Kauno technologijos universitetas

Informatikos fakultetas

Hibridinis sukčiavimo SMS žinučių aptikimo metodas

Baigiamasis magistro projektas

Ignas Rutkauskas

Projekto autorius

prof. dr. Agnius Liutkevičius

Vadovas

Kaunas, 2026

Kauno technologijos universitetas

Informatikos fakultetas

Hibridinis sukčiavimo SMS žinučių aptikimo metodas

Baigiamasis magistro projektas

Informacijos ir informacinių technologijų sauga (6211BX008)

Ignas Rutkauskas

Projekto autorius

prof. dr. Agnius Liutkevičius

Vadovas

prof. dr. Jevgenijus Toldinas

Recenzentas

Kaunas, 2026

Kauno technologijos universitetas

Informatikos fakultetas

Ignas Rutkauskas

Hibridinis sukčiavimo SMS žinučių aptikimo metodas

Akademinio sąžiningumo deklaracija

Patvirtinu, kad:

1. baigiamąjį projektą parengiau savarankiškai ir sąžiningai, nepažeisdama(s) kitų asmenų autoriaus ar kitų teisių, laikydamasi(s) Lietuvos Respublikos autorių teisių ir gretutinių teisių įstatymo nuostatų, Kauno technologijos universiteto (toliau – Universitetas) intelektinės nuosavybės valdymo ir perdavimo nuostatų bei Universiteto akademinės etikos kodekse nustatytų etikos reikalavimų;

2. baigiamajame projekte visi pateikti duomenys ir tyrimų rezultatai yra teisingi ir gauti teisėtai, nei viena šio projekto dalis nėra plagijuota nuo jokių spausdintinių ar elektroninių šaltinių, visos baigiamojo projekto tekste pateiktos citatos ir nuorodos yra nurodytos literatūros sąraše;

3. įstatymų nenumatytų piniginių sumų už baigiamąjį projektą ar jo dalis niekam nesu mokėjęs (-usi);

4. suprantu, kad išaiškėjus nesąžiningumo ar kitų asmenų teisių pažeidimo faktui, man bus taikomos akademinės nuobaudos pagal Universitete galiojančią tvarką ir būsiu pašalinta(s) iš Universiteto, o baigiamasis projektas gali būti pateiktas Akademinės etikos ir procedūrų kontrolieriaus tarnybai nagrinėjant galimą akademinės etikos pažeidimą.

Ignas Rutkauskas

Patvirtinta elektroniniu būdu

Rutkauskas, Ignas. Hibridinis sukčiavimo SMS žinučių aptikimo metodas. Magistro baigiamasis projektas / vadovas prof. dr. Agnius Liutkevičius; Kauno technologijos universitetas, Informatikos fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): Informatikos inžinerija (Informatikos mokslai).

Reikšminiai žodžiai: sukčiavimo SMS žinutės, hibridinis metodas, mašininis mokymasis, taisyklėmis grįsta analizė, teksto klasifikavimas, kibernetinis saugumas

Kaunas, 2026. 75 p.

Santrauka

Šiame magistro baigiamajame darbe nagrinėjama sukčiavimo SMS žinučių aptikimo problema bei pristatomas hibridinis jų klasifikavimo metodas, pritaikytas lietuvių kalbos tekstams. Darbo aktualumą lemia sparčiai augantis SMS sukčiavimo atvejų skaičius bei ribotas lietuvių kalbai skirtų automatizuotų aptikimo sprendimų kiekis. Sukčiavimo SMS žinutės kelia reikšmingą grėsmę naudotojų informaciniam saugumui, todėl efektyvus jų aptikimas tampa svarbia kibernetinio saugumo užduotimi.

Darbo tikslas – sukurti efektyvų ir praktiškai pritaikomą metodą, leidžiantį automatiškai identifikuoti neteisėtas SMS žinutes lietuvių kalboje. Tyrimo metu atlikta mokslinės literatūros analizė, apžvelgiant dažniausiai taikomus SMS sukčiavimo aptikimo metodus, jų privalumus bei trūkumus. Taip pat išanalizuoti pagrindiniai sukčiavimo žinučių požymiai, tokie kaip nuorodų naudojimas, skubos kūrimas, apgaulingi raktažodžiai ar netipinės tekstinės struktūros.

Darbe sukurtas hibridinis metodas, jungiantis taisyklėmis grįstą analizę ir mašininio mokymosi modelį. Mašininio mokymosi dalyje buvo taikomas logistinės regresijos klasifikatorius su TF-IDF tekstų reprezentacija bei simbolių n-gramų analize. Taisyklių metodas paremtas specifinių sukčiavimo požymių identifikavimu ir jų svoriniu vertinimu. Galutinis klasifikavimo sprendimas priimamas derinant abiejų metodų rezultatus.

Eksperimentinio vertinimo metu buvo palyginti keli mašininio mokymosi modeliai, tarp jų logistinė regresija, Naive Bayes, Random Forest, Extra Trees ir Bi-LSTM. Geriausius rezultatus tarp pavienių modelių pasiekė logistinės regresijos metodas, kurio F1 įvertis siekė 89,32 %. Tuo tarpu sukurtas hibridinis metodas pasiekė 92 % tikslumą, 93,75 % preciziškumą, 90 % atkūrimo rodiklį bei 91,84 % F1 įvertį. Lyginant su vien tik mašininio mokymosi modeliu pagrįstu metodu, hibridinis sprendimas sumažino klaidingai teigiamų klasifikacijų skaičių, todėl pasižymi didesniu patikimumu praktiniam taikymui. Tyrimo metu taip pat buvo analizuojama klasifikavimo slenksčių įtaka modelio veikimui bei vertinamas skirtingų metodų jautrumas klaidingai teigiamoms ir klaidingai neigiamoms klasifikacijoms. Nustatyta, kad hibridinis metodas leidžia pasiekti geresnį balansą tarp sukčiavimo žinučių aptikimo ir teisėtų žinučių apsaugos nuo klaidingo pažymėjimo, todėl toks sprendimas yra tinkamas praktinėms SMS filtravimo sistemoms.

Gauti rezultatai parodė, kad taisyklėmis grįstos analizės integravimas leidžia pagerinti bendrą klasifikavimo kokybę ir sumažinti klaidingų perspėjimų skaičių. Sukurtas metodas yra pritaikytas lietuvių kalbos ypatumams ir gali būti naudojamas kaip pagrindas kuriant praktines SMS filtravimo sistemas, skirtas naudotojų apsaugai nuo sukčiavimo žinučių.

Rutkauskas, Ignas. A Hybrid Method for Detecting Fraudulent SMS Messages. Master's Final Degree / supervisor prof. Agnius Liutkevičius; Faculty of Informatics, Kaunas University of Technology.

Study field and area (study field group): Informatics Engineering (Computing).

Keywords: smishing, phishing via SMS, hybrid method, machine learning, rule-based analysis, text classification, cybersecurity

Kaunas, 2026. 75 p.

Summary

This master's thesis examines the problem of fraudulent SMS message detection and presents a hybrid classification method adapted for Lithuanian-language text messages. The relevance of the research is determined by the rapidly increasing number of SMS fraud cases and the limited availability of automated detection solutions designed specifically for the Lithuanian language. Fraudulent SMS messages pose a significant threat to users' information security; therefore, their effective detection has become an important cybersecurity task.

The aim of the thesis is to develop an efficient and practically applicable method capable of automatically identifying fraudulent SMS messages in Lithuanian. During the research, a review of scientific literature was conducted, analyzing the most commonly used SMS fraud detection methods, as well as their advantages and limitations. In addition, the main characteristics of fraudulent messages were examined, including the use of links, urgency creation, deceptive keywords, and atypical text structures.

A hybrid method combining rule-based analysis and a machine learning model was developed in this work. The machine learning component was based on a logistic regression classifier using TF-IDF text representation and character n-gram analysis. The rule-based method relied on the identification and weighted evaluation of specific fraud-related indicators. The final classification decision was obtained by combining the outputs of both methods.

During the experimental evaluation, several machine learning models were compared, including Logistic Regression, Naive Bayes, Random Forest, Extra Trees, and Bi-LSTM. Among the individual models, Logistic Regression achieved the best results with an F1-score of 89.32%. Meanwhile, the proposed hybrid method achieved 92% accuracy, 93.75% precision, 90% recall, and a 91.84% F1-score. Compared to the method based solely on the machine learning model, the hybrid approach reduced the number of false positive classifications, demonstrating higher reliability for practical applications. The study also examined the impact of classification thresholds on model performance and evaluated the sensitivity of different methods to false positive and false negative classifications. The results showed that the hybrid method provides a better balance between fraudulent message detection and the protection of legitimate messages from incorrect classification, making it suitable for practical SMS filtering systems.

The obtained results demonstrated that integrating rule-based analysis improves overall classification quality and reduces the number of false alerts. The developed method is adapted to the characteristics of the Lithuanian language and can serve as a basis for the development of practical SMS filtering systems intended to protect users from fraudulent messages.

Turinys

Lentelių sąrašas

Paveikslų sąrašas

Įvadas

Šis magistro baigiamasis projektas priklauso Informacijos ir informacinių technologijų saugos (6211BX008) studijų programai.

Projekto naujumas ir aktualumas

Šio darbo tema yra „Hibridinis sukčiavimo SMS žinučių aptikimo metodas“. Sukčiavimo SMS žinutėmis (angl. *smishing*) atvejai Lietuvoje tampa vis aktualesni ir sudėtingesni. Sukčiai, pasinaudodami patikimų įmonių ar institucijų vardais, siekia išgauti asmeninius ar finansinius duomenis, o šio pobūdžio atakos kelia didelę grėsmę tiek gyventojams, tiek organizacijoms. Be to, tokie incidentai neigiamai veikia ir įmonių reputaciją, kuriomis sukčiai apsimeta.

Darbo aktualumą lemia tai, kad Lietuvoje vis dar trūksta sprendimų, skirtų SMS žinučių sukčiavimo atpažinimui vietiniame kontekste. Dauguma egzistuojančių metodų yra orientuoti į anglų kalba vykdomas atakas, todėl jų efektyvumas mažėja taikant juos lietuviškoms žinutėms. Šiame darbe analizuojami realūs lietuviški SMS žinučių pavyzdžiai, todėl tai leidžia geriau atspindėti vietinį grėsmių kontekstą ir padidinti aptikimo tikslumą.

Projekto naujumas slypi kuriamame hibridiniame sukčiavimo atpažinimo metode, kuris jungia taisyklėmis grįstą analizę ir mašininio mokymosi modelį. Toks metodas leidžia efektyviau aptikti tiek aiškiai identifikuojamus sukčiavimo požymius, tiek sudėtingesnius ar mažiau akivaizdžius atvejus. Taisyklių modulis leidžia identifikuoti konkrečias rizikos priežastis, o mašininio mokymosi modelis padeda apdoroti didelius tekstinių duomenų kiekius ir aptikti sudėtingesnius šablonus.

Praktinė darbo reikšmė susijusi su sukurtu prototipu, kuris gali būti naudojamas SMS žinučių analizavimui ir sukčiavimo tikimybei įvertinti. Toks sprendimas gali būti pritaikomas kaip pagalbinė priemonė naudotojams ar organizacijoms, siekiančioms sumažinti sukčiavimo riziką.

Projekto tikslas –

sukurti sukčiavimo SMS žinučių atpažinimo metodą, pritaikytą lietuvių kalbai, ir jo pagrindu realizuoti sukčiavimo žinučių aptikimo sistemos prototipą.

Projekto uždaviniai:

1. išanalizuoti esamas sukčiavimo SMS žinutėmis problemas ir jų poveikį žmonėms;
2. išanalizuoti analogiškus mokslinius tyrimus ir egzistuojančias sistemas;
3. sukurti lietuviškų SMS sukčiavimo žinučių atpažinimo metodą, atsižvelgiantį į lietuvių kalbos specifiką;
4. apžvelgti duomenų rinkimo strategijas ir aptarti jų taikymo galimybes sukčiavimo SMS žinučių analizėje;
5. pasiūlyto metodo pagrindu realizuoti sukčiavimo žinučių atpažinimo sistemos prototipą;
6. įvertinti siūlomo metodo veiksmingumą, atliekant eksperimentinį sukurto prototipo vertinimą.

Dokumento struktūra

Magistro baigiamasis darbas sudarytas iš kelių pagrindinių dalių. Įvade pristatomas darbo aktualumas, naujumas, tikslas ir uždaviniai. Analitinėje dalyje nagrinėjama sukčiavimo SMS žinutėmis problematika bei apžvelgiami egzistuojantys aptikimo metodai ir sprendimai. Projektinėje

dalyje aprašomas kuriamas hibridinis sukčiavimo aptikimo metodas ir jo veikimo principai. Prototipo dalyje pateikiamas sukurto sprendimo realizavimas ir sistemos architektūra. Eksperimentinio vertinimo dalyje analizuojamas metodo veiksmingumas ir pateikiami tyrimo rezultatai. Darbo pabaigoje pateikiamos išvados.

1. Sukčiavimo SMS žinutėmis problemos apžvalga

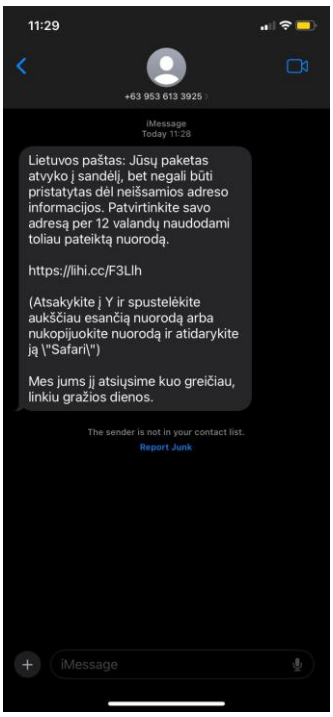
Sukčiavimas SMS žinutėmis (angl. *smishing*) būtent taip vadinamas angliškai dėl SMS bei „phishing“ žodžių derinio, yra vienas iš socialinės inžinerijos metodų, kuriuo sukčiai siekia apgauti aukas, naudodami mobiliuosius prietaisus. Šis sukčiavimo būdas, kilęs iš elektroninio pašto apgavysčių (angl. *phishing*), yra orientuotas į SMS žinučių siuntimą aukoms, siekiant pavogti jų asmeninę informaciją, paskyrų duomenis ar net bandyti išvilioti pinigus, bandant apsimesti patikimais šaltiniais [1].

Sukčiavimo žinutės dažniausiai ragina asmenis paspausti ant kenksmingos nuorodos ar atsakyti žinute, pateikiant savo asmeninę informaciją. Sukčiavimo žinutės gali paprašyti asmens jautrios informacijos, tokios kaip slaptažodžiai, PIN kodai ar tam tikri autentifikavimo kodai. Nuorodos dažniausiai būna nukreipiamos į suklastotus tinklalapius, kurie būna paruošti pavogti naudotojo asmeninę informaciją ar įdiegti kenksmingą programinę įrangą. Pačios nuorodos gali atrodyti teisėtos, tačiau gali turėti tam tikrą, mažai pastebimą variaciją arba gali būti naudojamos sutrumpintos nuorodos taip paslepiant tikrąjį nuorodos galutinį tašką [1].

Plačiai paplitusi piktnaudžiavimo registruojant klaidingus domenus praktika dar vadinama „typosquatting“ yra grindžiama spausdinimo klaidomis įvedant nuorodą į naršyklę. Piktavaliai registruoja domenus, kuriuose, lyginant su teisėtais domenais, yra rašybos klaidų, taip pasinaudodami neišvengiamais žmonių aplaidumais. Tokiose svetainėse gali būti skelbiamas įvairus turinys, kuriuo dažniausiai bandoma siekti pelno [2]. Būtent tokios nuorodos yra dažniausiai skelbiamos su SMS sukčiavimo žinute, taip bandant apgauti naudotoją, jį nukreipiant į kenksmingą svetainę.

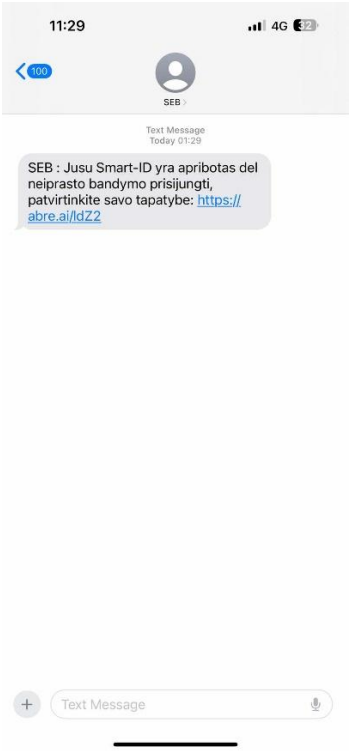
Be „typosquatting“ atvejų, kaip ir buvo minima, SMS sukčiavimo žinutėse taip pat būna nuorodos, sutrumpintos per specialias programas. Tai dar vienas iš būdų neatskleisti sutrumpintos nuorodos galutinio adreso, į kurį nukreips naudotoją. Dažniausiai tokios sutrumpintos nuorodos nukreipia žmogų į įvairiausias reklamas, taip siekiant pelno, tačiau taip pat jos gali nukreipti į kenksmingas svetaines. Net jei sutrumpintų nuorodų intencija yra nuvesti naudotoją į galinį interneto adresą, kuris yra saugus, pačios paslaugos, kurios atlieka adresų trumpinimą, gali tam tikru metu nukreipti naudotoją į nesaugią svetainę [3].

Be pateiktos SMS žinutės turinio, nuorodos esančios žinutėje, daug informacijos gavėjui gali suteikti siuntėjo telefono numeris. Piktavaliai taip pat gali naudoti įvairiausius užmaskavimo būdus, bandydami imituoti siuntimą iš organizacijų, taip norėdami išgauti kuo daugiau informacijos iš savo aukos, ar kad asmuo paspaustų kenksmingą nuorodą [4]. Dažnu atveju piktavaliai net nesistengia imituoti tos šalies numerio, kuri yra tikslinė auditorija, taip siekdami sumažinti savo kaštus ir naudotis žymiai pigesnėmis paslaugomis, siųsdami SMS sukčiavimo pranešimus. Pavyzdžiui, 1 paveiksle galima matyti pranešimą, kuris imituoja Lietuvos paštą. Kaip ir minėta anksčiau, kai kurie piktavaliai nesistengia siųsti pranešimą iš tikslinės auditorijos šalies numerio, taip mažindami kaštus. Pranešimas, kaip matoma, siųstas iš telefono numerio, prasidedančio „+63“, kuris indikuoja, kad tai yra filipinietiškas numeris.



1 pav. SMS sukčiavimo ataka nukreipta prieš lietuvišką auditoriją, pasitelkiant Filipinų numerį

Tačiau žemiau pavaizduotame 2 paveiksle galimam pastebėti, kad buvo apsimesta „SEB“ banku bei SMS sukčiavimo ataka siųsta būtent iš tokio numerio, kurį gali turėti ir „SEB“ bankas, kaip indikuojama nurodant siuntėjo pavadinimą mobiliajame įrenginyje. Iš pirmo žvilgsnio atrodo tikroviška žinutė, kol nepastebima vienos iš esminių detalių SMS sukčiavimo žinutėse – nuorodos. Tam tikrų kibernetinės saugos žinių turintis asmuo galėtų lengvai identifikuoti, kad „SEB“ bankas nesiųstų SMS žinučių, kuriuose būtų toki nepatikimai atrodantys interneto adresai, tačiau asmenį, kuris neturi tokių žinių, būtų pakankamai lengva apgauti.



2 pav. SMS sukčiavimo žinutė, apsimetant „SEB“ banku

1.1. Sukčiavimo metodai pasitelkiant SMS žinutes

Kaip jau buvo minėta, SMS sukčiavimo žinutės dažnai manipuliuoja gavėjų pasitikėjimu ir skatina impulsyvius veiksmus, ypač tarp tų, kurie neturi pakankamai kibernetinio saugumo žinių. Tokie asmenys ne tik dažniau spaudžia kenksmingas nuorodas ar pateikia savo asmeninę informaciją, bet ir tampa lengvu taikiniu, kurį piktavaliai gali nuolat išnaudoti. Piktavaliai dažniausiai remiasi psichologiniu žmonių pažeidžiamumu ir tokiomis emocijomis kaip baimė, nerimas ir susijaudinimas [5].

Buvo nustatyta, kad piktavaliui sukčiaujant su SMS žinutėmis, SMS žinutės gavėjui yra daromas poveikis produktyvumui, tapatybės vagystei, asmeninio įvaizdžio ar reputacijos žalai, emocinei žalai, pinigų praradimui [5].

Kaip rodo įvairūs tyrimai, SMS sukčiavimas yra ne tik technologiškai sudėtinga ataka, bet ir psichologiniu bei socialiniu požiūriu orientuota į aukų pažeidžiamumą. Vienas iš tyrimų nagrinėjo, kaip ir kodėl žmonės reaguoja į SMS sukčiavimo atakas. Tyrime dalyvavo 265 asmenys, kuriems buvo siunčiamos įvairios melagingos SMS žinutės, siekiant išanalizuoti jų veiksmus bei reakcijas [4].

Šis tyrimas ne tik patvirtino, kad sukčiavimas su SMS žinutėmis yra veiksmingas manipuliacijos įrankis, bet ir atskleidė, kokie veiksniai lemia asmenų polinkį į tokių atakų aukas. Pavyzdžiui, rezultatai parodė, kad 16,92% dalyvių atsakė į pirmojo bandymo sukčiavimo žinutes, o net ir po antrojo eksperimento ciklo 12,82% vis dar reagavo į sukčiavimo žinutes. Toks aukų pasikartojimo rodiklis atskleidžia ne tik žinių trūkumą, bet ir sisteminių apsaugos mechanizmų nepakankamumą [4].

Be to, tyrimas atskleidė, jog dalyvių reakcijos į SMS sukčiavimo žinutes priklausė nuo tam tikrų atributų, tokių kaip naudojamas scenarijus (baimės ar atlygio motyvacija), žinutėje paminėtos organizacijos patikimumas bei asmens mobiliojo telefono naudojimo įpročiai. Pavyzdžiui, žinutės, kuriuose buvo imituojamas prisijungimas prie „Facebook“ paskyros, pasirodė esančios itin paveikios, turinčios net 34,62% sėkmės rodiklį [4].

Dar vienas tyrimas, kuriame dalyvavo 187 asmenys, buvo pateikta 16 SMS žinučių ekrano nuotraukų su realiomis ir melagingomis žinutėmis. Dalyviai turėjo įvertinti jų tikrumą, o gauti rezultatai atskleidė kelias esmines tendencijas. Pavyzdžiui, dalyviai net 34,4% atvejų melagingas žinutes neteisingai identifikavo kaip tikras, o tik 43,6% realių žinučių buvo atpažintos teisingai. Tyrimas taip pat parodė, kad vartotojai linkę daugiau dėmesio skirti tam tikriems žinučių elementams, tokiems kaip siuntėjas ar URL adresai, tačiau dažnai jų analizė nebuvo pakankamai nuosekli ir tiksli [6].

Dar vienas tyrimas apžvelgė ne tik techninius aspektus, bet ir žmogaus elgsenos ypatumus, susijusius su SMS sukčiavimo atakomis. Tyrime buvo surinkti duomenys iš 28 dalyvių, daugiausia ne techninių sričių studentų, turinčių vidutiniškus kibernetinės higienos rodiklius. Tyrimas atskleidė, kad nors beveik visi dalyviai pripažino, jog nuorodų naudojimas iš nežinomų siuntėjų SMS žinučių gali kelti saugumo riziką, net pusė jų vis tiek spustelėdavo tokias nuorodas ir prisijungdavo prie savo paskyrų. Šis neatitikimas tarp žinių apie riziką ir faktinio elgesio pabrėžia esminį žmonių kibernetinio saugumo iššūkį. Tyrimo rezultatai taip pat atskleidė, kad dalyviai linkę labiausiai pasitikėti SMS žinučių siuntėjų numeriais, žinučių detalėmis ir gramatinėmis klaidomis, tačiau retai kreipia dėmesį į tokias detales kaip pristatymo laikas ar žinutės gavėjų sąrašas [7].

1.2. Sukčiavimo poveikis žmonėms ir įmonėms

Sukčiavimo atakos, tokios kaip SMS žinučių sukčiavimas, turi reikšmingą poveikį tiek žmonėms, tiek įmonėms. Šios atakos dažnai nukreiptos į individualius naudotojus, siekiant išvilioti jų asmeninius duomenis ar finansinius išteklius, ir gali sukelti ne tik ekonominius nuostolius, bet ir psichologinį stresą. Tyrimai parodė, kad daugelis naudotojų, ypač tie, kurie nesijaučia techniškai išprusę, nesuvokia galimų rizikų. Pavyzdžiui, iš Filipinuose atlikto tyrimo paaiškėjo, kad net 50 % apklaustųjų spaudė nuorodas iš gautų žinučių, nepaisant to, jog 75 % jų žinojo apie šių atakų keliamą pavojų [8]. SMS sukčiavimo atakos yra ypač pavojingos dėl mobiliojo ryšio naudotojų įpročių – žmonės linkę greitai reaguoti į žinutes, ypač jei jos atrodo susijusios su svarbiais klausimais, tokiais kaip banko paskyros patvirtinimas ar siūlomas prizas. Tokie sukčiavimai, kurie iš pirmo žvilgsnio atrodo nekalti, gali leisti įsilaužėliams įdiegti kenksmingą programinę įrangą, pavogti konfidencialius duomenis ar net pasisavinti finansinius išteklius.

Įmonėms šios atakos kelia dar didesnį pavojų, ypač kai jos taikomos strategiškai, siekiant destabilizuoti verslo veiklą ar pažeisti infrastruktūrą. Tyrimai rodo, kad organizacijos, kurios pasikliauja mobiliosiomis technologijomis, pavyzdžiui, energetikos sektorius, gali patirti SMS sukčiavimo atakų, kurios sukurtos manipuliuoti naudotojų elgesį. Pavyzdžiui, sukčiai gali pasinaudoti elektromobilių įkrovimo sistemomis, sukeldami neplanuotą įkrovimo piką, kuris destabilizuoja elektros tinklą ir lemia transformatorių perkrovas bei įtampas [9]. Tokios situacijos ne tik pakenkia tinklų stabilumui, bet ir mažina įmonių reputaciją bei sukelia klientų nepasitenkinimą.

Svarbu atkreipti dėmesį, kad SMS sukčiavimo atakų sudėtingumas nuolat auga dėl naujų technologijų, tokių kaip generatyvinio dirbtinio intelekto sistemos, kurios leidžia automatizuotai kurti įtikinamas ir sunkiai aptinkamas melagingas žinutes [10]. Šios technologijos leidžia sukčiams efektyviau manipuliuoti žmonėmis ir įmonėmis, apsunkinant tradicinių apsaugos priemonių veikimą. Kadangi mobiliosios platformos suteikia didesnę laisvę kibernetiniams nusikaltėliams, būtina diegti prevencines priemones.

SMS žinučių sukčiavimo prevencija reikalauja investicijų į pažangias technologijas, kurios leidžia automatiškai aptikti ir blokuoti pavojingus pranešimus. Vienas iš veiksmingų būdų – giluminio mokymosi algoritmų naudojimas. Tyrimai parodė, kad konvoliuciniai neuroniniai tinklai (angl. *convolutional neural network*) ir ilgalaikės atminties tinklai (angl. *long short-term memory networks*) gali būti labai efektyvūs, nes jie geba tiksliai analizuoti didelius duomenų rinkinius ir klasifikuoti žinutes į teisėtas arba pavojingas kategorijas [5]. Taip pat svarbu integruoti prevencines priemones, tokias kaip nuorodų analizė, juodųjų sąrašų naudojimas bei rizikų modeliavimas, leidžiantis numatyti galimus grėsmių scenarijus ir užtikrinti veiksnią apsaugą.

Žmogiškasis veiksnys taip pat yra kritiškai svarbus. SMS sukčiavimo atakos dažnai remiasi psichologinėmis manipuliavimo strategijomis, tokiomis kaip įtikinėjimas ir skubos jausmas, siekiant priversti naudotojus atlikti greitus ir neracionalius veiksmus. Todėl būtina ne tik taikyti technologinius sprendimus, bet ir didinti vartotojų sąmoningumą apie šių atakų metodus ir potencialų poveikį. Edukacinės iniciatyvos, tokios kaip informacinės kampanijos ir mokymai, gali ženkliai sumažinti naudotojų pažeidžiamumą [5].

1.3. Teoriniai metodai ir algoritmai sukčiavimo atpažinimui

Sukčiavimo atpažinimo teoriniai metodai ir algoritmai yra itin svarbi mokslinių tyrimų sritis, siekiant sukurti pažangias ir patikimas priemones mobiliojo ryšio saugumui užtikrinti. Ši sritis nagrinėja ne tik techninius, bet ir kontekstinius bei psichologinius aspektus, siekiant sukurti universalias ir efektyvias apsaugos priemones nuo įvairių sukčiavimo atakų, tokių kaip sukčiavimas SMS žinutėmis. Šiame kontekste pasitelkiami pažangūs dirbtinio intelekto metodai, apimantys giluminį mokymąsi (angl. *deep learning*), mašininio mokymosi modelius (angl. *machine learning*) bei natūralios kalbos apdorojimo algoritmus (angl. *natural language processing, NLP*) ar hibridinį atpažinimą, kuris jungia tiek taisyklėmis atliekamus patikrinimus, tiek kitą modelį.

