Vilniaus universiteto leidykla

Contents lists available at Vilnius University Press

Jaunųjų mokslininkų darbai
2022, vol. 52 (2), pp. 50–59

eISSN 1648-8776
DOI: https://doi.org/10.15388/JMD.2022.2.5

Daugiaklasių duomenų klasifikavimo metodų tyrimas

Researh of Multi-label Data Classification Solutions

Emilija Valujavičiūtė

Vilnius TECH, Fundamentinių mokslų fakultetas, Informacinių sistemų katedra
E. p. e.valujaviciute@gmail.com

Santrauka. Straipsnyje analizuojama, kokią įtaką lietuvių kalba rašytų tekstų, turinčių kelias klases, klasifikavimui turi pasirinktas modelio taikymo būdas. Pristatomas daugiaklasių lietuvių kalba rašytų duomenų klasifikavimo metodų tyrimas, kurio metu atlikta duomenų klasifikavimo metodų taikymo tikslumo lietuvių kalba rašytų daugiaklasių tekstų automatiniam klasifikavimui analizė. Apžvelgiami klasifikavimo metodai, vertinimo kriterijai, jų panaudojimo galimybės ir duomenų paruošimo klasifikavimui principai. Parengus tekstinius duomenis klasifikavimo uždaviniams spręsti, tyrimui buvo suformuoti 44 klasifikatorių deriniai ir atliktas klasifikavimas, naudojant tris skirtingus daugiaklasių duomenų klasifikacijos metodus: kategorijų nustatymo, priklausymo kiekvienai kategorijai ir kategorijų kombinacijos nustatymo. Rezultatai lyginami laiko ir klasifikavimo tikslumo aspektais, nustatant geriausių rezultatų pasiekusius klasifikatorius ir įvardijant naudotų klasifikavimo būdų skirtumus bei privalumus.

Prasminiai žodžiai: daugiaklasis klasifikavimas, lietuvių kalba, daugiaklasiai tekstiniai duomenys, teksto klasifikacija, kategorijų nustatymo metodas, priklausymo kiekvienai kategorijai metodas, kategorijų kombinacijos nustatymo metodas.

Summary. The article analyzes the impact of the chosen method of model application on the classification of multi-label texts written in the Lithuanian language. The article presents a study of mult-label data classification methods in Lithuanian, which includes an analysis of the accuracy of the application of data classification methods for the automatic classification of multiclass texts written in Lithuanian. The classification methods, evaluation criteria, their applicability and the principles of data preparation for classification are reviewed. After preparing the text data for classification tasks, 44 combinations of classifiers were formed for the study and classification was performed using 3 different methods of multi-label data classification: category detection, category membership and category combination detection. The results obtained are compared in terms of time and classification accuracy, identifying the best performing classifiers and identifying the differences and advantages of the classification methods used.

Keywords: multi-label classification, the Lithuanian language, multi-label text data, text classification, category detection method, category membership method, category combination detection method.

Received: 2023-05-06. **Accepted:** 2023-11-13
Copyright © 2022 Emilija Valujavičiūtė. Published by Vilnius University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Įvadas

Šiuo metu vis daugiau duomenų teikiama ne popieriniu, bet skaitmeniniu formatu. Skaitmenizacijos procesas populiarėja dėl galimybės apdoroti informaciją įvairiomis programomis ir greitai gauti rezultatus, taip sutaupant dirbančiojo laiką. Šio proceso neįmanoma įsivaizduoti be sistemingo duomenų apdorojimo[1]. Didėjant tvarkomų duomenų kiekiams, taip pat sudėtingėja ir jų raktinis, t. y. svarbiausių informacinių laukų, apdorojimas. Dėl šios priežasties automatinis duomenų tvarkymas dar sparčiau keičia popierinį formatą.

Automatinis duomenų apdorojimas apima ir duomenų klasifikavimą – procesą, kai duomenims pagal jų tipą ar tam tikrą reikšmę yra priskiriama klasė, padedanti sudaryti lygiaverčių duomenų aibes tolesniam darbui su tekstu. Tačiau ganėtinai dažnai vienas įrašas gali priklausyti ir kelioms kategorijoms. Pavyzdžiui, klasifikuojant filmus pagal žanrus, vienas filmas gali būti priskirtas ir romantiniam, ir komediniam žanrui. Tokiais atvejais reikia taikyti įvairius sprendimus, kurie leistų įvertinti kelioms kategorijoms ir būtent kokioms įrašas (duomenų rinkinio vienetas) priklauso.

Tyrimo objektas – automatinio teksto klasifikavimo sprendimai, skirti kelių klasių priskyrimui vienam įrašui.

