ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

ҚАЗАҚСТАН РЕСПУБЛИКАСЫ
ҰЛТТЫҚ ҒЫЛЫМ АКАДЕМИЯСЫ

әл-Фараби атындағы Қазақ ұлттық университетінің

ХАБАРЛАРЫ

ИЗВЕСТИЯ

НАЦИОНАЛЬНОЙ АКАДЕМИИ
НАУК РЕСПУБЛИКИ КАЗАХСТАН
Казахский национальный
университет имени аль-Фараби

N E W S

OF THE ACADEMY OF SCIENCES
OF THE REPUBLIC OF
KAZAKHSTAN
al-Farabi Kazakh National University

PHYSICO-MATHEMATICAL SERIES

4 (344)

OCTOBER – DECEMBER 2022

PUBLISHED SINCE JANUARY 1963

PUBLISHED 4 TIMES A YEAR

ALMATY, NAS RK

«ҚР ҰҒА Хабарлары. Физика-математикалық сериясы».
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

Меншіктеуші: «Қазақстан Республикасының Ұлттық ғылым академиясы» РҚБ (Алматы қ.). Қазақстан Республикасының Ақпарат және қоғамдық даму министрлігінің Ақпарат комитетінде 14.02.2018 ж. берілген **No 16906-Ж** мерзімдік басылым тіркеуіне қойылу туралы куәлік.
Тақырыптық бағыты: *физика-математикалық сериясы».*
Қазіргі уақытта: *«ақпараттық технологиялар»бағыты бойынша ҚР БҒМ БҒСБК ұсынған журналдар тізіміне енді.*
Мерзімділігі: *жылына 4 рет.*
Тиражы: *300 дана.*
Редакцияның мекен-жайы: *050010, Алматы қ., Шевченко көш., 28, 219 бөл., тел.: 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

Типографияның мекен-жайы: «Аруна» ЖК, Алматы қ., Мұратбаев көш., 75.

«Известия НАН РК. Серия физико-математическая».
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

Собственник: *Республиканское общественное объединение «Национальная академия наук Республики Казахстан» (г. Алматы).*
Свидетельство о постановке на учет периодического печатного издания в Комитете информации Министерства информации и общественного развития Республики Казахстан **No 16906-Ж** выданное 14.02.2018 г.
Тематическая направленность: *серия физика-математическая.*
В настоящее время: *вошел в список журналов, рекомендованных ККСОН МОН РК по направлению «информационные коммуникационные технологии».*
Периодичность: *4 раз в год.*
Тираж: *300 экземпляров.*
Адрес редакции: *050010, г. Алматы, ул. Шевченко, 28, оф. 219, тел.: 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

Адрес типографии: ИП «Аруна», г. Алматы, ул. Муратбаева, 75.

News of the National Academy of Sciences of the Republic of Kazakhstan.
Physico-matematical series.
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

Owner: RPA «National Academy of Sciences of the Republic of Kazakhstan» (Almaty). The certificate of registration of a periodical printed publication in the Committee of information of the Ministry of Information and Social Development of the Republic of Kazakhstan **No. 16906-Ж,** issued 14.02.2018
Thematic scope: *physical-mathematical series.*
Currently: *included in the list of journals recommended by the CCSES MES RK in the direction of «information and communication technologies».*
Periodicity: *4 times a year.*
Circulation: *300 copies.*
Editorial address: *28, Shevchenko str., of. 219, Almaty, 050010, tel. 272-13-19*
http://www.physico-mathematical.kz/index.php/en/

Address of printing house: ST «Aruna», 75, Muratbayev str, Almaty.

NEWS OF THE NATIONAL ACADEMY OF SCIENCES
OF THE REPUBLIC OF KAZAKHSTAN
PHYSICO-MATHEMATICAL SERIES
ISSN 1991-346X
Volume 4, Namber 344 (2022), 16-29
https://doi.org/10.32014/2022.2518-1726.153

УДК 372.851.02., 372.800.4.02

М.А. Болатбек[1*], Ш.Ж. Мусиралиева[2], К. Багитова[2,3], А.Т. Нюсупов[2], Е. Абайұлы[2]

[1]Пассау университеті, Пасса, Германия;
[2]Әл-Фараби атындағы Қазақ ұлттық университеті, Алматы, Қазақстан;
[3]Х. Досмұхамедов атындағы Атырау университеті, Атырау, Қазақстан.
E-mail: *bolatbek.milana@gmail.com*

ВЕБ-РЕСУРСТАРДАҒЫ ФИШИНГТІК ХАБАРЛАМАЛАР ЖӘНЕ ОЛАРДЫ МАШИНАЛЫҚ ОҚЫТУ ӘДІСТЕРІ АРҚЫЛЫ АНЫҚТАУ

Аннотация. Қазіргі таңда веб-ресурстардың қолданысының артуына байланысты киберқылмыс түрлері де артып келеді. Солардың ішінде жиі кездесетін қылмыс түрлерінің бірі фишинг болып табылады, фишинг – пайдаланушыларды алдау және оның құпиясөзін, несие картасының нөмірімен қатар басқа да құпия түрдегі ақпараттарды алу әдістерінің жиынтығы. Көбінесе шабуылдаушылар өздерін электрондық поштадағы қызметкер немесе телефон қоңырауларының танымал ұйымдардың өкілі ретінде көрсетеді. Фишинг – кибершабуылдардың ең қарапайым тәсілі, бірақ ең қауіпті және тиімді әдістердің бірі. Фишингтік хабарламалар қарапайым заңды хабарламаларға ұқсас болуы мүмкін, шабуылдаушы Интернет қолданушысына бұрмаланған веб-сайттарды, сілтемелерді ұсыну арқылы өз мақсатына жетуі мүмкін. Қазіргі уақытта фишингке қарсы тәсілдер сарапшылардан фишингтік сайттардың белгілерін анықтауды және фишингтік веб-сайттарды анықтау үшін үшінші тарап қызметтерін пайдалануды талап етеді. Берілген жұмыста фишинг түсінігіне, оның түрлеріне анықтама беріледі. Берілген бағыттағы соңғы ғылыми мақалаларға шолу жасалады. Сонымен қатар авторлар фишингтік мәліметтерді анықтау үшін машиналық оқыту әдістерін оқытуға және тестілеуге арналған фишингтік мәтіндер жинағын құрастырып, аталған жинақ бойынша әр түрлі машиналық оқыту әдістерін қолданған.