Vienas pagrindinių iššūkių sukčiavimo atpažinimo srityje yra daugiaprasmė žinučių turinio prigimtis, kurioje piktavaliai dažnai pasitelkia subtilias manipuliavimo technikas. Sukčiavimo žinutės, turinčios trumpus tekstus ar nuorodas, tampa sunkiai atpažįstamos tradiciniais metodais. Šią problemą sprendžiant, svarbu analizuoti žinučių turinį, struktūrą bei elgsenos modelius. Viename iš tyrimų buvo pasiūlyta inovatyvi strategija, jungiant reguliariąsias išraiškas (angl. *regular expression*), giluminio mokymosi modelius, tokius kaip ilgalaikės atminties tinklai (angl. *long short-term memory*) ir jų išplėstinės formos – dvikrypčiai ilgalaikės atminties tinklai (angl. *bidirectional long short-term memory*) [11]. Šių modelių gebėjimas analizuoti duomenų ilgojo nuotolio priklausomybes leidžia efektyviau atpažinti paslėptą sukčiavimo modelį, pasiekiant aukštą tikslumą.

Natūralios kalbos apdorojimo metodai, tokie kaip „TF-IDF“ (angl. *term frequency-inverse document frequency*), užima ypatingai svarbią vietą šiame procese, nes leidžia identifikuoti retai pasitaikančius, tačiau reikšmingus terminus. Viename iš tyrimų buvo išskirtas šio metodo veiksmingumas kartu su atraminių vektorių klasifikatoriumi (angl. *support vector machine*), pasiekiant net 98,39 % tikslumą [12]. Šiame kontekste svarbus ir tinkamas funkcijų atrankos procesas, kuris leidžia optimizuoti modelio mokymosi procesą ir sumažinti klaidingų teigiamų (angl. *false positive*) rezultatų skaičių.

Sukčiavimo atpažinimui taip pat naudojamos modulinės sistemos, kurios analizuoją turinį skirtingais lygiais. Viename iš tyrimų yra siūloma keturių modulių sistemą, kuri apima turinio analizę, nuorodos tikrinimą, šaltinio kodo analizę ir „APK“ (angl. *Android package kit*) tipo failų atsisiuntimo stebėjimą [13]. Toks hierarchinis požiūris leidžia atsekti sukčiavimo veiksmus skirtinguose jų etapuose, o šio tyrimo eksperimentiniai rezultatai parodė 96,29% tikslumą. Tai patvirtina, kad modulinė architektūra, kurioje integruojami įvairūs analitiniai metodai, gali taip pat veiksmingai pagauti sukčiavimo atvejus.

Kombinuota turinio ir nuorodų analizė yra dar vienas svarbus požiūris. Viename iš tyrimų buvo naudojami keli mašininio mokymosi algoritmai, tokie kaip K artimiausių kaimynų metodas (angl. *k-nearest neighbor*), atsitiktiniai miškai (angl. *random forest*) ir ypač atsitiktinių medžių klasifikatorius (angl. *extremely randomized tree classifier*) [14]. Šių metodų derinys pasiekė įspūdingą 99,03% tikslumą. Šis rezultatas pabrėžia kompleksinių metodų integracijos naudą, ypač siekiant sukurti lankstų ir efektyvų sukčiavimo aptikimo algoritmą.

Giluminis mokymasis, ypatingai jungtinis konvoliucinis neuroninis tinklas (angl. *convolutional neural network*) ir „LSTM“ (angl. *long short-term memory*) modelis, pasižymi ypatingai dideliu potencialu šioje srityje. Viename iš tyrimų buvo pasiūlyta tokia architektūra, kuri leidžia konvoliuciniam neuroninio tinklo modeliams aptikti duomenų hierarchines struktūras, o „LSTM“ sluoksniams – analizuoti laiko priklausomybes [15]. Ši kombinacija pasiekė ne tik 99,74 % tikslumą,

bet ir ženkliai sumažino klaidingų teigiamų rezultatų skaičių, kas itin svarbu realaus pasaulio taikomosiose sistemose.

Apibendrinant, teoriniai metodai ir algoritmai, taikomi sukčiavimo atpažinimui, atspindi visapusišką požiūrį, kuris apima giluminio mokymosi algoritmų taikymą, turinio bei nuorodų analizę, modulinę architektūrą. Šių metodų efektyvumas atsiskleidžia ne tik dideliame tikslume, bet ir gebėjime pritaikyti juos įvairiose kontekstinėse situacijose, mažinant klaidingų teigiamų rezultatų skaičių. Ateities tyrimai galėtų sutelkti dėmesį į šių metodų pritaikomumo plėtrą realaus laiko aplinkose bei jų gebėjimą prisitaikyti prie nuolat kintančių sukčiavimo technikų.

1.4. Mašininio mokymosi taikymas sukčiavimo atpažinimui

Mašininio mokymosi metodai tapo reikšminga priemone kovojant su sukčiavimo SMS žinutėmis problema. Šios technologijos pagrindas yra gebėjimas analizuoti ir atpažinti modelius bei anomalijas dideliuose duomenų rinkiniuose. Sukčiavimo žinučių aptikimui pasitelkiama teksto analizė, kurioje identifikuojami modeliai, remiantis specifinėmis savybėmis, išskiriamomis iš žinučių turinio ir nuorodų adresų. Tokios savybės, kaip žodžių dažnumas, retumas ar jų tarpusavio ryšiai, padeda sukurti efektyvius klasifikatorius, skirtus atskirti teisėtas žinutes nuo sukčiavimo žinučių. Vienas plačiausiai naudojamų metodų savybių reprezentavimui yra „TF-IDF“ metodas, kuris nustato svarbiausius žodžius, remiantis jų dažniu ir retumu visame duomenų rinkinyje. Šis metodas buvo taikytas viename tyrime, siekiant sukurti tikslius ir patikimus savybių rinkinius SMS klasifikacijai [12].

Pasirinkus tinkamas savybes, taikomi įvairūs klasifikavimo algoritmai, tokie kaip „atraminių vektorių mašinos“ (angl. *support vector machine, SVM*), „atsitiktiniai miškai“ (angl. *random forest, RF*) ar „naivusis Bajesas“ (angl. *naive Bayes, NB*). „SVM“ pasižymi gebėjimu efektyviai atskirti klases net esant sudėtingiems modeliams, kai sukčiavimo žinutės pasižymi įvairiomis lingvistinėmis ar semantinėmis manipuliacijomis. Tyrimas parodė, kad „SVM“ algoritmas pasiekė aukščiausią tikslumą – 98,39%, lyginant su kitais algoritmais [12].

Logistinės regresijos metodas taip pat plačiai taikomas sukčiavimo SMS žinučių atpažinimo užduotyse ir pasižymi geru tikslumo bei interpretabilumo balansu. Naujesni tyrimai rodo, kad šis modelis gali pasiekti stabilius rezultatus, ypač derinant jį su TF-IDF požymių reprezentacija ir papildomais semantiniais indikatoriais. Pavyzdžiui, viename tyrime nustatyta, kad į modelį įtraukus emocinius požymius ir frazių lygmens rizikos indikatorius, logistinės regresijos tikslumas padidėjo nuo 91,86 % iki 92,21 %, kas patvirtina šio metodo tinkamumą trumpų ir emociškai manipuliatyvių SMS žinučių klasifikavimui [16]. Tuo tarpu kitame darbe siūlomas turinio ir URL analizės derinimas parodė, kad net ir paprastesni klasifikatoriai, įskaitant logistinės regresijos modelį, gali efektyviai prisidėti prie bendro sistemos tikslumo, ypač kai derinami keli skirtingi požymių tipai, tokie kaip tekstiniai ir nuorodų atributai [17]. Šie rezultatai rodo, kad logistinė regresija išlieka praktiškai vertingu sprendimu dėl nedidelių skaičiavimo sąnaudų, gero interpretavimo galimybių ir gebėjimo efektyviai veikti tekstų klasifikavimo užduotyse.

Kitas svarbus tyrimas pasiūlė integruotą metodą, apimantį kelis analizės modulius: teksto analizatorių, nuorodų filtrą, šaltinio kodo analizavimą ir „APK“ tipo failų atsisiuntimo analizavimą [13]. Šis metodas ne tik analizavo SMS žinučių turinį, bet ir įvertino nuorodos adreso patikimumą, nuorodos šaltinio kodą ir galimą kenksmingų failų atsisiuntimą. „NB“ (angl. *naive Bayes, NB*) klasifikatorius, taikomas šioje sistemoje, leido tiksliai atpažinti sukčiavimo žinučių dalis pagal teksto

turinį, o papildoma nuorodų ir šaltinio kodo analizė reikšmingai sumažino klaidingų teigiamų (angl. *false positive*) rezultatų skaičių. Tyrimo metu buvo pasiektas 96,29% tikslumas, kuris parodė šio metodo veiksmingumą.

Svarbu pažymėti, kad ne visi mašininio mokymosi algoritmai yra vienodai efektyvūs visais atvejais. Jų veikimas priklauso nuo duomenų kokybės, savybių rinkinio ir pačių algoritmų gebėjimo pritaikyti duomenų šablonus.

Nepaisant pasiektų rezultatų, mašininio mokymosi metodai turi tam tikrų apribojimų. Vienas pagrindinių iššūkių yra statinis savybių rinkinių pobūdis, kuris gali būti nepakankamas dinamiškoms ir nuolat kintančioms sukčiavimo strategijoms. Piktavaliai greitai adaptuojasi prie egzistuojančių apsaugos sistemų, todėl mašininio mokymosi modeliai turi būti reguliariai atnaujinami ir treniruojami su naujausiais duomenimis.

Apibendrinant galima teigti, kad mašininio mokymosi metodai teikia tvirtą pagrindą sukčiavimo SMS žinutėmis aptikimui, tačiau jų efektyvumas priklauso nuo tinkamai parinkto savybių rinkinio, algoritmų ir duomenų kokybės. Šių metodų tobulinimas ir integracija su kitomis technologijomis, tokiomis kaip giluminio mokymosi modeliai, gali žymiai pagerinti kovos su sukčiavimo atakomis efektyvumą.

1.5. Hibridinių analizės metodų taikymas sukčiavimo atpažinimui

Hibridinių analizės metodų taikymas sukčiavimo SMS aptikimui tapo svarbia sritimi, siekiant užtikrinti efektyvesnę apsaugą nuo nuolat tobulėjančių sukčiavimo strategijų. Šie metodai jungia skirtingas technologijas ir analizės technikas, siekiant integruoti įvairių modelių stipriąsias puses. Hibridinių metodų pagrindas yra daugelio analizės etapų derinimas, įtraukiant teksto turinio, nuorodų adresų ir kitų šaltinių analizę. Toks daugiasluoksnis požiūris leidžia pagerinti aptikimo tikslumą, sumažinti klaidingų teigiamų rezultatų skaičių ir prisitaikyti prie sudėtingesnių grėsmių.

Viename iš tyrimų buvo pasiūlytas inovatyvus požiūris, kuriame naudojami hibridiniai analizės moduliai, siekiant tiksliau atskirti teisėtas žinutes nuo sukčiavimo žinučių [13]. Ši sistema susideda iš keturių pagrindinių modulių: teksto analizatoriaus, nuorodų filtro, šaltinio kodo analizatoriaus ir „APK“ tipo failų atsisiuntimo detektoriaus. Teksto analizatorius naudoja „naivųjį Bajesą“ kaip pagrindinį klasifikatorių, kuris leidžia nustatyti sukčiavimo elementus žinučių turinyje, remiantis iš anksto apibrėžtomis savybėmis. Nuorodų filtras tiria nuorodos patikimumą, įvertindamas domenų savybes ir galimas grėsmes, susijusias su nuorodomis, o šaltinio kodo analizatorius identifikuoja kenksmingą kodą, įterptą tinklalapyje. Šis daugiasluoksnis požiūris užtikrina didesnį tikslumą, sumažinant klaidingų teigiamų rezultatų skaičių, ir parodė 96,29% tikslumą.

Kitame tyrime buvo nagrinėjama, kaip teksto analizė ir nuorodų adresų klasifikavimas gali būti efektyviai sujungti naudojant klasifikatorių, apimantį algoritmus, tokius kaip „k artimiausių kaimynų“ metodas (angl. *k-nearest neighbors, KNN*), „atsitiktiniai miškai“ (angl. *random forest, RF*) ir „ekstremaliai atsitiktiniai medžiai“ (angl. *extremely randomized trees, ETC*) [14]. Šiame modelyje, teksto ir nuorodų savybės buvo sujungtos, siekiant pagerinti aptikimo tikslumą ir sumažinti klaidingų rezultatų skaičių. Tyrime buvo naudojama „TF-IDF“ technika, siekiant išryškinti svarbiausius terminus teksto analizei, o duomenų balansas buvo pasiektas naudojant sintetinio mažumos parinkimo metodą (angl. *Synthetic Minority Oversampling Technique, SMOTE*). Eksperimentiniai rezultatai parodė, kad klasifikatorius pasiekė 99,03% tikslumą.

Svarbu pažymėti, kad hibridiniai metodai neapsiriboja tik tradiciniais mašininio mokymosi algoritmais bei kitomis analizės formomis. Viename iš tyrimų buvo nagrinėta giluminio mokymosi ir reguliariųjų išraiškų (angl. *regular expressions, Regex*) derinimo efektyvumas sukčiavimo žinučių aptikimui [11]. Pasiūlytas metodas siekia spręsti dvi pagrindines problemas: nepakankamą teksto konteksto suvokimą tradiciniuose klasifikavimo algoritmuose ir duomenų disbalansą tarp sukčiavimo ir teisėtų žinučių klasių. Naudojant reguliariųjų išraiškų taisykles kaip pradinį filtrą, sukčiavimo žinutės buvo išgrynintos prieš jų apdorojimą giluminio mokymosi modeliuose, įskaitant „ilgosios-trumposios atminties tinklus“ (angl. *long short-term memory, LSTM*) ir jų pažangesnę versiją – dvisluoksnius „ilgosios-trumposios atminties tinklus“ (angl. *bidirectional LSTM, Bi-LSTM*). Tyrimo metu „Bi-LSTM“ modeliai, derinami su reguliariųjų išraiškų filtru, parodė aukščiausią tikslumą, viršijantį tradicinius mašininio mokymo (angl. *machine learning*) modelius, tokius kaip „SVM“ ar „NB“.

Reguliariųjų išraiškų įtraukimas į duomenų paruošimą pasirodė esąs svarbus veiksnys, nes jis sumažino „triukšmo“ lygį ir padėjo geriau išskirti svarbiausias teksto savybes. Tai užtikrino geresnį giluminio mokymosi modelių jautrumą, kuris buvo vertinamas pagal įvairius rodiklius. „LSTM“ grįsti modeliai pasižymėjo ypatingu tikslumu, siekiančiu 98%, o reguliariųjų išraiškų taisyklės leido gerokai sumažinti klaidingai teigiamų (angl. *false positive*) rezultatų skaičių. Toks požiūris rodo didelį potencialą tobulinant SMS sukčiavimo aptikimo technologijas.

Tačiau hibridinių metodų įgyvendinimas susiduria su tam tikrais iššūkiais. Pirma, šie metodai dažnai reikalauja didesnio skaičiavimo resursų dėl sudėtingesnių modelių ir kelių analizės etapų integravimo. Antra, nuolat besikeičiantys sukčiavimo būdai reikalauja, kad hibridinės sistemos būtų reguliariai atnaujinamos ir pritaikomos naujausioms grėsmėms. Be to, šių metodų veiksmingumas labai priklauso nuo kokybiškų duomenų ir tinkamai sukonstruotų savybių rinkinių.

Apibendrinant galima teigti, kad hibridinių analizės metodų taikymas sukčiavimo SMS aptikimui teikia didelį potencialą, derinant skirtingų technologijų stiprybes. Jie užtikrina aukštesnį tikslumą, mažesnį klaidingų rezultatų skaičių ir didesnį atsparumą dinamiškai kintančioms grėsmėms. Tačiau norint maksimaliai išnaudoti šių metodų privalumus, būtina užtikrinti, kad jie būtų taikomi kartu su reguliaria priežiūra, naujausių duomenų integracija ir tinkamu išteklių valdymu. Šis požiūris sudaro tvirtą pagrindą tolimesniam giluminiam sukčiavimo aptikimo technologijų vystymui.

1.6. Modelių validacijos bei efektyvumo patikrinimo metodai sukčiavimo atpažinimui

Modelių validacija ir efektyvumo patikrinimas yra esminė dalis kuriant veiksmingus sukčiavimo SMS aptikimo algoritmus. Šių modelių efektyvumas vertinamas pagal tam tikras metrikas, kurios leidžia nustatyti jų gebėjimą teisingai klasifikuoti žinutes. Tokios metrikos kaip tikslumas, jautrumas, preciziškumas ar „F1“ balas suteikia galimybę analizuoti skirtingų modelių stipriąsias bei silpnąsias puses. Šios metrikos yra kritiškai svarbios, ypač kai algoritmai pritaikomi praktiškai, nes realiame pasaulyje modeliai dažnai susiduria su nesubalansuotais ir „triukšmingais“ duomenimis. Jų formulės bei aprašymai yra aprašomi vienodai daugelio tyrimų [11,15]:

Tikslumas yra viena pagrindinių metrikų, naudojamų modelių efektyvumui vertinti. Šis rodiklis apskaičiuojamas kaip teisingai klasifikuotų žinučių dalis iš visų žinučių:

$$Tikslumas\ (angl.accuracy) = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

čia:

- TP (angl. *true positive*) – teisingai atpažintos sukčiavimo žinutės;
- TN (angl. *true negative*) – teisingai atpažintos teisėtos žinutės;
- FP (angl. *false positive*) – klaidingai sukčiavimu priskirtos teisėtos žinutės;
- FN (angl. *false negative*) – klaidingai teisėtomis priskirtos sukčiavimo žinutės.

Jautrumas (angl. *recall*) apskaičiuojamas kaip santykis tarp teisingai atpažintų sukčiavimo žinučių ir visų faktinių sukčiavimo žinučių:

$$Jautrumas\ (angl. sensitivity\ arba\ recall) = \frac{TP}{TP + FN} \quad (2)$$

čia:

- TP (angl. *true positive*) – teisingai atpažintos sukčiavimo žinutės;
- FN (angl. *false negative*) – klaidingai teisėtomis priskirtos sukčiavimo žinutės.

Šis rodiklis pabrėžia modelio gebėjimą aptikti visas sukčiavimo žinutes, o jo svarba ypač išryškėja, kai klaidingai neaptiktos sukčiavimo žinutės gali sukelti didelę žalą.

Preciziškumas (angl. precision) parodo, kokia dalis žinučių, klasifikuotų kaip sukčiavimo, iš tikrųjų buvo sukčiavimo žinutės. Jis yra apskaičiuojamas:

$$Preciziškumas\ (angl. precision) = \frac{TP}{TP + FP} \quad (3)$$

čia:

- TP (angl. *true positive*) – teisingai atpažintos sukčiavimo žinutės;
- FP (angl. *false positive*) – klaidingai sukčiavimu priskirtos teisėtos žinutės.

F1 balas, kuris yra harmoninis jautrumo ir tikslumo vidurkis, suteikia subalansuotą vertinimą, ypač kai duomenys yra nesubalansuoti:

$$F1\ balas = 2 \times \frac{Jautrumas \times Preciziškumas}{Jautrumas + Preciziškumas} \quad (4)$$

čia:

- Preciziškumas – preciziškumo rodiklis;
- Jautrumas – jautrumo rodiklis.

Klaidų matrica yra vizualus modelio klaidų ir teisingų prognozių atvaizdavimas. Ji leidžia detaliai suprasti modelio veikimą. Kad padėti tai suprasti, galima pažvelgti į 1 lentelę esančią žemiau.

1 lentelė. Klaidų matricos atvaizdavimo pavyzdys.

	Prognozė: Tikra žinutė	Prognozė: Sukčiavimo žinutė
Realybė: Tikra žinutė	TN (angl. *true negative*)	FP (angl. *false positive*)
Realybė: Sukčiavimo žinutė	FN (angl. *false negative*)	TP (angl. *true positive*)

Šioje matricoje lengva nustatyti, kiek klaidų yra padaryta ir kokios jos rūšys „FP“ ar „FN“.

Toliau pateikiama 2 lentelė, kuri apibendrina svarbiausias metrikas, gautas įvairiuose apžvelgtuose tyrimuose:

2 lentelė. Svarbiausios metrikos, gautos įvairiuose tyrimuose.

Tyrimo pavadinimas	**Metodologijos tipas**	**Modelis**	**Modulių skaičius**	**Tikslumas (%)**	**Jautrumas (%)**	**Preciziškumas (%)**	**F1 balas**
Deep learning-based smishing message identification using regular expression feature generation [11]	Giluminis mokymasis	„Bi-LSTM“ (dvipusis ilgalaikės atminties tinklas) + „Regex“ taisyklių generavimas duomenų išankstiniam apdorojimu	1	99,09	94,75	99,57	0,9710
Smishing Detector: A security model to detect smishing through SMS content analysis and URL behavior analysis [13]	Hibridinis	„NB“ + Nuorodų filtras + Šaltinio kodo analizė + APK failų analizatorius	4	96,29	92,00	93,00	0,92
Detecting Smishing Attacks Using Feature Extraction and Classification Techniques [12]	Mašininis mokymasis	Atraminių vektorių mašinos (angl. *support vector machine, SVM*)	1	98,39	99,79 tikrų žinučių, 89,26 sukčiavimo žinučių	98,37 tikrų žinučių, 98,52 sukčiavimo žinučių	0,9908 tikrų žinučių, 0,9366 sukčiavimo žinučių
A content and URL analysis-based efficient approach to detect smishing SMS in intelligent systems [14]	Hibridinis	Klasifikatorius (KNN, RF, ETC) + TF-IDF tekstų savybėms išgauti	2	99,03	-	98,94	-
Enhancing Smishing Detection: A Deep Learning Approach for Improved Accuracy and Reduced False Positives [15]	Giluminis mokymasis	CNN-LSTM (konvoliuciniai neuroniniai tinklai + ilgalaikės atminties tinklai)	1	99,74	99,00	99,00	0,99

1.7. Duomenų apdorojimo problemos ir metodai mašininiame mokyme

Mašininio mokymosi modeliai yra tiesiogiai priklausomi nuo duomenų kokybės, todėl prieš juos naudojant būtina atlikti duomenų paruošimą. Neapdoroti duomenys dažnai turi trūkstamų reikšmių, klaidų ar nereikalingos informacijos, o tai gali lemti netikslius modelio rezultatus [18].

Viena iš dažniausiai pasitaikančių problemų – trūkstami duomenys. Jie gali atsirasti dėl įvairių priežasčių, pavyzdžiui, netinkamo duomenų rinkimo ar techninių klaidų. Trūkstamos reikšmės gali būti pakeistos paprastais būdais, pavyzdžiui, įrašant vidutinę ar dažniausiai pasitaikančią reikšmę. Jei duomenų trūksta per daug, gali būti tikslinga juos pašalinti, kad modelis nebūtų klaidinamas [18].

Kita problema – duomenų netikslumai ir triukšmas. Tai gali būti neteisingos reikšmės ar netiksli informacija, kuri trukdo modeliui mokytis. Tokiu atveju gali būti naudojami filtravimo metodai, leidžiantys išvalyti duomenis ir palikti tik reikšmingą informaciją [18].

Kartais duomenų reikšmės būna skirtingo mastelio, pavyzdžiui, vienos gali būti didelės, kitos labai mažos. Dėl to modelis gali vienoms duomenų dalims suteikti didesnę svarbą nei kitoms. Norint to išvengti, duomenys gali būti suvienodinami, kad visi rodikliai turėtų panašią svarbą [18].

Kai duomenų yra per daug, dalis jų gali būti nereikalinga arba jie gali dubliuotis. Tai gali apsunkinti modelio darbą ir pailginti skaičiavimus. Tokiais atvejais galima pašalinti nereikšmingus duomenis arba sumažinti jų kiekį, paliekant tik svarbiausią informaciją [18].

Priešingai, kai duomenų yra per mažai, modelis gali tapti nepatikimas ir per daug prisitaikyti prie esamų duomenų, todėl prastai veiks su naujais duomenimis. Tokiais atvejais galima dirbtinai padidinti duomenų kiekį (angl. *data augmentation*), pavyzdžiui, sukuriant naujus duomenis pagal esamus, keičiant jų padėtį ar formą, kad modelis galėtų išmokti įvairesnius pavyzdžius [18].

Jei duomenys surinkti iš skirtingų šaltinių, jie gali būti skirtingo formato ar struktūros, todėl svarbu juos suderinti ir sujungti į vieną bendrą duomenų rinkinį. Tai padeda išvengti dublikatų ar nesuderinamumo problemų ir užtikrinti, kad modelis dirbtų su vieninga, aiškia informacija [18].

Tinkamai apdoroti duomenys leidžia modeliams pasiekti geresnių rezultatų, veikti greičiau ir tiksliau. Tai ne tik sumažina klaidų tikimybę, bet ir padeda užtikrinti, kad modelis gebės pritaikyti savo išmoktas žinias naujiems duomenims. Be šio žingsnio net pažangiausi algoritmai negalėtų veikti efektyviai, todėl duomenų paruošimas yra būtinas kiekviename mašininio mokymosi procese.

1.8. Diakritinių ženklų įtaka mašininio mokymosi modelių efektyvumui

Diakritiniai ženklai yra svarbus rašytinės kalbos elementas, lemiantis ne tik jos skaitomumą, bet ir tikslumą. Daugelio natūralios kalbos apdorojimo (angl. *natural language processing*) sistemų efektyvumas priklauso nuo teksto kokybės, o diakritinių ženklų trūkumas gali neigiamai paveikti įvairias taikymo sritis, tokias kaip automatinis vertimas, informacijos paieška ar sintaksinė analizė. Ne tik lietuvių, bet ir kitų kalbų, turinčių diakritinius ženklus, automatinis apdorojimas susiduria su problema, kai tekste šie ženklai yra praleisti dėl naudotojų rašymo įpročių arba technologinių apribojimų [19].

Diakritinių ženklų atkūrimas dažnai laikomas specialiu rašybos korekcijos atveju, tačiau tai sudėtingesnė užduotis, nes gali kilti daugiaprasmiškumo problemų. Pavyzdžiui, be diakritinių ženklų užrašytas lietuviškas žodis "rastas" gali reikšti tiek "rąstas" (medžio kamieno dalis), tiek "raštas"

(dokumentas), tiek "rastas" (būtojo laiko veiksmažodis). Šios problemos sprendimui taikomi du pagrindiniai metodai: simbolių lygmens mašininis mokymasis ir kalbos modeliavimas. Simbolių lygmens metodai išmoksta atkurti diakritinius ženklus remdamiesi atskiromis raidėmis bei jų kontekstu, tuo tarpu kalbos modeliai įvertina platesnį kontekstą, dažniausiai n-gramų principu [19].

Lietuvių kalbai atlikti tyrimai parodė, kad trigraminis kalbos modelis pasiekia ~99,5 % tikslumą atkuriant simbolius ir ~98,4 % tikslumą atkuriant žodžius, lenkdamas simbolių lygmens metodus atitinkamai ~1,4 % ir ~3,8 %. Šie rezultatai rodo, kad įtraukus platesnį kontekstą galima ženkliai padidinti diakritinių ženklų atkūrimo tikslumą. Be to, kalbos modeliavimas geriau susidoroja su atvejais, kai žodžio forma yra daugiaprasmė, nes įvertina greta esančius žodžius ir jų tarpusavio ryšius. Tai ypač svarbu apdorojant realaus pasaulio tekstus, tokius kaip interneto komentarai ar socialinių tinklų įrašai, kuriuose dažnai pasitaiko įvairių neatitikimų ir klaidų [19].

Vienas iš aktualių praktinių pritaikymų yra SMS sukčiavimo atakų aptikimas lietuvių kalboje, kur kyla klausimas, kaip optimaliai tvarkyti diakritinius ženklus. Galimos trys strategijos: visiškai išlaikyti diakritinius ženklus, visiškai jų atsisakyti arba taikyti hibridinį metodą, priklausomai nuo turimų duomenų pobūdžio.

Pirmoji strategija – visų diakritinių ženklų išsaugojimas – yra naudinga, kai analizuojami formalūs tekstai, kur rašyba atitinka norminės kalbos reikalavimus. Tai leidžia modeliams išnaudoti kontekstinę informaciją ir tiksliau atpažinti struktūrinius teksto elementus. Tačiau šis metodas gali būti mažiau veiksmingas realiomis sąlygomis, kur diakritiniai ženklai dažnai būna praleidžiami ar naudojami netaisyklingai.

Antroji strategija – visiškas diakritinių ženklų pašalinimas – supaprastina duomenų apdorojimą ir sumažina modelio priklausomybę nuo skirtingų rašybos variantų. Tai gali būti efektyvu sistemose, kurios veikia su dideliais tekstų kiekiais, tačiau kyla rizika, kad modelis negebės atskirti homonimų ar tiksliai suprasti semantinių skirtumų tarp žodžių, pavyzdžiui, „karstas“ ir „karštas“.

Trečioji strategija – hibridinis metodas – leidžia išlaikyti diakritinius ženklus, jei jie yra pateikiami tekste, tačiau taip pat leidžia modeliams apdoroti jų neturinčias alternatyvas. Tai ypač svarbu SMS sukčiavimo atakų aptikime, kur sukčiai gali naudoti tiek taisyklingus, tiek modifikuotus tekstus. Pavyzdžiui, jei sistemoje aptinkami žodžiai „demesio“ ir „dėmesio“, abi versijos turėtų būti vertinamos kaip galimos grėsmės indikacijos. Toks požiūris užtikrina didesnį lankstumą ir prisitaikymą prie realių duomenų, išlaikant semantinį tikslumą.

1.9. SMS sukčiavimo žinučių atpažinimo metodų pritaikymas lietuvių kalbai

Sukčiavimo SMS žinučių aptikimo tyrimai Lietuvoje šiuo metu yra itin riboti – jų paprasčiausiai nėra, o lietuvių kalbos specifika šiame kontekste kelia reikšmingus iššūkius. Daugelis aptartų metodų, tokių kaip „TF-IDF“, giluminio mokymosi ar hibridiniai sprendimai, buvo pritaikyti pasauliniu mastu anglų kalbos ar kitų plačiai naudojamų kalbų tekstams. Tačiau lietuvių kalba, turinti sudėtingą gramatinę struktūrą, didelę žodžių formų variaciją ir specifinį kontekstą, reikalauja kitokio požiūrio. Tokios kalbinės savybės gali daryti tiesioginę įtaką mašininio mokymosi modelių veikimui ir tikslumui, todėl būtina adaptuoti arba kurti naujus metodus, pritaikytus šiai kalbai.