Tyrimo tikslas – atlikti duomenų klasifikavimo metodų taikymo tikslumo lietuvių kalba rašytų daugiaklasių tekstų automatiniam klasifikavimui tyrimą.

Tyrimo uždaviniai:

1. Susipažinti su tekstinių duomenų klasifikavimo metodais ir nustatyti jų taikymo galimybes.
2. Apžvelgti daugiaklasių duomenų klasifikavimui naudojamus sprendimus.
3. Parengti tyrimo metodiką, skirtą įvertinti lietuvių kalba rašytų daugiaklasių tekstų klasifikavimo tikslumą, taikant skirtingus sprendimus.
4. Įvertinti lietuvių kalba rašytų tekstinių daugiaklasių duomenų klasifikavimo rezultatus ir pateikti išvadas.

Tyrimo metodai – sisteminė literatūros apžvalga, lyginamoji analizė ir modeliavimas.

Klasifikavimo kriterijai

Atliekant bet kokį klasifikavimą – daugiaklasį ar ne – klasifikavimo kriterijai leidžia įvertinti nuspėjamąjį modelį. Klasifikacijos atveju, metrikos lygina numatomą (angl. *expected*) klasės etiketę su spėjama (angl. *predicted*) arba interpretuoja numatomas klasei priskirtų etikečių tikimybes (angl. *label*).

Vertinimo priemonės atlieka lemiamą vaidmenį vertinant klasifikavimo efektyvumą ir modeliuojant klasifikatorių. Jas galima suskirstyti į tris grupes: 1) slenkstines metrikas (angl. *threshold*), kurios apibrėš jautrumo, specifiškumo, preciziškumo, priminimo, tikslumo ir klaidingumo kriterijus, 2) reitingo metrikas (angl. *ranking*), kurios apims IVC ir

[1] Duomenų apdorojimas – tai neapdorotų duomenų konvertavimas į mašininio skaitymo formą. Duomenų srautas yra perduodamas per centrinį procesorių ir atmintį į išvesties įrenginius ir išvesties formatavimą arba transformavimą. Bet kokios operacijos, atliekamos su duomenimis naudojantis kompiuteriu, gali būti apibrėžtos „duomenų apdorojimo“ terminu (Data Processing | Definition & Facts, 2021).

PP kreives, ir 3) tikimybės metrikas (angl. *probability*), kurioms priskiriamos kryžminės entropijos ir Brierio balo charakteristikos.

- *Tikslumas* yra teisingų spėjimų ir visų spėjimų santykis, šiai metrikai atvirkštinis kriterijus – *klaidingumas*, nusakantis neteisingų spėjimų ir visų spėjimų santykį.
- *Jautrumas* nurodo tikrąjį teigiamą rodiklį ir apibendrina, kaip gerai buvo prognozuojama teigiama klasė. *Specifiškumas*, atvirkščiai jautrumui, apibendrina kaip gerai buvo prognozuojama neigiama klasė.
- *Preciziškumas* nurodo teigiamą klasei priskirtų pavyzdžių, priklausančių teigiamai klasei, dalį.
- *Priminimas* apibendrina kaip gerai buvo prognozuojama teigiama klasė.
- Imtuvo veikimo charakteristikų, arba *IVC kreivė*, apibendrina studijų lauką, skirtą analizuoti dvejetainius klasifikatorius pagal jų gebėjimą atskirti klases. Šios kreivės alternatyva – preciziškumo ir priminimo, arba *PP kreivė*, kuri padeda įvertinti mažumos klasės klasifikatorių.
- *Kryžminė entropija* yra viena iš dažniausiai naudojamų metrikų, siekiant įvertinti spėjamas tikimybes, kurios rezultatas apibendrina vidutinį skirtumą tarp dviejų tikimybių skirstinių: blogesnių klasifikatorių, kurių įverčiai yra teigiami ir artėja link begalybės, ir puikių klasifikatorių, kurių rezultatai lygūs nuliams.
- Kitas kriterijus, vertinantis spėjamas tikimybes, yra *Brierio balas*, kurio privalumas yra galimybė susitelkti į teigiamą nesubalansuotoje klasifikacijoje mažumą atstovaujančią klasę (Brownlee, 2021).

Pasirinkti tinkamą klasifikavimo kriterijų gali būti sudėtinga, dėl to iš pradžių yra svarbu išnagrinėti duomenų rinkinį, su kuriuo žadama dirbti, ir žinoti darbo tikslą – kokie duomenys ir rodikliai yra reikalingi bei kokie tarpusavio ryšiai yra tarp klasių.

Egzistuojantys dokumentų klasifikavimo algoritmai ir klasifikavimo metodai

Teksto klasifikacijos klausimas per pastaruosius dešimtmečius buvo plačiai nagrinėjamas ir sprendžiamas realiose programėlėse, naudojant tris pagrindinius algoritmus: *Rocchio,* didinimo ir maišymo algoritmus.