Түйін сөздер: веб-ресурс, киберқауіпсіздік, фишинг, мәліметтер жинағы, мәтінді жіктеу.

Жұмыс AP15473408 «Әлеуметтік желілердегі экстремистік мазмұнды анықтау модельдерін және әдістерін құрастыру» жобасының аясында орындалды.

М.А. Болатбек[1*], Ш.Ж. Мусиралиева[2], К. Багитова[2,3], А.Т. Нюсупов[2], Е. Абайулы[2]

[1]Университет Пассау, Пассау, Германия;
[2]Казахский национальный университет им. аль-Фараби, Алматы, Казахстан;
[3]Атырауский университет им. Х. Досмухамедова, Атырау, Казахстан.
E-mail: *bolatbek.milana@gmail.com*

ФИШИНГОВЫЕ СООБЩЕНИЯ НА ВЕБ-РЕСУРСАХ И ИХ ОПРЕДЕЛЕНИЕ МЕТОДАМИ МАШИННОГО ОБУЧЕНИЯ

Аннотация. В настоящее время в связи с увеличением использования веб-ресурсов увеличивается и количество видов киберпреступности. Среди них одним из наиболее распространенных видов преступлений является фишинг, фишинг – это совокупность способов обмана пользователей и получения их пароля, а также номера кредитной карты и другой конфиденциальной информации. Чаще всего злоумышленники позиционируют себя как сотрудников электронной почты или представителей известных организаций сотовых связей. Фишинг – один из самых простых, но самых опасных и эффективных способов кибератак. Фишинговые сообщения могут быть похожи на обычные легальные сообщения, когда злоумышленник может достичь своей цели, предоставляя пользователю Интернета искаженные веб-сайты, ссылки. В настоящее время антифишинговые подходы требуют от экспертов выявления признаков фишинговых сайтов и использования сторонних сервисов для идентификации фишинговых веб-ресурсов. В данной работе дается определение понятия фишинга, его видов. Дается обзор последних научных статей в данном направлении. Кроме того, авторы составили сборник фишинговых текстов для обучения и тестирования методов машинного обучения для определения фишинговых данных и использовали различные методы машинного обучения по данному набору данных.

Ключевые слова: веб-ресурс, кибербезопасность, фишинг, набор данных, классификация текста.

M. Bolatbek[1*], Sh. Musiralieva[2], K. Bagitova[2,3], А. Нюсупов [2], E. Abaiuly[2]

[1]University of Passau, Passau, Germany;
[2]Al-Farabi Kazakh National University, Almaty, Kazakhstan;
[3]Kh. Dosmukhamedov Atyrau University, Atyrau, Kazakhstan.
E-mail: *bolatbek.milana@gmail.com*

PHISHING MESSAGES ON WEB RESOURCES AND THEIR DETECTION BY MACHINE LEARNING METHODS

Abstract. Currently, due to the increase in the use of web resources, the number of types of cybercrime is also increasing. Among them, one of the most common types of crimes is phishing, which is a set of ways to deceive users and obtain their password, as well as credit card numbers and other confidential information. Most often, attackers position themselves as an email employee or a representative of well-known cellular communications organizations. Phishing is one of the simplest, but most dangerous and effective methods of cyber attacks. Phishing messages can be similar to ordinary legal messages, when an attacker can achieve his goal by providing the Internet user with distorted websites, links. Currently, anti-phishing approaches require experts to identify signs of phishing web-sites and use third-party services to identify phishing web resources. This paper defines the concept of phishing, its types. An overview of the latest scientific articles in this direction is given. In addition, the authors compiled a dataset of phishing texts for training and testing machine learning methods to identify phishing data and used various machine learning methods for this dataset.

Key words: web-resource, cybersecurity, phishing, dataset, text classification.

Кіріспе. COVID-19 пандемиясы біздің өмірімізге үлкен өзгерістер алып келді, адамдардың жұмыс істеу, сауда жасау, қарым-қатынас орнату тәртібі және т.б. өзгеріске ұшырады. Вирустың таралуын шектеу мақсатында адамдарды өз-өзін оқшаулауға және әлеуметтік қашықтықты сақтауға мәжбүр болды, ал бұл өз кезегінде бүкіл әлемдегі Интернеттің, онлайн медианың қолданысын бірнеше есе арттырды.

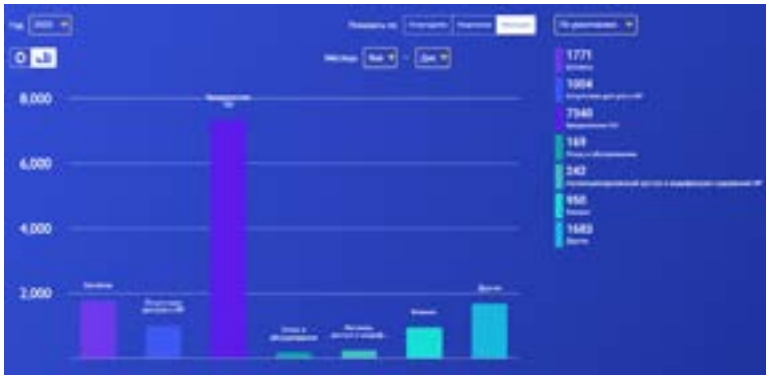
Ақпараттық-коммуникациялық технологиялар қолданысының артуының жетістіктері өте көп, десек те аталған мүмкіндікті алаяқтар да өз пайдасына асыру үстінде. Ресми деректерге сәйкес әлемдегі, соның ішінде біздің елімізде де ақпараттық қауіпсіздік инциденттерінің саны күрт өскен. Солардың ішінде ең жиі кездесетін шабуылдардың бірі фишинг болып табылады. Қазіргі таңда фишинг – әлемдегі ең көп таралған киберқылмыс түрлерінің бірі, оның көмегімен аккаунттықжәне банктік деректер ұрланады.