Lietuvoje šiuo metu nėra viešai prieinamų duomenų rinkinių, skirtų SMS sukčiavimo tyrimams lietuvių kalba. Šis trūkumas reiškia, kad vienas iš projekto tikslų yra sukurti ir parengti naują

duomenų rinkinį, kuriame būtų apimti realūs pavyzdžiai, sudaryti iš teisėtų ir sukčiavimo SMS žinučių. Tokio rinkinio kūrimas bus unikalus ir reikšmingas indėlis šioje srityje, užtikrinantis, kad naujai sukurti modeliai galėtų būti tinkamai apmokomi ir vertinami. Rinkinio sudarymui gali būti naudojami duomenų šaltiniai, tokie kaip viešai skelbiamos žinučių pavyzdžių ekrano kopijos, surinktos iš socialinių tinklų ar forumų, taip pat bus sukurta platforma, kur žmonės savanoriškai galės kelti SMS sukčiavimo nuotraukas, taip pildant archyvą bei įspėjant kitus apie panašius pavojus Lietuvoje.

Metodų pritaikymas lietuvių kalbai taip pat reikalauja kalbos apdorojimo technikų pritaikymo. Pavyzdžiui, „TF-IDF“ metodas, plačiai naudojamas kitose kalbose, lietuvių kalboje gali būti mažiau efektyvus dėl žodžių įvairovių. Todėl būtina analizuoti bei atrasti būdus, gebančius aptikti lietuviškus SMS sukčiavimo atvejus.

Taip pat svarbu paminėti, kad šiuo metu Lietuvoje nėra sistemų, skirtų SMS sukčiavimo aptikimui ir prevencijai. Tai reiškia, kad šis projektas ne tik užpildys šią spragą, bet ir taps pirmuoju tokio tipo sprendimu, skirtu Lietuvos rinkai. Sistemos unikalumas slypi ne tik kalbos specifikos adaptacijoje, bet ir integruotuose sprendimuose, apimančiuose realaus laiko analizę ir kontekstinį supratimą. Tokia sistema galėtų užtikrinti ne tik efektyvų sukčiavimo aptikimą, bet ir prisidėti prie visuomenės švietimo, informuojant visuomenę apie galimas grėsmes bei naujus sukčiavimo atvejus.

Aptarus turimus metodus, galima teigti, kad, nors kai kurie iš jų galėtų būti pritaikyti lietuvių kalbai su tam tikrais pakeitimais, tokio konteksto tyrimų stoka rodo, kad reikalingi nauji ir originalūs sprendimai. Todėl šis projektas yra ne tik inovatyvus, bet ir būtinas, siekiant sukurti efektyvią ir unikaliai lietuvių kalbai pritaikytą SMS sukčiavimo aptikimo sistemą. Ši sistema turėtų tapti pagrindu tolimesniems tyrimams ir galimybe integruoti pažangius metodus į Lietuvos kibernetinio saugumo ekosistemą.

1.10. Esamų sprendimų ir realizacijų apžvalga sukčiavimo atpažinimui

Vienas iš reikšmingų darbų, nagrinėjančių sukčiavimo SMS aptikimą yra „Implementation of ‘Smishing Detector’“ tyrimas [20]. Šiame darbe autoriai pateikė modelį, kuris naudoja dirbtinius neuroninius tinklus (angl. *artificial neural networks, ANN*), siekiant klasifikuoti SMS žinutes į teisėtas ir sukčiavimo kategorijas. Mokymo procese buvo taikomas atgalinės sklaidos algoritmas (angl. *backpropagation algorithm*), kuris optimizuoja svorius, mažindamas prognozės paklaidas. Eksperimento metu buvo naudojamas realių žinučių SMS rinkinys, sudarytas iš 538 sukčiavimo ir 5320 teisėtų žinučių. Modelis pasiekė 97,4% tikslumą, lenkdamas „naivųjį Bajeso“ ir „sprendimų medžio“ (angl. *decision tree*) algoritmus, kurių tikslumas atitinkamai buvo 96,29% ir 93,4%. Tyrimo metu taip pat buvo išskirtos septynios pagrindinės sukčiavimo žinučių savybės, kurios apima nuorodas (URL), sukčiavimo raktinius žodžius, specialius simbolius ir telefono numerius. Ypač reikšminga buvo nuorodų analizė, kuri viena pasiekė 94% tikslumą. Tyrimo rezultatai parodė, kad „ANN“ geba efektyviai identifikuoti sukčiavimo SMS, išryškindamas svarbiausias savybes, kurios užtikrina tikslų ir patikimą modelio veikimą.

Kitame darbe buvo pristatyta mobilioji programa, skirta automatiniam sukčiavimo SMS aptikimui [21]. Ši sistema integruoja įvairius komponentus, tokius kaip nuorodų analizatorius, turinio analizė ir dinaminis modelių mokymas, siekiant aptikti sukčiavimo žinutes realiuoju laiku. Mokymui buvo naudojamas platus duomenų rinkinys, kurį sudarė tiek teisėtos, tiek sukčiavimo žinutės, surinktos iš įvairių šaltinių. Tyrime pabrėžta svarba įtraukti realaus laiko scenarijus, kad sistema galėtų efektyviai

veikti mobiliųjų įrenginių aplinkoje, ypač atsižvelgiant į nuolat besikeičiančias sukčiavimo schemas. Rezultatai parodė, kad sistema pasiekė 98,42% tikslumą, išnaudodama pažangią analizę, kuri padėjo aptikti subtilius sukčiavimo bruožus, dažnai pasitaikančius tiek lokaliuose, tiek globaliuose sukčiavimo scenarijuose.

Trečiasis tyrimas pristato „DSmishSMS“ sistemą, kuri taip pat akcentavo pažangių metodų naudojimą sukčiavimo SMS aptikimui [22]. Tyrime buvo naudojamas hibridinis požiūris, apimantis nuorodų analizę, raktažodžių atitikmenų tikrinimą ir mašininio mokymosi modelius, kurie buvo kruopščiai pritaikyti pagal įvairius duomenų šaltinius. Sistemos veikimo pagrindas buvo duomenų rinkinys, sudarytas iš daugiau nei 5000 teisėtų ir sukčiavimo SMS. Be to, tyrime buvo atkreiptas dėmesys į piktavalių naudojamų metodų evoliuciją, o sistema buvo pritaikyta aptikti tiek paprastus, tiek sudėtingesnius sukčiavimo atvejus. Modelis, apjungęs kelias analizės sritis, pasiekė 97,93% tikslumą. Tyrimo rezultatai parodė, kad hibridinės sistemos, įtraukdamos kelių lygių analizę, yra ypač efektyvios aptinkant įvairaus pobūdžio grėsmes, ir šis metodas atskleidžia potencialą, skirtą platesniam taikymui mobiliojo ryšio saugumo srityje.

Šie tyrimai parodo skirtingų metodologijų, tokių kaip dirbtinių neuroninių tinklų, hibridinių metodų ir giluminio mokymosi, efektyvumą sukčiavimo SMS aptikime. Kiekvienas iš jų turi unikalių privalumų, priklausomai nuo konteksto ir tikslų, tačiau visi jie prisideda prie pažangesnių sprendimų kūrimo kovojant su kibernetinėmis grėsmėmis.

1.11. Esamų sprendimų metodų bei algoritmų pasirinkimo bei jų efektyvumo apžvalga

Šiuolaikiniame kibernetinio saugumo kontekste naudojami įvairūs metodai ir algoritmai, skirti sukčiavimo SMS žinučių aptikimui. Vienas iš pavyzdžių – „Smishing Detector“ modelis, kuris remiasi dirbtinių neuroninių tinklų (ANN) taikymu, klasifikuojant SMS žinutes į teisėtas ir sukčiavimo kategorijas [20]. Naudojant atgalinės sklaidos algoritmą optimizuojami svoriai, siekiant sumažinti prognozės paklaidas. Tyrime buvo pasiektas 97,4% tikslumas, lenkiant „naivųjį Bajeso“ (96,29%) ir „sprendimų medžio“ (93,4%) algoritmus. „F1“ rodiklis šioje sistemoje siekė 86,72%, preciziškumas siekė apytiksliai 84%, o jautrumas apytiksliai 93%.

„SMSProtect“ sistema įgyvendina realaus laiko sukčiavimo žinučių aptikimą, pasitelkdama nuorodų analizę, turinio analizę ir dinaminį modelių mokymą [21]. Šioje sistemoje taikyti metodai pasiekė 98,42% tikslumą, preciziškumas siekė 98,4%, jautrumas bei „F1“ rodiklis taipogi siekė 98,4%.

„DSmishSMS“ tyrime buvo akcentuotas hibridinių metodų naudojimas, apjungiant nuorodų analizę, raktažodžių tikrinimą ir mašininio mokymosi modelius [22]. Ši sistema, naudodama daugiau nei 5000 SMS žinučių duomenų rinkinį, pasiekė 97,93% tikslumą. Preciziškumas siekė 84%, o jautrumas siekė 94%. Nors „F1“ rodiklis yra nurodytas tik lentelėje, o ne tekste, galima spėti, kad jis siekė apie 91%.

Galiausiai dar vienas tyrimas nagrinėjo įvairių komercinių įrankių efektyvumą prieš šiuolaikines sukčiavimo grėsmes [23]. Šis eksperimentai parodė, kad mobiliojo tinklo operatorių sukčiavimo blokavimo lygis svyravo nuo 25% iki 35%. Aukščiausią – 35% blokavimo lygį pasiekė „T-Mobile“, o trečiųjų šalių aplikacijos, tokios kaip „Robokiller“ ir „Textkiller“, pasiekė atitinkamai 61,6% ir 53,9% sukčiavimo aptikimo lygį. Tačiau kai kurios aplikacijos, tokios kaip „Robokiller“, turėjo itin aukštą teisėtų žinučių blokavimo rodiklį, siekiantį net 100%. Tai rodo, kad daugeliui komercinių sprendimų dar reikia tobulinti algoritmus, siekiant sumažinti teisėtų žinučių klaidingą blokavimą.

3 lentelė. Esamų sprendimų efektyvumo apžvalga

Pavadinimas	**Kategorija**	**Tikslumas (accuracy)**	**Jautrumas (recall)**	**Preciziškumas (precision)**	**F1 rodiklis**
SMSProtect	Akademinis tyrimas	98,42%	98,4%	98,4%	98,4%
DSmishSMS	Akademinis tyrimas	97,93%	~94%	~84%	~91%
Smishing Detector	Akademinis tyrimas	97,4%	~93%	~84%	86,72%
Robokiller	Komercinė programėlė	61,6%	-	-	-
TextKiller	Komercinė programėlė	53,9%	-	-	-
T-Mobile	Mobiliojo ryšio operatorius	35%	-	-	-

Nors egzistuoja komerciniai sukčiavimo SMS aptikimo įrankiai ir mobiliojo tinklo operatorių teikiamos paslaugos, šių sprendimų aptikimo rodikliai yra ganėtinai žemi, o klaidingai teigiamų rezultatų skaičius yra didelis. Tai pabrėžia, kad tokie sprendimai dar nėra pakankamai išvystyti, kad galėtų efektyviai prisitaikyti prie greitai kintančių grėsmių ir užtikrinti aukštą tikslumo lygį. Todėl šioje srityje būtina pasikliauti mokslinių tyrimų rezultatais bei esančiais sprendimais, kuriuose pateikiamos pažangios metodikos ir gerosios praktikos, užtikrinančios efektyvesnį sukčiavimo SMS aptikimą ir prevenciją. Tokios sistemos kaip „SMSProtect“ ir panašios akademinių tyrimų metu sukurtos priemonės demonstruoja gerokai aukštesnį efektyvumą, kuris gali būti naudingas sprendžiant šiuolaikinius kibernetinio saugumo iššūkius.

1.12. Duomenų rinkimo strategijos

Efektyvus duomenų rinkimas yra esminis veiksnys, užtikrinantis duomenų rinkinių kokybę, patikimumą ir atitiktį saugumo reikalavimams. Šiame skyriuje analizuojamos bendruomenės pagrindo duomenų rinkimo metodikos ir aptariamos duomenų saugumo bei anonimiškumo užtikrinimo strategijos.

1.13. Bendruomenės pagrindo duomenų rinkimo metodika

Bendruomenės pagrindu grindžiama duomenų rinkimo (angl. *crowdsourcing arba community-sourced*) metodika yra vienas iš efektyviausių būdų gauti šviežius ir aktualius SMS sukčiavimo atakų pavyzdžius. Šis metodas leidžia naudotojams savanoriškai dalintis gautais įtartinais pranešimais per specialiai tam skirtą platformą. Tyrime aptariamas SMS sukčiavimo atakų pranešimų rinkimas naudojant platformą „smishtank.com“, kuri veikė kaip bendruomenės duomenų rinkimo centras [24]. Šios platformos dėka buvo surinkta 1062 SMS sukčiavimo pranešimų rinkinio pavyzdžių tyrimo publikavimo metu, tačiau šiuo metu, platformos dėka, pranešimų skaičius jau viršija 3000.

Platforma leidžia naudotojams pateikti gautus SMS sukčiavimo atakų pranešimus, kurie vėliau tampa prieinami tyrėjams ir kitiems duomenų analitikams. Šis metodas išsiskiria tuo, kad surinkti duomenys yra tiesiogiai perduodami iš naudotojų, taip užtikrinant informacijos aktualumą ir šviežumą. Tai ypač svarbu SMS sukčiavimo tyrimuose, nes šios grėsmės pobūdis sparčiai kinta, atsiranda naujų schemų, o seni duomenys tampa mažiau reikšmingi. Bendruomenės pagrindu grindžiamas duomenų rinkimas

ne tik leidžia gauti naujausius pranešimų pavyzdžius, bet ir užtikrina jų įvairovę, nes pranešimus teikia naudotojai iš skirtingų aplinkų.

Ši strategija taip pat suteikia galimybę greitai ir efektyviai auginti duomenų bazę. „Smishtank.com“, per skatinimo kampanijas socialiniuose tinkluose ir bendruomeninius kvietimus, įtraukė didelį kiekį naudotojų, kurie nuolat pateikia naujus pranešimus. Šį modelį būtų galima pritaikyti ir Lietuvoje, sukuriant analogišką platformą lietuvių kalba, kur naudotojai galėtų bendradarbiauti, dalintis SMS sukčiavimo pavyzdžiais ir prisidėti prie nacionalinio duomenų rinkimo proceso.

1.14. Duomenų saugumo ir anonimiškumo užtikrinimo strategijos

Siekiant apsaugoti tiek vartotojų pateiktą informaciją, tiek platformą nuo galimų kibernetinių grėsmių, svarbią vietą užima modernios technologijos, tokios kaip duomenų šifravimas ir patikimos vykdymo aplinkos (angl. *trusted execution environments, TEE*). Kaip pabrėžiama tyrime, nuo vieno galo iki kito galo šifravimas (angl. *end-to-end encryption, E2EE*) yra vienas veiksmingiausių būdų apsaugoti duomenis nuo neautorizuotos prieigos tiek išorės, tiek vidaus grėsmių kontekste [25].

Dar viena svarbi apsaugos sritis – naudotojo sąsajos (angl. *front-end*) saugumo stiprinimas. Remiantis tyrimu, dažnos naršyklių pažeidžiamumo formos, tokios kaip kenkėjiškos naršyklės plėtiniai, gali būti panaudotos duomenų nutekėjimui dar prieš jiems pasiekiant šifravimo mechanizmus [25]. Norint išvengti šių grėsmių, rekomenduojama naudoti saugius ir patikrintus naršyklės plėtinius.

Šios strategijos sudaro daugiapakopį duomenų apsaugos modelį, kurį būtų galima pritaikyti bet kur, tačiau reikia atkreipti dėmesį ir į patogumą bei panaudojimo atvejį. Jeigu kuriama panaši platforma į „smishtank.com“, kurios tikslas ir yra, kad visi naudotojai matytų žinutes, užtektų apsaugoti patį tinklalapį bei duomenų bazę, kad ji nebūtų ištrinta ar modifikuojama piktavalių naudai.

1.15. Analizės išvados

1. SMS sukčiavimas yra rimta grėsmė, paveikianti tiek individualius naudotojus, tiek įmones. Tyrimai atskleidė, kad sukčiavimo žinutės dažnai manipuliuoja gavėjų emocijomis, tokiomis kaip baimė ar skubumas, todėl vartotojai tampa lengvais taikiniais. Dažniausios pasekmės apima finansinius nuostolius, tapatybės vagystes ir reputacijos žalą.

2. Pažangūs mašininio mokymosi ir giluminio mokymosi modeliai, tokie kaip „LSTM“ ir „Bi-LSTM“, pasiekia aukštą tikslumą (daugiau nei 99%) atpažįstant sukčiavimo žinutes. Hibridiniai sprendimai, derinantys turinio analizę, nuorodų tikrinimą ir šaltinio kodo analizę, leidžia aptikti sudėtingesnius sukčiavimo atvejus, tačiau jų įgyvendinimas reikalauja didelių skaičiavimo išteklių.

3. Lietuvių kalbai specifinės gramatinės savybės ir žodžių formų įvairovė apsunkina pasauliniu mastu pritaikomų sukčiavimo aptikimo metodų veikimą. Šiuo metu Lietuvoje nėra tinkamų duomenų rinkinių ar sistemų, skirtų SMS sukčiavimo atpažinimui, todėl aktualu kurti naujus duomenų rinkinius ir metodus, pritaikytus vietiniam kontekstui.

4. Bendruomenės platformos, tokios kaip „Smishtank.com“, leidžia efektyviai rinkti naujausius sukčiavimo žinučių pavyzdžius. Tokios iniciatyvos gali būti pritaikytos ir Lietuvoje, užtikrinant, kad naudotojų pateikta informacija būtų saugiai saugoma ir anonimiška.

5. Be technologinių sprendimų, svarbu ugdyti vartotojų sąmoningumą apie sukčiavimo metodus. Edukacinės iniciatyvos, tokios kaip mokymai ir informacinės kampanijos, gali sumažinti naudotojų pažeidžiamumą ir skatinti atsargų elgesį su įtartinomis žinutėmis.

6. Nors daugelis akademinių sprendimų rodo puikius rezultatus laboratorinėmis sąlygomis, jų pritaikymas realioje aplinkoje dažnai susiduria su iššūkiais, tokiais kaip klaidingi teigiami rezultatai ar sukčiavimo modelių kaita. Todėl svarbu užtikrinti nuolatinį sistemų atnaujinimą ir modelių adaptaciją.

2. Sukčiavimo SMS žinutėmis aptikimo metodo kūrimo projektas

Ši dalis pristato sukčiavimo SMS žinučių aptikimo metodo projektavimą, aptariant sukurtą koncepcinę struktūrą, taikytinus modelius aptikti SMS sukčiavimo žinutes.

2.1. Projekto koncepcija

Šiame skyriuje pateikiama kuriamos hibridinės analizės sistemos, skirtos sukčiavimo SMS žinučių aptikimui Lietuvoje, koncepcija (žr. **3 pav.**). Pagrindinis šio metodo tikslas – efektyviai identifikuoti sukčiavimo pobūdžio SMS žinutes, derinant taisyklėmis grįstą analizę ir mašininio mokymosi modelį. Toks sprendimas leidžia sujungti interpretuojamą taisyklių logiką su duomenimis grįstu modelio gebėjimu atpažinti tekstinius dėsningumus.

Gavus SMS žinutę, sistema inicijuoja du lygiagrečius duomenų apdorojimo procesus. Pirmasis skirtas taisyklėmis grįstam sprendimų priėmimui, antrasis – mašininio mokymosi modeliui. Abu šie srautai veikia nepriklausomai vienas nuo kito ir analizuoja tą pačią žinutę skirtingais metodais.

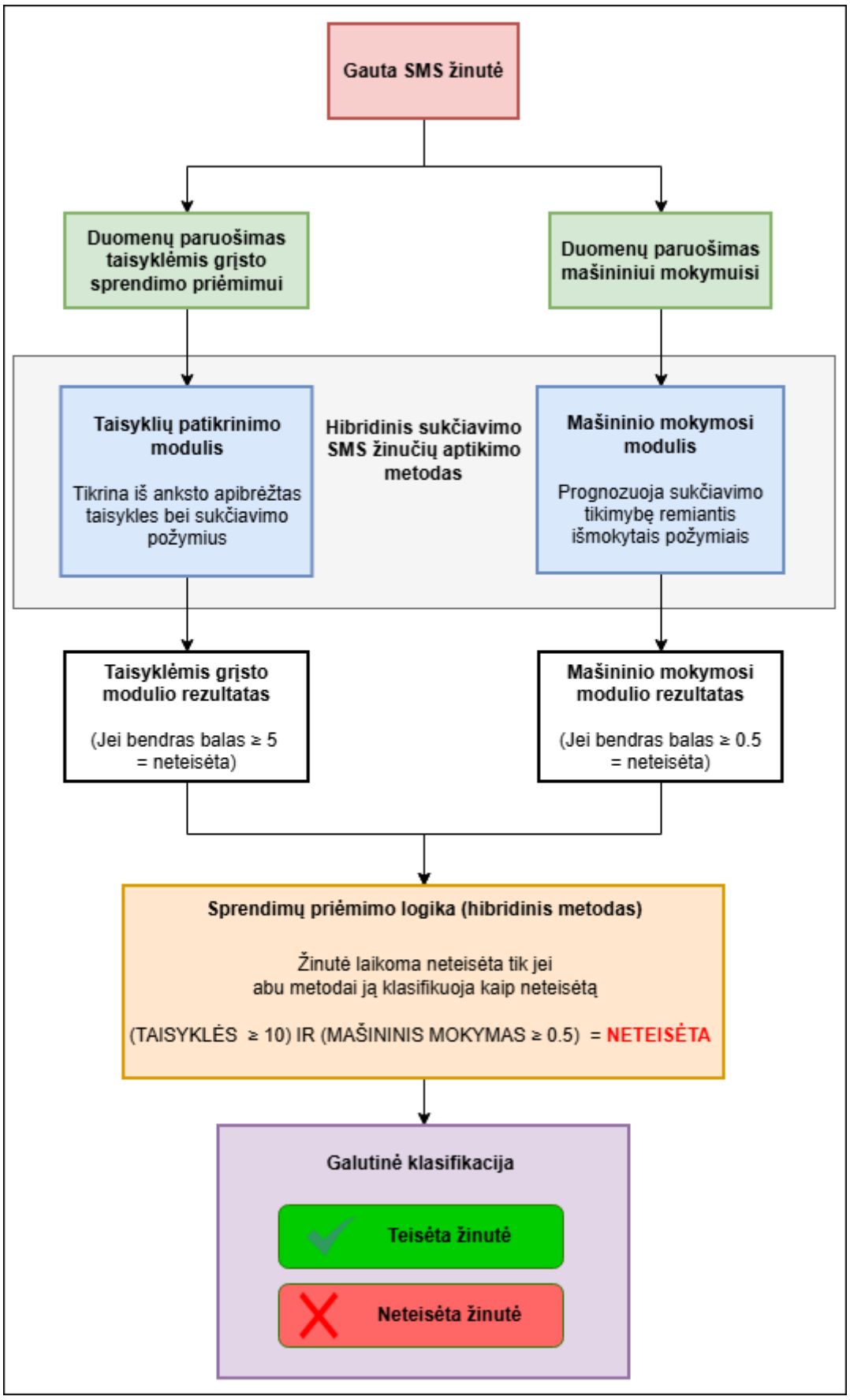
Taisyklėmis grįstas modulis tikrina žinutę pagal iš anksto apibrėžtus sukčiavimo požymius. Analizuojami tokie elementai kaip nuorodų buvimas, įtartinos domenų galūnės, sutrumpintos nuorodos, telefono numerio struktūra bei sukčiavimo žinutėms būdingi raktažodžiai ar frazės. Remiantis šiais kriterijais, modulis pateikia dvejetainį sprendimą – žinutė laikoma teisėta arba neteisėta.

Tuo pačiu metu mašininio mokymosi modulis analizuoja žinutės tekstą, naudodamas natūralios kalbos apdorojimo metodus. Tekstas paverčiamas skaitine reprezentacija (naudojant TF-IDF), o tuomet klasifikuojamas logistinės regresijos modeliu, kuris įvertina tikimybę, ar žinutė yra sukčiavimo pobūdžio. Literatūroje taip pat plačiai taikomi ir kiti metodai, tokie kaip BiLSTM [26] arba BERT [27], kurie naudojami teksto klasifikavimo užduotyse. Tačiau šiame darbe pasirinktas logistinės regresijos modelis su TF-IDF reprezentacija, nes toks derinys leidžia efektyviai apdoroti trumpus tekstus ir užtikrina stabilų veikimą nagrinėjamame duomenų rinkinyje. Pagal nustatytą klasifikavimo slenkstį modelis pateikia galutinį dvejetainį sprendimą.

Skirtingai nei kai kuriuose kituose sprendimuose, šiame metode nėra naudojamas balų jungimas ar kelių lygių klasifikacija. Vietoje to taikoma aiški sprendimo priėmimo logika: žinutė priskiriama neteisėtų žinučių klasei tik tuo atveju, jei abu metodai – tiek taisyklėmis grįstas, tiek mašininio mokymosi – ją identifikuoja kaip neteisėtą. Jei bent vienas iš metodų žinutę klasifikuoja kaip teisėtą, galutinis sprendimas laikomas įtartinu.

Tokia sprendimo priėmimo strategija leidžia sumažinti klaidingai teigiamų klasifikacijų skaičių, t. y. atvejus, kai teisėtos žinutės klaidingai pažymimos kaip sukčiavimo. Tai yra ypač svarbu praktinėse sistemose, kuriose siekiama išvengti nepagrįsto teisėtų žinučių blokavimo.

Galutiniame etape sistema pateikia dvejetainę klasifikaciją: žinutė pažymima kaip teisėta arba neteisėta (sukčiavimo pobūdžio). Tokia architektūra leidžia sukurti aiškų ir efektyvų sprendimą, pritaikytą realioms SMS žinučių analizės sąlygoms.



3 pav. Sukčiavimo SMS žinutėmis aptikimo metodo koncepcija

2.2. SMS sukčiavimo žinutėmis aptikimo metodo veikimo etapai

Toliau šiame skyriuje, pateiktuose 4–10 paveikslėliuose, bus išsamiau aprašomas sukčiavimo SMS žinučių atpažinimo metodo veikimas bei pagrindiniai vykstantys proceso etapai. Bus nuosekliai paaiškinta, kaip sistemoje atliekamas duomenų surinkimas, apdorojimas ir rizikos vertinimas.

2.2.1. Bendras sukčiavimo SMS žinučių atpažinimo metodas

4 paveiksle pateikiama bendra sukčiavimo SMS žinučių atpažinimo metodo veiklos diagrama. Šis procesas susideda iš kelių pagrindinių etapų, nuo naudotojo veiksmo iki galutinio rizikos įvertinimo ir atsakymo gavimo.

Naudotojas pradeda sąveiką atidarydamas sistemos puslapį, kur jam pateikiama galimybė įkelti gautą įtartiną SMS žinutę - tai turi būti žinutės tekstas bei telefono numeris, iš kurio žinutė buvo gauta. Kai naudotojas paspaudžia pateikimo mygtuką, žinutės duomenys perduodami į sistemą.

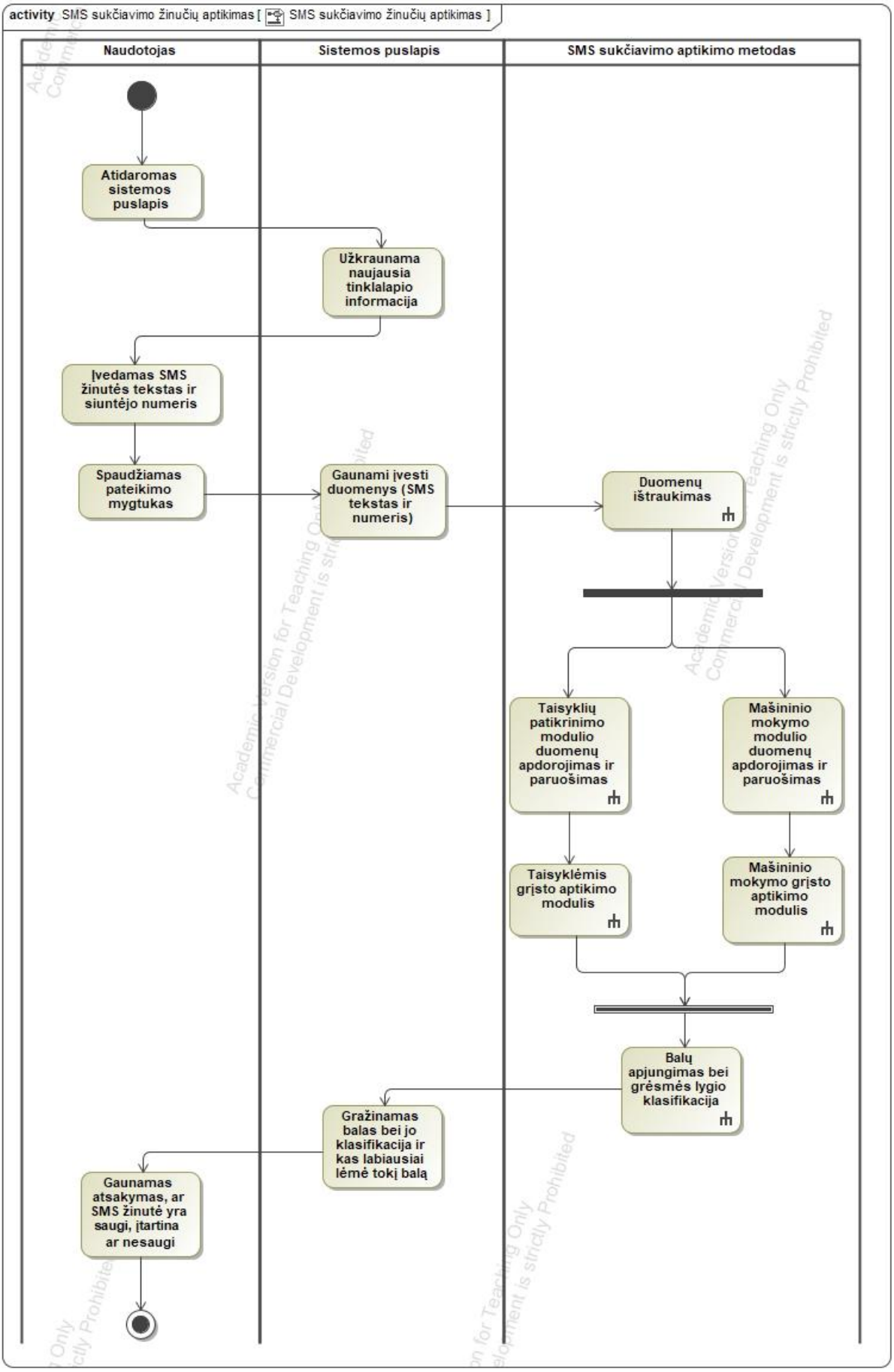
Sistemos puslapis gauna naudotojo pateiktą informaciją ir pradeda pradinį duomenų ištraukimą.

