Prekybos tinklų prekių klasifikavimas naudojant skenavimo duomenis

Ieva Marija Noreikaitė, Audrius Sabaitis, Rūta Levulienė

Vilniaus universitetas, Matematikos ir informatikos fakultetas,
Taikomosios matematikos institutas, Naugarduko g. 24, Vilnius, Lietuva
ieva.noreikaite@mif.stud.vu.lt, audrius.sabaitis@stat.gov.lt,
ruta.levuliene@mif.vu.lt

Santrauka. Skenavimo duomenų klasifikavimas pagal COICOP 2018 klasifikatorių yra svarbus etapas siekiant panaudoti šiuos duomenis vartotojų kainų indeksų skaičiavimams, todėl dėl didelės duomenų apimties aktualu sumažinti rankinio klasifikavimo poreikį. Šiame darbe analizuoti Lietuvos didžiųjų prekybos centrų skenavimo duomenys, kurie buvo paversti į vektorinę reprezentaciją taikant TF – IDF metodą, ir tada taikyti logistinės regresijos, atraminių vektorių bei fastText klasifikatoriai, atliktas jų palyginimas ir tinkamumas šiems duomenims klasifikuoti.

Raktiniai žodžiai: klasifikavimas, skenavimo duomenys, COICOP, atraminių vektorių klasifikatorius, logistinės regresijos klasifikatorius, fastText klasifikatorius.

1 Įvadas

Vartotojų kainų indeksas (toliau – VKI) yra vienas svarbiausių makroekonominių rodiklių, naudojamas kainų pokyčiams skaičiuoti bei infliacijos lygiui įvertinti. Šis rodiklis padeda vertinti šalies ekonominę situaciją ir priimti tinkamus sprendimus. Lietuvoje nuo 2026 metų pradžios į VKI skaičiavimą įtraukti prekybos centrų skenavimo duomenys, kurie pasižymi itin dideliu kiekiu trumpų tekstinių įrašų – prekių pavadinimų aprašymų ar jų santrumpų. Norint panaudoti šiuos duomenis VKI skaičiavimui, kiekvienas įrašas turi būti priskirtas atitinkamai individualaus vartojimo išlaidų pagal paskirtį (angl. *Classification of Individual Consumption According to Purpose*) (toliau – COICOP) 2018 klasifikatoriaus kategorijai, todėl svarbu užtikrinti tinkamą parengimą efektyvesniam šių duomenų panaudojimui.

COICOP – tarptautinis klasifikatorius, skirtas namų ūkio vartojimo ir paslaugų išlaidoms klasifikuoti. Minėtasis klasifikatorius sudarytas hierarchiniu pavidalu iki 5-ojo lygmens (žr. 1 lentelę) [1].

Nors įvairūs teksto klasifikavimo metodai yra plačiai ištirti įvairiose srityse [2], [3], skenavimo duomenų klasifikavimas pagal COICOP klasifikatorių

DOI: https://doi.org/10.15388/LMITT.2026.20

yra dar nauja ir nepakankamai išnagrinėta sritis, ypatingai Lietuvoje. Esamuose tyrimuose taikomi ir tradiciniai mašininio mokymo [4], ir *„state-of-the-art“* metodai [5], [6], tačiau nėra sutarta dėl vieno tinkamo metodo.

Šio darbo tikslas – nustatyti, kuris modelis geriau suklasifikuoja Lietuvos prekybos centrų skenavimo duomenis pagal COICOP 2018 klasifikatorių. Gauti rezultatai aktualūs sumažinant rankinį duomenų klasifikavimą bei prisideda prie skenavimo duomenų efektyvesnio panaudojimo.

1 lentelė. COICOP 2018 klasifikatoriaus pavyzdys.

COICOP 2018 kodas	Kodo lygmuo	Pavadinimas
01	1	Maistas ir nealkoholiniai gėrimai
01.1	2	Maistas
01.1.6	3	Vaisiai ir riešutai (ND)
01.1.6.4	4	Šviežios uogos
01.1.6.4.3	5	Šviežios avietės

2 Duomenys ir metodologija

Tyrime naudojami Valstybės duomenų agentūros gauti prekybos tinklų skenavimo duomenys. Klasifikavimui naudojami įrašai apibrėžiantys prekės pavadinimą, aprašymą, kategorijos aprašymą lietuvių bei originalo (anglų, vokiečių) kalbomis.

