Avtomatizacija kategoriziranja obstoječih učinkov uporabe odprtih podatkov glede na opise primerov uporabe

Nejc Čelik, Aljaž Ferencek
Univerza v Mariboru, Fakulteta za organizacijske vede
nejc.celik1@um.si, aljaz.ferencek1@student.um.si

Izvleček

Odprti podatki (OP) predstavljajo pomemben vir javno dostopnih podatkov, ki izhajajo iz javnega sektorja. Osrednji cilj OP je omogočanje transparentnosti, odgovornosti in ustvarjanje dodane vrednosti. Z naraščanjem količine podatkov, ki jih ustvarja javni sektor, rastejo tudi prizadevanja za zagotavljanje njihove dostopnosti javnosti. Raziskave kažejo, da so OP dostopni javnosti in tudi uporabljeni na področju ekonomije, kjer podjetja uporabljajo poslovno inteligenco v kompleksnem globalnem gospodarstvu. Vendar pa ekonomske koristi predstavljajo le en vidik učinka OP. Prepoznavanje in kvantificiranje učinka OP je oteženo zaradi njegove posredne narave. Študije, ki prepoznavajo učinek OP, obsegajo predhodne ocene iz anket, ki so omejene s strani osebja in financiranja za dejavnosti, povezane z OP. Izziv torej leži v prepoznavanju učinkov OP, za kar v literaturi zasledimo predloge uporabe tehnik podatkovnega rudarjenja in umetne inteligence. Namen te raziskave je potrditi že prepoznana področja učinkov OP s strani Evropske komisije in usmeriti nadaljnje raziskave s predlogom novih področij učinkov. V raziskavi smo se ravnali po metodi CRISP-DM, uporabili pa smo različne modele strojnega učenja za klasifikacijo primerov uporabe OP. Rezultati kažejo na potencial umetne inteligence pri prepoznavanju učinkov OP, a je potrebno izdelati končno in podrobnejšo taksonomijo prepoznanih področij učinka. Raziskava je prepoznala nove kategorije uporabe OP, ki bi lahko prispevale k bolj natančni in uporabni klasifikaciji učinkov uporabe OP.

Ključne besede: odprti podatki, podatki javnega sektorja, umetna inteligenca, nevronske mreže

Automating the Categorization of Existing Open Data Impacts Based on Use Case Descriptions

Abstract

Open Government Data (OGD) represents an important source of publicly accessible data originating from the public sector. The primary goal of OGD is to enable transparency, accountability, and the creation of added value. With the increasing volume of data generated by the public sector, there is a strong effort to ensure its accessibility to the public. Research shows that OGD is accessible to the public and also used in the field of economics, where companies utilize business intelligence in a complex global economy. However, economic benefits represent only one of the aspects of the impact of OGD. Recognizing and quantifying the impact of OGD is challenging due to its indirect nature. Studies assessing the impact of OGD include preliminary estimates from surveys, which are limited by staff and funding for OGD-related activities. The challenge lies in recognizing the impact of OGD, for which the literature suggests using data mining and artificial intelligence techniques. The purpose of this research is to confirm the already recognized areas of OGD impact by the European Commission and to guide further research with the proposal of new impact areas. The research followed the CRISP-DM method and utilized various machine learning models to classify OGD use cases. The results indicate the potential of artificial intelligence in recognizing the impacts of OGD, however, there is a need to develop a final and more detailed taxonomy of identified impact areas. The research identified new categories of OGD use that could contribute to a more precise and useful classification of OGD impacts.

Keywords: open data, open government data, artificial intelligence, neural networks

1 UVOD

Odprti podatki (OP) predstavljajo pomemben vir javno dostopnih podatkov, ki izhajajo iz javnega sektorja. Osrednji cilj OP je omogočanje transparentnosti, odgovornosti in ustvarjanje dodane vrednosti [1]. V zadnjih letih smo priča znatnemu porastu produkcije in analize podatkov v javnem sektorju. Ta trend je privedel do občutnega povečanja raziskav na področju odprtih podatkov [2], [3], [4]. Z naraščanjem količine podatkov, ki jih ustvarja javni sektor, rastejo tudi prizadevanja za zagotavljanje njihove dostopnosti javnosti. Ta prizadevanja so skladna s širšim dolgoročnim ciljem, ki je izboljšanje splošne transparentnosti vlade [5], [6].