Po duomenų ištraukimo vyksta sprendimas, kokiam analizės etapui bus ruošiami duomenys: taisyklių patikrinimo moduliui arba mašininio mokymosi moduliui. Šiame žingsnyje vyksta specifinis duomenų apdorojimas kiekvienam iš metodų – duomenys ruošiami pagal atitinkamas savybes ir formatus, reikalingus kiekvienam analizės būdui.

Taisyklių patikrinimo modulyje duomenys analizuojami remiantis iš anksto apibrėžtomis taisyklėmis, tokiomis kaip domenų, TLD (angl. *Top-Level Domain*), numerių ir žodžių šablonų tikrinimas. Tuo tarpu mašininio mokymosi modulyje atliekamas tekstų vektorizavimas, sekų analizė, diakritikų nustatymas ir kitos dalys, kurios vėliau bus išsamiau padetalizuotos.

Abi analizės kryptys savarankiškai pateikia rizikos vertinimus – taisyklių analizės balą ir mašininio mokymosi analizės balą. Šie rezultatai sujungiami balų apjungimo fazėje, kur taikomas svertinis balų jungimo algoritmas, klaidų kontrolė (false positive/false negative) ir adaptuojami slenksčiai, leidžiantys sistemiškai prisitaikyti prie situacijos.

Galiausiai naudotojui grąžinamas rezultatas, kuriame nurodoma, ar SMS žinutė yra saugi, įtartina ar akivaizdus sukčiavimo atvejis. Kartu pateikiamas galutinis balas bei trumpas paaiškinimas, kokie faktoriai labiausiai lėmė tokį klasifikacijos rezultatą (pvz., įtartinas domenas, grėsmingi žodžiai, siuntėjo neatitikimai).



4 pav. Bendra sukčiavimo SMS žinučių atpažinimo metodo veiklos diagrama

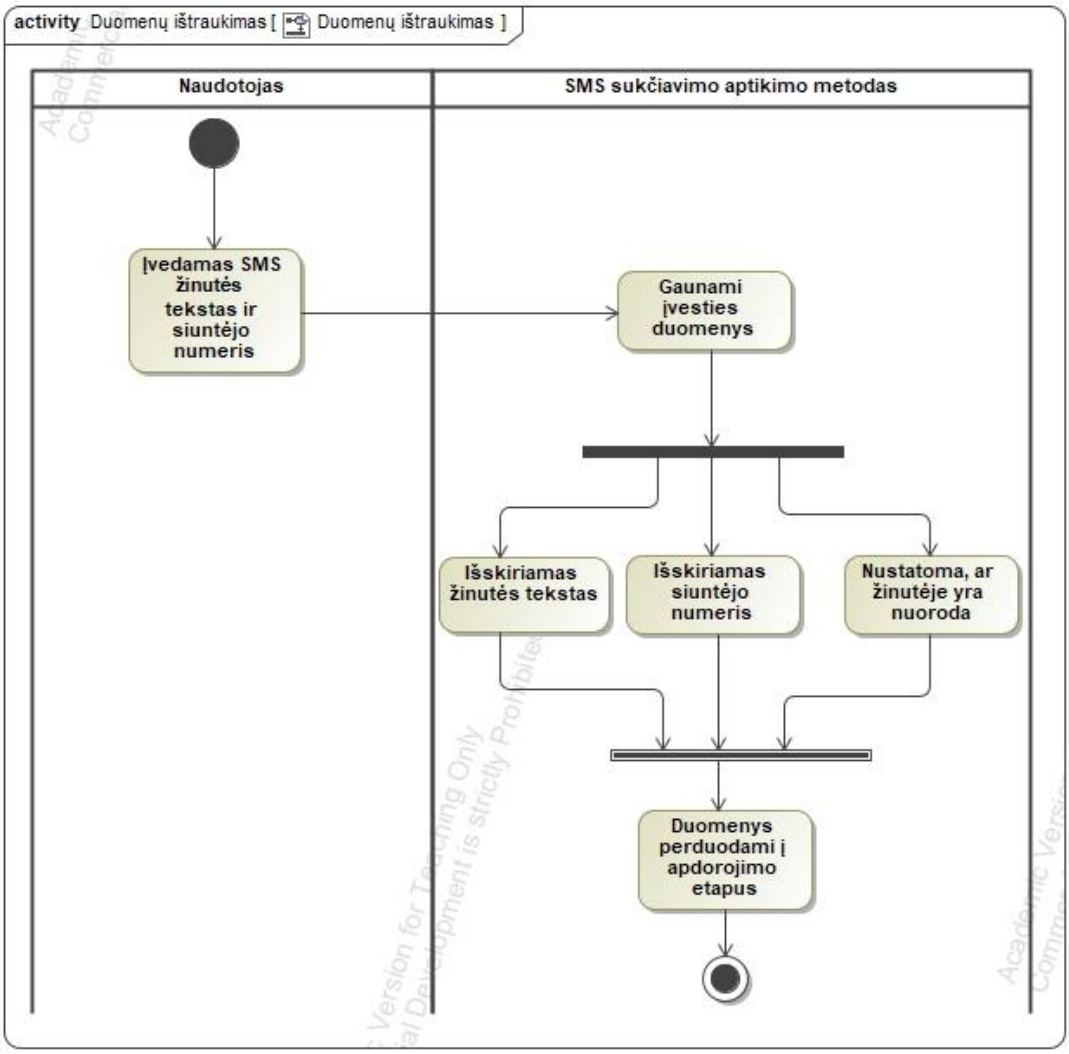
2.2.2. Duomenų ištraukimas

5 paveiksle pateikiama duomenų ištraukimo veiklos diagrama, kuri yra vienas iš pradinių etapų kuriamoje SMS sukčiavimo žinučių analizės sistemoje. Šiame etape sistema priima naudotojo pateiktus duomenis – SMS žinutės tekstą bei siuntėjo telefono numerį, kurie vėliau naudojami tolimesnei analizei.

Gavus įvesties duomenis, atliekamas pradinis jų apdorojimas, kurio metu informacija paruošiama tolimesniam nagrinėjimui. Šiame etape užtikrinamas duomenų vientisumas ir tinkamumas analizei, pašalinant nereikalingus simbolius ar atliekant kitus paruošimo veiksmus.

Toliau vykdomas duomenų išskaidymas į atskiras sudedamąsias dalis. Iš pateiktos informacijos išskiriamas žinutės tekstas, siuntėjo telefono numeris, taip pat nustatoma, ar žinutėje yra nuorodų. Jei nuorodos aptinkamos, jos papildomai identifikuojamos ir paruošiamos tolimesniam vertinimui. Toks duomenų išskaidymas leidžia efektyviau analizuoti skirtingus žinutės aspektus ir pritaikyti jiems atitinkamus analizės metodus.

Po atskirų duomenų elementų išskyrimo jie sujungiami į struktūrizuotą formą ir perduodami tolimesniems analizės etapams. Paruošti duomenys naudojami tiek taisyklių pagrindu veikiančiame modulyje, tiek mašininio mokymosi modelyje. Tokiu būdu užtikrinamas nuoseklus ir vieningas duomenų srautas visoje sistemoje bei sudaromos sąlygos efektyviam sukčiavimo žinučių aptikimui.



5 pav. Duomenų ištraukimo veiklos diagrama

2.2.3. Taisyklių patikrinimo modulio duomenų apdorojimas ir paruošimas

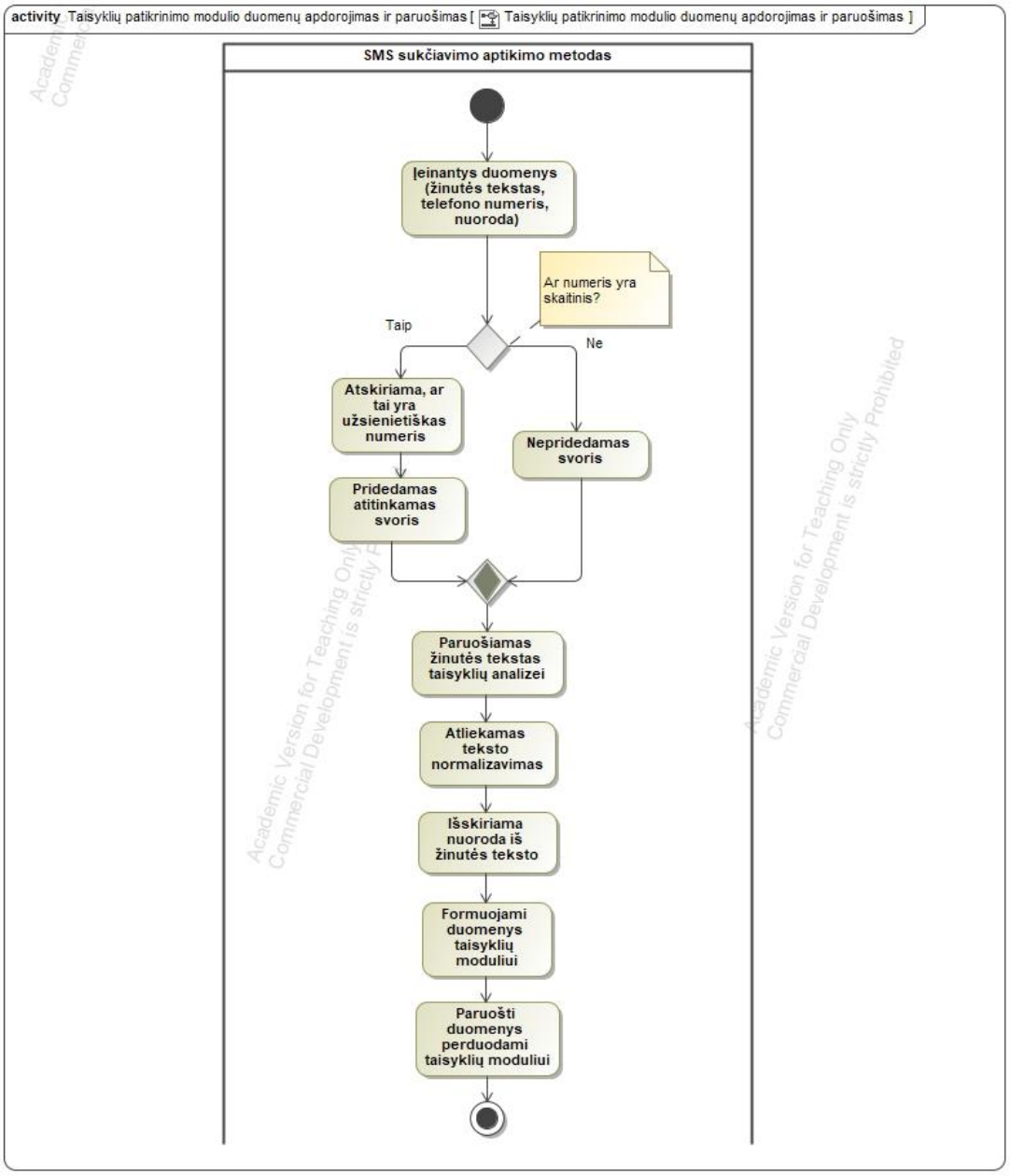
6 paveiksle pavaizduotoje veiklos diagramoje rodomas SMS žinutės duomenų apdorojimo ir paruošimo etapas prieš perduodant informaciją taisyklėmis grįstam analizės moduliui. Šiame etape atliekamas pradinis duomenų struktūravimas ir transformavimas, siekiant užtikrinti, kad tolimesnis taisyklių taikymas būtų nuoseklus ir patikimas.

Procesas prasideda nuo įeinančių duomenų gavimo – tai žinutės tekstas, siuntėjo telefono numeris bei galimai žinutėje esanti nuoroda. Pirmiausia analizuojamas telefono numeris, tikrinant, ar jis yra skaitmeninio tipo. Jei numeris yra skaitmeninis, papildomai nustatoma, ar jis priklauso užsienio tinklui (pvz., pagal prefiksą). Tokiu atveju gali būti pridedamas papildomas svoris, atspindintis didesnę riziką. Jei numeris nėra skaitmeninis (pvz., vardinis siuntėjas), papildomas svoris nėra pridedamas.

Toliau vykdomas žinutės teksto paruošimas taisyklių analizei. Šiame etape tekstas perduodamas į tolimesnius apdorojimo žingsnius, kuriuose atliekamas jo normalizavimas. Teksto normalizavimas apima teksto formos suvienodinimą, pavyzdžiui, didžiųjų ir mažųjų raidžių ignoravimą bei nereikšmingų simbolių įtakos sumažinimą, kad būtų galima patikimai taikyti taisykles nepriklausomai nuo teksto pateikimo formos.

Po normalizavimo iš žinutės teksto išskiriama nuoroda, jei tokia yra. Šiame žingsnyje identifikuojamos visos tekstinės nuorodos (pvz., prasidedančios „http://“ arba „https://“), kurios vėliau naudojamos domeno ir kitų požymių analizei.

Galiausiai suformuojamas struktūrizuotas duomenų rinkinys, skirtas taisyklių moduliui. Šis rinkinys apima paruoštą žinutės tekstą, siuntėjo numerį bei išskirtą nuorodą. Paruošti duomenys perduodami taisyklių patikrinimo moduliui, kuriame jau atliekamas konkrečių sukčiavimo požymių tikrinimas ir rizikos vertinimas.



6 pav. Taisyklių patikrinimo modulio duomenų apdorojimo ir paruošimo veiklos diagrama

2.2.4. Mašininio mokymo modulio duomenų apdorojimas ir paruošimas

7 paveiksle pateiktoje veiklos diagramoje pavaizduotas mašininio mokymo modulio duomenų apdorojimo ir paruošimo procesas. Šio etapo tikslas – paruošti įeinančius duomenis taip, kad jie galėtų būti tiesiogiai panaudoti mašininio mokymosi modelio prognozei gauti.

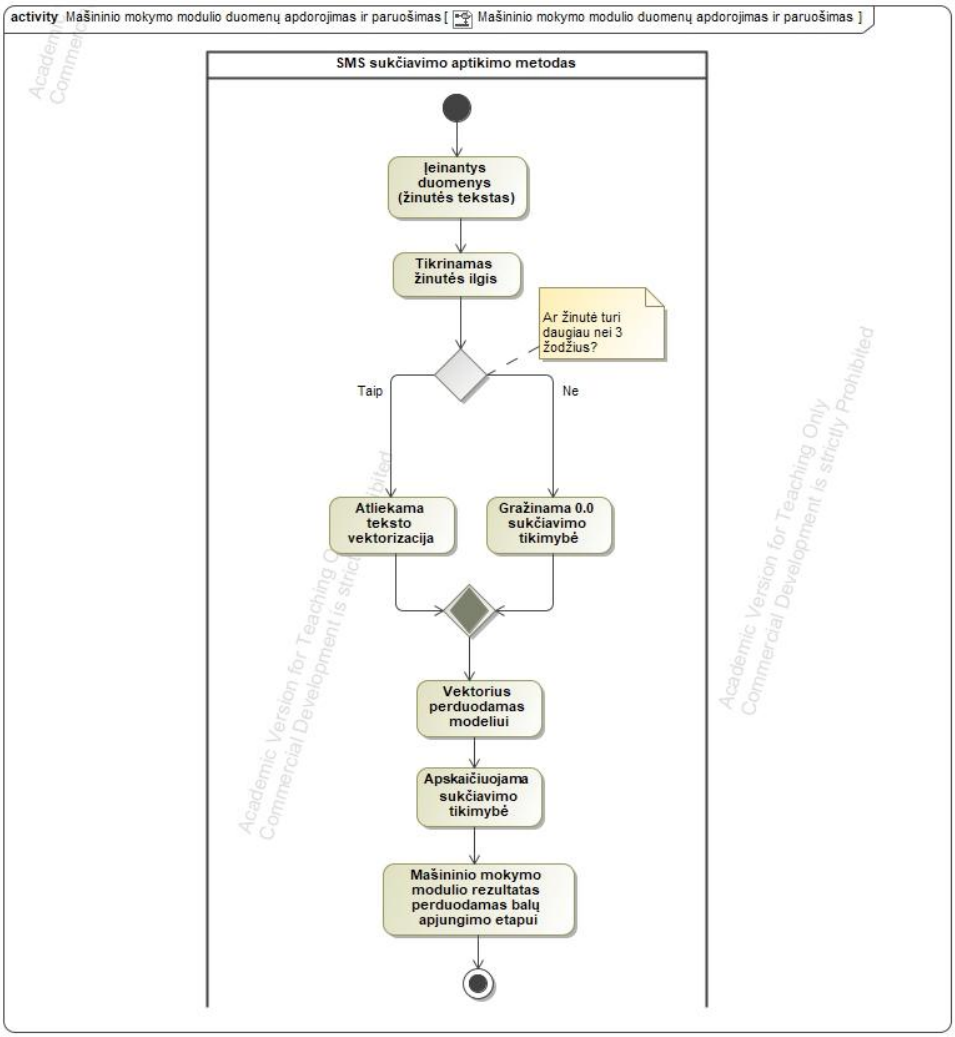
Procesas prasideda nuo įeinančių duomenų – analizuojamos SMS žinutės teksto. Skirtingai nei taisyklėmis grįstame modulyje, šiame etape papildoma informacija, tokia kaip siuntėjo numeris ar nuorodos, nėra tiesiogiai naudojama – pagrindinis dėmesys skiriamas pačiam žinutės turiniui.

Pirmiausia atliekamas žinutės ilgio patikrinimas. Nustatoma, ar žinutė sudaryta iš daugiau nei trijų žodžių. Šis žingsnis yra svarbus siekiant išvengti netikslių prognozių, nes labai trumpos žinutės neturi pakankamai informacijos patikimai klasifikacijai. Jei žinutė yra per trumpa, jai iš karto priskiriama 0,0 sukčiavimo tikimybė ir tolimesnis apdorojimas nevykdomas.

Jeigu žinutė atitinka minimalaus ilgio reikalavimą, ji perduodama vektorizacijos etapui. Šiame žingsnyje tekstas transformuojamas į skaitinę formą naudojant iš anksto apmokytą vektorizavimo modelį. Vektorizacija leidžia paversti tekstą į požymių rinkinį, kurį gali interpretuoti mašininio mokymosi algoritmas.

Toliau gautas vektorius perduodamas klasifikavimo modeliui, kuris apskaičiuoja tikimybę, kad žinutė yra sukčiavimo pobūdžio. Modelis grąžina reikšmę intervale nuo 0 iki 1, kuri atspindi sukčiavimo tikimybę.

Galiausiai mašininio mokymo modulio rezultatas perduodamas balų apjungimo etapui, kuriame jis bus integruotas su taisyklėmis grįsto modulio rezultatais ir panaudotas galutiniam sprendimui priimti.



7 pav. Mašininio mokymo modulio duomenų apdorojimo ir paruošimo veiklos diagrama

2.2.5. Taisyklėmis grįsto aptikimo modulis

8 paveiksle pateiktoje veiklos diagramoje pavaizduotas taisyklėmis grįsto aptikimo modulio veikimas. Šis modulis yra viena iš hibridinės SMS sukčiavimo žinučių analizės sistemos dalių ir remiasi iš anksto apibrėžtomis taisyklėmis bei požymiais, leidžiančiais įvertinti žinutės patikimumą.

Procesas prasideda nuo duomenų, gautų iš ankstesnio paruošimo etapo. Į modulį perduodamas žinutės tekstas, siuntėjo numeris arba vardinis identifikatorius bei, jei yra, žinutėje aptikta nuoroda. Taip pat gaunama informacija apie tai, ar siuntėjas yra skaitmeninis numeris ir ar jis gali būti užsienio kilmės.

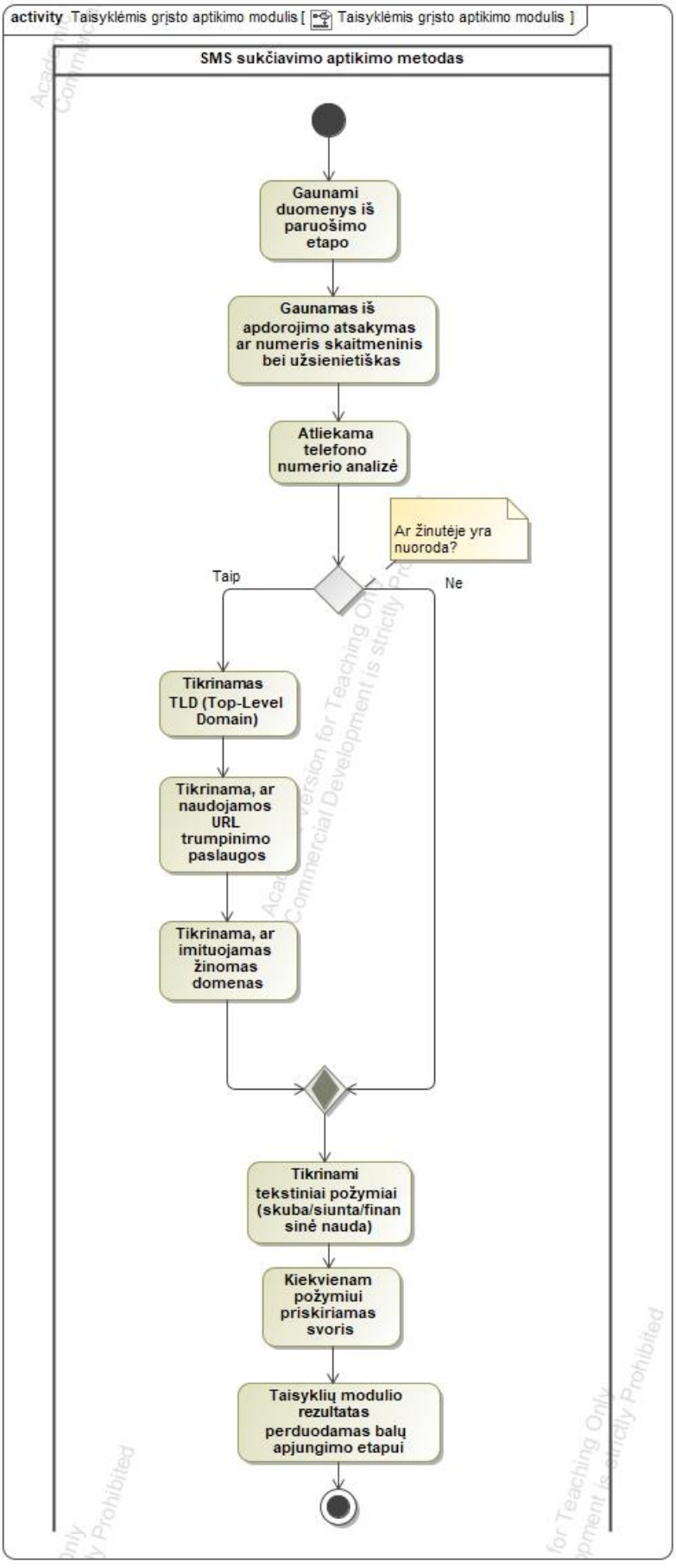
Toliau atliekama telefono numerio analizė. Šiame žingsnyje įvertinama, ar numeris yra įprasto formato, ar gali būti laikomas įtartinu (pvz., užsienietiškas numeris). Ši informacija vėliau naudojama kaip vienas iš požymių vertinant žinutę.

Po to tikrinama, ar žinutėje yra nuoroda. Jei nuoroda aptinkama, atliekama papildoma jos analizė. Pirmiausia tikrinama domeno galūnė (angl. *Top-Level Domain*), siekiant nustatyti, ar naudojamas įtartinas domeno plėtinys (pvz., .xyz, .top). Taip pat tikrinama, ar nuoroda nėra sutrumpinta naudojant URL trumpinimo paslaugas (pvz., bit.ly, tinyurl), nes tokios nuorodos dažnai naudojamos sukčiavimo atvejais siekiant paslėpti tikrąjį adresą. Be to, vertinama, ar domenas neimituoja žinomų organizacijų pavadinimų (pvz., bankų ar pristatymo tarnybų), kas yra dažnas sukčiavimo metodas.

Toliau vykdoma tekstinių požymių analizė. Tikrinama, ar žinutėje yra žodžių ar frazių, susijusių su skuba, finansine nauda ar siuntų pristatymu (pvz., „skubiai“, „atlygis“, „siunta“). Tokie raktažodžiai yra dažnai naudojami siekiant sukelti vartotojo emocinę reakciją ir paskatinti veiksmą.

Kiekvienam aptiktam požymiui priskiriamas tam tikras svoris. Šie svoriai atspindi požymio svarbą vertinant žinutės patikimumą – kuo požymis labiau būdingas sukčiavimui, tuo didesnį svorį jis turi.

Galutiniame etape taisyklių modulio rezultatas (aptikti požymiai ir jų svoriai) perduodamas balų apjungimo etapui, kuriame jis bus sujungtas su mašininio mokymosi modulio rezultatais ir panaudotas galutinei žinutės klasifikacijai.



8 pav. Taisyklėmis grįsto aptikimo modulio veiklos diagrama

2.2.6. Mašininio mokymo grįsto aptikimo modulis

9 paveiksle pavaizduotas mašininio mokymosi grįsto aptikimo modulio veikimo procesas. Šis modulis yra viena iš pagrindinių hibridinės sistemos dalių, atsakinga už automatizuotą SMS žinučių klasifikavimą, remiantis statistiniais ir tekstiniais požymiais.

Procesas prasideda nuo paruoštų SMS duomenų gavimo iš ankstesnio etapo. Šie duomenys apima išgrynintą žinutės tekstą, kuris jau yra tinkamas tolesniam apdorojimui mašininio mokymosi metodais.

Toliau duomenys padalijami į mokymo ir testavimo aibes. Toks padalijimas leidžia modelį apmokyti naudojant vieną duomenų dalį, o jo veikimą objektyviai įvertinti su kita, anksčiau nematyta duomenų aibe.

Kitame etape atliekama teksto vektorizacija taikant TF-IDF metodą. Šio proceso metu žinutės tekstas paverčiamas į skaitinę reprezentaciją, kuri atspindi simbolių sekų dažnius ir jų svarbą visame duomenų rinkinyje. Tai leidžia modeliuoti tekstinius duomenis kaip skaitinius požymius.

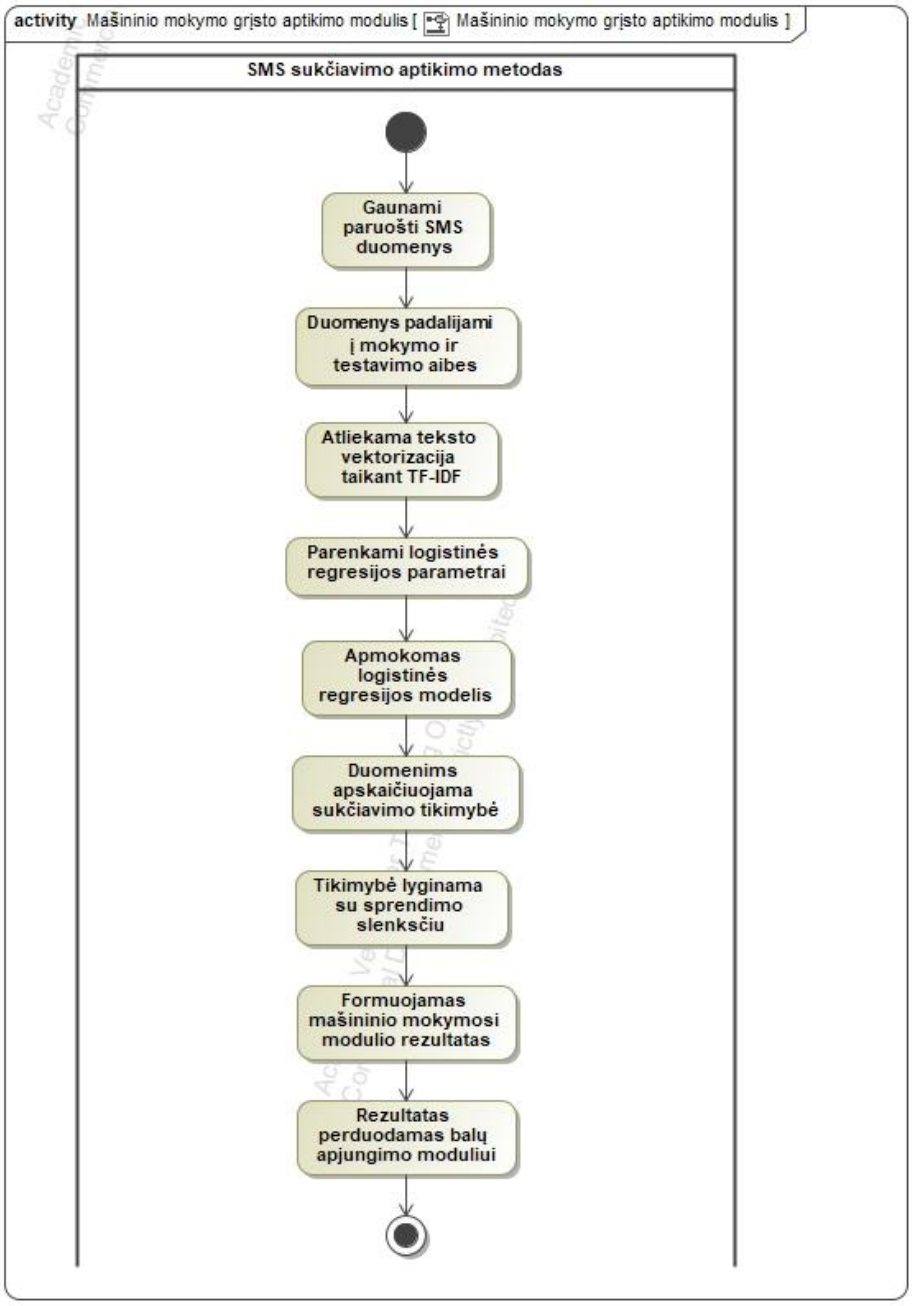
Po to vykdomas logistinės regresijos modelio parametrų parinkimas. Parametrai parenkami naudojant kryžminę validaciją, siekiant rasti optimalų modelio konfigūracijos variantą, kuris užtikrintų geriausią klasifikavimo kokybę.

