlari yordamida baholash fishing havolalarni aniqlash samaradorligini oshirish imkonini beradi.

Ushbu tadqiqotning maqsadi URL manzillaridan ajratib olinadigan strukturaviy xususiyatlar asosida fishing havolalarni aniqlash hamda Random Forest va XGBoost mashinali o‘qitish algoritmlarining samaradorligini qiyosiy baholashdan iborat. Buning uchun URLlardan strukturaviy belgilar ajratib olinadi, ular asosida belgilar vektori shakllantiriladi va klassifikatsiya modellari o‘qitiladi. Modellarning samaradorligi Accuracy, Precision, Recall va F1-score ko‘rsatkichlari asosida baholanadi.

Tadqiqot doirasida bir xil ma’lumotlar to‘plami va bir xil strukturaviy belgilar asosida Random Forest hamda XGBoost algoritmlarining natijalari solishtiriladi. Bunday yondashuv URL strukturaviy xususiyatlari asosida fishing havolalarni aniqlashda qaysi algoritm yuqoriroq samaradorlik ko‘rsatishini aniqlash va ushbu belgilarni avtomatik fishing aniqlash tizimlarida qo‘llash imkoniyatlarini baholashga xizmat qiladi.

Adabiyotlar tahlili

Fishing havolalarni avtomatik aniqlash masalasi kiberxavfsizlik sohasidagi dolzarb tadqiqot yo‘nalishlaridan biri hisoblanadi. Mavjud tadqiqotlarda fishing URLlarni aniqlash uchun qora ro‘yxatlarga asoslangan usullar, evristik yondashuvlar, URLning leksik va strukturaviy xususiyatlari hamda mashinali o‘qitish algoritmlaridan foydalanilgan. Ayniqsa, URL manzilining o‘zidan belgilar ajratib olishga asoslangan usullar veb-sahifa tarkibini yuklash zaruratini kamaytirishi sababli amaliy jihatdan muhim hisoblanadi.

[1] ishda fishing URLlarni mashinali o‘qitish asosida aniqlash masalasini tadqiq qilgan. Mualliflar 36 400 ta ishonchli va 37 175 ta fishing URLdan tashkil topgan ma’lumotlar to‘plamidan foydalangan. Tadqiqotda URLlardan NLP va leksik xususiyatlar ajratilib, yettita klassifikatsiya algoritmi qiyoslangan. Tajriba natijalarida Random Forest algoritmi NLP asosidagi belgilar bilan 97,98% aniqlik ko‘rsatgan. Ushbu tadqiqot URLning o‘zidan ajratiladigan belgilar fishing havolalarni aniqlashda yetarlicha yuqori samaradorlikka ega bo‘lishi mumkinligini ko‘rsatadi.

[2] tadqiqotda fishing URLlarni aniqlashda bir nechta mashinali o‘qitish algoritmlarini qo‘llagan. Tadqiqotda Random Forest, Decision Tree, LightGBM, Logistic Regression va Support Vector Machine algoritmlari o‘rganilgan. URLlardan tegishli belgilar ajratilib, keyinchalik ushbu belgilar klassifikatsiya modellariga uzatilgan. Mualliflar ma’lumotlarning 80 foizidan o‘qitish, 20 foizidan esa testlash uchun foydalangan. Olingan natijalar fishing URLlarni aniqlashda URL xususiyatlarini ajratish va mashinali o‘qitish algoritmlaridan foydalanish samarali yondashuv ekanligini ko‘rsatgan.

URL ma’lumotlari asosida fishing saytlarni aniqlashga bag‘ishlangan boshqa tadqiqotda [3] ishda Decision Tree, Random Forest, K-Nearest Neighbors, Gaussian Naive Bayes va XGBoost algoritmlari o‘zaro taqqoslangan. Klassifikatsiya uchun URL va veb-resursga tegishli muhim belgilar tanlab olingan. Tajriba natijalarida Random Forest 97,0% aniqlik bilan ko‘rib chiqilgan algoritmlar orasida yuqori natijani qayd etgan. Bu turli mashinali o‘qitish algoritmlarining samaradorligi foydalanilayotgan belgilar va ma’lumotlar to‘plamiga bog‘liq ekanligini ko‘rsatadi.