- *Rocchio* algoritmas iš dokumentų mokymo rinkinio sukuria vektoriaus prototipą kiekvienai klasei – šis prototipas yra tam tikrai klasei priklausančių mokymo dokumentų vektorių vidurkis. Tuomet algoritmas kiekvieną bandomąjį dokumentą priskiria tai klasei, kurioje yra didžiausias panašumas tarp bandomojo dokumento ir kiekvieno iš prototipo vektorių.
- *Didinimas* – tai viena iš populiariausių ansamblinio mokymosi algoritmų technikų. Iš pradžių dvimačių duomenų rinkinyje duomenys yra sužymimi (angl. *labeled*), tada apmokomi naudojant daugiamodelines (angl. *multi-model*) architektūras, kurios žinomos kaip ansamblinis mokymas (Kowarski et al, 2019).
- *Maišymas* – tai antroji populiariausia ansamblinio mokymosi algoritmų technika. Naudojant didinimo algoritmą, generuojamas analogiškas (angl. *uniform*) pavyzdys,

pvz., iš mokymo rinkinio. Jei turima N pradinių didinimo pavyzdžių B_1, B_2, ..., B_N, tuomet gaunama N klasifikatorių (C), kurių C_i yra sudaryti iš kiekvieno didinimo pavyzdžio B_i. Tada maišymo klasifikatoriuje C yra arba yra sukuriami C_1, C_2, ..., C_N pavyzdžiai, kurių išvestis yra klasė, ją dažniausiai prognozuoja jo subklasifikatoriai su savavališkai laužomais ryšiais (angl. *ties broken arbitrarily*).

Dažniausiai klasifikavimo uždaviniui spręsti yra naudojami devyni klasifikatoriai:

1. *Logistinė regresija* (angl. *logistic regression*) yra statistinis klasifikavimo modelis, geriausiai tinkantis dvejetainei klasifikacijai ir nusakantis tikimybę, jog stebėjimas priklauso konkrečiai klasei (Garg, 2018).
2. *K-artimiausi kaimynai* (angl. *K-Nearest neighbours*) yra tingaus mokymosi klasifikatorius, kuris klasifikavimo uždavinį sprendžia K artimiausių kiekvieno taško kaimynų balsų dauguma (ten pat).
3. *Atraminių vektorių mašina* (angl. *support vector machine*) yra prižiūrimo mašininio mokymo metodas, kuris apibrėžia sprendimo ribą kartu su maksimalia riba, leidžiančia beveik visus taškus atskirti į dvi klases (Bento, 2021a).
4. *Naiviojo Bajeso* (angl. *naive Bayes*) klasifikatorius yra paremtas Bajeso teorema ir yra apmokomas analizuojant mokomųjų duomenų, priskirtų konkrečioms klasėms, rinkinį (University of Helsinki, 2021).
5. *Sprendimų medis* (angl. *decision tree*) yra indukcijos ir genėjimo procesais paremtas hierarchinių žingsnių, kurie padeda priimti tam tikrus sprendimus, rezultatas (Team, 2021).
6. *Atsitiktinis miškas* (angl. *random forest*) yra meta-įvertis, kuris pritaiko daugybę sprendimų medžių įvairiuose duomenų rinkinių pavyzdžiuose ir naudoja vidurkį tam, kad pagreitintų nuspėjamąjį modelio tikslumą ir valdytų perteklinį pritaikymą (Garg, 2018).
7. *Stochastinis gradiento nusileidimas* (angl. *stochastic gradient descent*) yra metodas, kuris, ieškodamas efektyvaus būdo pasiekti minimalią funkcijos reikšmę, išrenka modelį, geriausiai atitinkantį mokymo duomenis (Bento, 2021b).
8. *Daugiasluoksnis perceptronas* (angl. *multilayer perceptron*) yra neuroninis tinklas, kuris išmoksta ryšį tarp linijinių ir nelinijinių duomenų (Bento, 2021c).
9. *Dirbtinių neuroninių tinklų* (angl. *artificial neural network*) klasifikatoriaus veikimo principą nusako neuronai, kurie yra prijungti prie įvesties sluoksnio ir induktyviai mokosi modelių iš mokymosi duomenų tam, kad būtų galima sudaryti bandomųjų duomenų spėjimus (Burns & Burke, 2021).

Šie klasifikatoriai pasižymi sudėtingais algoritmais, dėl to skaičiavimai neretai būna itin ilgi ir reikalaujantys nemažai laiko sprendimams. Tačiau, neskaitant didelių sąnaudų, klasifikatoriai naudojami keturioms pagrindinėms užduotims: nuotaikai analizuoti (tekstui ar jo segmentui priskiriama nuotaika / emocija), klasifikuoti elektroninio pašto šiukšles, klasifikuoti dokumentus ir klasifikuoti vaizdus (Wolff, 2020).