Parinkus parametrus, apmokomas logistinės regresijos modelis, kuris išmoksta atskirti teisėtas ir sukčiavimo žinutes pagal jų tekstinius požymius.

Apmokytam modeliui pateikus duomenis, apskaičiuojama kiekvienos žinutės priklausymo sukčiavimo klasei tikimybė. Ši reikšmė yra intervale nuo 0 iki 1, kur didesnės reikšmės rodo didesnę tikimybę, kad žinutė yra sukčiavimo pobūdžio.

Toliau gauta tikimybė lyginama su iš anksto nustatytu sprendimo slenksčiu. Jei tikimybė viršija šį slenkstį, žinutė laikoma įtartina arba sukčiavimo, priešingu atveju – teisėta.

Galiausiai suformuojamas mašininio mokymosi modulio rezultatas, kuris perduodamas balų apjungimo moduliui. Šiame etape jis bus sujungtas su taisyklėmis grįsto modulio rezultatais, siekiant priimti galutinį sprendimą dėl žinutės patikimumo.



9 pav. Mašininio mokymo grįsto aptikimo modulio veiklos diagrama

2.2.7. Balų apjungimo bei grėsmės lygio klasifikacija

10 paveiksle pavaizduotas galutinis SMS žinutės vertinimo etapas – balų apjungimas bei grėsmės lygio klasifikacija. Šiame etape sujungiami taisyklėmis grįsto aptikimo modulio ir mašininio mokymosi modulio rezultatai, siekiant priimti galutinį sprendimą dėl žinutės patikimumo.

Procesas prasideda nuo abiejų modulių rezultatų gavimo. Taisyklių modulis pateikia sprendimą, pagrįstą iš anksto apibrėžtomis heuristinėmis taisyklėmis, o mašininio mokymosi modulis – prognozę, gautą iš apmokyto klasifikavimo modelio. Abu rezultatai jau yra įvertinti pagal nustatytus slenksčius ankstesniuose etapuose, todėl šiame žingsnyje naudojami kaip dvejetainiai sprendimai (teisėta arba neteisėta).

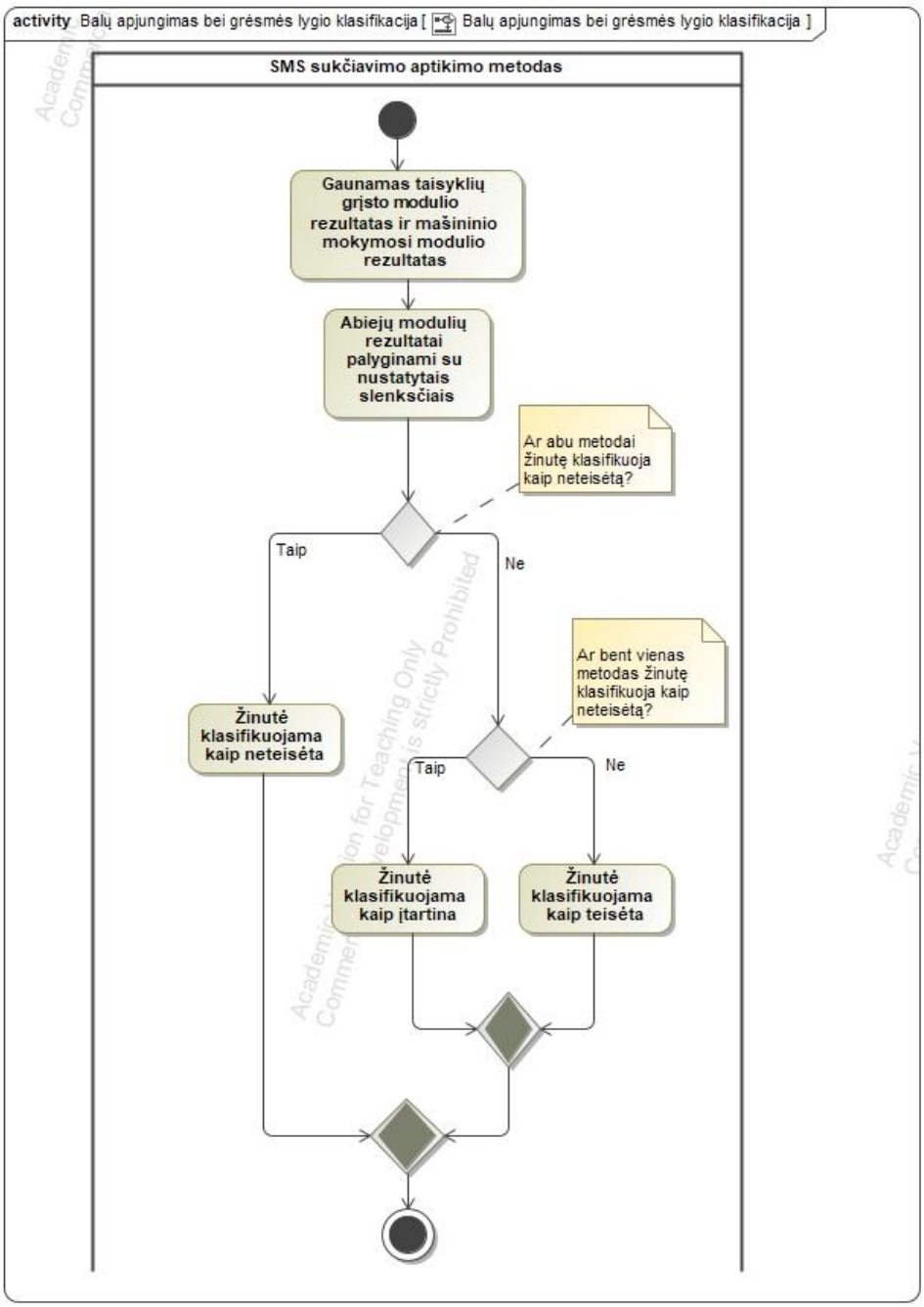
Toliau atliekamas abiejų metodų rezultatų palyginimas. Pirmiausia tikrinama, ar abu metodai žinutę klasifikuoja kaip neteisėtą. Jei taip, žinutė laikoma aukštos rizikos ir priskiriama „neteisėtos“ kategorijai.

Jeigu abu metodai nesutampa, tikrinama, ar bent vienas iš jų žinutę identifikuoja kaip neteisėtą. Tokiu atveju žinutė klasifikuojama kaip „įtartina“, nes egzistuoja bent vienas sukčiavimo požymis, tačiau nėra pakankamai pagrindo ją laikyti visiškai nesaugia.

Jei nei vienas metodas neaptinka sukčiavimo požymių, žinutė laikoma saugia ir priskiriama „teisėtos“ kategorijai.

Toks sprendimo priėmimo principas leidžia efektyviai sujungti dviejų skirtingų metodų privalumus: taisyklėmis grįstas metodas patikimai aptinka aiškius sukčiavimo požymius, o mašininio mokymosi modelis geba identifikuoti sudėtingesnius ar mažiau akivaizdžius atvejus.

Galiausiai, remiantis nustatyta klasifikacija, formuojamas naudotojui pateikiamas atsakymas. Jame nurodomas žinutės patikimumo lygis bei, esant poreikiui, pateikiamas paaiškinimas, kokie veiksniai turėjo įtakos galutiniam sprendimui.



10 pav. Balų apjungimo bei grėsmės lygio klasifikacijos veiklos diagrama

2.3. Grėsmės lygio klasifikacija

Sukčiavimo SMS aptikimo sprendimas šiame darbe grindžiamas dviem tarpusavyje papildančiais analizės komponentais: taisyklėmis grįstu grėsmės balu ir giluminio mokymosi modelio klasifikacija. Abu metodai realizuoti nepriklausomai, tačiau jų rezultatai sujungiami, siekiant padidinti sprendimo patikimumą ir sumažinti klaidingų teigiamų arba neigiamų klasifikacijų tikimybę. Šis derinys leidžia sudaryti hibridinę klasifikavimo sistemą, kurioje aiškiai išreikštas tiek loginis, tiek statistinis sprendimo pagrindas.

2.3.1. Taisyklėmis grįsto metodo klasifikavimo logika

Kaip aprašyta ankstesniuose skyriuose, kiekvienai žinutei taikoma 10 apibrėžtų požymių (taisyklių), apimančių struktūrinius, semantinius bei stilistinius aspektus. Jei žinutė atitinka konkrečią taisyklę, prie bendro grėsmės balo (GB) pridedamas tos taisyklės svoris (nuo 1 iki 5), nustatytas pagal požymio pasikartojimo dažnį apgaulingų žinučių rinkinyje.

Bendras grėsmės balas apskaičiuojamas taip:

$$Grėsmės\ balas\ (GB) = \sum_{i=1}^{n} W_i \tag{5}$$

čia:

- GB – bendras žinutės grėsmės balas;
- W_i – i-tojo aktyvaus požymio svoris (nuo 1 iki 5);
- n – visų aptiktų požymių skaičius žinutėje.

Pvz., jeigu žinutėje aptinkami 4 požymiai, kurių svoriai: 5, 3, 2 ir 1, tuomet:

$$Grėsmės\ balas\ (GB) = 5 + 3 + 2 + 1 = 11$$

Pagal eksperimentiškai nustatytus duomenis, optimalus slenkstinis balas (S) bus parinktas sekančiuose skyriuose:

$$S = Slenkstinis\ Balas$$

Tada taikoma ši klasifikavimo taisyklė:

$$\begin{cases} GB \geq S = Neteisėta\ žinutė, \\ GB < S = Teisėta\ žinutė \end{cases} \tag{6}$$

čia:

- GB – grėsmės balas;
- S – slenkstinis balas.

Šis metodas pasižymi aiškiu interpretavimo pranašumu - galima tiksliai nustatyti, kokie požymiai lėmė klasifikavimo rezultatą.

2.3.2. Mašininio mokymosi modelio sprendimas

Klasikinis mašininio mokymo modelis, paremtas TF-IDF (angl. *Term Frequency–Inverse Document Frequency*) vektorizacija ir logistinės regresijos klasifikatoriumi, kiekvienai analizuojamai SMS žinutei pateikia tikimybės įvertinimą P(žinutė), nusakantį, kokiu laipsniu pateikta žinutė priklauso neteisėtos („spam“) klasės kategorijai:

$$P(z) \in [0,1] \tag{7}$$

čia:

- $P(z)$ – tikimybė, kad žinutė z priklauso neteisėtų žinučių klasei.

Klasifikavimo sprendimas bus priimamas pagal slenkstį, kuris bus nustatytas sekančiuose skyriuose:

$$\begin{cases} P(z) \geq S = Neteisėta\ žinutė, \\ P(z) < S = Teisėta\ žinutė \end{cases} \quad (8)$$

čia:

- $P(z)$ – modelio apskaičiuota tikimybė;
- S – klasifikavimo slenkstis.

2.3.3. Hibridinis vertinimo modelis

Siekiant padidinti klasifikavimo patikimumą, abu metodai gali būti sugretinami hibridinėje sprendimų logikoje, kai klasifikacija laikoma teigiama tik abiem algoritmams sutapus:

$$(GB \geq S) \cap (P(z) \geq S) = Neteisėta\ žinutė \quad (9)$$

čia:

- GB – taisyklių metodo apskaičiuotas grėsmės balas;
- $P(z)$ – mašininio mokymosi modelio apskaičiuota tikimybė;
- S – klasifikavimo slenkstis.

Toks metodas sumažina klaidingų teigiamų klasifikacijų (angl. *false positives*) tikimybę, nes žinutė turi atitikti ir struktūrinius (taisyklių), ir semantinius (modelio) kriterijus. Nors šis požiūris gali šiek tiek sumažinti jautrumą, jis ypač tinkamas situacijose, kai prioritetas teikiamas tikslumui ir patikimumui, pavyzdžiui, automatizuotoje SMS filtravimo ar finansinių paslaugų aplinkoje.

2.4. Apibendrinimas

Apibendrinant šį projektinį skyrių, galima teigti, kad buvo suprojektuota išsami, hibridine analize paremta SMS sukčiavimo žinučių atpažinimo sistema, kuri apima tiek taisyklėmis grįstą, tiek mašininio mokymosi pagrindu veikiančią analizę. Išanalizuoti visi pagrindiniai proceso etapai – nuo naudotojo įvesties iki duomenų ištraukimo, apdorojimo, analizės, rizikos įvertinimo ir grėsmės klasifikacijos. Sistemos architektūra sukurta taip, kad užtikrintų tiek apdorojimo tikslumą, tiek lankstumą besikeičiančių grėsmių akivaizdoje. Detalios veiklos diagramos padeda aiškiai suprasti, kaip skirtingi komponentai bendradarbiauja, kad būtų pasiektas pagrindinis tikslas – laiku ir tiksliai identifikuoti galimą SMS sukčiavimą bei informuoti naudotoją suprantamu ir skaidriu būdu.

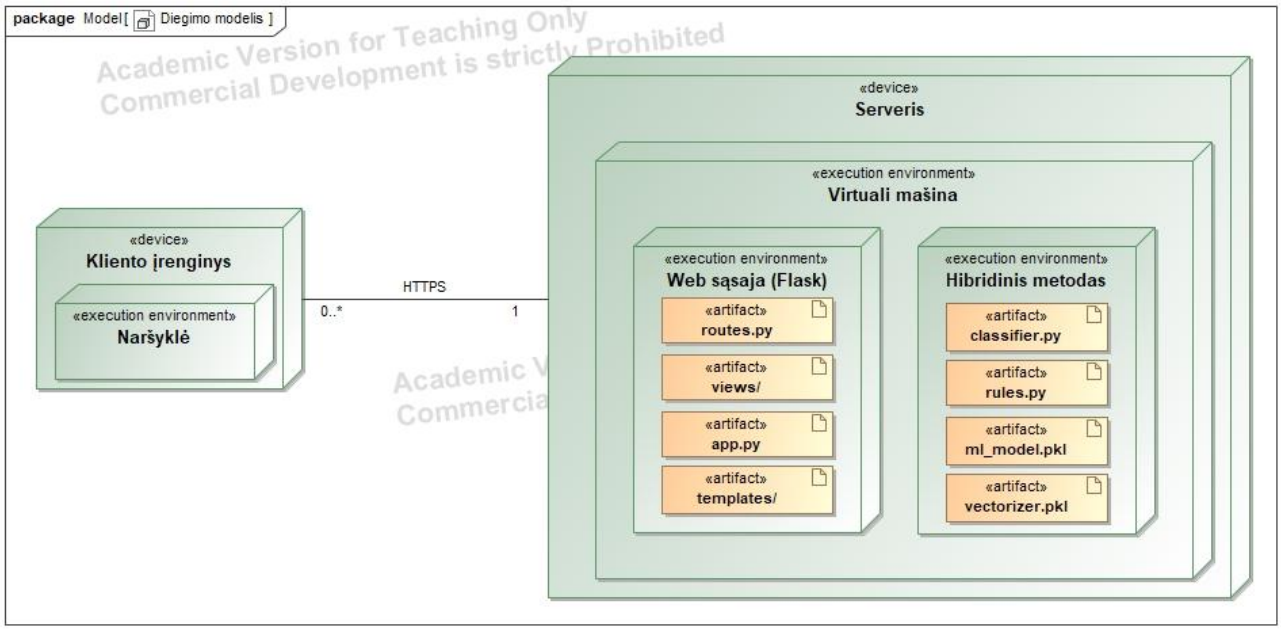
3. Sukčiavimo SMS žinutėmis aptikimo metodą realizuojančios sistemos prototipas

Šiame skyriuje pateikiamas sukurtos sukčiavimo SMS žinučių aptikimo sistemos prototipo realizavimas. Vadovaujantis ankstesniuose skyriuose suformuotu metodo projektu, šiame etape įgyvendinamos pagrindinės funkcijos: duomenų išgavimas, apdorojimas, analizė taisyklėmis grįstu bei mašininio mokymosi pagrindu realizuotu būdu, o taip pat galutinio rezultato pateikimas naudotojui. Prototipas leidžia praktiškai išbandyti parengtą metodą, įsitikinti jo veikimo logika ir pasiruošti tolimesniam sistemos efektyvumo vertinimui.

3.1. Sistemos architektūra ir diegimo modelis

Sukurtas SMS žinučių grėsmės vertinimo prototipas paremtas kliento-serverio architektūra, kur naudotojas sąveikauja su sistema per interneto naršyklę, o pagrindinis duomenų apdorojimas, analizė ir klasifikavimas atliekami serverio pusėje. Toks architektūrinis sprendimas leidžia centralizuotai valdyti klasifikavimo logiką, užtikrinti lengvesnį sistemos palaikymą bei sudaro prielaidas ateityje sistemą plėsti ar integruoti į kitas platformas.

Žemiau pateiktame paveiksle (11 pav.) pavaizduotas sistemos diegimo modelis, kuriame aiškiai matomi pagrindiniai komponentai, jų vykdymo aplinkos bei tarpusavio sąveika.



11 pav. Sistemos diegimo modelis

Naudotojo įrenginyje veikia interneto naršyklė, per kurią įvedama SMS žinutė ir siuntėjo identifikatorius. Naudotojo užklausos per saugų HTTPS protokolą perduodamos į serverį. Serverio pusėje sistema veikia virtualioje mašinoje, kurioje realizuotos dvi pagrindinės vykdymo aplinkos: web sąsaja ir hibridinis grėsmės klasifikavimo modulis.

Web sąsaja sukurta naudojant „Flask“ programinį karkasą ir atsakinga už naudotojo sąveiką su sistema. Šioje vykdymo aplinkoje veikia tokie artefaktai kaip app.py, routes.py bei HTML šablonai, esantys templates bei views kataloge. Web sąsajos funkcija - priimti naudotojo įvestį, perduoti ją klasifikavimo moduliui, gauti analizės rezultatus ir juos vizualiai pateikti naudotojui.

Hibridinis klasifikavimo modulis realizuotas atskiroje vykdymo aplinkoje ir apjungia du skirtingus sprendimų priėmimo metodus: taisyklėmis grįstą analizę ir mašininio mokymo modelį. Taisyklėmis grįsta logika įgyvendinta faile rules.py, kuriame aprašyti požymiai, jų svoriai bei balų skaičiavimo mechanizmas. Šis metodas leidžia identifikuoti aiškius ir iš anksto žinomus sukčiavimo požymius, tokius kaip įtartinos nuorodos, skubos terminai ar finansinių duomenų minėjimas.

Mašininio mokymo dalis realizuota naudojant klasikinius teksto klasifikavimo metodus. Ji susideda iš dviejų pagrindinių artefaktų: vectorizer.pkl, kuriame saugomas apmokytas TF-IDF vektorizatorius, ir ml_model.pkl, kuriame saugomas apmokytas logistinės regresijos klasifikatorius. Šie modeliai į sistemą įkeliami serverio paleidimo metu ir naudojami visų gaunamų SMS žinučių analizei be papildomo persikrovimo, kas užtikrina efektyvų ir greitą sistemos veikimą.

Vidinė komunikacija tarp web sąsajos ir hibridinio klasifikavimo modulio vykdoma Python aplinkoje per tiesioginius funkcijų kvietimus, todėl papildomi tinklo protokolai tarp modulių nėra reikalingi. Toks sprendimas sumažina sistemos sudėtingumą ir padidina jos našumą.

Apjungus taisyklėmis grįsto metodo balus ir mašininio mokymo modelio pateikiamą tikimybę, hibridinis modulis priima galutinį sprendimą dėl SMS žinutės rizikos lygio - ar ji laikytina teisėta, įtartina ar neteisėta. Šis sprendimas perduodamas web sąsajai ir pateikiamas naudotojui suprantama forma.

Toks sistemos architektūros ir diegimo modelis leidžia aiškiai atskirti naudotojo sąsajos ir analizės logikos atsakomybes, užtikrina sistemos lankstumą bei sudaro tvirtą pagrindą tolimesniam funkcionalumo plėtimui ir eksperimentiniams tyrimams.

3.2. Naudotos technologijos ir įrankiai

Sistemos realizavimas atliktas taikant modernius programinius įrankius bei bibliotekas, pritaikytas teksto analizės ir neuroninių tinklų kūrimo užduotims. Žemiau pateikiamos pagrindinės naudotos technologijos:

Programavimo aplinka:

- **Python 3.10** [28] - lanksti, plačiai naudojama programavimo kalba, pasirinkta dėl savo galimybių dirbtinio intelekto srityje bei didelės atvirojo kodo bibliotekų ekosistemos.
- **Jupyter Notebook** [29] - interaktyvi aplinka, leidžianti vienoje vietoje rašyti kodą, vykdyti eksperimentus ir pateikti komentaruose analizę bei rezultatus.

Duomenų analizė ir apdorojimas:

- **Pandas** [30] - struktūrizuotų duomenų tvarkymui, CSV rinkinio nuskaitymui, filtravimui, grupavimui ir konvertavimui į tinkamą formatą.
- **NumPy** [31] - efektyviam matricų skaičiavimui, tekstinių sekų transformacijai į skaitinius vektorius bei papildomiems statistiniams veiksmams.

Modelio kūrimas ir treniravimas:

- **Scikit-learn** [32] - naudotas duomenų skaidymui į treniravimo ir testavimo dalis, klasifikacijos tikslumo skaičiavimui, metrikų generavimui bei papildomoms pagalbinėms funkcijoms.

Rezultatų vizualizavimas:

- **Matplotlib** [33] - grafinių diagramų kūrimui, mokymo eigos (pvz., nuostolių ir tikslumo kitimo) vizualizacijai.
- **Seaborn** [34] - papildomas įrankis, skirtas vizualiai patrauklesniam duomenų pasiskirstymo ir rezultatų atvaizdavimui.

3.3. Duomenų struktūra

Šis darbas taiko SMS žinučių klasifikavimo metodiką, kurios tikslas - sukurti hibridinį modelį, gebantį automatiškai atpažinti teisėtas (ham) ir potencialiai apgaulingas (spam) žinutes. Siekiant tai įgyvendinti, pirmiausia buvo suformuotas eksperimentinis duomenų rinkinys, kuriuo remiantis buvo vykdoma analizė ir kuriami klasifikavimo sprendimai.

Duomenys buvo surinkti iš realių šaltinių – socialinio tinklo „Facebook“ [35], viešai prieinamų grupių bei forumų [36], kuriuose vartotojai dalijosi gautomis įtartinomis trumposiomis žinutėmis. Visos žinutės buvo surinktos rankiniu būdu, atmetant pasikartojančius, neaiškios kilmės ar techniniu požiūriu netinkamus įrašus. Gauta informacija buvo struktūrizuota ir suvienodinta, siekiant užtikrinti vientisą tolesnio apdorojimo procesą.

Tokie šaltiniai pasirinkti dėl jų autentiškumo bei realios praktinės vertės - surinktos žinutės dažniausiai atspindi aktualias socialinės inžinerijos tendencijas, dažniausiai pasitaikančius sukčiavimo scenarijus, kalbines struktūras. Tokiu būdu formuojamas duomenų rinkinys yra artimas realioms situacijoms, su kuriomis susiduria vartotojai, o tai leidžia testuoti sukurtą sistemą realistiškame kontekste.

Sukaupti duomenys buvo apdoroti ir suvienodinti į struktūrizuotą lentelę, kuri išsaugota CSV formatu (angl. *Comma-Separated Values*). Kiekviena lentelės eilutė (žr. **4 lentelė**) atitinka vieną SMS žinutę su papildomais metaduomenimis - siuntėju, data, priskyrimu klasei ir pan. Toks struktūrizavimas leidžia vieningai taikyti tiek taisyklių analizę, tiek neuroninių tinklų metodus.

4 lentelė. Duomenų rinkinio struktūros aprašas

Nr.	Pavadinimas	Reikšmė
1.	Numeris	Eilutės identifikatorius (numeracija)
2.	Žinutė	Tikslus SMS tekstas (turinys), kuris bus klasifikuojamas
3.	Telefono numeris	Siuntėjo identifikatorius - telefono numeris arba pavadinimas
4.	URL	Nuoroda, pateikta žinutėje (jei yra)
5.	Nuotraukos pavadinimas	Ekrano kopijos failo pavadinimas
6.	Tipas	Turinys suskirstytas į temines kategorijas (pvz., siuntos, darbas)
7.	Data	Žinutės gavimo ar fiksavimo data (formatu YYYY-MM-DD)
8.	Teisėta ar apgaulinga	Klasė, nurodanti ar žinutė teisėta (Ham) ar apgaulinga (Spam)

3.4. Hibridinės analizės metodas

Siekiant padidinti SMS žinučių sukčiavimo aptikimo sistemos efektyvumą, darbe taikomas hibridinis analizės metodas, apjungiantis dvi skirtingas, tačiau viena kitą papildančias strategijas - taisyklėmis grįstą analizę ir giluminio mokymosi pagrindu veikiantį neuroninį tinklą. Tokia kombinacija leidžia

išnaudoti tiek iš anksto žinomų rizikos indikatorių atpažinimo pranašumus, tiek gebėjimą automatiškai išmokti sudėtingesnius semantinius dėsningumus, kurie nebūtinai yra aiškiai apibrėžiami.

Taisyklėmis grįstas metodas leidžia greitai identifikuoti žinomas rizikos charakteristikas, tokias kaip nuorodų struktūra, siuntėjo identifikatorius ar dažnai pasitaikantys raktiniai žodžiai. Tuo tarpu neuroninio tinklo modelis išmoksta atpažinti sudėtingesnius, kontekste pasireiškiančius požymius, remdamasis ankstesne mokymosi patirtimi.

Hibridinis požiūris pasirenkamas siekiant sukurti patikimesnę ir universalesnę sistemą, kuri ne tik pasikliauja iš anksto nustatytais bruožais, bet ir geba prisitaikyti prie besikeičiančių sukčiavimo strategijų. Tolesnėse skiltyse pateikiamas išsamesnis abiejų komponentų aprašymas.

3.4.1. Taisyklėmis grįstas aptikimas

Taisyklėmis grįstas SMS žinučių sukčiavimo aptikimo metodas sukurtas remiantis iš anksto apibrėžtu požymių (angl. *features*) rinkiniu, kuris leidžia automatiškai įvertinti trumpųjų žinučių turinį ir priimti preliminarų sprendimą dėl jų teisėtumo. Kiekvienas požymis atspindi konkretų sukčiavimo indikatorių, kuris dažnai pasitaiko realiose apgaulės schemose. Šiems požymiams priskirti skirtingi svoriai (balai nuo 1 iki 5), priklausomai nuo jų reikšmingumo vertinant riziką. Rezultatas išreiškiamas bendru balu, o viršijus iš anksto nustatytą slenkstį – 10 balų, žinutė klasifikuojama kaip galimai apgaulinga (spam), kitaip - kaip teisėta (ham).