URLlarni tahlil qilish asosidagi yana bir tadqiqotda [4] 6000 ta URLdan iborat ma’lumotlar to‘plamidan o‘nta belgi ajratilib, sakkizta mashinali o‘qitish algoritmi sinovdan o‘tkazilgan. Ushbu tajribada eng yuqori natija Multilayer Perceptron algoritmida kuzatilib, 85,41% Accuracy va 85,17% F1-score qayd etilgan. Natijalarning boshqa tadqiqotlardagi ko‘rsatkichlardan farq qilishi ma’lumotlar to‘plami hajmi, tanlangan belgilar va klassifikatsiya algoritmining fishingni aniqlash sifatiga sezilarli ta’sir qilishini ko‘rsatadi.

URL belgilarining sonini kamaytirish masalasi ham alohida tadqiq qilingan. Ayrim ishlarda ko‘p sonli belgilarni hisoblash o‘rniga URLdan olinadigan kam sonli leksik xususiyatlardan foydalanish taklif etilgan [5]. Bunday yondashuvda URL uzunligi, maxsus belgilar, raqamlar va URL tarkibiga oid boshqa xususiyatlardan foydalanib, hisoblash murakkabligini kamaytirgan holda fishing havolalarni aniqlash mumkinligi ko‘rsatilgan. Bu real vaqt rejimida ishlaydigan tizimlarni yaratishda URLdan tez ajratib olinadigan belgilarning ahamiyatini oshiradi.

Fishing URLlarni aniqlashda Random Forest bilan bir qatorda gradient boosting asosidagi algoritmlar, jumladan XGBoost ham keng qo‘llanmoqda. XGBoost bir nechta qaror daraxtlarini ketma-ket shakllantirib, oldingi bosqichlarda yuzaga kelgan klassifikatsiya xatolarini keyingi bosqichlarda kamaytirishga harakat qiladi. So‘nggi tadqiqotlarda XGBoost URLlardan ajratilgan belgilar asosida turli ma’lumotlar to‘plamlarida tekshirilgan va modelning fishing URLlarni klassifikatsiya qilishda barqaror natija bera olishi ko‘rsatilgan [6].

Tahlil qilingan ishlardan ko‘rinadiki, URL asosidagi fishingni aniqlashda natija faqat tanlangan klassifikatorga emas, balki URLdan qanday belgilar ajratib olinishi, ma’lumotlar to‘plamining hajmi va tarkibi hamda o‘qitish va testlash shartlariga ham bog‘liq. Shu sababli bir xil ma’lumotlar va bir xil belgilar to‘plamida turli algoritmlarni baholash ularning samaradorligini obyektiv taqqoslash imkonini beradi.

Ushbu tadqiqotda URLning umumiy uzunligi, domen va yo‘l uzunligi, subdomenlar soni, raqamli va maxsus belgilar soni, HTTPS protokolidan foydalanish hamda shubhali tokenlarning mavjudligi kabi strukturaviy xususiyatlardan foydalaniladi. Bir xil belgilar to‘plami asosida Random Forest va XGBoost algoritmlari o‘qitilib, ularning fishing havolalarni aniqlash samaradorligi Accuracy, Precision, Recall va F1-score mezonlari bo‘yicha qiyosiy baholanadi.

Taklif etilayotgan usul

Fishing havolalarni aniqlashda URL manzilining o‘zida mavjud bo‘lgan strukturaviy belgilar muhim axborot manbai hisoblanadi. Fishing uchun yaratilgan URLlar ko‘pincha haqiqiy veb-manzilga o‘xshatib shakllantiriladi, biroq ularning tarkibida URL uzunligining ortishi, ko‘p sonli subdomenlar, raqamlar va maxsus belgilar, shuningdek, shubhali so‘zlarning uchrashi kabi farqlovchi xususiyatlar mavjud bo‘lishi mumkin. Shu sababli ushbu tadqiqotda veb-sahifa tarkibini tahlil qilmasdan, URL manzilidan bevosita ajratib olinadigan strukturaviy xususiyatlar asosida fishing havolalarni aniqlash yondashuvi qo‘llaniladi.