Tyrimo metodologija ir rezultatai

Duomenų klasifikacijai naudoti metodai: 1) tekstinių duomenų paruošimui naudotos penkios technikos (jos aptartos plačiau žemiau), 2) daugiaklasių duomenų klasifikavimo uždaviniui spręsti pasirinkta lyginamoji keturiasdešimt keturių kombinacijų, sudarytų iš aštuonių klasifikatorių, analizė ir trys klasifikavimo metodai.

Tyrimo duomenys. Tyrimui atlikti pasirinktas daugiaklasių lietuvių kalba rašytų duomenų rinkinys, kurį sudaro iš viešos prieigos svetainių surinkti su finansais susijusių žinių tekstiniai duomenys. Šiems duomenims pagal teksto tematiką priskiriama viena arba dvi kategorijos iš dešimties galimų, t. y. *finansai, inovatyvumas, kolektyvas, pandemija, patikimumas, plėtra, politika, pramonė, tarptautinis* ir *teisėsauga*. Duomenų rinkinį iš viso sudaro 12 485 duomenų eilutės ir 13 stulpelių (*tekstas, kategorija 1, kategorija 2* ir kiekvienos konkrečios kategorijos stulpelis), iš kurių 12 484 yra įrašai. Stulpelyje *Tekstas* pateikiami tekstiniai duomenys, kur ilgiausią tekstą sudaro 2 122 simboliai, trumpiausią – vos 19, o vidutinis įrašo ilgis yra ~269 simboliai. Duomenų rinkinyje priskiriamų įrašams kategorijų pasiskirstymas nėra tolygus: daugiausiai kartų priskiriama kategorija *Finansai* – 4 517 kartų, o mažiausiai – kategorija *Politika*, t. y. 814 kartų. Vidutiniškai viena kategorija turi ~1 894 ją atitinkančius įrašus.

Tekstinių duomenų paruošimas tyrimui. Tekstinių duomenų sutvarkymas apima procesą, kurio rezultatas yra suprantamas programinių įrangų ir jos gali su tuo rezultatu dirbti. Tyrimui atlikti naudotos penkios technikos, siekiant paruošti duomenis darbui su jais:

1) *teksto vertimas į mažąsias raides* išsprendžia neteisingo interpretavimo problemą, kai skirtingas raidžių dydis (mažosios ar didžiosios) lemia tų pačių žodžių deklaravimą kaip skirtingų kintamųjų;
2) *žodžių lematizacija* (angl. *lemmatization*) yra viena iš svarbesnių technikų dirbant su lietuvių kalbos tekstiniais duomenimis, leidžianti skirtingų linksnių ir laikų žodžius paversti į pagrindines formas;
3) *skyrybos bei ne raidinių ir ne skaitinių simbolių šalinimo* technika, garantuojanti, jog skyrybos ženklai neturės įtakos saugant kintamuosius ir šie nebus deklaruojami kaip skirtingi dėl skyrybos ar ne raidinių / ne skaitinių simbolių;
4) *stabdymo žodžiai* – dažnai pasikartojantys tarnybiniai žodžiai (*ir, kad, bet, o, su*), turintys didžiausius svorius žodžių, kurių šalinimas panaikina jų daromą įtaką teksto klasifikacijoje, sąrašuose;
5) *tokenizacija* (angl. *tokenization*) yra technika, kai tekstas iš vientisos struktūros kintamojo yra padalijamas į atskirus eilutės (angl. *string*) tipo elementus (žodžius), kurie sujungiami į sąrašo (angl. *list*) kintamąjį, sudarytą būtent iš jų (University of Sydney, 2022).

Pasiruošimas tyrimui. Daugiaklasių duomenų klasifikavimo uždaviniui spręsti iš viso buvo pasirinkti aštuoni klasifikatoriai, ir jiems pritaikius skirtingus nustatymus buvo gautos 44 kombinacijos, su kuriomis atliktas klasifikavimo tyrimas.