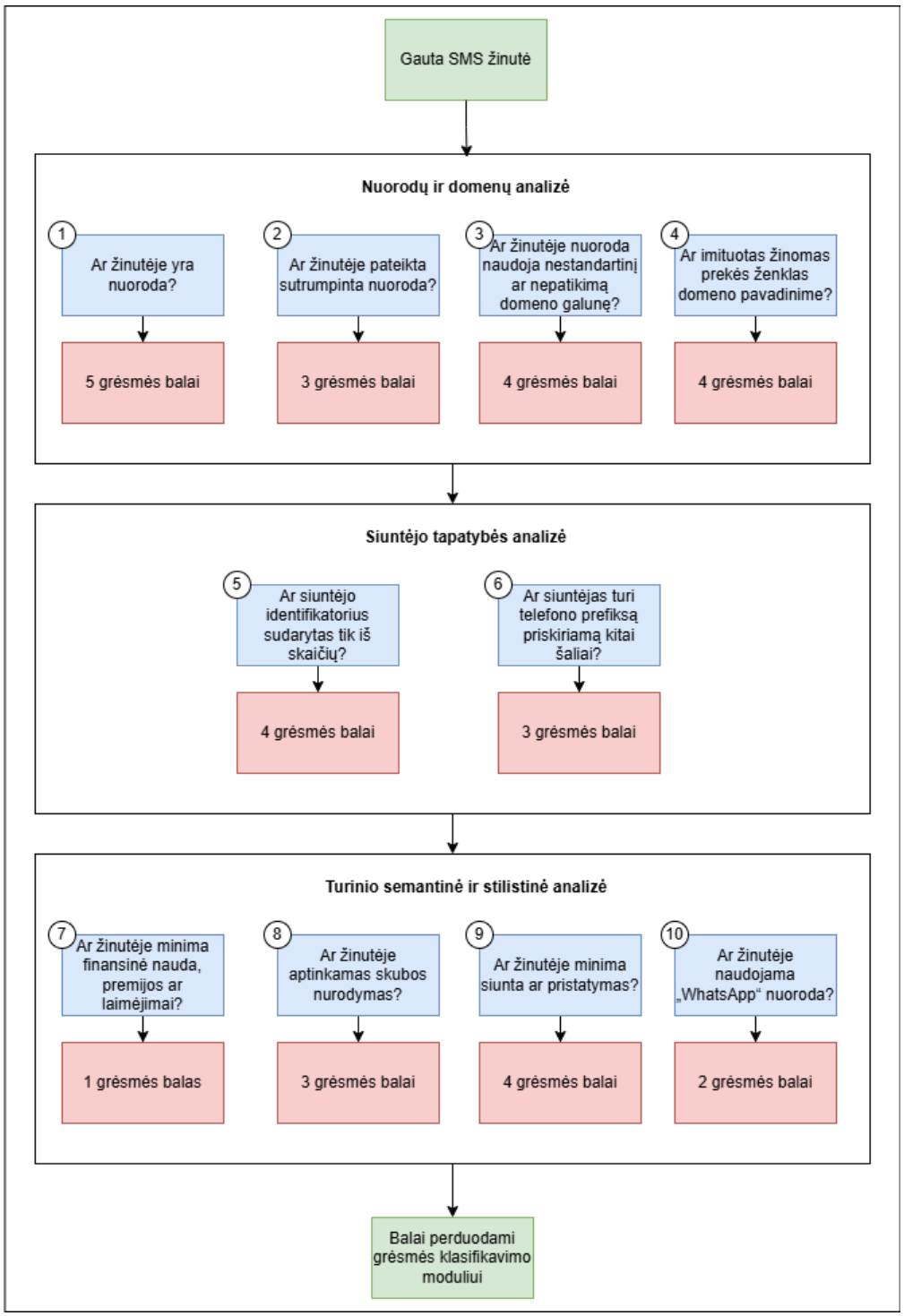
Siekiant išlaikyti struktūrinį aiškumą ir nuoseklumą, požymiai suskirstyti į tris pagrindines grupes: nuorodų ir domenų analizė, siuntėjo tapatybės analizė ir turinio semantinė bei stilistinė analizė.

3.4.2. Taisyklėmis grįsto aptikimo realizacija

Žemiau pateiktoje schemoje pavaizduotas taisyklėmis grįsto aptikimo modulio veikimo principas nuo gautos SMS žinutės iki galutinio rezultato perdavimo klasifikavimo moduliui. Gavus naują žinutę, jos turinys perkeliamas į analizės procesą, kuris suskaidytas į tris logines analizės sritis: nuorodų ir domenų analizę, siuntėjo tapatybės analizę ir turinio semantinę bei stilistinę analizę. Kiekvienoje iš šių sričių žinutė tikrinama pagal atitinkamus požymius (iš viso - 10), kurių kiekvienas apibrėžtas remiantis realių apgaulingų žinučių analize.

Kiekvienas požymis vertinamas dvejetainiu principu – jis arba atitinka, arba ne. Jei požymis aptinkamas, jam priskiriamas svoris (nuo 1 iki 5), kuris prisideda prie bendro žinutės grėsmės balo. Šis procesas leidžia kiekybiškai įvertinti rizikos lygį, remiantis aiškiai apibrėžtais požymiais, kurių buvimas dažnai koreliuoja su apgaulės tikimybe.

Kai visos trys analizės sritys įvykdytos, susumuotas grėsmės balas perduodamas grėsmės klasifikavimo moduliui. Pastarasis, remdamasis hibridine logika (kartu su mašininio mokymo modeliu), priima galutinį sprendimą dėl žinutės teisėtumo.



12 pav. Taisyklėmis grįsto metodo realizuojančio algoritmo etapai

3.4.2.1. Nuorodų ir domenų analizė

Ši kategorija apima požymius, susijusius su interneto nuorodų buvimu, jų struktūra, domeno kilme bei semantiniu panašumu į žinomus prekių ženklus. Nuorodos dažniausiai yra pagrindinis elementas, leidžiantis sukčiams perimti naudotojo duomenis, todėl jos analizė yra itin reikšminga. Žemiau pateikta lentelė aprašo pagrindines taisykles, susijusias su nuorodų buvimu ir jų charakteristikomis. Taisyklės parinktos remiantis pradiniu duomenų rinkiniu (139 SMS žinučių), kuriame pastebėta, kad didžioji dalis apgaulingų pranešimų turėjo aktyvių arba paslėptų nuorodų, vedančių į nepatikimus domenus.

5 lentelė. Nuorodų ir domenų analizės požymiai

Nr.	Požymis	Aprašymas	Pavyzdys	Svoris (1-5)	Pagrindimas
1.	Nuorodos buvimas žinutėje	Ar žinutėje yra aktyvi nuoroda („http“, „www“).	https://post.ltxh.homes/lt	5	98,3 % spam žinučių turėjo nuorodą; retai pasitaiko teisėtose. Labai stiprus indikatorius.
2.	Trumpintų nuorodų naudojimas	Žinutėje pateikta sutrumpinta nuoroda, kuri slepia tikrąjį nukreipimo adresą	https://bit.ly/3abc	3	Iš apgaulingų žinučių, turinčių URL, net 37,9 % naudoja nuorodų trumpinimus, tokius kaip bit.ly, cutt.ly. Šis požymis retas teisėtose žinutėse ir ženkliai padidina riziką, nes slepia tikslinį adresą.
3.	Įtartino aukščiausio lygio domeno (TLD) naudojimas	URL naudoja nestandartinį arba nepatikimą domeno galūnę	https://posttrack.cfd	4	Iš spam žinučių su URL, net 69,0 % turėjo neįprastus TLD, tokius kaip .top, .xyz, .site, .cfd. Tai dažnai pasirenkama sukčiavimui dėl žemų kainų ir minimalių registracijos reikalavimų.
4.	Imituotas žinomas ženklas domeno pavadinime	Domeno pavadinime naudojami vizualiai ar fonetiškai panašūs pavadinimai, imituojantys žinomus paslaugų teikėjus ar institucijas	postltad.top, vmi-secure.site, venipak-track.cfd	4	Iš spam žinučių su URL, net 56,9 % domenų imitavo gerai žinomus prekių ženklus ar organizacijas, tokias kaip „Lietuvos paštas“, „VMI“, „Venipak“, „Smart-ID“ ar „Swedbank“. Tai vienas dažniausių socialinės inžinerijos būdų, todėl šiam požymiui suteikiamas maksimalus svoris.

3.4.2.2. Siuntėjo tapatybės analizė

Ši analizės kategorija orientuota į SMS siuntėjo identifikatoriaus ypatumus, kurie gali signalizuoti apie neautentišką ar automatinį žinutės šaltinį. Skirtingai nei teisėtuose pranešimuose, sukčiai dažnai pasitelkia užsienietiškus numerius arba bendrinius automatinius siuntėjus, kuriuos sunku atsekti ar identifikuoti. Būtent siuntėjo kilmė ir formatas gali būti vieni iš pirmųjų signalų, rodančių, jog žinutė potencialiai nėra patikima.

Remiantis pradine duomenų imtimi (139 SMS, iš kurių 58 priskirtos neteisėtai kategorijai), buvo nustatyta, kad didžioji dalis apgaulingų žinučių buvo siunčiamos iš neaiškių skaitmeninių numerių, o reikšminga dalis - iš užsienietiškų tinklų su prefiksais, nesusijusiais su Lietuva. Šie požymiai leidžia efektyviai identifikuoti automatizuotą ar užsienio kilmės siuntimą ir yra vertingi klasifikavimo modelio dalis.

Žemiau pateiktoje lentelėje apibendrinamos dvi pagrindinės taisyklės, susijusios su siuntėjo tapatybės vertinimu.

6 lentelė. Siuntėjo tapatybės analizės požymiai

Nr.	Požymis	Aprašymas	Pavyzdys	Svoris (1-5)	Pagrindimas

1.	Skaitmeninio formato siuntėjo ID	Siuntėjo identifikatorius sudarytas tik iš skaičių, be vardo ar prekės ženklo	+63 963 306 4080, +212 6 20 23 68 21...	4	Net 65,5 % neteisėtų žinučių siųstos iš skaitmeninių numerių. Teisėtos žinutės dažniau naudoja aiškius, atpažįstamus pavadinimus (pvz., „Swedbank“). Tai stiprus požymis, būdingas automatizuotiems siuntėjams.
2.	Užsienio numerio prefikso naudojimas	Siuntėjas turi telefono prefiksą, priskiriamą kitai šaliai (ne +370)	+44..., +60..., +91...	3	46,6 % neteisėtų žinučių siųstos iš užsienietiškų numerių. Nors ne visada tai reiškia sukčiavimą, šis požymis dažnai rodo siuntimą iš žemo patikimumo platformų ar šalių už ES ribų.

3.4.2.3. Turinio semantinė ir stilistinė analizė

Ši analizės kategorija orientuota į pačios SMS žinutės turinį - t. y. nagrinėjama ne siuntimo kilmė ar techniniai požymiai, o tai, ką bando pasakyti žinutės autorius. Analizuojami raktažodžiai, frazės, stilistiniai bruožai ir emociniai ar psichologiniai signalai, dažnai pasitaikantys sukčiavimo kontekstuose. Toks požiūris leidžia atskleisti tipinius apgaulingų žinučių komunikacijos šablonus, padedančius manipuliuoti vartotojo elgsena, sukelti spaudimą, nerimą ar netikėtą pasitikėjimą.

Taisyklės šioje kategorijoje buvo formuojamos analizuojant pirminį duomenų rinkinį (58 apgaulingas SMS žinutes) ir identifikuojant dažniausiai pasitaikančius semantinius elementus, kurie galėtų būti automatiškai aptinkami ir įvertinami klasifikavimo metu.

7 lentelė. Semantinės ir stilistinės analizės požymiai

Nr.	Požymis	Aprašymas	Pavyzdys	Svoris (1-5)	Pagrindimas
1.	Minima finansinė nauda	Žinutėje minima finansinė nauda, premijos ar laimėjimai	Laimėjote 950.000 €, atsiimkite premiją!	1	Tik 8,6 % visų neteisėtų žinučių mini finansinę naudą. Nors tai gali būti stiprus emocinis indikatorius, jis pasitaiko retai, todėl priskiriamas mažas svoris.
2.	Skubos nurodymas	Naudojamos frazės, verčiančios veikti greitai, pabrėžiant terminus ar ribotą laiką	Reaguokite per 12 val., arba paskyra bus užblokuota	3	31,0 % žinučių turi skubos elementų. Tai vidutinio stiprumo manipuliacinis signalas, pasitaikantis sukčiavimo žinutėse.
3.	Minima siunta ar pristatymas	Žinutė susijusi su siuntų tema - nurodomas paketas, pristatymo data, atnaujinimas	Pristatymo bandymas 1/2- Kestutis, jusu ekspreso siunta #CS894389743LT bus grazinta siandien, jei jos nepatvirtinsite per 2 val.: ajuyip.com/YLxt10S	4	Net 60,3 % neteisėtų žinučių imituoja logistikos bendroves ar mini siuntas. Tai dažniausias teminis sukčiavimo scenarijus, todėl suteikiamas aukštas svoris.

4.	Naudojama WhatsApp nuoroda	Žinutėje pateikiama nuoroda į komunikaciją per WhatsApp platformą (pvz., wa.me)	Sveiki, mes esame Berkshire Hathaway TSQ Group projekto brokeriai ir šiuo metu ieškome ne visą darbo dieną dirbančios komandos. Prisijungę iš karto gausite papildomą 2800 eurų premiją. Spustelėkite nuorodą, kad pradėtumėte pokalbį per „„WhatsApp". https://wa.me/14014834630?ts=mLolO	2	Tik 10,3 % žinučių naudojo WhatsApp. Nors tai gali padėti išvengti filtrų, pasitaiko retai, todėl priskiriamas mažesnis svoris.

3.4.2.4. Svorio reikšmės interpretacija

Taisyklėmis grįstam metodui sukurti buvo pritaikytas svorio (balų) modelis, leidžiantis kiekvienam aptiktam požymiui priskirti tam tikrą reikšmę, remiantis jo dažniu tarp apgaulingų žinučių. Šis metodas suteikia galimybę kiekybiškai vertinti rizikos lygį ir priimti sprendimą dėl SMS žinutės teisėtumo.

Remiantis pirminio duomenų rinkinio analize, kiekvienam požymiui suteiktas balas nuo 1 iki 5, atsižvelgiant į tai, kokiai daliai žinučių tas požymis būdingas:

8 lentelė. Taisyklėmis grįsto metodo požymių svorių nustatymo kriterijai

Dažnio intervalas tarp neteisėtų žinučių	**Svoris (Balai)**	**Pagrindimas**
≥ 70 %	5	Kritinis požymis - būdingas daugumai žinučių, laikomas stipriu indikatoriumi
50 - 69 %	4	Labai svarbus požymis - pasitaiko daugiau nei pusėje atvejų
30 - 49 %	3	Vidutinio stiprumo indikatorius - veikia geriau kartu su kitais požymiais
10 - 29 %	2	Silpnesnis požymis - prideda papildomos informacijos, bet nepakankamas vienas
< 10 %	1	Retas požymis - laikomas silpnu signalu, bet vis tiek vertingas detekcijoje

Kiekvienai gaunamai žinutei taikomos visos 10 nustatytų taisyklių. Jei tam tikra taisyklė atitinka - jos svoris yra pridedamas prie bendro grėsmės balo.

3.4.3. Mašininio mokymo modelis

Kuriant SMS žinučių grėsmės vertinimo sistemą, šiame darbe buvo pasirinktas klasikinis mašininio mokymo metodas, paremtas tekstinių požymių išgavimu naudojant TF-IDF vektorizaciją ir logistinės regresijos klasifikatorių. Skirtingai nei taisyklėmis grįstas metodas, kuris remiasi iš anksto apibrėžtais

kriterijais ir jų svoriais, mašininio mokymo modelis leidžia automatiškai įvertinti žodžių ir jų junginių svarbą visame duomenų rinkinyje, mokantis iš realių pavyzdžių.

TF-IDF metodas leidžia kiekvieną žinutę paversti skaitinių požymių vektoriumi, kuriame atsispindi ne tik žodžio pasikartojimo dažnis konkrečioje žinutėje, bet ir jo reikšmingumas visame duomenų rinkinyje. Tokiu būdu dažni, tačiau mažai informatyvūs žodžiai gauna mažesnį svorį, o retesni, bet potencialiai svarbūs terminai - didesnį. Šis požiūris yra ypač tinkamas trumpų tekstinių pranešimų, tokių kaip SMS žinutės, analizei.

Modelio realizavimas atliktas naudojant Python programavimo kalbą ir interaktyvią Jupyter Notebook aplinką, kuri leidžia patogiai derinti programinį kodą, duomenų analizę bei tarpinius eksperimentų rezultatus. Tai sudarė sąlygas greitai keisti parametrus, vertinti skirtingas konfigūracijas ir stebėti modelio veikimą testiniuose duomenyse.

Modeliui sukurti ir treniruoti naudotos šios pagrindinės Python programavimo kalbos bibliotekos:

- **Pandas** [30] - biblioteka, skirta struktūrizuotų duomenų (pvz., lentelių, CSV, Excel failų) įkėlimui, analizavimui ir transformacijai. Šiame darbe „Pandas“ naudojama nuskaityti duomenų rinkinį (SMS žinutes), atlikti pradinius duomenų apdorojimo žingsnius bei pasirengti treniravimui.

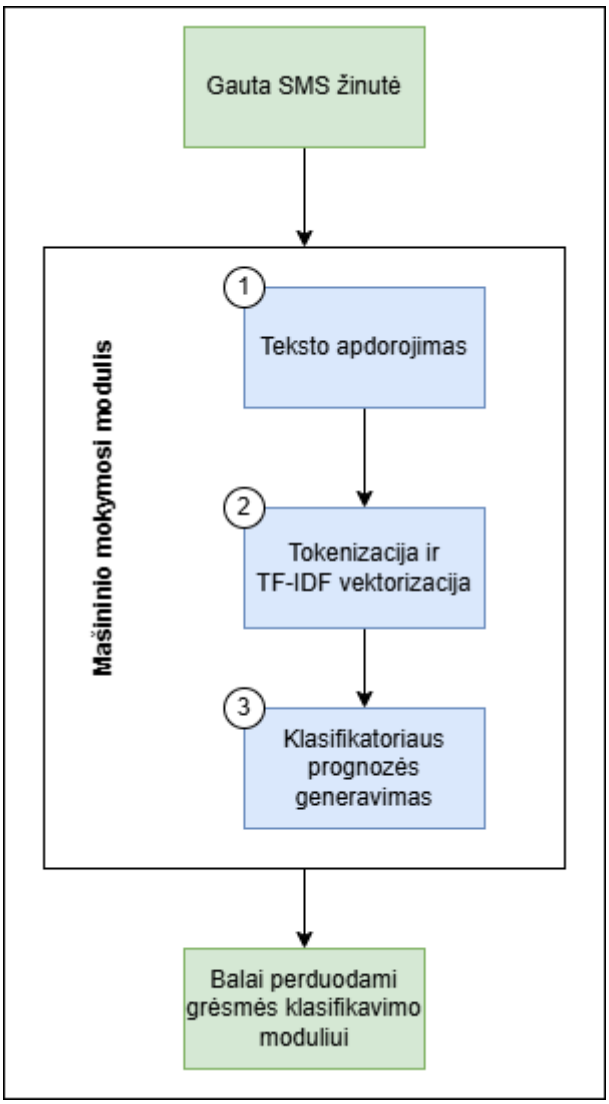
- **NumPy** [31] - skaitinių skaičiavimų biblioteka, kuri užtikrina efektyvų matricų, vektorių ir kitų daugiamačių duomenų struktūrų apdorojimą. Naudojama tekstinių duomenų pavertimui į skaitines sekas.

- **Scikit-learn** [32] - tai populiari duomenų analizės ir mašininio mokymo biblioteka, kuri šiame projekte naudojama duomenų skaidymui į treniravimo ir testavimo rinkinius, klasifikavimo rezultatų įvertinimui bei kai kurioms papildomoms operacijoms, pvz., teksto vektorizavimui naudojant TF-IDF.

- **Matplotlib** [33] ir **Seaborn** [34] - duomenų vizualizacijos bibliotekos, naudojamos duomenų pasiskirstymo, mokymo eigos ar modelio rezultatų pateikimui grafiniu formatu.

Mašininio mokymo modelis generuoja tikimybinę prognozę, kuri nusako, kokia yra tikimybė, kad pateikta SMS žinutė priklauso neteisėtos („spam“) kategorijai. Ši reikšmė yra intervale nuo 0 iki 1 ir vėliau naudojama hibridiniame sprendimų priėmimo mechanizme, kuriame ji apjungiama su taisyklėmis grįsto metodo rezultatais. Tokiu būdu užtikrinamas lankstesnis ir patikimesnis galutinio žinutės rizikos lygio nustatymas.

Žemiau pateiktame paveiksle pavaizduota mašininio mokymo modulio struktūros schema, atspindinti pagrindinius SMS žinutės analizės etapus - nuo pradinio teksto apdorojimo iki klasifikavimo rezultato generavimo.



13 pav. Mašininio mokymo modulio struktūros schema

3.4.4. Mašininio mokymo modulio realizacijos etapai

Mašininio mokymo modulis, įdiegtas šio darbo prototipe, atlieka SMS žinučių analizę pasitelkdamas iš anksto apmokytą klasikinį mašininio mokymo modelį, paremtą TF-IDF vektorizacija ir logistinės regresijos klasifikatoriumi. Apmokytas modelis bei teksto vektorizavimo komponentas yra įkeliami į sistemą kaip serializuoti „.pkl“ formato objektai ir į atmintį įkeliami tik vieną kartą - aplikacijos arba serverio paleidimo metu. Tai leidžia efektyviai apdoroti didelį kiekį žinučių be papildomų skaičiavimų kiekvienos užklausos metu.

Kai sistema gauna naują SMS žinutę, ji pereina tris pagrindinius analizės etapus, kurie pavaizduoti 13 paveiksle. Kiekvienas etapas yra būtinas norint transformuoti neapdorotą tekstą į skaitinę formą, tinkamą klasifikavimo algoritmui, ir gauti tikimybinę prognozę.

Pirmasis etapas - teksto apdorojimas. Pirmojo etapo metu atliekamas pirminis teksto apdorojimas, kurio tikslas - sumažinti triukšmą ir suvienodinti duomenų formatą. SMS žinutės tekstas konvertuojamas į žemąsias raides, pašalinami pertekliniai simboliai, o URL adresai, skaitinės reikšmės ar valiutos žymėjimai normalizuojami į bendrinius žymenis. Šis žingsnis padeda sumažinti skirtingų formų variacijas ir pagerina modelio gebėjimą atpažinti bendrus tekstinius dėsningumus.

Antrasis etapas - tokenizacija ir TF-IDF vektorizacija. Antrajame etape vykdoma teksto tokenizacija ir vektorizacija naudojant TF-IDF metodą. Skirtingai nei neuroniniuose tinkluose taikomas sekų kodavimas, šiame sprendime kiekviena SMS žinutė yra paverčiama į fiksuoto ilgio skaitinį vektorių, atspindintį žodžių bei jų junginių (n-gramų) svarbą.

TF-IDF metodas apskaičiuoja kiekvieno termino svorį atsižvelgiant į jo pasikartojimo dažnį konkrečioje žinutėje bei jo paplitimą visame duomenų rinkinyje. Tokiu būdu dažni, bet mažai informatyvūs žodžiai (pvz., bendriniai jungtukai) gauna mažesnį svorį, o retesni, tačiau su sukčiavimu susiję terminai - didesnį. Ši savybė yra ypač naudinga analizuojant trumpas SMS žinutes, kuriose svarbūs būtent pavieniai raktažodžiai ar jų kombinacijos.

Trečiasis etapas - klasifikatoriaus prognozės generavimas. Trečiajame etape gautas TF-IDF požymių vektorius perduodamas logistinės regresijos klasifikatoriui. Šis modelis apskaičiuoja tikimybę, kad analizuojama SMS žinutė priklauso neteisėto brukalo (angl. *spam*) kategorijai. Klasifikatoriaus išvestis yra realioji reikšmė intervale nuo 0 iki 1, kuri interpretuojama kaip sukčiavimo tikimybė.

Gauta tikimybė nėra naudojama izoliuotai - ji perduodama grėsmės klasifikavimo moduliui, kuriame vėliau apjungiama su taisyklėmis grįsto metodo rezultatais. Toks hibridinis sprendimų priėmimo mechanizmas leidžia pasiekti didesnį sistemos patikimumą, sumažinti klaidingų teigiamų atvejų skaičių ir lanksčiai identifikuoti tiek akivaizdžiai pavojingas, tiek potencialiai įtartinas SMS žinutes.

3.4.4.1. Modelio architektūra

Mašininio mokymo modelio architektūra šiame darbe paremta klasikiniu tekstų klasifikavimo principu, kuriame derinami statistiniai tekstinių požymių išgavimo metodai ir prižiūrimo mokymo klasifikatorius. Skirtingai nei giluminio mokymo sprendimuose, čia nenaudojami neuroniniai tinklai ar sekų apdorojimo sluoksniai - visa analizė grindžiama fiksuoto ilgio požymių vektoriais, kurie leidžia užtikrinti greitą ir stabilų veikimą realaus laiko sistemoje.

Pagrindinis mašininio mokymo modelio komponentas yra TF-IDF vektorizatorius, kuris neapdorotą SMS žinutės tekstą paverčia į skaitinį požymių vektorių. Šiame etape analizuojamas žodžių ir jų junginių (n-gramų) pasikartojimo dažnis konkrečioje žinutėje bei jų reikšmingumas visame duomenų rinkinyje. Tokiu būdu dažni, tačiau mažai informatyvūs terminai gauna mažesnį svorį, o retesni, tačiau su sukčiavimu susiję raktažodžiai - didesnį.

Gauti TF-IDF požymiai perduodami logistinės regresijos klasifikatoriui, kuris apskaičiuoja tikimybę, kad pateikta SMS žinutė priklauso neteisėto brukalo kategorijai. Logistinė regresija pasirinkta dėl savo paprastos struktūros, greito skaičiavimo bei aiškiai interpretuojamos išvesties, kuri yra ypač tinkama hibridiniam sprendimų priėmimo mechanizmui.

Modelio išvestis yra realioji reikšmė intervale nuo 0 iki 1, interpretuojama kaip sukčiavimo tikimybė. Ši reikšmė nėra naudojama izoliuotai – ji vėliau sujungiama su taisyklėmis grįsto metodo rezultatais, siekiant priimti galutinį sprendimą apie SMS žinutės rizikos lygį.

Mašininio mokymo modelio architektūrą sudarantys komponentai pateikti 9 lentelėje.

9 lentelė. Mašininio mokymo modelio architektūros komponentai

Komponentas	**Tipas**	**Paskirtis**
Teksto normalizavimas	Pirminis teksto apdorojimas	SMS žinutės teksto suvienodinimas ir nereikalingo triukšmo sumažinimas
TF-IDF vektorizatorius	Tekstinių požymių išgavimas	Teksto pavertimas į skaitinį požymių vektorių
N-gramų analizė	Požymių išplėtimas	Žodžių ir trumpų žodžių junginių reikšmingumo įvertinimas
Logistinės regresijos klasifikatorius	Mašininio mokymo algoritmas	Sukčiavimo (brukalo) tikimybės apskaičiavimas
Tikimybinė išvestis	Tikimybė įvertinimas	Naudojama hibridiniame sprendimų priėmimo mechanizme

Toks architektūrinis sprendimas leidžia pasiekti gerą balansą tarp klasifikavimo tikslumo, skaičiavimo efektyvumo ir modelio aiškinamumo, tai yra itin svarbu analizuojant trumpus tekstinius pranešimus, tokius kaip SMS žinutės.

3.4.4.2. Modelio parametrizavimas

Mašininio mokymo modelio parametrizavimas buvo atliekamas atsižvelgiant į analizuojamų SMS žinučių trumpą tekstinę struktūrą, ribotą duomenų rinkinio dydį bei sistemos realaus laiko veikimo reikalavimus. Kadangi šiame darbe taikomas klasikinis mašininio mokymo sprendimas, parametrai parinkti taip, kad būtų užtikrintas stabilus modelio veikimas, geras apibendrinimas ir nedidelės skaičiavimo sąnaudos. Pagrindiniai modelio ir teksto vektorizacijos parametrai, naudoti šiame darbe, pateikti 10 lentelėje.

Tekstinių požymių išgavimui pasirinktas TF-IDF vektorizacijos metodas, kuris leidžia statistiškai įvertinti žodžių svarbą kiekvienoje SMS žinutėje viso duomenų rinkinio kontekste. Šis metodas sumažina dažnų, tačiau mažai informatyvių žodžių įtaką ir išryškina terminus, kurie yra labiau susiję su sukčiavimo turiniu, todėl yra tinkamas trumpų tekstinių pranešimų analizei.

Vektorizacijos etape taikomas n-gramų diapazonas nuo 3 iki 5, leidžiantis analizuoti ne tik pavienius žodžius, bet ir trumpas žodžių kombinacijas. Tai ypač svarbu SMS sukčiavimo atpažinimo užduotyje, nes tokiose žinutėse dažnai pasikartoja specifinės frazės, susijusios su mokėjimais, siuntomis ar skubiais veiksmais.

Siekiant sumažinti triukšmą ir pagerinti modelio apibendrinimo gebėjimus, buvo nustatytos dokumentų dažnio ribos. Minimali dokumentų dažnio riba (min_df = 2) leidžia pašalinti itin retus terminus, kurie pasitaiko tik pavienėse žinutėse ir neturi statistinės reikšmės modelio mokymuisi. Maksimali dokumentų dažnio riba (max_df = 0,9) naudojama siekiant sumažinti labai dažnų, tačiau mažai informatyvių žodžių įtaką, kurie pasitaiko didžiojoje dalyje žinučių.

Klasifikavimo etape naudojamas logistinės regresijos algoritmas, kuris yra prižiūrimo mokymo metodas, tinkamas dvejetainės klasifikacijos užduotims. Šis algoritmas apskaičiuoja tikimybę, kad pateikta SMS žinutė priklauso sukčiavimo („spam“) klasei, ir pasižymi aiškiai interpretuojama išvestimi. Dėl savo paprastos struktūros ir greito veikimo logistinės regresijos modelis yra tinkamas realaus laiko sistemoms.

Kadangi duomenų rinkinyje pastebimas klasių disbalansas tarp teisėtų ir sukčiavimo žinučių, modelyje taikomi subalansuoti klasių svoriai (class_weight = balanced). Šis sprendimas leidžia kompensuoti klasių disproporciją ir sumažinti riziką, kad modelis bus per daug linkęs prognozuoti dažniau pasitaikančią klasę.

Logistinės regresijos modelio maksimalus iteracijų skaičius nustatytas į 1000, siekiant užtikrinti stabilų optimizavimo proceso konvergavimą. Tai leidžia modeliui pasiekti optimalius svorius net ir esant didelės dimensijos požymių erdvei, kuri susidaro taikant TF-IDF vektorizaciją.

Duomenų rinkinys buvo padalintas į treniruojamąją ir testinę dalis santykiu 70 % / 30 %, siekiant objektyviai įvertinti modelio veikimą su nematytais duomenimis ir kartu išlaikyti pakankamą treniravimo imtį. Toks pasiskirstymas yra plačiai taikomas mašininio mokymo praktikoje.

10 lentelėje pateikti parametrų pasirinkimai atspindi subalansuotą kompromisą tarp modelio tikslumo, skaičiavimo efektyvumo ir praktinio pritaikomumo realaus laiko SMS žinučių grėsmės vertinimo sistemoje.