Dastlab kiruvchi URL manzillari qayta ishlanib, har bir URL uchun strukturaviy belgilar ajratib olinadi. Tadqiqotda quyidagi asosiy xususiyatlardan foydalaniladi: URLning umumiy uzunligi, domen uzunligi, URL yo‘lining uzunligi, subdomenlar soni, raqamli belgilar soni, nuqtalar, defislar, qiya chiziqlar va boshqa maxsus belgilar soni, HTTPS protokolidan foydalanish holati hamda URL tarkibida shubhali tokenlarning mavjudligi.

Har bir URL uchun ajratib olingan xususiyatlar yagona belgilar vektori shaklida ifodalanadi:

$$F_{URL} = [f_1, f_2, ..., f_n],$$

bu yerda f_i — URLdan ajratib olingan alohida strukturaviy xususiyat, n esa foydalanilgan xususiyatlarning umumiy sonini bildiradi.

Masalan, URL uzunligi sonli qiymat sifatida, HTTPS protokolidan foydalanish esa ikkilik belgi ko‘rinishida ifodalanishi mumkin. Xuddi shuningdek, URL tarkibidagi raqamlar, nuqtalar, defislar va boshqa maxsus belgilar soni alohida qiymatlar sifatida belgilar vektoriga kiritiladi. Natijada har bir URL mashinali o‘qitish algoritmlari qayta ishlashi mumkin bo‘lgan sonli ko‘rinishga keltiriladi.

Shakllantirilgan F_{URL} belgilar vektori klassifikatsiya bosqichiga uzatiladi. Ushbu tadqiqotda klassifikatsiya uchun Random Forest va XGBoost algoritmlaridan foydalaniladi. Har ikkala algoritmga bir xil strukturaviy belgilar berilishi ularning samaradorligini bir xil tajriba sharoitida taqqoslash imkonini beradi.

Random Forest algoritmi bir nechta qaror daraxtlaridan tashkil topgan ansambl model hisoblanadi. Har bir daraxt kiruvchi URLning xususiyatlari asosida alohida klassifikatsiya

natijasini shakllantiradi. Yakuniy sinf daraxtlar tomonidan olingan natijalarni umumlashtirish asosida aniqlanadi. Ushbu jarayon quyidagicha ifodalanadi:

$y_{RF} = RF(F_{URL})$,

bu yerda F_{URL} — URLning strukturaviy belgilar vektori, y_{RF} esa Random Forest modeli tomonidan aniqlangan sinfdir.

Ikkinchi model sifatida XGBoost algoritmidan foydalaniladi. XGBoost qaror daraxtlarini ketma-ket shakllantirishga asoslangan bo‘lib, har bir keyingi model oldingi bosqichlarda yuzaga kelgan xatolarni kamaytirishga yo‘naltiriladi. URLning strukturaviy belgilariga asoslangan klassifikatsiya quyidagicha ifodalanadi:

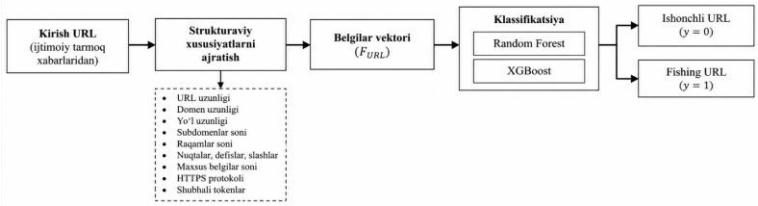
$y_{XGB} = XGBoost(F_{URL})$.