2021 жылдың қаңтар-желтоқсан аралығында елімізде тіркелген кибер-қауіпсіздік инциденттерінің саны 1-суретте келтірілген. Киберқауіпсіздік инциденттері ботнеттер, интернет ресурсқа қол жетімділіктің болмауы,

зиянды бағдарламалық жабдықтама, қызмет көрсетуден бас тарту, интернет ресурсқа рұқсатсыз қол жеткізу және оның мазмұнын модификациялау, фишинг және т.б. санаттарға жіктелген. Соның ішінде 658 фишинг шабуылы тіркелген.

1-сурет. 2021 жылы орын алған киберқауіпсіздік инциденттерінің статистикасы
(KZ-CERT деректері негізінде)

KZ-CERT қызметінің деректері бойынша 2022 жылы алдыңғы жылғы көрсеткішпен салыстырғанда басқа киберқауіпсіздік инциденттенің саны азайса, фишингтік шабуылдардың саны күрт өскен, 2022 жылдың қаңтар айы мен осы уақытқа дейінгі аралықта 950фишинг шабуылы анықталған, бұл бір жыл бұрынғыға қарағанда 1,4 есе көп (2-сурет).



2-сурет. 2022 жылға арналған ақпараттық қауіпсіздік инцидент түрлерінің динамикасы
(KZ-CERT деректері негізінде)

Қазіргі таңда фишингтік веб-сайттар оларды анықтау дәлдігінің төмендігімен қауіпті болып табылады. Сол себепті фишингтік хабарламаларды уақытылы және дұрыс анықтауды киберқауіпсіздіктің маңызды саласы ретінде қарастыруға болады. Фишинг – бұл әлеуметтік инженерияның бір түрі, онда шабуылдаушы алаяқтық хабарлама жібереді, ол интернет пайдаланушысын

алдау арқылы шабуылдаушыға құпия ақпаратты ашуға мәжбүр етеді немесе зиянды бағдарламалық жасақтаманы пайдаланушының инфрақұрылымына, мысалы, ransomware сияқты орналастыруға арналады. 2020 жылдан бастап фишинг киберқылмыскерлер жасаған ең көп таралған шабуыл болып табылады, Федералды іздеу бюросы интернет-қылмыс туралы шағымдар орталығы компьютерлік қылмыстың кез-келген түріне қарағанда екі есе көп фишинг жағдайларын тіркейді.

Фишингтік оқиғалардың алдын алу немесе азайту әрекеттері заңнаманы білуді, пайдаланушыларды оқытуды, қоғамды ақпараттандыруды және техникалық қауіпсіздік шараларын қамтиды. Фишинг туралы хабардар болу үйде де, жұмыс орнында да маңызды болып табылады. Мысалы, 2017 жылдан бастап 2020 жылға дейін кәсіпорындар арасында фишингтік шабуылдар саны 72%-дан 86%-ға дейін өсті. Ғалымдар фишингтің бірнеше түрін ажыратады, олардың бірқатарына төменде қысқаша тоқталып өтеміз.

Электрондық пошта арқылы жүргізілетін фишинг

Фишингтік хабарламалардың көпшілігі электрондық пошта спамы арқылы жеткізіледі және жекелендірілмеген немесе белгілі бір жеке тұлғаға немесе компанияға бағытталған – бұл "жаппай" фишинг деп аталады. Жаппай фишингтік хабарламаның мазмұны шабуылдаушының мақсатына байланысты әр түрлі болады – банктер мен қаржы қызметтері, электрондық пошта мен бұлтты қызмет провайдерлері және ағындық қызметтер жалпы мақсаттар болып табылады. Шабуылдаушылар деректерді жәбірленушіден ақшаны тікелей ұрлау үшін пайдалана алады, ұрланған ағындық қызмет шоттары әдетте қараңғы желі нарықтарындағы тұтынушыларға тікелей сатылады.

Тікелей (spear)фишинг Тікелей фишинг шабуылдаушы арнайы таңдалған фишингтік хабарламалар арқылы белгілі бір ұйымға немесе адамға тікелей бағытталған деп болжайды. Бұл электрондық поштаны заңды деп ойлау үшін белгілі бір адамға электрондық хаттар жіберуді білдіреді. Жаппай фишингтен айырмашылығы, тікелей фишинг арқылы шабуылдаушылар шабуылдың сәтті болу ықтималдығын арттыру үшін көп жағдайда өз мақсаттары туралы жеке ақпарат жинайды және пайдаланады.

Whaling және CEO қылмысы.

Фишингтің бұл түрі жоғары деңгейдегі басшыларға және басқа да жоғары деңгейдегі мақсаттарға бағытталған тікелей фишингтік шабуыл болып табылады.Бұл ұйымның басқа қызметкерлерін белгілі бір әрекетті орындауға мәжбүрлеу мақсатында, әдетте оффшорлық шотқа ақша аудару мақсатында жоғары деңгейдегі басшылардан жалған электрондық хаттар жасауды қамтиды.

Клон фишингі.

Клон фишингі – бұл фишинг шабуылының бір түрі, мұнда бірдей немесе клондалған электрондық поштаны құру үшін тіркеме немесе сілтеме бар заңды және бұрын жеткізілген электрондық поштаның мазмұны және алушылардың мекен-жайлары алынып, пайдаланылады.

Дыбыстық фишинг.

Дыбыстық фишинг немесе вишинг– фишингтік шабуылдарды жүргізу үшін телефонияны (көбінесе IP-телефония арқылы дауыстық байланысты) пайдалану. Зиянкестер көптеген телефон нөмірлерін тереді және көбінесе мәтінді сөйлеу синтезаторларын қолдана отырып жасалған автоматты жазбаларды ойнатады, олар банктік шоттармен немесе жәбірленушінің несие карталарымен алаяқтық әрекеттері туралы жалған мәлімдемелер жасайды. Жәбірленушіге шабуылдаушылар бақылайтын нөмірге қоңырау шалу ұсынылады, ол автоматты түрде алаяқтыққа "рұқсат беру" үшін құпия ақпаратты енгізуді немесе оларды ақпарат алу үшін әлеуметтік инженерияны қолдануға тырысатын адаммен байланыстыруды ұсынады.