Iz literature je moč zaznati, da so OP dostopni javnosti in tudi uporabljeni, kot na primer na področju ekonomije, saj podjetja vse bolj izkoriščajo odprte podatke in uporabljajo metode poslovne inteligence za poslovanje v kompleksnem globalnem gospodarstvu [7]. Čeprav se ekonomske koristi morda lažje kvantificirajo, vseeno predstavljajo le en vidik prednosti, ki jih ponujajo OP [8]. Zapletenost prepoznavanja in kvantificiranja učinka OP je še dodatno otežena zaradi posredne narave koristi, ki jih OP prinašajo [9]. Poleg tega študije, ki ocenjujejo učinek OP, večinoma obsegajo predhodne ocene, pridobljene iz anket [10]. Medtem ko anketne ocene ponujajo koristne vpoglede, so rezultati ali njihova koristnost omejeni s strani osebja in financiranja za dejavnosti povezanimi z odprtimi podatki na strani vladnih služb, saj javni uslužbenci pogosto prevzemajo druge, bolj prioritetne projekte [11].

Izziv torej leži v prepoznavanju učinka odprtih podatkov, za reševanje katerega pa Ferencek in Kljajić Borštnar [12], [13], [14], [15] predlagata uporabo tehnik podatkovnega rudarjenja in umetne inteligence na primerih uporabe, ki so objavljeni s strani Urada za publikacije Evropske unije [16]. Njune raziskave zaenkrat kažejo na potencial uporabe tehnik umetne inteligence za prepoznavanje učinka OP, a je pred vsesplošno uporabo predlaganih pristopov potrebno izdelati taksonomijo prepoznanih področij OP, ki pa se lahko razlikuje ali sovpada s področji učinka, ki se v anketah članic Evropske Unije (EU) uporabljajo za izdelavo Ocene zrelosti odprtih podatkov [17]. Slednja se izvaja za merjenje napredka evropskih držav pri spodbujanju in omogočanju razpoložljivosti in ponovne uporabe informacij javnega sektorja, zajema pa štiri razsežnosti zrelosti odprtih podatkov: politike (stopnja razvoja nacionalnih politik), ki spodbujajo odprte podatke; portali (značilnosti in podatki, ki so na voljo na nacionalnih podatkovnih portalih); kakovost (metapodatkov na nacionalnih podatkovnih portalih) in učinki (pobude za spremljanje ponovne uporabe in učinka odprtih podatkov) [18]. Ker v tej raziskavi preučujemo učinek OP, smo se zato posebej osredotočili na razsežnost »učinki«, ki spremlja pobude za spremljanje ponovne uporabe in učinka odprtih podatkov. Omenjena razsežnost v Oceni zrelosti odprtih podatkov, glede na OECD (Organisation for Economic Co-operation and Development) [17] definira štiri glavna področja učinka, ki so družbeno (angl. social), okoljsko (angl. environmental), vladno (angl. governmental) in ekonomsko (angl. economic).