Klassifikatsiya natijasida har bir URL ikkita sinfdan biriga tegishli deb aniqlanadi:

$y \in \{0, 1\}$,

bu yerda 0 — ishonchli (legitimate) URL, 1 — fishing URL sifatida belgilanadi.

Taklif etilayotgan yondashuvning umumiy ishlash ketma-ketligi quyidagi bosqichlardan iborat:



1-rasm — “Fishing URLlarni strukturaviy xususiyatlar asosida aniqlash jarayonining blok-sxemasi

Ushbu yondashuvda URL yo‘naltirayotgan veb-sahifaning matni yoki HTML tarkibini yuklash talab qilinmaydi. Klassifikatsiya URL satridan ajratib olinadigan strukturaviy xususiyatlar asosida amalga oshiriladi. Bu esa usulni ko‘p sonli URLlarni avtomatik tahlil qilish va shubhali havolalarni dastlabki bosqichda aniqlash uchun qo‘llash imkonini beradi.

Tajriba ma’lumotlari va tajriba shartlari

Taklif etilgan yondashuvning samaradorligini baholash uchun 20 000 ta URL manzilidan tashkil topgan ma’lumotlar to‘plamidan foydalanildi. Ma’lumotlar to‘plamida 16 000 ta ishonchli (legitimate) va 4 000 ta fishing URL mavjud bo‘lib, ular mos ravishda 0 va 1 sinflar bilan belgilandi.

Har bir URL manzilidan klassifikatsiya uchun strukturaviy xususiyatlar ajratib olindi. Ushbu xususiyatlarga URLning umumiy uzunligi, domen va yo‘l uzunligi, subdomenlar soni, raqamli belgilar soni, nuqtalar, defislar, qiya chiziqlar, savol va tenglik belgilarining soni, boshqa maxsus belgilar, HTTPS protokolidan foydalanish holati hamda shubhali tokenlarning mavjudligi kiritildi.

Ajratib olingan xususiyatlar asosida har bir URL uchun sonli belgilar vektori shakllantirildi. Bir xil URL takroran o‘qitish va test ma’lumotlariga tushib qolishining oldini olish maqsadida ma’lumotlar bo‘linishidan oldin takroriy URLlar tekshirildi.

Ma’lumotlar to‘plami o‘qitish va testlash uchun 80:20 nisbatda ajratildi. Natijada 16 000 ta URL o‘qitish, 4 000 ta URL esa testlash uchun foydalanildi. Sinflarning dastlabki nisbatini saqlab qolish maqsadida stratifikatsiyalangan bo‘linish usuli qo‘llanildi.

Tajriba davomida Random Forest va XGBoost algoritmlari bir xil o‘qitish va test ma’lumotlarida sinovdan o‘tkazildi. Har ikkala modelga ham bir xil URL strukturaviy xususiyatlari kirish ma’lumotlari sifatida berildi. Bu algoritmlarning natijalarini bir xil tajriba sharoitida qiyosiy baholash imkonini berdi.

Modellarning samaradorligini baholash uchun Accuracy, Precision, Recall va F1-score ko‘rsatkichlaridan foydalanildi. Accuracy model tomonidan to‘g‘ri klassifikatsiya qilingan URLlarning umumiy ulushini, Precision fishing deb aniqlangan URLlar orasida haqiqiy fishing havolalarning ulushini, Recall mavjud fishing URLlarning qanchasi model tomonidan to‘g‘ri aniqlanganini, F1-score esa Precision va Recall ko‘rsatkichlarining muvozanatlashgan qiymatini ifodalaydi.

Shunday qilib, tajriba sharoitlarining barcha modellar uchun bir xil saqlanishi Random Forest va XGBoost algoritmlarining URL strukturaviy xususiyatlari asosida fishing havolalarni aniqlash samaradorligini qiyosiy baholash imkonini berdi.