Pirmasis buvo *naiviojo Bajeso (NB)* klasifikatorius, kuriam nebuvo taikytos nustatymų kombinacijos ir palikti numatytieji parametrai. Antrasis – *atraminių vektorių mašinų*

(AVM) klasifikatorius, kuriam buvo sudarytos trys kombinacijos, keičiant jo branduolio tipus: 1) atraminės vektorių mašinos su linijiniu (angl. *linear*) branduoliu, 2) atraminės vektorių mašinos su daugianariu (angl. *poly*) branduoliu ir 3) atraminės vektorių mašinos su *rbf* (angl. *radial basis function* – spindulinio pagrindo funkcija) branduoliu. Trečiasis klasifikatorius – *sprendimų medis (SM)*, kuriam, kaip ir NB klasifikatoriui, nepritaikyti parametrų keitimai ir buvo palikti numatytieji nustatymai. Ketvirtasis klasifikatorius – *logistinė regresija (LR)*, kuriai buvo sudarytos trys kombinacijos, pritaikant skirtingas daugiaklasiškumo (angl. *multiclass*) parinktis. Pirmoji buvo logistinei regresijai pritaikius *vienas prieš visus* (angl. *one versus rest, OvR*) daugiaklasiškumo reikšmę, antroji – pritaikius automatinę reikšmę, trečioji – pritaikius daugianario (angl. *multinomial*) reikšmę. Penktasis klasifikatorius – *atsitiktinis miškas (AM)*. Šiam klasifikatoriui pritaikyta dešimt kombinacijų, siekiant išbandyti skirtingų kriterijų veikimą su skirtingais vertintojų (arba medžių) kiekiais, kur kriterijus buvo parenkamas kaip Gini indeksas arba kaip entropija, o kiekiai būdavo nustatomi 10, 20, 50, 100 arba 200. Šeštasis klasifikatorius – *k artimiausi kaimynai (KAK)*, kurių parametruose keičiant artimiausių kaimynų kiekį į 5, 2, 1, 10, 15, 20, 50 ir 100 buvo sudarytos aštuonios kombinacijos klasifikavimo uždaviniui įgyvendinti. Septintajam klasifikatoriui – *stochastinis gradiento nusileidimas (SGN)* – sudarytos šešios kombinacijos, parametruose keičiant nuostolio funkciją (*Modified huber, Epsilon-insensitive* arba *Hinge*) ir keičiant baudos funkciją (*L2* arba *Elasticnet*). Paskutinis, aštuntasis, klasifikatorius – *daugiasluoksnis perceptronas (DP)*[2]. Su šiuo klasifikatoriumi sudaryta dvylika DP kombinacijų, kuriose paslėptų sluoksnių skaičius (angl. *hidden layer size*) buvo arba 150 100 50, arba 100 100 100, parinkta viena iš *relu, logistic* ir *tanh* aktyvacijos funkcijų ir išrinktas vienas iš *adam* ir *lbfgs* sprendimų algoritmų.

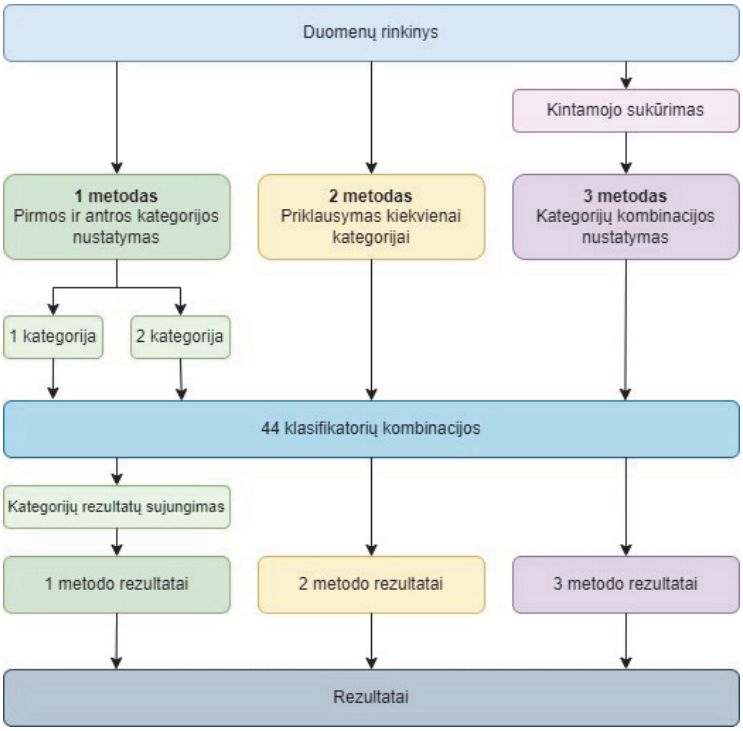
Naudojant šias keturiasdešimt keturias klasifikatorių kombinacijas ir siekiant atlikti daugiaklasių duomenų klasifikavimą (žr. 1 pav.), numatyti trys metodai su kiekviena kombinacija: 1) kiekvienos kategorijos atskiras klasifikavimas[3], kai buvo nustatomi kategorijose geriausią tikslumą pasiekiantys klasifikatoriai, 2) abiejų kategorijų klasifikavimas naudojant vieną klasifikatorių, t. y. kiekvienos kategorijos atveju buvo nustatytas to klasifikatoriaus tikslumas ir 3) klasifikavimas sujungus abi kategorijas į vieną kintamąjį ir pritaikius vieną klasifikatorių.