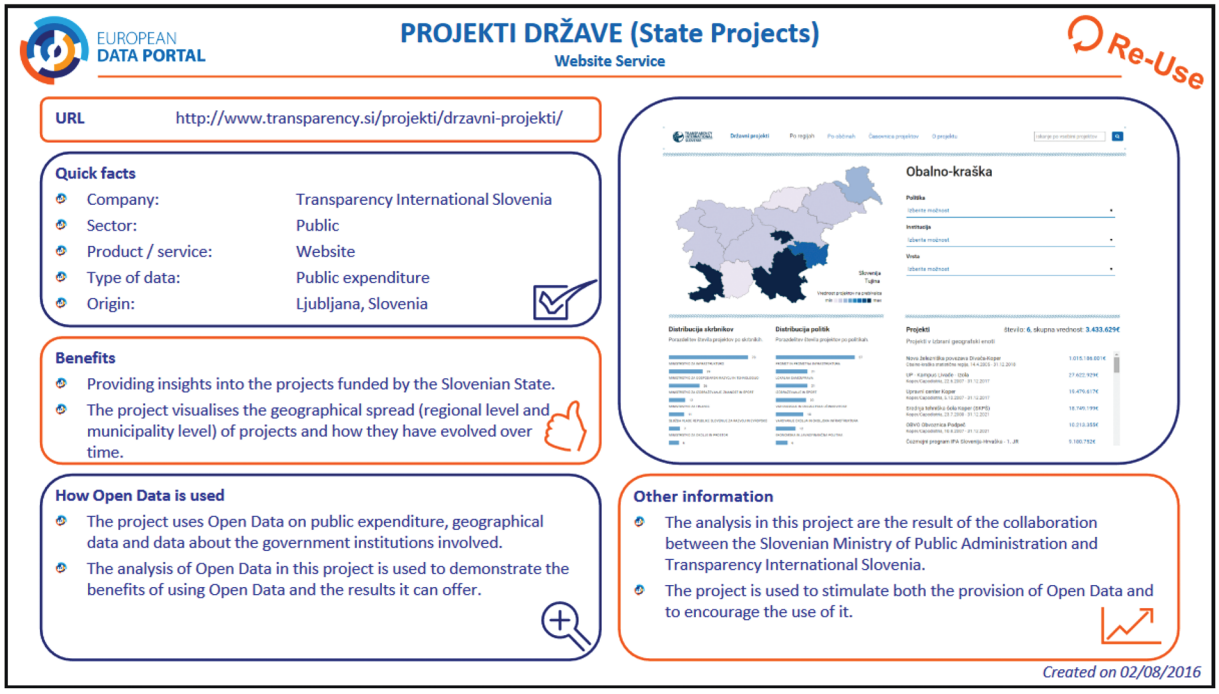
Namen te raziskave je torej potrditi že prepoznana in uveljavljena področja učinka OP s strani Evropske Komisije na podlagi podatkovne zbirke, ki jo uporabljata Ferencek in Kljajić Borštnar [13], [14] v svojih raziskavah za prepoznavo bolj podrobnih področij učinka in izdelavo taksonomije področij OP z metodami umetne inteligence. V prejšnjih raziskavah so bile za avtomatsko prepoznavanje področji učinkov uporabljene preprostejše metode, kot sta npr. TD-IDF [19] in Yake! [20], ki pa nista prinesli želenih rezultatov [14]. V tej raziskavi smo zato uporabili tudi model globoke nevronske mreže [21].

2 METODOLOGIJA

V prispevku naslavljamo problem razvoja klasifikacijskega problema za avtomatizirano določanje kategorije učinkov aplikacij odprtih podatkov.

Pri tem smo sledili metodologiji načrtovanja in razvoja (Design Science Research) [22], ki je sestavljena iz treh glavnih ciklov (opredelitev problema, razvoj in vrednotenje rezultatov). V ciklu razvoja smo uporabili CRISP-DM [23] za razvoj in evalvacijo modela za klasifikacijo učinkov aplikacij odprtih podatkov. CRISP-DM vključuje faze od poslovnega razumevanja, priprave in razumevanja podatkov, do modeliranja, evalvacije in implementacije. S kombinacijo teh pristopov smo sistematično zbirali in preprocesirali podatke, razvijali modele strojnega učenja ter jih iterativno izboljševali, kar je omogočilo robustno klasifikacijo učinkov aplikacij odprtih podatkov.

Ideja prispevka je, da lahko iz opisov primerov uporabe razvijemo model, s katerim bi lahko avtomatsko uvrstili primere uporabe odprtih podatkov glede na področje učinka uporabe.



Slika 1: **Zajeta slika PDF dokumenta enega od primerov uporabe [16], ki smo jih uporabili v tej raziskavi.**

Zbrali smo 697 opisov primerov uporabe, dostopnih na European data portal [16]. Ti opisi so shranjeni v PDF datotekah z osnovnimi podatki o primeru uporabe in krajšim opisom uporabe podatkov ter morebitnimi dodatnimi informacijami ali načrti za nadaljnji razvoj (Slika 1). V PDF datotekah je običajen tudi okvir, ki vsebuje sliko, ki občasno prikazuje izdelan produkt (npr. uporabniški vmesnik) pogosto pa je na sliki le logotip projekta, zato smo se odločili, da v okviru tega članka teh slik ne bomo uporabljali.