Duomenys padalinti į mokymo (75 %) ir testavimo (25 %) aibes, kur mokymo aibę sudaro 70626 įrašai, o testavimo – 23542.

Darbe duomenys į vektorinę reprezentaciją paversti naudojant TF – IDF (angl. *Term Frequency-Inverse Document Frequency*) [7] metodą, o vėliau taikomi atraminių vektorių (angl. *support vector machine*) (AVK) [8] ir fastText [9] klasifikatoriai. Logistinės regresijos klasifikatorius (LR) [10] pasirinktas baziniu metodu palyginimui. Geriausi parametrai tradiciniams klasifikavimo modeliams parinkti mokymo aibėje naudojant parametrų gardelę su kryžmine validacija atsižvelgiant į F1 metriką. Gauta, kad logistinės regresijos klasifikatoriui optimaliausias parametras C = 2 su reguliarizacija (angl. *penalty*) = „l2“ ir sprendimo algoritmu (angl. *solver)* = „saga“ , atraminių vektorių C = 2 su numatytais parametrais. fastText parametrai parinkti pačiame modelyje – mokymosi greitis (angl. *learning rate*) = 0.2, epochų skaičius (angl. *epoch)* = 25, maksimalus žodžių N–gramų ilgis (angl. *wordNgrams)* = 2. Modelių gerumas vertinamas testavimo aibėje jautrumo, preciziškumo

bei F1 metrikomis COICOP skyriaus lygmeniui. F1 rodiklis apibrėžiamas kaip jautrumo ir preciziškumo harmoninis vidurkis ir yra tinkantis nevienodai pasiskirsčiusiems duomenims vertinti.

3 Rezultatai

Modelių metrikų reikšmės buvo apskaičiuotos testavimo aibei. Gauta (žr. 2 lentelę), kad jautrumas svyruoja nuo 0,7955 iki 0,9992, preciziškumas nuo 0,8530 iki 1, o F1 reikšmės svyruoja nuo 0,8839 iki 0,9657.

2 lentelė. Klasifikatorių metrikos COICOP skyriams.

Skyrius	Metodas	Preciziškumas	Jautrumas	F1
01	LR	0,9341	0,9991	**0,9657**
	AVK	**0,9642**	**0,9992**	**0,9657**
	fastText	0,9640	0,9986	0,9652
02	LR	**0,9983**	**0,9033**	0,9484
	AVK	0,9980	0,9027	**0,9487**
	fastText	**0,9983**	0,9027	0,9481
03	LR	**1**	**0,8150**	**0,8980**
	AVK	**1**	**0,8150**	**0,8980**
	fastText	0,9994	**0,8150**	0,8978
05	LR	0,9944	0,7955	0,8839
	AVK	**0,9958**	0,7989	**0,8865**
	fastText	0,9861	**0,8011**	0,8840
06	LR	0,9342	0,9726	0,9530
	AVK	0,9351	**0,9863**	**0,9600**
	fastText	**1**	0,9041	0,9496
07	LR	**1**	**0,9048**	**0,9500**
	AVK	0,9828	**0,9048**	0,9421
	fastText	**1**	**0,9048**	**0,9500**
08	LR	**1**	0,9333	**0,9655**
	AVK	**1**	**0,9667**	0,9631
	fastText	0,8530	**0,9667**	0,9063
09	LR	0,9952	0,9046	0,9477
	AVK	**1**	0,9024	**0,9487**
	fastText	0,9929	**0,9067**	0,9478
13	LR	0,9942	**0,9311**	0,9616
	AVK	**0,9954**	**0,9311**	**0,9622**
	fastText	**0,9954**	**0,9311**	**0,9622**

Analizuojant modelių tikslumus, pastebėta, kad didžiausias jautrumas daugeliu atvejų (6 atvejai) yra pasiektas taikant atraminių vektorių klasifikatorių, o didžiausios preciziškumo reikšmės buvo pasiektos panašų sykį kartų visų klasifikatorių. Taip pat daugiausia atvejų didžiausia F1 reikšmė pasiekta taikant AVK. Nepaisant to, daugeliu atvejų, visi trys modeliai pasiekdavo panašius rezultatus, dažnai su šimtųjų skirtumu, todėl galutinis modelio pasirinkimas gali priklausyti nuo turimų skaičiavimo resursų.