SMS фишинг.

SMS фишингнемесе смишингэлектрондық пошта фишингіне ұқсас, тек шабуылдаушылар мәтіндік хабарламаларды жеткізу үшін ұялы телефон нөмірлерін пайдаланады. Фишингтік шабуылдар әдетте пайдаланушыны сілтемені басуға, телефон нөміріне қоңырау шалуға немесе SMS хабарлама арқылы шабуылдаушы ұсынған электрондық пошта мекен-жайына хабарласуға шақырады. Содан кейін жәбірленушіге жеке мәліметтерін беру ұсынылады; бұл көбінесе басқа веб-сайттар немесе қызметтер үшін тіркелгі деректері болып табылады.

Парақшаны басып алу.

Бұл пайдаланушыларды зиянды веб-сайтқа немесе сайтаралық сценарий арқылы эксплуатациялар жиынтығына бағыттау мақсатында заңды веб-парақшаларға нұқсан келтіруді қамтиды. Хакер веб-сайтты бұзып, бұзылған веб-серверге кіретін заңды пайдаланушыларды шабуылдау үшін *MPack* сияқты эксплуатациялар жиынтығын енгізе алады. Парақшаны басып алу көбінесе корпоративті нысандарға шабуыл жасаумен бірге қолданылады.

Күнтізбелік фишинг.

Мұндай жағдайда фишингтік сілтемелер күнтізбелік шақыртулар арқылы жеткізіледі. Күнтізбеден шақыртулар жіберіледі, олар көптеген күнтізбелерге автоматты түрде қосылады. Бұл шақыртулар көбінесе жауапты жиналысқа шақыру және басқа да кең таралған іс-шаралар сұранысы түрінде болуы мүмкін.

Әдебиеттерге шолу. Блокчейн технологиясының дамуы криптовалюта нарығының дамуына жол ашты және блокчейнді қылмыс әлемінде қолданылуы мүмкін платформалардың біріне айналдырды. Ең жиі кездесетін қылмыс түрлерінің бірі ретінде фишинг блокчейн платформалары мен қолданушыларының үлкен экономикалық шығынға ұшырауына алып келеді. Бұл жұмыста фишингтік аккаунттарды анықтауға арналған гибридті терең оқыту желілеріне негізделетін модель ұсынылады және оның тиімділігі эфириум ортасында дәлелденеді. Ұсынылатын модель транзакциялық жазбалардан алынған белгілер арасындағы өзара байланысты алу үшін нейрондық желілерге және мақсатты аккаунттың транзакцияларын талдауға негізделетін жаңа әдіс болып табылады. Эксперимент нәтижесінде ұсынылатын әдістің

фишингтік алаяқтық аккаунттарды жоғары дәлдікпен табатындығы анықталған (Wen және т.б., 2023).

Бұл жұмыста фишингтік электронды хаттарды анықтау және жіктеу үшін биогеографияға негізделетін терең оқыту арқылы оңтайландыру моделі ұсынылады. Ұсынылатын модельдің басты мақсаты – заңды және фишингтік электронды хаттарды ажырату. Электронды поштаны жіктеу үшін терең пайымдаулардың тиімді желісінің моделі пайдаланылады және оның тиімділігі гиперпараметрлерді баптау арқылы арттырылады. Ұсынылатын модель бірнеше бағалау параметрлері бойынша ағымдағы қолданыстағы әдістерден жоғары көрсеткіш көрсеткен (Dutta және т.б., 2023).

Бұл жұмыста фишингтік хабарламаларды анықтауға арналған үш семантикалық модель ұсынылады: мәліметтер қабатын біріктірудің көп масштабты моделі (MDF), нысандар қабатын біріктірудің көп масштабты моделі (MFF) және тереңдетіп біріктірудің көп масштабты моделі (MIF). Эксперименттер нәтижесінде үш модельдің де жоғары дәлдікпен жұмыс істейтіні анықталған. 6 айға созылған 3016 фишингтік веб-сайттарды белсенді бақылау нәтижесінде ұсынылатын модельдің нақты жағдайда анықтау тиімділігі дәлелденген (Lui және т.б., 2022).

Бұл мақалада фишинг мәселесін шешуге арналған жаңа әдіс ұсынылады. Алдымен берілген веб-парақшаның HTML және ашық мәтін бөлігіндегі терминдер жиілігі символдар деңгейінде TF-IDF әдісі арқылы анықталады. Сонымен қатар, ұсынылатын гиперсілтемелік белгілер веб-парақшаның контенті мен URL мекен-жайы арасындағы өзара байланысты білдіреді. Үлкен көлемдегі дайын мәтіндік корпустың болмауына байланысты зерттеушілер 60 252 веб-парақшадан тұратын мәліметтер жинағын құрастырған. Эмпирикалық нәтижелер ұсынылатын әрбір жеке белгінің фишингтік хабарламаларды анықтауда маңызды екендігін көрсетті, десе де барлық белгілерді біріктіру фишингтік сайттарды анықтау дәлдігін арттыратындығы анықталды (Aljofey және т.б., 2022).

Бұл жұмыста көп қабатты персептрон (MLP)Hybrid Salp Swarm Jaya (HSSJAYA) көмегімен оқытылады және веб-сайттардың күмәнді, заңды немесе фишингтік екендігін анықтау үшін қолданылады. Гибридті алгоритмдер көмегімен оқытылған MLP-ді тестілеу үшін Salp Swarm (SSA) және Jaya алгоритмдері Cuckoo (CS), генетикалық алгоритм (GA) және Firefly алгоритмі арқылы оқытылған MLP-мен салыстырылған. Эксперименттер нәтижесінде HSSJAYA көмегімен оқытылған MLP-дің басқа алгоритмдермен салыстырғанда веб-сайттың фишингтік техникасын жақсырақ табатындығы анықталған (Erdemir және т.б., 2022).