Galiausiai, siekiant palyginti metoduose gautus klasifikavimo rezultatus, pirmojo metodo rezultatai per skirtingas kategorijas naudojant tą patį klasifikatorių buvo sujungti pagal formulę (1):

1 kategorijos klasifikavimo tikslumas*2 kategorijos klasifikavimo tikslumas 100 (l)

[2] Daugiasluoksnis perceptronas yra dirbtinio neuroninio tinklo klasė arba, kitaip tariant, – tai supaprastintas dirbtinio neuroninio tinklo modelis. Atlikus klasifikavimo uždavinį, naudojant DP, nebuvo gauti geriausi rezultatai, dėl to nuspręsta dirbtinių neuroninių tinklų, kaip atskiro klasifikatoriaus, nebandyti.

[3] Pirmojo metodo rezultatai per skirtingas kategorijas naudojant tą patį klasifikatorių buvo sujungti sudauginus kategorijose gautus klasifikavimo tikslumus ir juos padalinus iš 100, kad būtų gauti realūs klasifikavimo tikslumo skaičiai, atsižvelgiant, jog duomenyse naudojamos ne viena, o dvi kategorijos. Šio skaičiavimo nereikėjo pritaikyti antrajam ir trečiajam metodui, kadangi antrajame metode klasifikatorius identifikuoja kiekvienos galimos klasės tikslumą ir išveda pasvertąjį vidurkį, o trečiajame metode sudaromi nauji kintamieji, kuriuose užfiksuojamos visos galimos kategorijų parinkimo kombinacijos.