Vertinant rezultatus pagal COICOP skyrius, pastebima, jog prasčiausi rezultatai 05 (Būsto Apstatymo, Namų Ūkio Įranga ir Kasdienė Namų Priežiūra) skyriaus, geriausi – 01 (Maistas ir Nealkoholiniai Gėrimai). Taip gali būti dėl nevienodo įrašų kiekio tarp skyrių (01 – 14889 įrašai, o 05 – 885) [11] ir skirtingo tekstinių įrašų informatyvumo bei ilgio. Atsižvelgiant į tai, galima teigti, kad modelių rezultatams įtakos gali turėti nevienodas duomenų pasiskirstymas bei jų kokybė, todėl norint taikyti klasifikatorius prasmingai analizei, naudinga mokymo aibėje turėti įvairesnius bei informatyvesnius duomenis COICOP skyrių atžvilgiu.

4 Išvados

Atsižvelgiant į gautus rezultatus galima teigti, kad geriausiai veikė atraminių vektorių klasifikatorius, su jautrumu nuo 0,7989 iki 0,9992 ir F1 metrika nuo 0,8865 iki 0,9657. Vis dėlto, nustatyta, kad daugeliu atvejų rezultatai tarp klasifikatorių yra panašūs, todėl galutinis modelio pasirinkimas priklauso ir nuo turimų skaičiavimo resursų. Taip pat pastebėta, kad klasifikavimo kokybei įtakos turi duomenų pasiskirstymas pagal skyrius ir tekstinių įrašų informatyvumas, todėl siekiant geresnių rezultatų, tikslinga praplėsti duomenų aibės įvairovę COICOP skyriaus lygmens atžvilgiu.

Literatūra

[1] Oficialiosios statistikos portalas. Individualaus vartojimo išlaidų pagal paskirtį klasifikatorius (COICOP). https://osp.stat.gov.lt/individualaus-vartojimo-islaidu-pagal-paskirti-klasifikatorius-coicop , žiūrėta 2026-04-18

[2] A. Palanivinayagam, C. Z. El-Bayeh & R. Damaševičius. (2023). Twenty Years of Machine Learning-Based Text Classification: A Systematic Review. Algorithms, 16(5), 236. https://doi.org/10.3390/a16050236

[3] Kowsari, K., Jafari Meimandi, K., Heidarysafa, M., Mendu, S., Barnes, L., & Brown, D. (2019). Text Classification Algorithms: A Survey. Information, 10(4), 150. https://doi.org/10.3390/info10040150

[4] D. Muller, B. Toth and S. Jentoft. (2022). From Manual to Machine: challenges in machine learning for COICOP coding. Nordic Statistical Meeting, Process and Analyze. https://static1.squarespace.com/static/606f36b890215d7048ddaac0/t/62ed23ac0fefd90598690723/1659708333795/FROM+MANUAL+TO+MACHINE+-+CHALLENGES+IN+MACHINE+LEARNING+FOR+COICOP+CODING.pdf

[5] L. Benedikt, C. Joshi, L. Nolan, N. de Wolf N and B. Schouten. (2020). Optical character recognition and machine learning classification of shopping receipts. Report. HBS An app-assisted approach for the Household Budget Survey. https://ec.europa.eu/eurostat/documents/54431/11489222/6+Receipt+scan+analysis.pdf

[6] T. Seimandi, T. Leroy, L. Malherbe & E. Coudin. (2022). Machine learning for coding occupations in the Census: first lessons from experiments to production. https://www.researchgate.net/publication/361638454_Machine_learning_for_coding_occupations_in_the_Census_first_lessons_from_experiments_to_production

[7] C. Sammut, G. I. Webb. (2011). TF–IDF. Encyclopedia of Machine Learning. https://doi.org/10.1007/978-0-387-30164-8_832

[8] C. Cortes, V. Vapnik. Support-vector networks. *Mach Learn* **20**, 273–297 (1995). https://doi.org/10.1007/BF00994018

[9] A. Joulin, E. Grave, P. Bojanowski, T. Mikolov, Bag of Tricks for Efficient Text Classification https://arxiv.org/abs/1607.01759

[10] P. McCullagh and J. A. Nelder. *Generalized linear models*. Vol. 37. CRC press, 1989. https://jhanley.biostat.mcgill.ca/bios602/b-d-II-ch-1-2-3/GLM-McCullagh-Nelder-toc.pdf

[11] J. Leevy, T. Khoshgoftaar, R. Bauder and N. Seliya. (2018). A survey on addressing high-class imbalance in big data. Journal of Big Data. 5. 10.1186/s40537-018-0151-6.