Tajriba natijalari

URLlarning strukturaviy xususiyatlari asosida fishing havolalarni aniqlash samaradorligini baholash uchun Random Forest va XGBoost algoritmlari bir xil tajriba sharoitida sinovdan o‘tkazildi. Har ikkala model uchun bir xil o‘qitish va test ma’lumotlari hamda bir xil strukturaviy belgilar to‘plamidan foydalanildi.

Olingan natijalar 1-jadvalda keltirilgan.

1-jadval. Random Forest va XGBoost algoritmlarining qiyosiy natijalari

Model	Accuracy, %	Precision, %	Recall, %	F1-score, %
Random Forest	89,4	89,1	88,7	88,9
XGBoost	92,1	91,8	92,4	92,1

Random Forest algoritmi test ma’lumotlarida 89,4% Accuracy ko‘rsatdi. Modelning Precision ko‘rsatkichi 89,1%, Recall ko‘rsatkichi 88,7% va F1-score qiymati 88,9% ni tashkil etdi. Bu natijalar URLdan ajratib olingan strukturaviy belgilar fishing va ishonchli havolalarni ajratish uchun yetarlicha axborot bera olishini ko‘rsatadi.

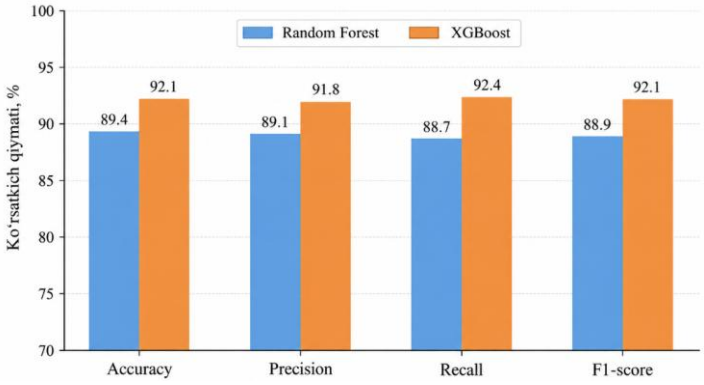
XGBoost algoritmida esa barcha asosiy baholash ko‘rsatkichlari yuqoriroq natija berdi. Modelning Accuracy ko‘rsatkichi 92,1%, Precision — 91,8%, Recall — 92,4%, F1-score esa 92,1% ni tashkil etdi.

Ikki algoritm natijalarini taqqoslaganda, XGBoost modelining Accuracy qiymati Random Forestga nisbatan 2,7 foiz punktga yuqori ekanligi aniqlandi. Precision ko‘rsatkichi 2,7 foiz punktga, Recall 3,7 foiz punktga va F1-score 3,2 foiz punktga oshgan.

Ayniqsa, Recall ko‘rsatkichining 88,7% dan 92,4% gacha oshishi muhim hisoblanadi. Fishing havolalarni aniqlash masalasida Recall qiymatining yuqori bo‘lishi mavjud zararli havolalarning ko‘proq qismini to‘g‘ri aniqlash imkonini beradi. Past Recall qiymatida esa ayrim fishing URLlar ishonchli havola sifatida noto‘g‘ri klassifikatsiya qilinishi mumkin.

Olingan natijalarga ko‘ra, har ikkala algoritm URLning strukturaviy xususiyatlari asosida fishing havolalarni aniqlash imkoniyatiga ega. Biroq ushbu tajriba sharoitida XGBoost algoritmi Random Forestga nisbatan yuqoriroq natija ko‘rsatdi. XGBoostning 92,1% Accuracy va 92,1% F1-score ko‘rsatkichlariga erishishi URL strukturaviy belgilaridan foydalanishda gradient boosting yondashuvining samaradorligini ko‘rsatadi.

Natijalarni yanada aniq ko‘rsatish uchun Random Forest va XGBoost algoritmlarining Accuracy, Precision, Recall va F1-score qiymatlarini grafik ko‘rinishda taqqoslash mumkin.