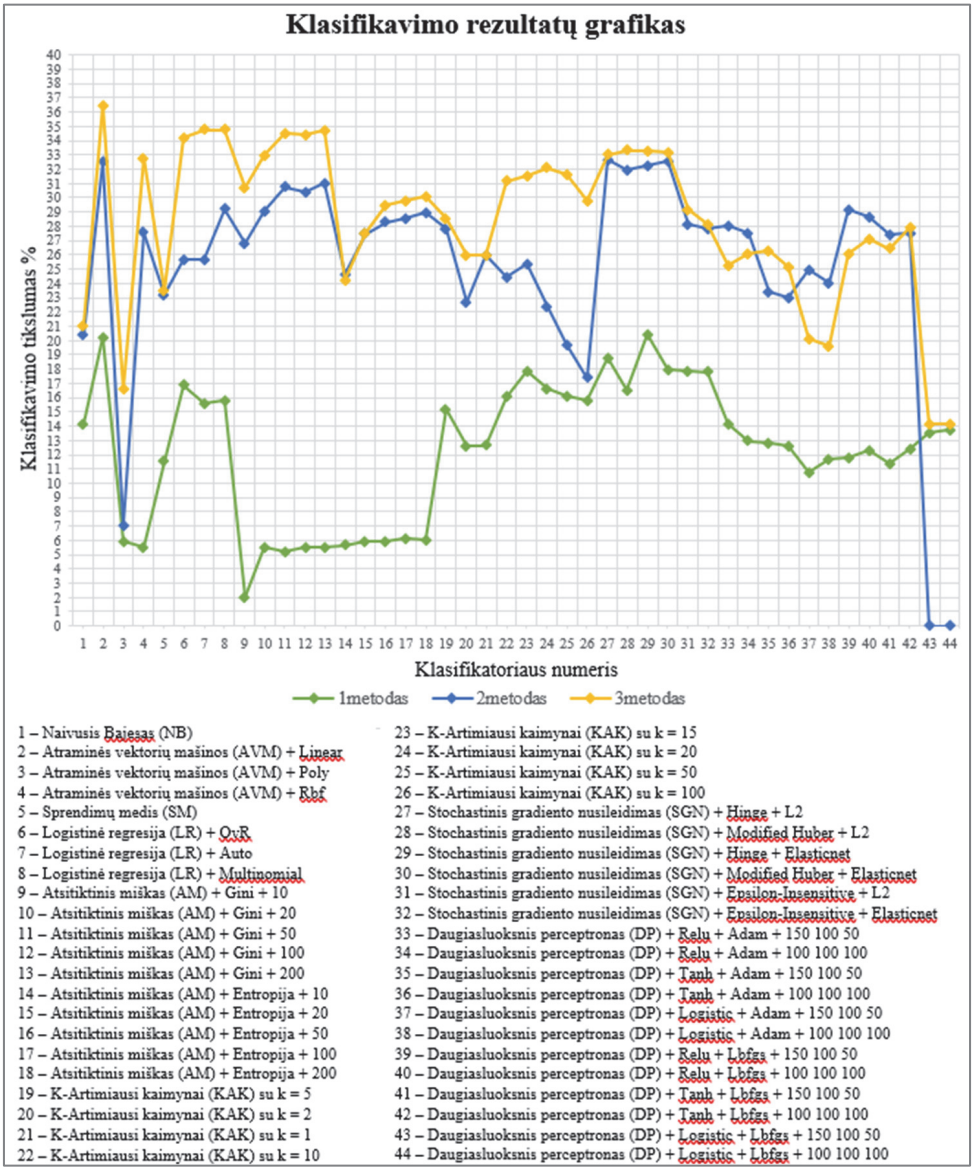
1 pav. **Tyrimo atlikimo schema**

Toks skaičiavimas atliktas siekiant gauti realius klasifikavimo tikslumo skaičius atsižvelgiant, jog duomenyse naudojamos ne viena, o dvi kategorijos. Šio skaičiavimo nereikėjo pritaikyti antrajam ir trečiajam metodui, kadangi antruoju metodu klasifikatorius identifikuoja kiekvienos galimos klasės tikslumą ir išveda pasvertąjį vidurkį, o trečiuoju metodu sudaromi nauji kintamieji, kuriuose užfiksuojamos visos galimos kategorijų parinkimo kombinacijos.

Tyrimo metodų rezultatai

Atlikus klasifikavimą su keturiasdešimt keturiomis klasifikatorių kombinacijomis, naudojant tris skirtingus metodus, buvo gauti rezultatai, iš kurių pagal klasifikavimo tikslumus buvo nubraižyta palyginamoji diagrama (žr. 2 pav.).

Kaip galima matyti 2 paveiksle, geriausią rezultatą pasiekia trečiojo metodo atraminių vektorių mašinų su linijiniu branduoliu klasifikatorius, kurio tikslumas yra 36,47 %, o veikimo laikas – 12 minučių ir 45 sekundės. Blogiausiai klasifikavimą atliko antrojo metodo daugiasluoksnio perceptrono klasifikatorius su logistine aktyvacija ir *lbfgs* optimizatoriumi neatsižvelgiant į sluoksnių kiekį – šio klasifikatoriaus kombinacijos klasifikavimo tikslumas buvo nulinis.



2 pav. **Metodų klasifikavimo tikslumo rezultatų palyginimas**

Iš 2 paveikslo duomenų matyti, jog trečiojo metodo klasifikatoriai pasiekė geresnių rezultatų nei pirmojo ar antrojo metodų, iš kurių pirmasis metodas turėjo mažiausius klasifikavimo tikslumus. Tačiau, čia verta pastebėti, jog pirmasis metodas veikė trumpiausiai – ilgiausiai veikęs trečiojo metodo klasifikatorius buvo daugiasluoksnio perceptrono su logistine aktyvacija, *adam* optimizatoriumi ir 150 100 50 sluoksniais, kurio trukmė buvo 56 sekundės. Tuo tarpu antrajame metode ilgiausiai veikė daugiasluoksnis percep-

tronas su *relu* aktyvacijos funkcija, *lbfgs* optimizatoriumi ir 150 100 50 sluoksniais – jo trukmė buvo 24 minutės ir 15 sekundžių, o trečiajame metode daugiausiai laiko prireikė atraminių vektorių mašinų su daugiasluoksniu branduoliu klasifikatoriui – jis veikė 15 minučių ir 46 sekundes.