Za razvrstitev primerov uporabe v različne kategorije učinkov uporabe smo analizirali njihove opise. Domenski ekspert je razvrstil 697 primerov uporabe v eno od štirih kategorij učinkov (družbena, okoljska, vladna in ekonomska). Prepoznanih je bilo 421 družbenih, 94 okoljskih, 96 vladnih in 86 ekonomskih primerov uporabe (Tabela 1).

Tabela 1: **Tabelarični prikaz kategorij učinkov in pripadajoče število primerov uporabe.**

Kategorija primera uporabe	Število primerov uporabe
DRUŽBENI	421
OKOLJSKI	94
VLADNI	96
EKONOMSKI	86

Za klasifikacijo primerov uporabe smo najprej pretvorili besedila v vektorski prostor. Za pretvorbo smo uporabili več metod in sicer TF-IDF metodo [19] in model globoke nevronske mreže [21] s transformer arhitekturo [24]. Model nevronske mreže ki smo ga uporabili je imenovan General-purpose Text Embeddings v1.5 (GTE) [25], [26], ki je prilagojen BERT model [27] za vektorizacijo besedila. Za obe metodi smo uporabili surovo obliko besedila izluščenega iz pdf datotek in preprocesirano obliko besedila, kjer smo iz besedila ostranili »stop-words« besede, pretvorili besedilo v male črke, izvedli lematizacijo in odstranili razne šume, kot so ločila, posebni znaki, polni URL-ji, e-poštni naslovi, podvojeni presledki ipd.

Za lažje razumevanje podatkov smo dobljene vektorje vizualizirali s pomočjo tehnike UMAP [28]. UMAP (Uniform Manifold Approximation and Projection) je tehnika za zmanjšanje dimenzionalnosti podatkov, kar omogoča vizualizacijo večdimenzionalnih podatkov v dvo- ali tridimenzionalnem prostoru. Ta tehnika je še posebej uporabna za vizualizacijo kompleksnih podatkovnih nizov, kot so besedilni vektorji, saj omogoča enostavno prepoznavanje vzorcev in skupin.

Ker je bilo primerov uporabe v kategoriji »Družbeni« bistveno več kot v drugih kategorijah (421),

Tabela 2: **Tabelarični prikaz elementov hiperparametrizacije ter njihovih vrednosti**

Hiperparametri	Vrednost
Velikost paketa (ang. batch size)	32
Število epoh (ang. epoch)	10
Začetna stopnja učenja (ang. initial learning rate)	5e-4 (»cosine decay« [35] do 0 v korakih)
Naključno izpuščanje (ang. dropout) [29]	50 %
Skriti sloj	1024 nevronov (»GELU« aktivacijska funkcija [32])
Izhodni sloj	4 nevroni (»softmax« aktivacijska funkcija [33])

smo za uravnoteženje nabora podatkov pri klasifikaciji naključno izbrali 100 primerov uporabe iz te kategorije [29]. Preostale primere uporabe iz te kategorije smo odstranili iz nabora podatkov.

Uravnotežene podatke smo naključno razdelili na učno in testno množico v razmerju 66 % učni in 33 % testni. Za klasifikacijo v kategorije smo uporabili naslednje metode: nevronska mreža [21], naključni gozd (random forest) [30] in metoda podpornih vektorjev (Support Vector Machine - SVM) [31]. Velikost in globino naključnega gozda smo določili s testiranjem naključnih kombinacij, kjer smo globino variirali med 1 in 5, velikost pa med 10 in 500.

Nevronska mreža [29], ki smo jo uporabili za klasifikacijo, je sestavljena iz enega skritega sloja s 1024 nevroni z »GELU« [32] aktivacijsko funkcijo in izhodnim slojem z 4 nevroni s »softmax« [33] aktivacijsko funkcijo. Med vhodi in prvim (skritim) slojem ter med prvim in izhodnim slojem smo med učenjem uporabili naključno izpuščanje (ang. dropout) [34] z verjetnostjo 50 %. V tabeli (Tabela 2) so prikazani hiperparametri učenja.