2-rasm — Random Forest va XGBoost uchun Accuracy, Precision, Recall, F1-score qiyosiy diagrammasi

Natijalarni muhokamasi

O‘tkazilgan tajribalar URLning strukturaviy xususiyatlari fishing havolalarni avtomatik aniqlashda muhim axborot manbai bo‘la olishini ko‘rsatdi. Random Forest va XGBoost algoritmlarining har ikkisi 88% dan yuqori qiymatlarni qayd etdi. Bu URLning matni yoki veb-

sahifa tarkibini qo‘shimcha tahlil qilmasdan ham strukturaviy belgilar asosida fishing havolalarni ma’lum darajada aniqlash mumkinligini ko‘rsatadi.

Tajriba natijalarida Random Forest algoritmining Accuracy ko‘rsatkichi 89,4% ni tashkil etgan bo‘lsa, XGBoost algoritmida ushbu qiymat 92,1% ga yetdi. Shunday qilib, bir xil ma’lumotlar va bir xil strukturaviy xususiyatlardan foydalanilgan holda XGBoost Random Forestga nisbatan 2,7 foiz punkt yuqori aniqlik ko‘rsatdi.

XGBoost algoritmining ustunligi Precision, Recall va F1-score ko‘rsatkichlarida ham kuzatildi. Ayniqsa, Recall qiymatining Random Forestdagi 88,7% dan XGBoostda 92,4% gacha oshishi fishing havolalarni aniqlash masalasi uchun muhim natija hisoblanadi. Chunki fishing URLning ishonchli URL sifatida noto‘g‘ri aniqlanishi foydalanuvchining zararli manbaga kirishiga olib kelishi mumkin.

XGBoost modelining yuqoriroq natija ko‘rsatishini uning o‘qitish jarayonining xususiyati bilan izohlash mumkin. Random Forestda bir nechta qaror daraxtlari nisbatan mustaqil shakllantiriladi va ularning natijalari umumlashtiriladi. XGBoostda esa daraxtlar ketma-ket qurilib, navbatdagi bosqich oldingi bosqichlarda yuzaga kelgan xatolarni kamaytirishga yo‘naltiriladi. Natijada URLning bir nechta strukturaviy xususiyatlari o‘rtasidagi murakkab bog‘lanishlarni aniqlash imkoniyati yuzaga keladi.

Shu bilan birga, olingan natijalar strukturaviy xususiyatlarning imkoniyatlari ma’lum chegaraga ega ekanligini ham ko‘rsatadi. Zamonaviy fishing havolalar tashqi ko‘rinishi va strukturasi bo‘yicha ishonchli URLlarga o‘xshatib yaratilishi mumkin. Bunday holatda faqat URL uzunligi, domen tuzilishi yoki maxsus belgilar kabi xususiyatlar fishing va ishonchli havolalarni to‘liq ajratish uchun yetarli bo‘lmasligi mumkin.

Umuman olganda, tajriba natijalari URLning strukturaviy xususiyatlari asosida fishing havolalarni aniqlashda XGBoost algoritmi Random Forestga nisbatan samaraliroq ekanligini ko‘rsatdi. Shu bilan birga, strukturaviy xususiyatlarga asoslangan yondashuv veb-sahifa tarkibini yuklashni talab qilmasligi sababli fishing havolalarni dastlabki avtomatik filtrlash bosqichida qo‘llash uchun mos keladi.

Xulosa

Ushbu tadqiqotda URL manzillarining strukturaviy xususiyatlari asosida fishing havolalarni mashinali o‘qitish algoritmlari yordamida aniqlash masalasi ko‘rib chiqildi. URLning umumiy uzunligi, domen va yo‘l uzunligi, subdomenlar soni, raqamlar, nuqtalar, defislar, qiya chiziqlar va boshqa maxsus belgilar soni, HTTPS protokolidan foydalanish hamda shubhali tokenlarning mavjudligi klassifikatsiya belgilari sifatida qo‘llanildi.