10 lentelė. Mašininio mokymo modelio parametrizavimo reikšmės

Parametras	**Reikšmė**	**Paaiškinimas**
Vektorizacijos metodas	TF-IDF	Tekstinių požymių išgavimas, įvertinant žodžių svarbą visame duomenų rinkinyje
N-gramų diapozonas	3-5	Analizuojami pavieniai žodžiai ir trumpos žodžių kombinacijos
Minimali dokumentų dažnio riba (min_df)	2	Pašalinami itin reti ir mažai informatyvūs terminai
Maksimali dokumentų dažnio riba (max_df)	0,9	Sumažinama labai dažnų, bet mažai informatyvių žodžių įtaka
Klasifikatorius	Logistinė regresija	Prižiūrimo mokymo algoritmas dvejetainės klasifikacijos užduočiai
Klasės svoriai	Subalansuoti	Kompensuojamas klasių disbalansas duomenų rinkinyje
Duomenų skaidymas	70% / 30%	Duomenų rinkinio padalijimas į treniravimo ir testavimo dalis

3.5. Apibendrinimas

- Šiame skyriuje buvo įgyvendintas sukčiavimo SMS žinutėmis aptikimo metodo prototipas, kuriame praktiškai realizuoti ankstesniuose skyriuose aprašyti teoriniai sprendimai. Prototipas leidžia automatizuotai analizuoti SMS žinutes ir priskirti jas teisėtos, įtartinos arba neteisėtos kategorijoms.

- Darbo metu buvo surinktas, apdorotas ir struktūrizuotas realių SMS žinučių duomenų rinkinys, kuris panaudotas tiek taisyklėmis grįsto metodo kūrimui, tiek mašininio mokymo modelio treniravimui ir testavimui. Duomenų analizė leido identifikuoti dažniausiai pasitaikančius sukčiavimo scenarijus ir kalbinius požymius.

– Buvo sukurtas taisyklėmis grįstas analizės modulis, kuriame apibrėžti konkretūs sukčiavimo požymiai, jų svoriai ir bendro rizikos balo skaičiavimo mechanizmas. Šis metodas leidžia aiškiai interpretuoti sprendimus ir pagrįsti, kodėl konkreti žinutė laikoma rizikinga.

– Papildomai realizuotas mašininio mokymo modelis, paremtas TF-IDF vektorizacija ir logistinės regresijos klasifikatoriumi, kuris leidžia automatiškai įvertinti SMS žinučių turinį ir pateikti tikimybinę sukčiavimo prognozę, nepriklausomai nuo iš anksto apibrėžtų taisyklių.

– Galiausiai, šiame skyriuje buvo sujungti abu analizės metodai į vieningą hibridinį sprendimų priėmimo mechanizmą, kuris leidžia sumažinti klaidingų klasifikacijų tikimybę ir tiksliau identifikuoti potencialiai pavojingas SMS žinutes, sudarant pagrindą tolimesniam eksperimentiniam sistemos vertinimui.

4. Modelio veikimo analizė ir eksperimentinis įvertinimas

Šiame skyriuje pateikiamas sukurtos SMS žinučių klasifikavimo sistemos eksperimentinis įvertinimas. Atliekama kelių mašininio mokymo algoritmų palyginamoji analizė, siekiant nustatyti, kuris modelis geriausiai tinka sukčiavimo žinučių atpažinimo užduočiai. Taip pat vertinamas pasirinkto klasifikatoriaus veikimas bei analizuojamas kombinuoto analizės metodo poveikis galutiniams klasifikavimo rezultatams. Gauti rezultatai leis pagrįsti pasirinkto sprendimo tinkamumą praktiniam taikymui.

4.1. Tyrimo uždaviniai ir vertinimo kriterijai

Šio tyrimo tikslas – įvertinti sukurtos SMS žinučių klasifikavimo sistemos veikimą bei nustatyti, kuris iš nagrinėtų metodų yra tinkamiausias sukčiavimo žinučių atpažinimo užduočiai. Eksperimentinio vertinimo metu analizuojamas skirtingų mašininio mokymo algoritmų veikimas taikant vienodus duomenų paruošimo ir testavimo principus. Taip siekiama objektyviai palyginti modelių gebėjimą atskirti teisėtas ir neteisėtas SMS žinutes bei įvertinti, kuris metodas pasižymi geriausiomis klasifikavimo savybėmis nagrinėjamame duomenų rinkinyje. Be to, tyrime nagrinėjamas ir hibridinis klasifikavimo metodas, kuriame mašininio mokymo modelio rezultatai derinami su taisyklėmis grįstu vertinimu. Toks metodas leidžia įvertinti, ar papildomas taisyklių taikymas gali pagerinti sistemos patikimumą, ypač mažinant klaidingai teigiamų klasifikacijų skaičių.

Eksperimentinio vertinimo metu siekiama atsakyti į šiuos pagrindinius tyrimo klausimus:

- kuris iš taikomų mašininio mokymo algoritmų geriausiai atpažįsta sukčiavimo pobūdžio SMS žinutes nagrinėjamame duomenų rinkinyje;
- ar hibridinis metodas, jungiantis mašininio mokymo modelį ir taisyklėmis pagrįstą analizę, gali sumažinti klasifikavimo klaidų skaičių ir pagerinti sistemos patikimumą praktiniame naudojime.

Modelių veikimas vertinamas taikant standartines dvejetainės klasifikacijos metrikas [37]. Pagrindiniai vertinimo rodikliai yra tikslumas (angl. *accuracy*), preciziškumas (angl. *precision*), atkūrimas (angl. *recall*) ir F1 įvertis (angl. *F1-score*). Tikslumas parodo bendrą teisingai klasifikuotų žinučių dalį visame testavimo rinkinyje, tačiau vien šio rodiklio nepakanka modelio veikimui įvertinti. Preciziškumas leidžia nustatyti, kokia dalis žinučių, pažymėtų kaip neteisėtos, iš tikrųjų yra tokios, todėl šis rodiklis ypač svarbus siekiant sumažinti teisėtų žinučių klaidingą klasifikavimą. Atkūrimo rodiklis parodo, kokią dalį visų realių neteisėtų žinučių sistema sugeba aptikti, o F1 įvertis apjungia preciziškumo ir atkūrimo rodiklius į vieną bendrą balansinį rodiklį.

4.2. Duomenų paruošimas ir vertinimo metodika

Eksperimentiniam vertinimui naudotas SMS žinučių duomenų rinkinys (žr. 3.3 skyrių), kuriame kiekviena žinutė priskirta vienai iš dviejų klasių: teisėta žinutė arba neteisėta (sukčiavimo pobūdžio) žinutė. Prieš atliekant eksperimentus duomenys buvo apdoroti ir suvienodinti – pašalintos trūkstamos reikšmės, suvienodintas žinučių tekstų formatas, o klasės reikšmės konvertuotos į dvejetainį skaitinį pavidalą, naudojamą klasifikavimo algoritmams.

Siekiant objektyviai įvertinti modelių veikimą, duomenų rinkinys buvo padalintas į mokymo ir testavimo dalis. Testavimo rinkinys sudarytas taip, kad jame būtų po 25% kiekvienos klasės žinučių,

todėl galutinis testavimo rinkinys sudarė 50% žinučių. Likusieji duomenys buvo naudojami modelių mokymui.

Svarbus vertinimo metodikos aspektas buvo šablonų pasikartojimo kontrolė. SMS sukčiavimo žinutės dažnai generuojamos pagal pasikartojančius tekstinius šablonus, todėl egzistuoja rizika, kad identiškos ar labai panašios žinutės pateks tiek į mokymo, tiek į testavimo rinkinį. Tokia situacija galėtų dirbtinai pagerinti modelio rezultatus, nes modelis iš esmės matytų jau anksčiau matytus tekstinius modelius. Siekiant to išvengti, kiekvienai žinutei buvo apskaičiuojama normalizuota šablono reprezentacija, kurioje URL adresai, el. pašto adresai, telefonų numeriai ir skaitinės reikšmės pakeičiamos bendrais žymekliais. Formuojant testavimo rinkinį buvo užtikrinta, kad jame esantys šablonai nepatektų į mokymo rinkinį. Tokiu būdu buvo pašalinta vadinamoji duomenų nutekėjimo (angl. *data leakage*) problema.

Be pagrindinio treniravimo ir testavimo padalijimo, modelių stabilumui įvertinti taip pat buvo taikytas penkių dalių kryžminė validacija (angl. *5-fold cross-validation*) [38]. Šio metodo metu mokymo duomenys padalijami į penkias dalis, o modelis kiekvieną kartą treniruojamas naudojant keturias dalis ir vertinamas su likusia dalimi. Toks procesas pakartojamas penkis kartus, kiekvieną kartą keičiant validacijos dalį. Galutinis modelio veikimo įvertinimas apskaičiuojamas imant metrikų vidurkį per visus validacijos kartus. Taip sumažinama atsitiktinio duomenų padalijimo įtaka rezultatams ir gaunamas stabilesnis modelių veikimo įvertinimas.

Eksperimentų pakartojamumui užtikrinti visame eksperimento procese buvo naudojamas fiksuotas atsitiktinumo pradinis parametras (angl. *random seed*). Šis parametras taikytas duomenų maišymui, kryžminei validacijai bei modelių mokymo procesams. Toks sprendimas leidžia užtikrinti, kad pakartotinai vykdant eksperimentus su tais pačiais duomenimis būtų gaunami identiški rezultatai, todėl eksperimentai tampa pakartojami ir lengviau patikrinami.

4.3. Klasifikavimo algoritmų palyginamoji analizė

Siekiant nustatyti, kuris klasifikavimo metodas geriausiai tinka sukčiavimo SMS žinučių atpažinimo užduočiai, eksperimento metu buvo palyginti keli skirtingi mašininio mokymo modeliai. Modelių pasirinkimas buvo pagrįstas ankstesniame darbo skyriuje atlikta mokslinės literatūros analize, kurioje nagrinėjami įvairūs SMS sukčiavimo aptikimo metodai. Literatūroje dažniausiai taikomi tiek klasikiniai mašininio mokymo algoritmai, tokie kaip naivusis Bajeso klasifikatorius ar atsitiktinių miškų metodas, tiek pažangesni giluminio mokymosi modeliai, pavyzdžiui, dvipusiai ilgalaikės atminties neuroniniai tinklai.

Atsižvelgiant į šią analizę, eksperimento metu buvo įvertinti keli skirtingi klasifikavimo metodai: logistinė regresija, naivusis Bajeso klasifikatorius, atsitiktinių miškų metodas (Random Forest), Extra Trees klasifikatorius bei giluminio mokymosi modelis Bi-LSTM (žr. 2 lentelę). Tokia modelių įvairovė leidžia palyginti skirtingų metodologinių principų pagrindu veikiančius algoritmus ir įvertinti jų tinkamumą SMS žinučių klasifikavimo užduočiai.

Klasikiniams mašininio mokymo modeliams tekstinių duomenų reprezentacijai buvo taikomas TF-IDF metodas, naudojant simbolių n-gramų (3–5 simbolių) reprezentaciją. Toks tekstų reprezentavimo būdas leidžia efektyviai išgauti būdingus SMS žinučių struktūrinius ir kalbinius fragmentus, kurie gali būti naudojami klasifikavimo modeliams treniruoti.

Visi modeliai buvo treniruojami naudojant tą patį mokymo duomenų rinkinį ir vertinami su tuo pačiu testavimo rinkiniu, siekiant užtikrinti rezultatų palyginamumą. Modelių veikimas buvo vertinamas naudojant tikslumo (Accuracy), preciziškumo (Precision), atkūrimo (Recall) ir F1 įverčio metrikas.

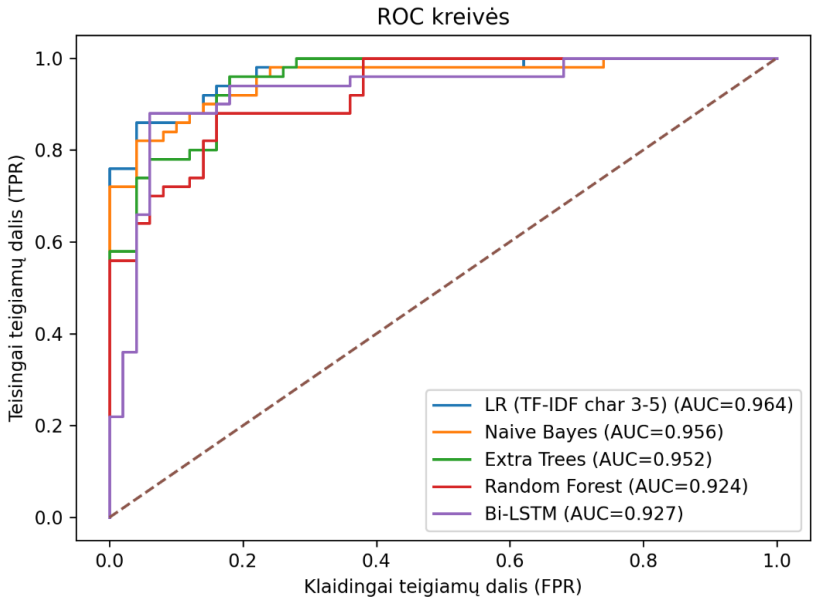
Toliau pateikiamoje 11 lentelėje pateikiami visų testuotų klasifikavimo modelių rezultatai, gauti naudojant testavimo duomenų rinkinį.

11 lentelė. Testuotų klasifikavimo modelių veikimo rodikliai

Naudotas modelis	**Tikslumas (angl. *accuracy*)**	**Preciziškumas (angl. *precision*)**	**Atkūrimas (angl. *recall*)**	**F1 įvertis (angl. *F1 score*)**
Logistic Regression	89%	86,79%	92%	89,32%
Naive Bayes	86%	80%	96%	87,27%
Extra Trees	84%	84%	84%	84%
Random Forest	84%	84%	84%	84%
Bi-LSTM	65%	59,26%	96%	73,28%

Remiantis gautais rezultatais galima pastebėti, kad skirtingi klasifikavimo modeliai pasižymi skirtingomis stiprybėmis. Kai kurie modeliai pasiekia aukštesnį atkūrimo rodiklį, tačiau tuo pačiu sukuria daugiau klaidingai teigiamų klasifikacijų, tuo tarpu kiti modeliai pasižymi didesniu preciziškumu, tačiau aptinka mažesnę dalį sukčiavimo žinučių. F1 įvertis leidžia subalansuotai įvertinti šių dviejų rodiklių santykį ir yra naudojamas kaip vienas pagrindinių modelių palyginimo kriterijų.

Siekiant vizualiai palyginti modelių gebėjimą atskirti teisėtas ir neteisėtas žinutes, papildomai buvo sudarytos ROC (angl. *Receiver Operating Characteristic*) kreivės, kurios pavaizduotos 14 paveiksle, jos parodo teisingai teigiamų klasifikacijų dalies ir klaidingai teigiamų klasifikacijų dalies santykį skirtingais klasifikavimo slenksčiais. Šiame kontekste klasifikavimo slenkstis reiškia modelio prognozuojamos tikimybės ribą, nuo kurios žinutė priskiriama neteisėtų žinučių klasei (pavyzdžiui, jei modelio prognozuota tikimybė viršija 0,5, žinutė laikoma neteisėta).



14 pav. ROC kreivės skirtingiems klasifikavimo modeliams

ROC kreivės leidžia įvertinti, kaip gerai klasifikavimo modelis sugeba atskirti teisėtas ir neteisėtas žinutes nepriklausomai nuo pasirinkto klasifikavimo slenksčio. Modelio veikimas dažnai apibūdinamas naudojant AUC rodiklį (angl. *Area Under the Curve* - plotas po kreive). Kuo ši reikšmė didesnė, tuo geriau modelis geba atskirti skirtingas klases.

Gauti rezultatai rodo, kad vienus geriausių rezultatų tarp testuotų modelių pademonstravo logistinės regresijos modelis su TF-IDF tekstų reprezentacija, kurio ROC kreivės plotas (AUC) yra didžiausias tarp nagrinėtų metodų. Panašius rezultatus pasiekė ir naivaus Bajeso bei Extra Trees modeliai, tačiau jų AUC rodikliai buvo šiek tiek mažesni. Random Forest ir Bi-LSTM modeliai parodė kiek prastesnius rezultatus, nors jų klasifikavimo gebėjimas vis dar išlieka pakankamai aukštas.

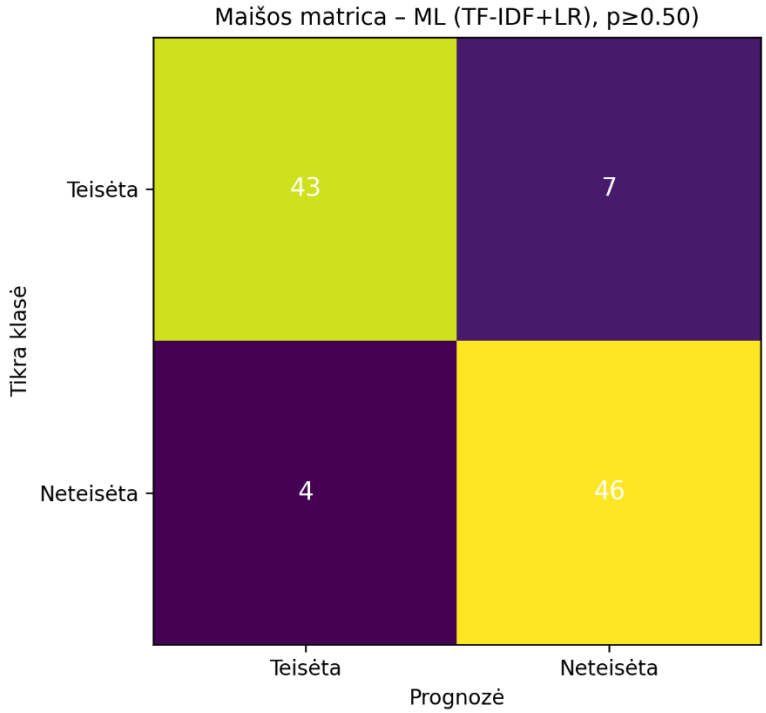
Logistinės regresijos modelis su TF-IDF tekstų reprezentacija šiame darbe buvo pasirinktas kaip pagrindinis klasifikavimo metodas. Eksperimentų rezultatai patvirtina, kad šis modelis yra tinkamas nagrinėjamai užduočiai, nes jis pasižymi geru balansu tarp preciziškumo ir atkūrimo rodiklių bei demonstruoja aukščiausią AUC reikšmę tarp visų testuotų modelių.

4.4. Pasirinkto modelio veikimo charakteristikos

Remiantis ankstesniame skyriuje pateikta modelių palyginamąja analize, šiame darbe pagrindiniu klasifikavimo modeliu pasirinktas logistinės regresijos modelis, kuriame tekstinių duomenų reprezentacijai naudojamas TF-IDF metodas. Šis metodas leidžia kiekvieną SMS žinutę paversti skaitinių požymių vektoriumi, kuris vėliau naudojamas klasifikavimo algoritmui treniruoti.

Tekstų reprezentacijai šiame tyrime buvo taikomi simbolių n-gramai, kurių ilgis svyruoja nuo 3 iki 5 simbolių. Tai reiškia, kad modelis analizuoja visus galimus trijų, keturių ir penkių simbolių fragmentus žinutės tekste. Toks požymių išgavimo būdas leidžia efektyviau aptikti būdingus SMS žinučių struktūrinius fragmentus, pavyzdžiui, nuorodas, numerių struktūras ar dažnai pasitaikančius žodžių junginius. Simbolių n-gramų metodas yra ypač tinkamas trumpiems tekstams, tokiems kaip SMS žinutės, nes leidžia aptikti informatyvius tekstinius modelius net ir esant nedideliam teksto kiekiui.

Siekiant įvertinti pasirinkto modelio klasifikavimo rezultatus, buvo sudaryta maišos matrica (angl. *confusion matrix*), kuri leidžia detaliai analizuoti teisingų ir klaidingų klasifikacijų skaičių. Ją galima pamatyti 15 paveiksle.

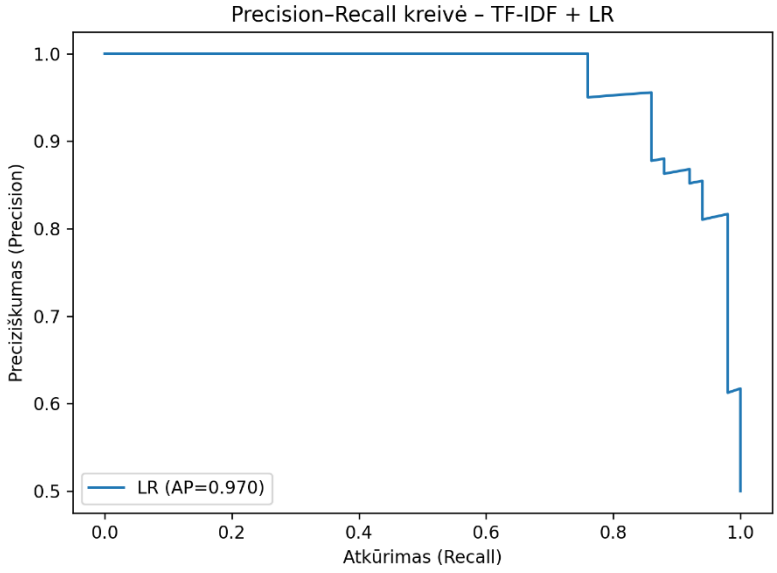


15 pav. Logistinės regresijos modelio maišos matrica

Maišos matrica parodo, kad modelis teisingai klasifikavo 43 teisėtas SMS žinutes ir 46 neteisėtas žinutes. Tuo tarpu 7 teisėtos žinutės buvo klaidingai priskirtos neteisėtų žinučių klasei, o 4 neteisėtos žinutės buvo klaidingai klasifikuotos kaip teisėtos.

Šie rezultatai rodo, kad modelis gana tiksliai atpažįsta sukčiavimo žinutes ir pasižymi aukštu atkūrimo rodikliu. Tačiau dalis teisėtų žinučių vis dar klaidingai pažymimos kaip neteisėtos.

Papildomai buvo sudaryta preciziškumo ir atkūrimo kreivė (angl. *Precision-Recall curve*), pavaizduota 16 paveiksle, leidžianti įvertinti modelio veikimą skirtingais klasifikavimo slenksčiais.

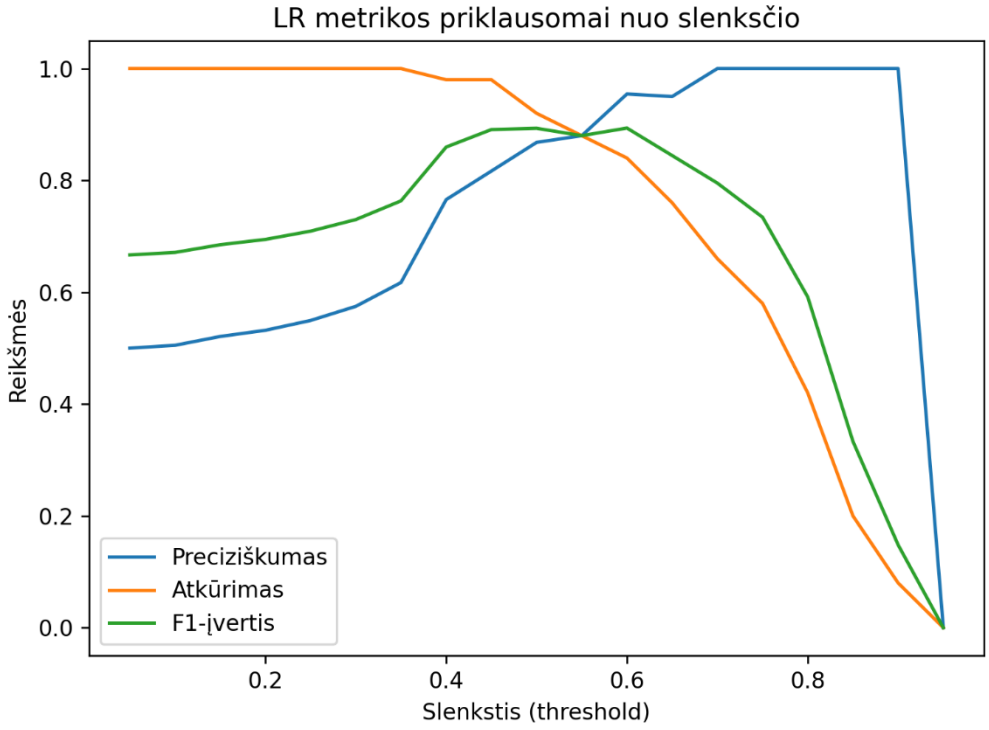


16 pav. Preciziškumo ir atkūrimo kreivė logistinės regresijos modeliui.

Precision-Recall kreivė parodo preciziškumo ir atkūrimo rodiklių santykį skirtingais klasifikavimo slenksčiais. Ši kreivė yra ypač naudinga vertinant klasifikavimo modelius užduotyse, kuriose svarbu sumažinti klaidingai teigiamų klasifikacijų skaičių. Kuo kreivė yra arčiau viršutinio dešiniojo grafiko kampo, tuo geresni modelio rezultatai.

Grafike matomas vidutinio preciziškumo rodiklis AP (angl. *Average Precision*), kuris apibendrina visos Precision-Recall kreivės plotą. Ši reikšmė leidžia įvertinti bendrą modelio gebėjimą išlaikyti aukštą preciziškumo ir atkūrimo balansą skirtingais klasifikavimo slenksčiais. Šiame tyrime gauta AP = 0,970 reikšmė rodo, kad logistinės regresijos modelis pasižymi labai geru gebėjimu atskirti teisėtas ir neteisėtas SMS žinutes.

Siekiant detaliau įvertinti modelio veikimą, taip pat buvo analizuojama, kaip keičiasi pagrindiniai klasifikavimo rodikliai priklausomai nuo pasirinkto klasifikavimo slenksčio. Rodiklių priklausomybė pavaizduota 17 paveiksle.



17 pav. Logistinės regresijos modelio metrikų priklausomybė nuo klasifikavimo slenksčio

Grafike pateikiama, kaip keičiasi pagrindiniai klasifikavimo rodikliai - preciziškumas, atkūrimas ir F1 įvertis - priklausomai nuo pasirinkto klasifikavimo slenksčio. Analizuojant grafiką matyti, kad mažesnių slenksčio reikšmių srityje modelis pasižymi labai aukštu atkūrimo rodikliu, kuris siekia beveik 1,0. Tai reiškia, kad tokiu atveju modelis aptinka beveik visas neteisėtas SMS žinutes, tačiau preciziškumo reikšmė yra santykinai mažesnė, nes dalis teisėtų žinučių taip pat priskiriamos neteisėtų žinučių klasei.

Didėjant slenksčio reikšmei, pastebimas priešingas efektas - preciziškumo rodiklis palaipsniui didėja ir aukštesnių slenksčių srityje artėja prie maksimalios reikšmės. Tai rodo, kad modelis tampa konservatyvesnis ir rečiau pažymi teisėtas žinutes kaip neteisėtas. Tačiau tuo pačiu metu atkūrimo rodiklis mažėja, nes dalis realių neteisėtų žinučių lieka neaptiktos.

Grafike taip pat matyti, kad geriausias preciziškumo ir atkūrimo balansas pasiekiamas vidutinėse slenksčio reikšmėse, kur F1 įvertis pasiekia didžiausią reikšmę. Ši sritis rodo optimalų kompromisą tarp neteisėtų žinučių aptikimo ir klaidingų klasifikacijų skaičiaus. Atsižvelgiant į šią analizę, eksperimento metu pasirinktas 0,5 klasifikavimo slenkstis, kuris užtikrina pakankamai gerą preciziškumo ir atkūrimo balansą.

Tokie rezultatai rodo, kad klasifikavimo slenkstis turi tiesioginę įtaką modelio veikimui ir gali būti pritaikomas priklausomai nuo sistemos naudojimo scenarijaus. Pavyzdžiui, sistemose, kuriose svarbiausia aptikti kuo daugiau sukčiavimo žinučių, galima naudoti mažesnį slenkstį, o sistemose, kuriose svarbiau sumažinti klaidingai pažymėtų teisėtų žinučių skaičių, gali būti pasirinktas didesnis slenkstis.

4.5. Taisyklių metodo veikimo charakteristikos

Prieš vertinant hibridinio metodo efektyvumą buvo atlikta taisyklių metodo slenksčio jautrumo analizė. Taisyklių metodo slenkstis buvo keičiamas, siekiant nustatyti, kokia taisyklių balo reikšmė yra tinkamiausia hibridiniam sprendimui.

Analizė atlikta naudojant tą patį testavimo duomenų rinkinį, kuris buvo taikytas ir ankstesniuose eksperimentuose. Hibridinis metodas vertintas taikant pasirinktą sprendimo taisyklę: žinutė priskiriama neteisėtų žinučių klasei tik tuo atveju, jei tiek mašininio mokymo modelis, tiek taisyklių metodas ją klasifikuoja kaip neteisėtą. Gauti rezultatai pateikiami 12 lentelėje.

12 lentelė. Taisyklių metodo slenksčio jautrumo analizė

Mašininio mokymosi slenkstis	**Taisyklių grįstos analizės balas**	**Tikslumas (angl. accuracy)**	**Preciziškumas (angl. precision)**	**Atkūrimas (angl. recall)**	**F1 įvertis (angl. F1 score)**
0,5	1	92%	93,75%	90%	91,84%
0,5	2	92%	93,75%	90%	91,84%
0,5	3	92%	93,75%	90%	91,84%
0,5	4	92%	93,75%	90%	91,84%
0,5	5	92%	93,75%	90%	91,84%
0,5	6	86%	95%	76%	84,44%
0,5	7	85%	94,87%	74%	83,15%
0,5	8	79%	93,94%	62%	74,70%

Gauti rezultatai rodo, kad taisyklių metodo slenksčiai nuo 1 iki 5 pateikė vienodus geriausius rezultatus pagal tikslumo, atkūrimo ir F1 įverčio rodiklius. Šiuo atveju hibridinis metodas pasiekė 92 % tikslumą, 93,75 % preciziškumą, 90 % atkūrimą ir 91,84 % F1 įvertį. Didinant taisyklių slenkstį nuo 6, pastebimas ryškesnis atkūrimo rodiklio mažėjimas, nes vis daugiau realių neteisėtų žinučių nebeatitinka taisyklių metodo kriterijaus ir todėl nėra priskiriamos neteisėtų žinučių klasei.