Бұл жұмыстар авторлар сертификаттардан алынған жалпыға мәлім ақпарат, сонымен қатар фишингтік веб-сайттардың суреттері мен ресурстары негізінде 133 667 фишингтік веб-сайттардың өмірлік циклына талдау жүргізген. Зерттеушілер фишингтік веб-сайттарды олардың домендік атауларындағы шаблондар бойынша науқандарға топтастыру арқылы анықталған науқан-

дардың көп жағдайда бастапқы бірнеше күн ішінде, орта есеппен 12 күн ішінде жүргізілетінін анықтаған. Алынған нәтижелер фишингтік веб-сайттарды ерте анықтаудың артықшылықтары мен шектеулерін көрсетеді, атап айтқанда, веб-сайтты құру мен оны бұғаттау тізіміне енгізу арасындағы уақыт аралығына және мақалада талданған фишингтік науқандарда бірнеше аптаға дейін домендік атаулардағы үлгілердің өзгермейтініне қатысты қорытындылар алынған (Drury және т.б., 2022).

Берілген жұмыста авторлар фишинг пен спам хабарламаларын анықтау үшін терең оқыту және табиғи тілді өңдеу әдістері қолданылады. Ұсынылған әдістің тиімділігі мәтіндік жинақ үшін LSTM моделі көмегімен жіктеу барысында орташа дәлдіктің 99%, MLP моделі үшін 94% болуымен дәлелденеді (Dewis және т.б., 2022).

Бұл жұмыста фишингтік веб-сайттарды анықтау және жіктеу үшін терең автокодтау (ODAE-WPDC) желісіне негізделетін оңтайлы модель ұсынылады. Ұсынылатын модель алдымен артық шудан тазарту үшін мәтіндерге алдын ала өңдеу процесін жүргізеді. Әрі қарай белгілер анықталады, олардың көмегімен жіктеу есебі орындалады, модельдің жұмыс өнімділігін арттыру үшін оңтайландыру алгоритмі қолданылады. Нәтижесінде ұсынылатын модельдің тиімділігі Kaggle репозиторийындағы фишингтік URL мәліметтер жинағы арқылы тексеріледі. Алынған нәтижелер ұсынылатын модельдің жұмыс өнімділігінің 99,28% екендігін көрсеткен (Alqahtani және т.б., 2022).

Бұл жұмыста зерттеушілер фишингтік веб-сайттарды анықтауда URL-дің қандай сипаттамаларының пайдалы болатындығын анықтау мақсатында белгілердің корреляциясы мен рекурсивті алынуын пайдаланады. Мақалада 48 және 87 нысаннан тұратын екі мәліметтер жинағы пайдаланылған. Бірінші сценарий қуатты болжаудың ұпайлық корреляциясы мен белгілердің рекурсивті алынуын, екінші сценарий максималды ақпараттың корреляция коэффиценті мен белгілердің рекурсивті алынуын, ал үшінші сценарий Спирмен корреляциясы мен белгілердің рекурсивті алынуын біріктіреді. Үш сценарий де белгілер жиыны кіші болған жағдайда да жоғары дәлдікті қамтамасыз етеді (Moedjahedy және т.б., 2022).

Әдетте фишингті анықтаудың бірінші тапсырмасы шынайы сайтқа ұқсайтын күмәнді парақшаның пайда болғандығын растау болып табылады. Содан кейін URL мекен-жайларды талдау механизмдері арқылы күмәнді веб-парақша туралы қорытынды жасауға болады. Шынайы веб-сайт пен оның күмәнді көшірмесін салыстыру логотип, тақырыптық жолақ, қаріп түсі және стилі арқылы жүргізіледі. Берілген жұмыста фишингті анықтау үшін логотипке негізделетін жаңа әдіс ұсынылады. Жаңа әдіс негізінде ұсынылатын механизм күмәнді логотипті барынша шынайы брендтің логотипіне жіктейді. Ұсыналатын әдістің тиімділігі Оңтүстік Азия аймағындағы танымал фишингтік брендтерге негізделетін жаңадан құрастырылған мәтіндік жинақ негізінде зерттеледі. Зерттеу нәтижелері пайдаланылған 5 машиналық оқыту әдістерінің ішінде ең жоғары тиімділік кездейсоқ орман ансамблі әдісін пайдалану барысында алынатындығын көрсеткен(Panda және т.б., 2022).

Берілген жұмыста зиянды мобильдік веб-парақшаларды анықтауға арналған тиімді әдіс ұсынылған. Авторлар ұсынатын әдіс белгілер векторын құру үшін берілген URL мекен-жайдан статикалық сипаттамалар мен сайттың танымалдық көрсеткіштерін алады. Содан кейін машиналық оқыту әдістері арқылы мәтіндер жинағындағы мәліметтерге жіктеу жүргізеледі. Кездейсоқ орман жіктеуіші басқа әдістермен салыстырғанда ең жоғарғы дәлдікті көрсеткен. Сонымен қатар, зерттеушілер қолданушыларға өз мобильді құрылғылары арқылы ұсынылатын жүйемен өзара әрекеттесуге мүмкіндік беретін қосымша құрастырған (Jain және т.б., 2022).

Материалдар мен әдістер. Веб-ресурстардағы фишингтік хабарламаларды машиналық оқыту әдістері арқылы анықтау үшін ең алдымен машиналық оқыту әдістерін оқыту және тестілеуге арналған мәтіндік корпус қажет болады. Берілген зерттеу жұмысында өзіндік фишингтік мәтіндер корпусы құрастырылып, пайдаланылды. Корпусты құрастыру барысында фишингтік бағыттағы және бейтарап мәтіндер заманауи бағдарламаларды қолдану арқылы жинақталды. Корпусты құрастыру барысында келесі әрекеттер орындалды: жинақталған корпустағы мәтіндерде фишингтік және бейтарап түрдегі мәтіндер үлесі шамалас болуын қадағалау; корпустағы мәліметтерге препроцессинг алгоритмдерін орындау; құрастырылған корпус бойынша машиналық оқыту алгоритмдерін оқыту; таңдалған әдіс бойынша жаңа ақпараттық ресурстарға талдау жүргізу.