Tajriba 20 000 ta URLdan tashkil topgan ma’lumotlar to‘plamida Random Forest va XGBoost algoritmlari yordamida amalga oshirildi. Natijalarga ko‘ra, Random Forest algoritmi 89,4% Accuracy, 89,1% Precision, 88,7% Recall va 88,9% F1-score ko‘rsatkichlariga erishdi. XGBoost algoritmida esa Accuracy 92,1%, Precision 91,8%, Recall 92,4% va F1-score 92,1% ni tashkil etdi.

Qiyosiy tahlil natijasida XGBoost algoritmi Random Forestga nisbatan Accuracy bo‘yicha 2,7 foiz punkt, Recall bo‘yicha 3,7 foiz punkt va F1-score bo‘yicha 3,2 foiz punkt yuqori natija ko‘rsatdi. Bu URLning strukturaviy belgilariga asoslangan fishing havolalarni klassifikatsiya qilishda XGBoost algoritmining samaraliroq ekanligini ko‘rsatadi.

Olingan natijalar URL strukturaviy xususiyatlaridan fishing havolalarni dastlabki avtomatik aniqlash va filtrlashda foydalanish mumkinligini ko‘rsatdi. Bunday yondashuvda veb-sahifa tarkibini yuklash talab etilmasligi URLlarni tezkor tahlil qilish imkonini beradi.

Keyingi tadqiqotlarda URL strukturaviy xususiyatlarini havola tarqatilgan matnning semantik xususiyatlari bilan birlashtirish hamda gibrid modellar yordamida fishing havolalarni aniqlash samaradorligini oshirish ko‘zda tutiladi.

Foydalanilgan adabiyotlar

1- Sahingoz O.K., Buber E., Demir O., Diri B. Machine learning based phishing detection from URLs // Expert Systems with Applications. – 2019. – Vol. 117. – P. 345–357. DOI: 10.1016/j.eswa.2018.09.029.

2- Hasane Ahammad S.K., Kale S.D., Upadhye G.D., Pande S.D., Venkatesh Babu E., Dhumane A.V., Bahadur D.K.J. Phishing URL detection using machine learning methods // Advances in Engineering Software. – 2022. – Vol. 173. – Article 103288. DOI: 10.1016/j.advengsoft.2022.103288.
3- A Comparative Analysis of Machine Learning-Based Website Phishing Detection Using URL Information // 2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI). – IEEE, 2022. DOI: 10.1109/PRAI55851.2022.9904055.
4- A high-accuracy phishing website detection method based on machine learning // Journal of Information Security and Applications. – 2023. – Vol. 77. – Article 103553. DOI: 10.1016/j.jisa.2023.103553.
5- Marchal S., François J., State R., Engel T. PhishStorm: Detecting Phishing With Streaming Analytics // IEEE Transactions on Network and Service Management. – 2014. – Vol. 11. – P. 458–471. DOI: 10.1109/TNSM.2014.2377295.
6- Xu P. A Transformer-based Model to Detect Phishing URLs. – 2021.
7- Otieno D.O., Abri F., Namin A.S., Jones K.S. Detecting Phishing URLs using the BERT Transformer Model // 2023 IEEE International Conference on Big Data (BigData). – 2023. – P. 2483–2492. DOI: 10.1109/BigData59044.2023.10386782.
8- PhishTransformer: A Novel Approach to Detect Phishing Attacks Using URL Collection and Transformer // Electronics. – 2024. – Vol. 13, No. 1. – Article 30. DOI: 10.3390/electronics13010030.
9- Wang C., Chen Y. TCURL: Exploring hybrid transformer and convolutional neural network on phishing URL detection // Knowledge-Based Systems. – 2022. – Vol. 258. – Article 109955. DOI: 10.1016/j.knosys.2022.109955.
10- Maneriker P., Stokes J.W., Lazo E.G., Carutasu D., Tajaddodianfar F., Gururajan A. URLTran: Improving Phishing URL Detection Using Transformers // MILCOM 2021 – IEEE Military Communications Conference. – 2021. – P. 197–204. DOI: 10.1109/MILCOM52596.2021.9653028.