Išvados

1. Tekstinių duomenų klasifikavimui galima naudoti įvairius algoritmus ir klasifikatorius, tačiau algoritmais grįstoje klasifikacijoje dažnai prarandamos interpretacijos, kadangi požymių svarbos nėra visapusiškai atskleidžiamos. Klasifikatoriuose dažniausiai pasikartojanti problema yra sudėtingi skaičiavimai, kuriems atlikti reikalingi ir dideli laiko resursai.
2. Dažniausiai daugiaklasių duomenų klasifikavimo uždaviniui spręsti yra naudojami naiviojo Bajeso, stochastinio gradiento nusileidimo, K artimiausių kaimynų, sprendimų medžio, atsitiktinio miško, logistinės regresijos ir atraminių vektorių mašinos klasifikatoriai.
3. Daugiaklasiams tekstiniams duomenims klasifikuoti gali būti naudojami skirtingi metodai: kiekvienos kategorijos nustatymo metodas, priklausymo kiekvienai kategorijai metodas ir kategorijų kombinacijos nustatymo metodas.
4. Kategorijų nustatymo (pirmajame) metode klasifikatoriai nepasiekia aukštų klasifikavimo tikslumų, tačiau jų veikimo laikas yra trumpiausias. Priklausymo kiekvienai kategorijai (antrajame) metode klasifikatoriai pasiekia aukštesnių klasifikavimo tikslumų nei pirmajame metode, tačiau nepagerina trečiojo metodo tikslumų. Vis dėlto antrojo metodo laiko sąnaudos yra mažesnės nei trečiojo. Kategorijų kombinacijos nustatymo (trečiajame) metode klasifikatoriai pasiekia geriausių klasifikavimo tikslumų, tačiau jų laiko resursai yra didžiausi.
5. Jeigu atliekant klasifikavimo uždavinį yra svarbus tikslumas, geriausiai būtų rinktis kategorijų kombinacijos nustatymo metodą, tačiau jei tikslumas nėra svarbus ir svarbiau atlikti sprendimą realiu laiku, tada geriau rinktis kategorijų nustatymo metodą.
6. Stochastinio gradiento nusileidimo su praradimo funkcija *Hinge* ir baudos funkcija *Elasticnet* arba *L2* klasifikatorius ir atraminių vektorių mašinos su linijiniu branduoliu klasifikatorius visuose trijuose metoduose yra geriausiai (arba vieni iš geriausiai) veikiančių klasifikatorių su šiuo lietuvių kalba rašytų daugiaklasių duomenų rinkiniu.
7. Blogiausiai lietuvių kalba rašytus daugiaklasius duomenis apdoroja atsitiktinio miško pirmajame metode ir daugiasluoksnio perceptrono su logistine aktyvacija ir *lbfgs* optimizatoriumi trečiajame metode klasifikatoriai. Taip pat aukštais rezultatais nepasižymi atraminių vektorių mašinų su *poly* branduoliu klasifikatorius. Dėl šios priežasties tolesniems šio duomenų rinkinio tyrimams nebūtų rekomenduotinos šios klasifikatorių kombinacijos (ar klasifikatoriai).

8. Lietuvių kalba rašytų tekstinių duomenų klasifikavimo uždavinio sprendimą apsunkina lietuvių kalboje esantys linksniai ir bibliotekų, pagal kurias atliekamas lematizavimas, nebuvimas arba jų mažumas / siaurumas. Siekiant gauti tikslesnius klasifikavimo rezultatus, darbe naudotą duomenų rinkinį būtų galima išversti į anglų kalbą ir, pritaikius tekstinių duomenų paruošimo klasifikacijai technikas, atlikti patį klasifikavimą.

Literatūra

1. Bento, C., 2021a, *Support Vector Machines explained with Python examples*. Medium. https://towardsdatascience.com/support-vector-machines-explained-with-python-examples-cb65e8172c85
2. Bento, C., 2021b, Stochastic Gradient Descent explained in real life - Towards Data Science. Medium. https://towardsdatascience.com/stochastic-gradient-descent-explained-in-real-life-predicting-your-pizzas-cooking-time-b7639d5e6a32
3. Bento, C., 2021c, Multilayer Perceptron Explained with a Real-Life Example and Python Code: Sentiment Analysis. Medium. https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-sentiment-analysis-cb408ee93141
4. Brownlee, J., 2021, Tour of Evaluation Metrics for Imbalanced Classification. Machine Learning Mastery. https://machinelearningmastery.com/tour-of-evaluation-metrics-for-imbalanced-classification/
5. Burns, E., & Burke, J., 2021, What is a neural network? Explanation and examples. SearchEnterpriseAI. https://www.techtarget.com/searchenterpriseai/definition/neural-network
6. Data processing | Definition & Facts, 2021, *Encyclopedia Britannica*. https://www.britannica.com/technology/data-processing
7. Garg, R., 2018, 7 Types of Classification Algorithms. Developers Corner. https://analyticsindiamag.com/7-types-classification-algorithms/
8. Kowsari, K., Meimandi, K. J., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D., 2019, Text Classification Algorithms: A Survey. *Information*, *10*(4), 150. https://doi.org/10.3390/info10040150
9. Team, D., 2021, Machine Learning Classification – 8 Algorithms for Data Science Aspirants. DataFlair. https://data-flair.training/blogs/machine-learning-classification-algorithms/
10. The University of Sydney, 2022, Text and data mining. Library of the University of Sydney. https://libguides.library.usyd.edu.au/text_data_mining/cleaning
11. University of Helsinki, 2021, Naive Bayes classification. Elements of AI. https://course.elementsofai.com/3/3
12. Wolff, R., 2020, 5 Types of Classification Algorithms in Machine Learning. MonkeyLearn Blog. https://monkeylearn.com/blog/classification-algorithms/