Мәліметтерді жинау. Ең алдымен фишингтік ақпараттар жинақталады, әлеуметтік желілерден белгілі бір адамдардың назарын аудартуға арналған фишингтік хабарламалар жинақталды. Фишингтік мәтіндер адамдарды тегін ұтыстар мен жеңіл ақшаға шақыратын сайттар мен әлеуметтік желілердегі пікірлерден тұрады. Корпус жинақтау барысында «Youtube», «Вконтакте» және де басқа да веб-ресурстар қолданылды.

Веб-ресурстардағы мәліметтерді жинақтау үшін дайын парсер бағдарламалары қолданылды, олар бірнеше кезеңдер бойынша жұмыс атқарады және әрбір кілттік сөз бойынша жеке мәліметтер іздестіреді. Аталған бағдарламаға қандай іздеу жүйесі арқылы (Google, Yandex) мәлімметтер жинақтау керек екенін, сонымен қатар қала, аймақ, қандайда бір мерзімді, іздеу тілін енгізе аламыз.Содан кейін белгілі бір алгоритм бойынша мәліметтер ізделіп, сүзгілеуден өткізіледі.Дайын есептерді қажетті форматта жүктеп алу мүмкіндігі бар.

Ақпарат жинау барысында «Youtube», «Вконтакте» секілді әлеуметтік желілермен басқа да веб-ресурстар таңдалды. Миллиондаған пайдаланушылар әлеуметтік желіде өздері туралы үлкен ақпаратпен бөліседі, статистикаға сәйкес, адамдар өздері туралы шындықты айтуды жөн көреді, сондықтан қандай да бір пайыздық қатынаста әлеуметтік желі парақшаларында көрсетілген ақпаратқа сенуге болады. Бұл ақпаратқа API арқылы қол жеткізуге болады. Vkontakte желісінде API алу арқылы өзімізге қажетті парақша түрін таңдау арқылы сол парақшадағы мәліметтер алынды (3-сурет).

3-сурет. «Вконтакте» әлеуметтік желісінен мәліметтер жинауға дайындық кезеңі

Препроцессинг кезеңі. Машиналық оқыту әдістерін оқытуға және тестілеуге арналған мәліметтер жинақталғаннан кейін ол мәліметтерді тазалау,токенизация, стоп сөздерді өшіру сияқтыалдын ала өңдеу процестері орындалады.

Жоғарыда келтірілген функциялар арқылы барлық мәліметтер төменгі регистргеәкелінеді,соданкейін«шулы»таңбаларбосорындарғаауыстырылады. Осыдан кейін сілтемелер «URL», ал сандар «NUM»тіркестеріне ауыстырады.

Токенизация дегеніміз – бұл мәтіндерді токенге (жеке бөліктерге) бөлудің негізгі процесі.

Көбінесе күнделікті өмірде басқа сөздерге қарағанда қарағанда жиі кездесетін, сонымен қатар сөз тіркестерінің семантикалық көлеңкесіне ешқандай әсерін тигізбейтін сөздерді стоп-сөздер деп аталады. Бұндай сөздерді негізінен мәтіннен алып тастауға да болады,себебі бұл сөздер талдау барысында белгілі бір мағына бермейді.Құрастырылған корпустың фишингтікмәліметтері үшін құрастырылған сөздер бұлтын 4-суреттен көруге болады.

4-сурет. Корпустағы фишингтік мәліметтер классының сөздер бұлты

Құрастырылған корпустағы мәтіндер мысалы мен пайыздық үлесін келесі суреттерден көруге болады (5-6 суреттер):

	label	message
0	phishing	Поздравляем, вы выиграли! или Что скрывается за этой лотереей..
1	phishing	Поздравляем Вы стали победителем конкурса фото дня ..
2	phishing	Поздравляем вы стали победителем автомобиля на конкурсе..
3	phishing	Выиграли приз (Айфон) в конкурсе..
4	phishing	Вы стали призером Системы быстрых платежей..

5-сурет. Құрастырылған корпустың фишингтік мәліметтерінің мысалы

	label	message
403	neutral	Здесь я спрятало ему зерно таким голосом только комарного писка. Неужели ты меня не видишь.
404	neutral	Бубаш повёл плавник в ту сторону откуда раздавался писк но ничего не увидел и тяжело вздохнул.
405	neutral	Зерно покатилось по земле подхваченное его дыханием и упало в маленькое отверстие вырытое муравьями.
406	neutral	Скатившись вниз по тонкому муравьиному ходу оно притихло и подумало Хорошо что подул сильный ветер и отнёс меня в сторону
407	neutral	А то бы я рассердилось и стукнуло бубашка. Вот было бы смеху если бы этот важный надутый бубаш покатился кубарем по дороге.

6-сурет. Құрастырылған корпустың бейтарап мәліметтерінің мысалы

Қазіргі таңда көптеген зерттеулерде фишингтік мәліметтерді анықтау үшін машиналық оқыту әдістеріне негізделетін алгоритмдер құрылған. Соған сәйкес берілген жұмыста бірнеше машиналық оқыту әдістері жиналған мәліметтер арқылы оқытылды және әр алгоритм бойынша жіктеу мәндері алынды. Мәтінді фишингтік және бейтарап класстарға жіктеу есебін шешу үшін Stochastic Gradient Descent, Random Forest, k-Nearest Neighbor, AdaBoost, Gaussian Naïve Bayes, Decision Tree сияқты алты түрлі машиналық оқыту әдістері пайдаланалды. Аталған әдістерді қолдану арқылы алынған жіктеу нәтижелері келесі бөлімде келтірілген.

Нәтижелер. Алдыңғы бөлімде айтылғандай, фишингтік және бейтарап мәтіндерді машиналық оқыту әдістері арқылы жіктеу барысында 6 түрлі әдіс пайдаланылды. Пайдаланылған әдістердің тиімділіктерін анықтау үшін Accuracy, Precision, Recall және F1-Score сияқты машиналық оқыту әдістерінің жұмыстарын бағалау параметрлері қолданылды.

Accuracy. Қарапайым жағдайда, мұндай метрика жіктеуіш дұрыс шешім қабылдаған құжаттардың үлесі бола алады.