Galutiniame hibridiniame metode pasirinktas 5 balų taisyklių slenkstis. Nors 1–5 slenksčių intervale gauti vienodi rezultatai, 5 balų reikšmė pasirinkta kaip konservatyvesnė ir stabilesnė alternatyva. Ji leidžia išlaikyti aukščiausius eksperimento metu gautus vertinimo rodiklius, tačiau kartu sumažina pavienių silpnų taisyklių požymių įtaką galutiniam sprendimui. Toks pasirinkimas yra tinkamesnis

praktiniam taikymui, nes reikalauja stipresnio taisyklių metodo pagrindimo prieš žinutę klasifikuojant kaip neteisėtą.

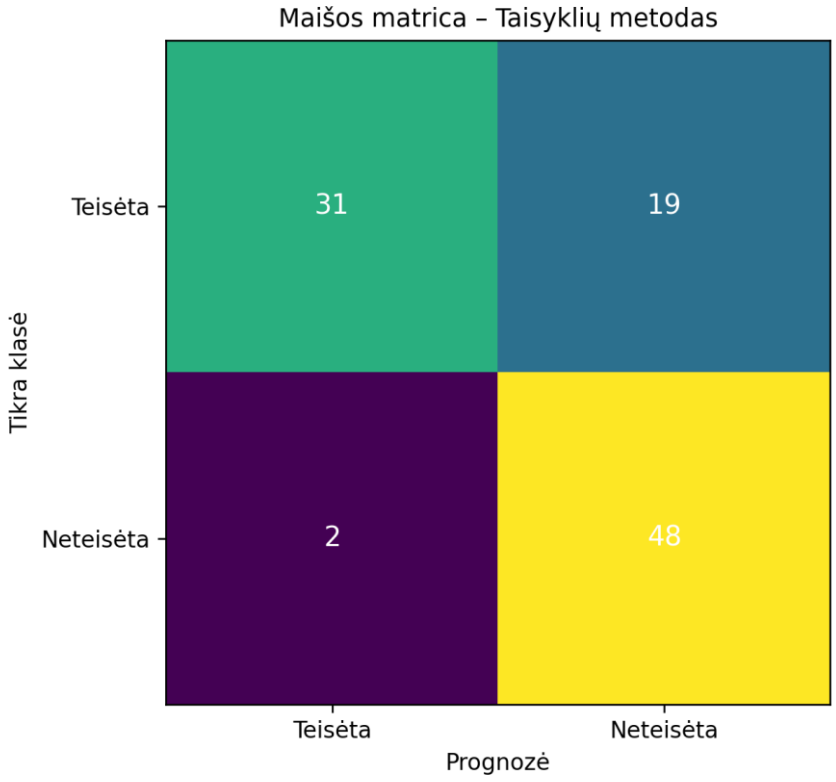
4.6. Hibridinio metodo efektyvumo įvertinimas

Šiame tyrime, be atskirų mašininio mokymo ir taisyklėmis grįstų metodų, taip pat buvo išbandytas hibridinis klasifikavimo metodas. Hibridinio metodo veikimo principas ir jo sudarymo logika detaliau aprašyti ankstesniame darbo skyriuje, kuriame pristatytas sukurtos sistemos architektūrinis sprendimas (žr. 3.4 skyrių). Šio metodo tikslas – sujungti mašininio mokymo modelio ir taisyklėmis grįstos analizės privalumus, siekiant sumažinti klaidingai klasifikuojamų žinučių skaičių ir padidinti sistemos patikimumą.

Vertinant hibridinio metodo veikimą buvo naudojamas tas pats testavimo duomenų rinkinys, kuris buvo taikytas ir atskirų mašininio mokymo modelių vertinimui. Toks sprendimas leidžia tiesiogiai palyginti skirtingų metodų rezultatus ir objektyviai įvertinti, ar hibridinis sprendimas iš tiesų pagerina klasifikavimo sistemos veikimą.

Taisyklių pagrindu veikiantis metodas analizuoja SMS žinučių turinį ir priskiria žinutei tam tikrą balą, priklausomai nuo aptiktų įtartinų požymių, tokių kaip nuorodos, specifiniai raktažodžiai ar kiti sukčiavimo žinutėms būdingi elementai. Jei žinutės balas viršija nustatytą slenkstį, ji priskiriama neteisėtų žinučių klasei.

Siekiant įvertinti taisyklėmis grįsto metodo veikimą, buvo sudaryta atskira maišos matrica, kuri pavaizduota 18 paveiksle.

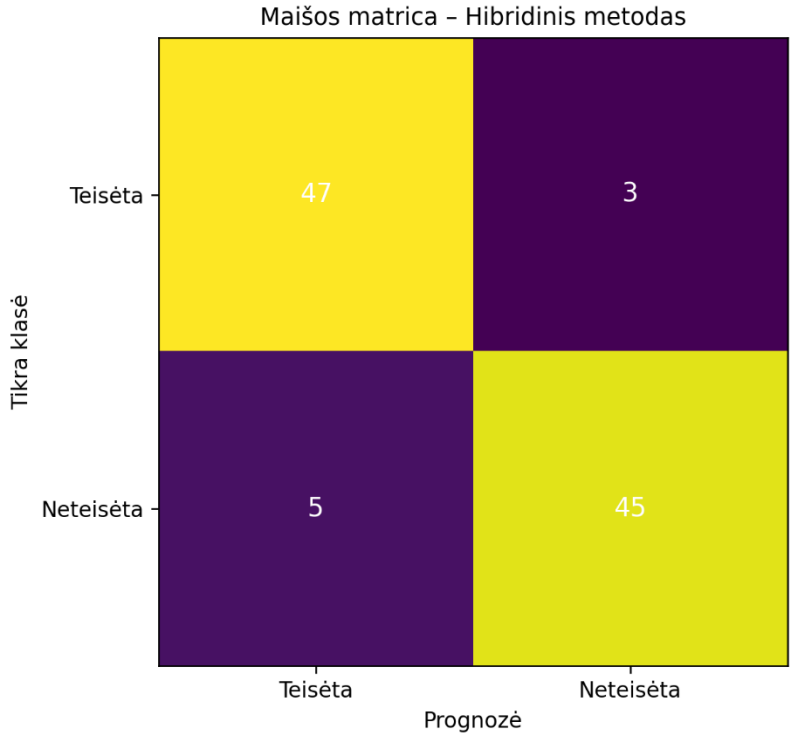


18 pav. Maišos matrica taisyklėmis grįstam metodui

Rezultatai rodo, kad taisyklėmis grįstas metodas teisingai klasifikavo 31 teisėtą žinutę ir 48 neteisėtas žinutes. Tuo tarpu 19 teisėtų žinučių buvo klaidingai priskirtos neteisėtų žinučių klasei, o 2 neteisėtos žinutės liko neaptiktos. Tai rodo, kad taisyklių metodas pasižymi aukštu atkūrimo rodikliu, tačiau kartu sukuria daugiau klaidingai teigiamų klasifikacijų.

Toliau buvo išbandytas sukurtas hibridinis metodas, kuriame mašininio mokymo modelio ir taisyklėmis grįsto metodo rezultatai buvo derinami tarpusavyje. Šiame tyrime taikyta sekanti sprendimo taisyklė: žinutė priskiriama neteisėtų žinučių klasei tik tuo atveju, jei tiek mašininio mokymo modelis, tiek taisyklėmis pagrįstas metodas ją identifikuoja kaip neteisėtą. Jei bent vienas iš metodų žinutę klasifikuoja kaip teisėtą, galutinė klasifikacija laikoma teisėta.

Tokiu būdu siekiama sumažinti klaidingai teigiamų klasifikacijų skaičių, nes žinutė laikoma neteisėta tik tada, kai abu metodai pateikia vienodą sprendimą. Hibridinio metodo rezultatai pavaizduoti 19 paveiksle.



19 pav. Maišos matrica hibridiniam metodui

Hibridinio metodo rezultatai rodo, kad sistema teisingai klasifikavo 47 teisėtas žinutes ir 45 neteisėtas žinutes. Tuo tarpu 3 teisėtos žinutės buvo klaidingai pažymėtos kaip neteisėtos, o 5 neteisėtos žinutės liko neaptiktos. Visų metodų suvestinė pateikiama 13 lentelėje.

13 lentelė. Skirtingų klasifikavimo metodų rezultatų palyginimas pagal maišos matricą

Metodas	**Teisingai klasifikuotos teisėtos žinutės (angl. *true positive*)**	**Klaidingai kaip neteisėtos pažymėtos teisėtos žinutės (angl. *false positive*)**	**Nepastebėtos neteisėtos žinutės (angl. *false negative*)**	**Teisingai aptiktos neteisėtos žinutės (angl. *true positive*)**
Mašininio mokymo metodas	43	7	4	46

Taisyklėmis grįstas metodas	31	19	2	48
Hibridinis metodas	47	3	5	45

Lyginant maišos matricų rezultatus matyti, kad kiekvienas metodas pasižymi skirtingu klaidų pobūdžiu. Mašininio mokymo metodas teisingai aptiko 46 iš 50 neteisėtų žinučių, tačiau 7 teisėtas žinutes klaidingai priskyrė neteisėtų žinučių klasei. Tai rodo, kad modelis gerai aptinka sukčiavimo žinutes, tačiau dalį teisėtų pranešimų vis dar pažymi kaip rizikingus.

Taisyklėmis grįstas metodas aptiko daugiausia neteisėtų žinučių – 48 iš 50, todėl pagal maišos matricą jis yra jautriausias sukčiavimo požymiams. Vis dėlto šis metodas klaidingai pažymėjo 19 teisėtų žinučių kaip neteisėtas. Tai rodo, kad taisyklių metodas veikia agresyviau: jis efektyviai aptinka įtartinus požymius, tačiau vien tik jo taikymas gali sukelti per daug klaidingų perspėjimų.

Hibridinis metodas sumažino šią problemą. Jis teisingai klasifikavo 47 teisėtas žinutes ir 45 neteisėtas žinutes. Klaidingai kaip neteisėtos buvo pažymėtos tik 3 teisėtos žinutės. Lyginant su mašininio mokymo metodu, šis skaičius sumažėjo nuo 7 iki 3. Lyginant su taisyklių metodu, klaidingai teigiamų klasifikacijų skaičius sumažėjo nuo 19 iki 3.

Toks rezultatas rodo, kad hibridinis metodas efektyviai sumažina klaidingų perspėjimų skaičių, nes žinutė laikoma neteisėta tik tada, kai ją kaip neteisėtą identifikuoja abu metodai. Praktiniu požiūriu tai yra svarbu, nes teisėtų SMS žinučių nepagrįstas pažymėjimas kaip sukčiavimo gali mažinti naudotojų pasitikėjimą sistema.

Kartu pastebima, kad hibridinis metodas, palyginti su atskirais metodais, praleidžia šiek tiek daugiau neteisėtų žinučių nei mašininio mokymo ar taisyklių metodas atskirai. Mašininio mokymo metodas nepastebėjo 4 neteisėtų žinučių, taisyklių metodas - 2, o hibridinis metodas - 5. Vis dėlto šis sumažėjimas yra nedidelis, o mainais pasiekiamas ryškus klaidingai teigiamų klasifikacijų sumažėjimas. Todėl hibridinis metodas gali būti laikomas labiau subalansuotu sprendimu praktiniam taikymui.

14 lentelė. Skirtingų klasifikavimo metodų rezultatų palyginimas pagal vertinimo metrikas

Metodas	**Tikslumas (angl. *accuracy*)**	**Preciziškumas (angl. *precision*)**	**Atkūrimas (angl. *recall*)**	**F1 įvertis (angl. *F1 score*)**
Mašininio mokymo metodas	89%	86,79%	92%	89,32%
Taisyklėmis grįstas metodas	79%	71,64%	96%	82,05%
Hibridinis metodas	92%	93,75%	90%	91,84%

Svarbu pažymėti, kad visi metodai šiame eksperimente buvo vertinami naudojant tą patį testavimo duomenų rinkinį. Tai leidžia objektyviai palyginti skirtingų metodų veikimą ir užtikrina, kad gauti rezultatai nėra susiję su skirtingais duomenų rinkiniais.

4.7. Apibendrinimas

– Šiame skyriuje buvo atliktas sukurtos SMS žinučių klasifikavimo sistemos eksperimentinis įvertinimas, kurio metu palyginti keli mašininio mokymo modeliai bei įvertintas hibridinio

klasifikavimo metodo efektyvumas. Visi modeliai buvo vertinami naudojant tą patį 100 SMS žinučių testavimo rinkinį, sudarytą iš 50 teisėtų ir 50 neteisėtų žinučių.

– Klasifikavimo algoritmų palyginimo rezultatai parodė, kad geriausią bendrą rezultatą pasiekė logistinės regresijos modelis su TF-IDF tekstų reprezentacija. Šis modelis pasiekė 89 % tikslumą, 86,79 % preciziškumą, 92 % atkūrimo rodiklį ir 89,32 % F1 įvertį. Jo F1 įvertis buvo apie 2 procentiniais punktais didesnis nei *naivaus Bajeso* modelio (87,27 %) ir daugiau nei 5 procentiniais punktais didesnis nei *Extra Trees* bei *Random Forest* modelių (84 %).

– Detalesnė modelio analizė parodė, kad logistinės regresijos modelis teisingai klasifikavo 89 iš 100 testuotų žinučių. Buvo teisingai atpažintos 43 teisėtos ir 46 neteisėtos žinutės, tačiau 7 teisėtos žinutės buvo klaidingai pažymėtos kaip neteisėtos, o 4 neteisėtos žinutės liko neaptiktos. *Precision-Recall* analizėje gautas aukštas vidutinio preciziškumo rodiklis AP, kuris siekė 0,970.

– Hibridinio metodo taikymas leido pagerinti bendrą klasifikavimo rezultatą ir sumažinti klaidingai teigiamų klasifikacijų skaičių. Hibridinis metodas pasiekė 92 % tikslumą, 93,75 % preciziškumą, 90 % atkūrimo rodiklį ir 91,84 % F1 įvertį. Lyginant su logistinės regresijos modeliu, F1 įvertis padidėjo nuo 89,32 % iki 91,84 %, o klaidingai teigiamų klasifikacijų skaičius sumažėjo nuo 7 iki 3. Lyginant su taisyklių metodu, klaidingai teigiamų klasifikacijų skaičius sumažėjo nuo 19 iki 3.

– Apibendrinant galima teigti, kad logistinės regresijos modelis su TF-IDF tekstų reprezentacija yra efektyvus sprendimas SMS sukčiavimo žinučių aptikimo užduočiai, o papildomas taisyklėmis grįstos analizės integravimas leidžia sumažinti klaidingų perspėjimų skaičių ir padidinti sistemos patikimumą praktinėse SMS filtravimo sistemose.

Išvados

1. Atlikta analizė parodė, kad SMS sukčiavimas yra plačiai paplitusi ir efektyvi socialinės inžinerijos forma, kuriai būdingas reikšmingas naudotojų pažeidžiamumas – tyrimų duomenimis, iki ~34 % naudotojų neteisingai identifikuoja sukčiavimo žinutes, o dalis jų pakartotinai tampa aukomis. Tai patvirtina, kad problema yra ne tik technologinė, bet ir susijusi su žmogiškuoju faktoriumi bei nepakankamu kibernetinio saugumo suvokimu.
2. Mokslinių tyrimų analizė parodė, kad pažangiausi mašininio mokymosi metodai pasiekia labai aukštą tikslumą – pavyzdžiui, giluminio mokymosi modeliai siekia iki ~99,74 % tikslumą, o hibridiniai metodai – apie ~96–99 % tikslumą. Vis dėlto dauguma šių sprendimų yra orientuoti į anglų kalbos duomenis ir nepritaikyti lietuvių kalbos specifikai, todėl jų tiesioginis taikymas lietuvių kalbai yra ribotas ir reikalauja papildomos adaptacijos.
3. Sukurtas sukčiavimo SMS žinučių hibridinis metodas, jungiantis taisyklėmis grįstą analizę ir mašininio mokymosi modelį, leidžia efektyviai išnaudoti abiejų metodų privalumus – taisyklės padeda identifikuoti aiškius sukčiavimo požymius (pvz., nuorodas ar raktinius žodžius), o modelis – sudėtingesnius lingvistinius ir semantinius dėsningumus. Toks derinys leidžia sumažinti klaidingų teigiamų ir klaidingų neigiamų rezultatų skaičių, lyginant su pavieniais metodais.
4. Eksperimentinis vertinimas parodė, kad hibridinis metodas pasiekė geriausią bendrą rezultatą – 92 % tikslumą, 93,75 % preciziškumą, 90 % atkūrimą ir 91,84 % F1 įvertį. Lyginant su vien tik mašininio mokymosi pagrįstu modeliu, hibridinis metodas sumažino klaidingai teigiamų klasifikacijų skaičių nuo 7 iki 3, todėl pasižymi didesniu patikimumu praktiniam taikymui ir geresniu balansu tarp sukčiavimo žinučių aptikimo bei klaidingų perspėjimų mažinimo.
5. Darbo metu buvo konceptualiai apibrėžta duomenų rinkimo ir valdymo metodika, orientuota į bendruomenės įsitraukimą bei realių sukčiavimo atvejų kaupimą, kuri gali būti laikoma pagrindu tolimesniam sistemos vystymui ir praktiniam įgyvendinimui ateityje.
6. Pasiekti rezultatai rodo, kad sukurtas hibridinis metodas geba efektyviai aptikti sukčiavimo SMS žinutes ir sumažinti klaidingų klasifikacijų skaičių. Darbas taip pat pasižymi naujumu, nes orientuojasi į lietuviškų SMS sukčiavimo žinučių analizę, kuriai šiuo metu nėra plačiai taikomų ar išsamiai ištirtų sprendimų.

Literatūros sąrašas

1. CALEB, A. Phishing and Smishing Attacks. In [interaktyvus]. 2021. [žiūrėta 2024-12-02]. Prieiga per internetą: <https://www.researchgate.net/publication/381583123_Phishing_and_Smishing_Attacks>.

2. BLEFARI, F. ir kt. Combining Anti-typosquatting Techniques. In STEFANIDIS, K. ir kt.Sud. *Web Engineering* [interaktyvus]. Cham: Springer Nature Switzerland, 2024. p. 246–254. [žiūrėta 2024-12-03]. ISBN 978-3-031-62361-5. Prieiga per internetą: <https://link.springer.com/10.1007/978-3-031-62362-2_17>.

3. FUKUSHI, N. ir kt. Understanding Security Risks of Ad-based URL Shortening Services Caused by Users' Behaviors. In *Journal of Information Processing* [interaktyvus]. 2022. Vol. 30, no. 0, p. 865–877. [žiūrėta 2024-12-03]. Prieiga per internetą: <https://www.jstage.jst.go.jp/article/ipsjjip/30/0/30_865/_article>.

4. RAHMAN, M.L. ir kt. Users Really Do Respond To Smishing. In *Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy* [interaktyvus]. Charlotte NC USA: ACM, 2023. p. 49–60. [žiūrėta 2024-12-02]. Prieiga per internetą: <https://dl.acm.org/doi/10.1145/3577923.3583640>.

5. BARRERA, D. ir kt. Literature Review of SMS Phishing Attacks: Lessons, Addresses, and Future Challenges. In GUARDA, T. ir kt.Sud. *Advanced Research in Technologies, Information, Innovation and Sustainability* [interaktyvus]. Cham: Springer Nature Switzerland, 2024. p. 191–204. [žiūrėta 2024-12-02]. ISBN 978-3-031-48854-2. Prieiga per internetą: <https://link.springer.com/10.1007/978-3-031-48855-9_15>.

6. TIMKO, D. ir kt. A Quantitative Study of SMS Phishing Detection [interaktyvus]. [s.l.]: arXiv, 2023. [žiūrėta 2024-12-02]. Prieiga per internetą: <https://arxiv.org/abs/2311.06911>.

7. EDWARDS, M. ir kt. SMiShing Attack Vector: Surveying End-User Behavior, Experience, and Knowledge. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* [interaktyvus]. 2023. Vol. 67, no. 1, p. 1911–1915. [žiūrėta 2024-12-03]. Prieiga per internetą: <https://journals.sagepub.com/doi/10.1177/21695067231192193>.

8. BLANCAFLOR, E. ir kt. A Case Study on Smishing: An Assessment of Threats against Mobile Devices. In *Proceedings of the 2023 9th International Conference on Computer Technology Applications* [interaktyvus]. Vienna Austria: ACM, 2023. p. 172–178. [žiūrėta 2024-12-04]. Prieiga per internetą: <https://dl.acm.org/doi/10.1145/3605423.3605446>.

9. SOYKAN, E.U. ir kt. Disrupting the power grid via EV charging: The impact of the SMS Phishing attacks. In *Sustainable Energy, Grids and Networks* [interaktyvus]. 2021. Vol. 26, p. 100477. [žiūrėta 2024-12-04]. Prieiga per internetą: <https://linkinghub.elsevier.com/retrieve/pii/S2352467721000485>.

10. PUTRA, F.P.E. ir kt. Analysis of Phishing Attack Trends, Impacts and Prevention Methods: Literature Study. In *Brilliance: Research of Artificial Intelligence* [interaktyvus]. 2024. Vol. 4, no. 1, p. 413–421. [žiūrėta 2024-12-02]. Prieiga per internetą: <https://jurnal.itscience.org/index.php/brilliance/article/view/4357>.

11. SHARAFF, A. ir kt. Deep learning-based smishing message identification using regular expression feature generation. In *Expert Systems* [interaktyvus]. 2023. Vol. 40, no. 4, p. e13153. [žiūrėta 2024-12-02]. Prieiga per internetą: <https://onlinelibrary.wiley.com/doi/10.1111/exsy.13153>.

12. ULFATH, R.E. ir kt. Detecting Smishing Attacks Using Feature Extraction and Classification Techniques. In AREFIN, M.S. ir kt.Sud. *Proceedings of the International Conference on Big Data, IoT, and Machine Learning* [interaktyvus]. Singapore: Springer Singapore, 2022. p. 677–689. [žiūrėta 2024-12-02]. ISBN 978-981-16-6635-3. Prieiga per internetą: <https://link.springer.com/chapter/10.1007/978-981-16-6636-0_51>.

13. MISHRA, S. - SONI, D. Smishing Detector: A security model to detect smishing through SMS content analysis and URL behavior analysis. In *Future Generation Computer Systems* [interaktyvus]. 2020. Vol. 108, p. 803–815. [žiūrėta 2024-12-02]. Prieiga per internetą: <https://linkinghub.elsevier.com/retrieve/pii/S0167739X19318758>.

14. JAIN, A.K. ir kt. A content and URL analysis-based efficient approach to detect smishing SMS in intelligent systems. In *International Journal of Intelligent Systems* [interaktyvus]. 2022. Vol. 37, no. 12, p. 11117–11141. [žiūrėta 2024-12-02]. Prieiga per internetą: <https://onlinelibrary.wiley.com/doi/10.1002/int.23035>.

15. MEHMOOD, M.K. ir kt. Enhancing Smishing Detection: A Deep Learning Approach for Improved Accuracy and Reduced False Positives. In *IEEE Access* [interaktyvus]. 2024. Vol. 12, p. 137176–137193. [žiūrėta 2024-12-02]. Prieiga per internetą: <https://ieeexplore.ieee.org/document/10684195/>.

16. KRAWCZYK, N. ir kt. Emotion and Phrase-Based Patterns in Smishing: A Feature-Driven Detection Framework. In *Procedia Computer Science* [interaktyvus]. 2025. Vol. 270, p. 4421–4430. [žiūrėta 2026-04-12]. Prieiga per internetą: <https://linkinghub.elsevier.com/retrieve/pii/S1877050925032405>.

17. JAIN, A.K. ir kt. A content and URL analysis-based efficient approach to detect smishing SMS in intelligent systems. In *International Journal of Intelligent Systems* [interaktyvus]. 2022. Vol. 37, no. 12, p. 11117–11141. [žiūrėta 2026-04-12]. Prieiga per internetą: <https://onlinelibrary.wiley.com/doi/10.1002/int.23035>.

18. MAHARANA, K. ir kt. A review: Data pre-processing and data augmentation techniques. In *Global Transitions Proceedings* [interaktyvus]. 2022. Vol. 3, no. 1, p. 91–99. [žiūrėta 2025-02-24]. Prieiga per internetą: <https://linkinghub.elsevier.com/retrieve/pii/S2666285X22000565>.

19. KAPOČIŪTĖ-DZIKIENĖ, J. ir kt. Character-Based Machine Learning vs. Language Modeling for Diacritics Restoration. In *Information Technology And Control* [interaktyvus]. 2017. Vol. 46, no. 4, p. 508–520. [žiūrėta 2025-02-24]. Prieiga per internetą: <http://itc.ktu.lt/index.php/ITC/article/view/18066>.

20. MISHRA, S. - SONI, D. Implementation of ‘Smishing Detector’: An Efficient Model for Smishing Detection Using Neural Network. In *SN Computer Science* [interaktyvus]. 2022. Vol. 3, no. 3, p. 189. [žiūrėta 2024-12-02]. Prieiga per internetą: <https://link.springer.com/10.1007/s42979-022-01078-0>.

21. AKANDE, O.N. ir kt. SMSPROTECT: An automatic smishing detection mobile application. In *ICT Express* [interaktyvus]. 2023. Vol. 9, no. 2, p. 168–176. [žiūrėta 2024-12-02]. Prieiga per internetą: <https://linkinghub.elsevier.com/retrieve/pii/S2405959522000868>.

22. MISHRA, S. - SONI, D. DSmishSMS-A System to Detect Smishing SMS. In *Neural Computing and Applications* [interaktyvus]. 2023. Vol. 35, no. 7, p. 4975–4992. [žiūrėta 2024-12-02]. Prieiga per internetą: <https://link.springer.com/10.1007/s00521-021-06305-y>.

23. TIMKO, D. - RAHMAN, M.L. Commercial Anti-Smishing Tools and Their Comparative Effectiveness Against Modern Threats. In *Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks* [interaktyvus]. Guildford United Kingdom: ACM, 2023. p. 1–12. [žiūrėta 2024-12-02]. Prieiga per internetą: <https://dl.acm.org/doi/10.1145/3558482.3590173>.

24. TIMKO, D. - RAHMAN, M.L. Smishing Dataset I: Phishing SMS Dataset from Smishtank.com. In [interaktyvus]. 2024. [žiūrėta 2024-12-18]. Prieiga per internetą: <https://arxiv.org/abs/2402.18430>.

25. LIAN, R. ir kt. Towards secure and trustworthy crowdsourcing: challenges, existing landscape, and future directions. In *Wireless Networks* [interaktyvus]. 2024. Vol. 30, no. 5, p. 4329–4341. [žiūrėta 2024-12-02]. Prieiga per internetą: <https://link.springer.com/10.1007/s11276-022-03015-8>.

26. GHOSH, S. ir kt. Natural language processing and sentiment analysis: perspectives from computational intelligence. In *Computational Intelligence Applications for Text and Sentiment Data Analysis* [interaktyvus]. [s.l.]: Elsevier, 2023. p. 17–47. [žiūrėta 2025-06-11]. ISBN 978-0-323-90535-0. Prieiga per internetą: <https://linkinghub.elsevier.com/retrieve/pii/B9780323905350000070>.

27. CESARIO, E. ir kt. A survey of the recent trends in deep learning for literature based discovery in the biomedical domain. In *Neurocomputing* [interaktyvus]. 2024. Vol. 568, p. 127079. [žiūrėta 2025-06-11]. Prieiga per internetą: <https://linkinghub.elsevier.com/retrieve/pii/S092523122301202X>.

28. Python. [Interaktyvus]. [žiūrėta 2025-11-15]. Prieiga per internetą: <https://www.python.org/>.

29. Jupyter Notebook. [Interaktyvus]. [žiūrėta 2025-11-15]. Prieiga per internetą: <https://jupyter.org/>.

30. Pandas. [Interaktyvus]. [žiūrėta 2025-11-15]. Prieiga per internetą: <https://pandas.pydata.org/>.

31. NumPy. [Interaktyvus]. [žiūrėta 2025-11-15]. Prieiga per internetą: <https://numpy.org/>.

32. Scikit-learn. [Interaktyvus]. [žiūrėta 2025-11-15]. Prieiga per internetą: <https://scikit-learn.org/stable/>.

33. Matplotlib. [Interaktyvus]. [žiūrėta 2025-11-15]. Prieiga per internetą: <https://matplotlib.org/>.

34. Seaborn. [Interaktyvus]. [žiūrėta 2025-11-15]. Prieiga per internetą: <https://seaborn.pydata.org/>.

35. Facebook. [Interaktyvus]. [žiūrėta 2025-03-18]. Prieiga per internetą: <https://www.facebook.com>.

36. Facebook grupė. [Interaktyvus]. [žiūrėta 2025-03-19]. Prieiga per internetą: <https://www.facebook.com/groups/aferos>.

37. SUJON, K.M. ir kt. Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models. In *Journal of Big Data* [interaktyvus]. 2025. Vol. 12, no. 1, p. 268. [žiūrėta 2026-03-15]. Prieiga per internetą: <https://journalofbigdata.springeropen.com/articles/10.1186/s40537-025-01313-4>.

38. BATES, S. ir kt. Cross-Validation: What Does It Estimate and How Well Does It Do It? In *Journal of the American Statistical Association* [interaktyvus]. 2024. Vol. 119, no. 546, p. 1434–1445. [žiūrėta 2026-03-15]. Prieiga per internetą: <https://www.tandfonline.com/doi/full/10.1080/01621459.2023.2197686>.