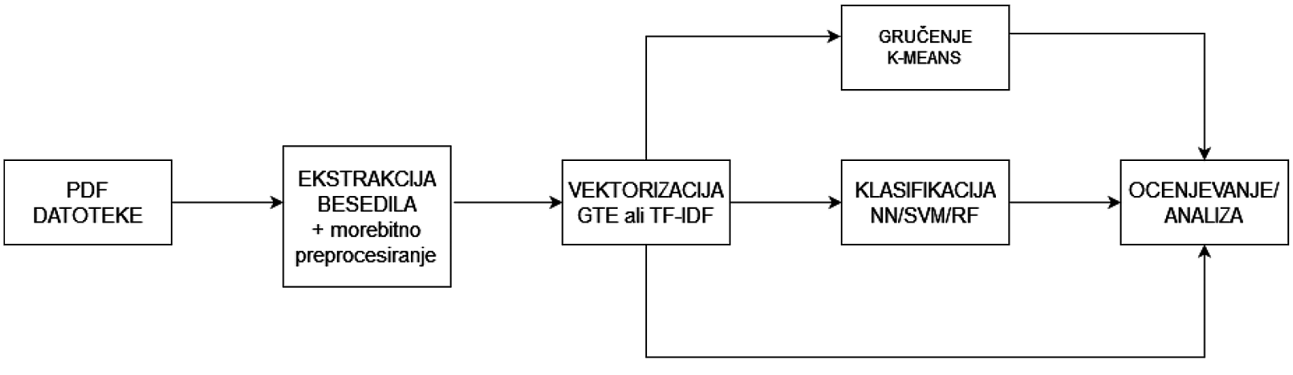
V okviru naše raziskave smo še želeli ugotoviti tudi, ali je trenutna kategorizacija učinkov uporabe odprtih podatkov ustrezna. Naš cilj je bil ugotoviti, ali bi lahko identificirali nove kategorije uporabe odprtih podatkov, kar bi lahko prispevalo k izboljšanju natančnosti, uporabnosti ter razumevanju kategorizacij učinkov. Glede na rezultate pri klasifikaciji smo določili najustreznejšo metodo za vektorizacijo opisov primerov uporabe za nadaljnjo analizo. Na sliki 2 so prikazane faze procesa klasifikacije in analize podatkov, ki smo jih izvajali.

Pri nadaljnji analizi pa smo uporabili metodo K-means - gručenja [36], [37], ki nam omogoča ločevanje na nove kategorije. Ustreznost novih kategorij smo ugotavljali glede na sovpadanje s že obstoječimi kategorijami in pregledom primerov uporabe v posameznih skupinah. Za določanje ustreznega števila skupin (clustrov) smo si pomagali z uporabilo elbow metode [38] in Silhouette analize [39].