Accuracy=P/N

мұндағы, P – жіктеуіш дұрыс шешім қабылдаған құжаттар саны, ал N – оқыту таңдамасының мөлшері.

Precision және Recall

Дәлдік пен толықтық (Precision және Recall)– бұл ақпарат алудың көптеген алгоритмдерін бағалауда қолданылатын өлшемдер. Кейде олар өздігінен қолданылады, кейде F-шара немесе R-Precision сияқты туынды метрикаларға негіз болады. Дәлдік пен толықтықтың мәні өте қарапайым болып табылады.

Жүйенің дәлдігі (Precision) дегеніміз – жүйенің берілген санатқа тағайындаған барлық құжаттарының шын мәнінде сол санатқа жататын құжаттар ішіндегі үлесі.

Жүйенің толықтығы (Recall) дегеніміз – жіктеуіш берілген санатқа жатады деп анықтаған құжаттардың тесттік жинақтағы сол санаттың барлық құжаттарына қатынасы.

F-өлшем

F өлшем – дәлдік пен толықтық арасындағы гармоникалық орташа мән. Егер дәлдік немесе толықтық нөлге ұмтылса, ол да нөлге ұмтылады.

$$F = 2\frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}}$$

Бұл формула дәлдік пен толықтыққа бірдей салмақ береді, сондықтан F өлшемі дәлдік пен толықтықтың төмендеуімен бірдей болады.

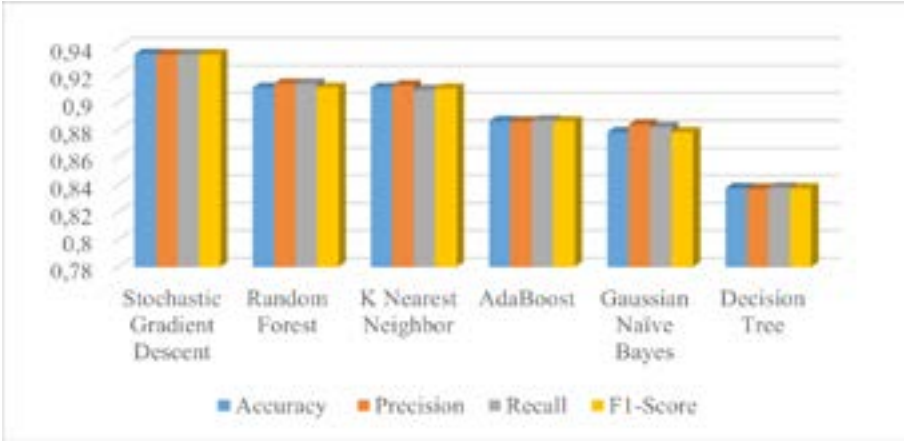
$$F = (\beta^2 + 1)\frac{\text{Precision x Recall}}{\beta^2\text{Precision} + \text{Recall}}$$

Дәлдікке басымдық бергіңіз келсе мұндағы β, 0 <β <1 диапазонындағы мәндерді қабылдайды, ал β> 1 толықтыққа басымдық береді. β = 1 болған кезде формула алдыңғы формулаға келеді және теңдестірілген F өлшемін аламыз (оны F1 деп те атайды).

Stochastic Gradient Descent, Random Forest, k-Nearest Neighbor, AdaBoost, Gaussian Naïve Bayes, Decision Treeсияқты машиналық оқыту әдістері арқылы мәтінді фишингтік және бейтарап санаттарға жіктеу нәтижесі төмендегі сурет пен диаграммада көрсетілген (7-8 сурет).

model_name	accuracy_score	precision_score	recall_score	f1_score
Stochastic Gradient Descent	0.934959	0.934748	0.934748	0.934748
Random Forest	0.910569	0.913528	0.913528	0.910568
K Nearest Neighbor	0.910569	0.912166	0.908856	0.909874
AdaBoost	0.886179	0.885714	0.886737	0.88599
Gaussian Naïve Bayes	0.878049	0.883662	0.88183	0.878017
Decision Tree	0.837398	0.836906	0.837798	0.837129

7-сурет. Машиналық оқыту әдістері арқылы мәтінді жіктеу нәтижесі



8-сурет. Машиналық оқыту әдістерінің нәтижелерін салыстыру

Кесте мен суреттен көріп тұрғанымыздай, фишингтік хабарламаларды анықтауда пайдаланылған әдістердің ішінде барлық бағалау параметрлері бойынша ең жоғары көрсеткішті көрсеткен Stochastic Gradient Descent әдісі болып табылады, яғни веб-ресурстардағы фишингтік хабарламаларды анықтау есебін аталған әдісті қолдану арқылы шешуге болады деген қорытындыға келеміз.

Қорытынды. Берілген жұмыста қазіргі таңда ең жиі кездесетін және анықталуы жағынан бірқатар ерекшеліктері бар фишингтік хабарламаларды анықтау тапсырмасы қарастырылады. Зерттеу барысында машиналық оқыту әдістерін оқыту мен тестілеуге арналған өзіндік корпус құрастырылды. Фишингтік хабарламаларды анықтау есебі құрастырылған корпус арқылы машиналық оқыту әдістерін пайдалана отырып шешіледі және бірнеше әдіс ішінен тиімдісі таңдалады. Келешекте аталған мәтіндік корпусты кеңейтіп, фишингтік хабарламаларды терең оқыту әдістері арқылы анықтау есебін шешу қарастырылады.

Information about the authors:

M. Bolatbek – University of Passau, researcher, Passau, Germany. Senior Lecturer, Department of Information Systems, Al-Farabi Kazakh National University, PhD, Almaty, Kazakhstan, E-mail: *bolatbek.milana@gmail.com*, *https://orcid.org/0000-0002-2153-180X;*

Sh. Mussiraliyeva – Candidate of Physical and Mathematical Sciences, Head of the Department of Information Systems of the Al-Farabi Kazakh National University, Almaty, Kazakhstan. E-mail: *mussiraliyevash@gmail.com*, *https://orcid.org/0000-0001-5794-3649;*

K. Bagitova – Doctoral student of the Department of Information Systems of the Al-Farabi Kazakh National University, Almaty, Kazakhstan. Senior Lecturer, Department of Computer science, Kh. Dosmukhamedov Atyrau University, Atyrau, Kazakhstan, E-mail: *kbbagitova@gmail.com*, *https://orcid.org/0000-0003-1587-1995;*

A. Nyussupov – Doctoral student of the Department of Information Systems of the Al-Farabi Kazakh National University, Almaty, Kazakhstan. E-mail: *moniumverse@gmail.com*, *https://orcid.org/0000-0003-3254-0901;*

Y. Abaiuly – Doctoral student of the Department of Information Systems of the Al-Farabi Kazakh National University, Almaty, Kazakhstan. E-mail:*erulan_97@mail.ru*, *https://orcid.org/0000-0003-2248-3819.*

REFERENCES:

Aljofey A., Jiang Q., Rasool A. et al. An effective detection approach for phishing websites using URL and HTML features. Sci Rep 12, 8842 (2022). *https://doi.org/10.1038/s41598-022-10841-5.*

Alqahtani H., Alotaibi S.S., Alrayes F.S., Al-Turaiki I., Alissa K.A., Aziz A.S.A., Maray M., Al Duhayyim M. Evolutionary Algorithm with Deep Auto Encoder Network Based Website Phishing

Detection and Classification. Applied Sciences. 2022; 12(15):7441. *https://doi.org/10.3390/app12157441*.

Dewis M., Viana T. Phish Responder: A Hybrid Machine Learning Approach to Detect Phishing and Spam Emails. Applied System Innovation. 2022; 5(4):73. *https://doi.org/10.3390/asi5040073*.

Dong-Jie Liu, Guang-Gang Geng, Xin-Chang Zhang. Multi-scale semantic deep fusion models for phishing website detection,Expert Systems with Applications,Volume 209,2022, *https://doi.org/10.1016/j.eswa.2022.118305*.

Erdemir E. i Altun A.A. (2022). Website Phishing Technique Classification Detection with HSSJAYA Based MLP Training. Tehnički vjesnik, 29 (5), 1696-1705. *https://doi.org/10.17559/TV-20211227132418*.

Jain A.K., Debnath N., Jain A.K. APuML: An Efficient Approach to Detect Mobile Phishing Webpages using Machine Learning. Wirel Pers Commun. 2022;125(4):3227-3248. doi: 10.1007/s11277-022-09707-w. Epub 2022 May 2. PMID: 35529800; PMCID: PMC9059682.

Kumar Dutta T. Meyyappan, Basit Qureshi, Majed Alsanea, Anas Waleed Abulfaraj, Manal M. Al Faraj, Abdul Rahaman Wahab Sait. Optimal Deep Belief Network Enabled Cybersecurity Phishing Email Classification Ashit, Computer Systems Science & Engineering DOI: 10.32604/csse.2023.028984.

Moedjahedy J., Setyanto A., Alarfaj F.K., Alreshoodi M. CCrFS: Combine Correlation Features Selection for Detecting Phishing Websites Using Machine Learning. Future Internet. 2022; 14(8):229. *https://doi.org/10.3390/fi14080229*.

Panda P., Mishra A.K., Puthal D. A Novel Logo Identification Technique for Logo-Based Phishing Detection in Cyber-Physical Systems. Future Internet. 2022; 14(8):241. *https://doi.org/10.3390/fi14080241*.

Tingke Wen, Yuanxing Xiao, Anqi Wang, Haizhou Wang.A novel hybrid feature fusion model for detecting phishing scam on Ethereum using deep neural network,Expert Systems with Applications,Volume 211,2023,118463,*https://doi.org/10.1016/j.eswa.2022.118463*.

Vincent Drury, Luisa Lux, and Ulrike Meyer. Dating Phish: An Analysis of the Life Cycles of Phishing Attacks and Campaigns. In Proceedings of the 17th International Conference on Availability, Reliability and Security (ARES '22). Association for Computing Machinery, New York, NY, USA, Article 14, 1–11. 2022.

https://doi.org/10.1145/3538969.3538997.

https://www.researchgate.net/publication/346808880_Internet_use_and_well-being_during_the_COVID-19_outbreak_Examining_the_role_of_gender_age_motives_for_using_the_internet_and_relational_resources_in_an_Italian_adult_sample.

https://www.cert.gov.kz/press_club/infographics.

https://en.wikipedia.org/wiki/Phishing.

МАЗМҰНЫ

СОДЕРЖАНИЕ

CONTENTS

Publication Ethics and Publication Malpracticein the journals of the National Academy of Sciences of the Republic of Kazakhstan

For information on Ethics in publishing and Ethicalguidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the National Academy of Sciences of the Republic of Kazakhstan implies that the described work has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see http://www.elsevier.com/postingpolicy), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any otherlanguage, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The National Academy of Sciences of the Republic of Kazakhstan follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (http://publicationethics.org/files/u2/New_Code.pdf). To verify originality, your article may be checked by the Cross Check originality detection service http://www.elsevier.com/editors/plagdetect.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the National Academy of Sciences of the Republic of Kazakhstan.

The Editorial Board of the National Academy of Sciences of the Republic of Kazakhstan will monitor and safeguard publishing ethics.

Правила оформления статьи для публикации в журнале смотреть на сайтах:
www:nauka-nanrk.kz
http://physics-mathematics.kz/index.php/en/archive
ISSN2518-1726 (Online),
ISSN 1991-346X (Print)

Директор отдела издания научных журналов НАН РК *А. Ботанқызы*
Заместитель директор отдела издания научных журналов НАН РК *Р. Жәлиқызы*
Редакторы: *М.С. Ахметова, Д.С. Аленов*
Верстка на компьютере *Г.Д.Жадыранова*

Подписано в печать 15.09.2022.
Формат 60х881/8. Бумага офсетная. Печать – ризограф.
10,5 п.л. Тираж 300. Заказ 4.

Национальная академия наук РК
050010, Алматы, ул. Шевченко, 28, т. 272-13-19