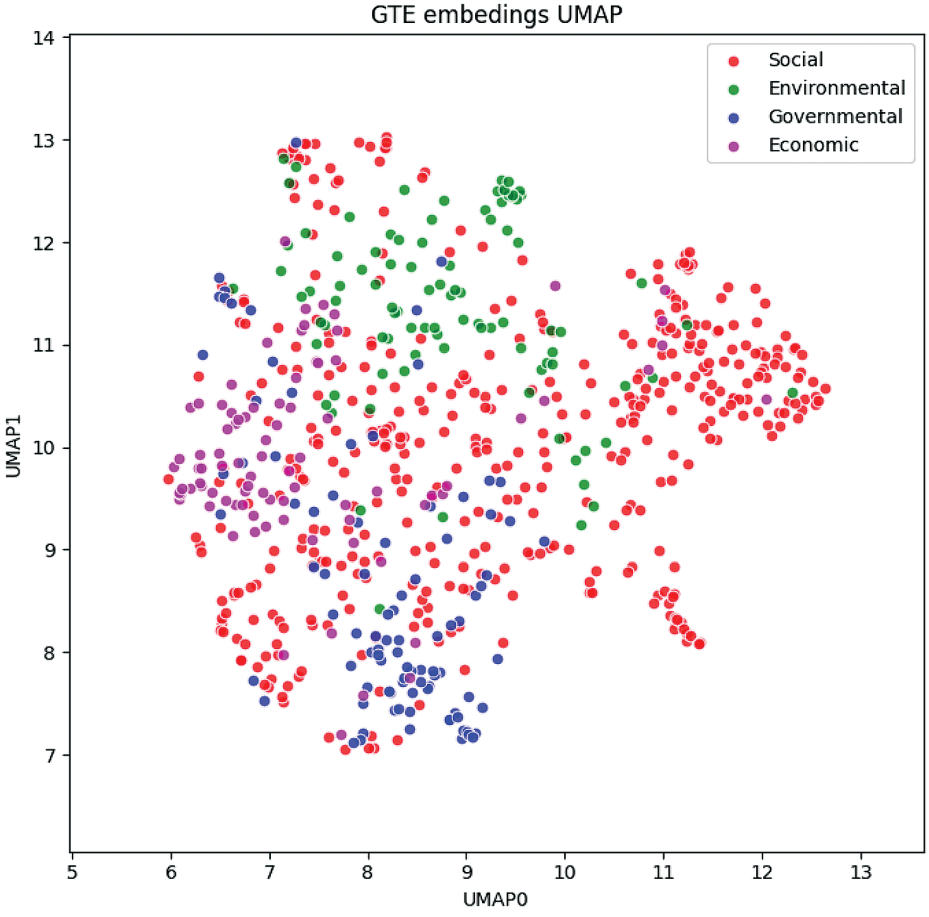
3 REZULTATI

Slika 3 prikazuje UMAP [28] projekcijo vektoriziranih besedil primerov uporabe odprtih podatkov, ki so bili vgrajeni z modelom GTE v1.5 [25], [26] brez preprocesiranja. Besedila so obarvana glede na kategorijo učinka uporabe, ki je bila določena s strani domenskega eksperta. Iz slike je razvidno, da med kategorijami prihaja do prekrivanj, ter da ni jasno razvidnih mej med kategorijami.

Pri klasifikaciji smo najboljše rezultate dosegli z uporabo vektorjev GTE v1.5 [25], [26] modela brez



Slika 2: **Grafični prikaz postopka izvedbe analize.**



Slika 3: **UMAP [28] projekcija vektoriziranih besedil primerov uporabe odprtih podatkov, ki so bili vgrajeni z modelom GTE v1.5 [25] [26] brez preprocesiranja.**

dodatnega preprocesiranja besedila in uporabo nevronske mreže za klasifikacijo teh vektorjev v posamezne kategorije. Rezultati so bili ocenjeni glede na klasifikacijsko točnost (classification accuracy ACC) [40], AUC oceno [41] in F1 oceno [42] (Tabela 3).

Tabela 3: **Primerjava rezultatov uporabe modelov GTE v1.5 [25], [26] ter TF-IDF [19] s preprocesiranjem podatkov in brez preprocesiranja podatkov.**

GTE v1.5.	**Brez preprocesiranja.**			**S preprocesiranjem**		
	ACC	AUC	F1	ACC	AUC	F1
NN	0,80	0,94	0,80	0,72	0,91	0,71
SVN	0,792	0,86	0,79	0,752	0,84	0,76
RF	0,736	0,82	0,74	0,728	0,82	0,72
TF-IDI	**Brez preprocesiranja.**			**S preprocesiranjem**		
	ACC	AUC	F1	ACC	AUC	F1
NN	0,64	0,90	0,63	0,56	0,90	0,55
SVN	0,272	0,51	0,18	0,336	0,57	0,26
RF	0,64	0,80	0,70	0,648	0,75	0,61

Na sliki 4 je prikazana UMAP [28] vizualizacija aktivacij zadnjega skritega sloja klasifikatorja za določanje učinkov uporabe odprtih podatkov glede na opise primerov uporabe. Vizualizacija je razdeljena na dva dela:

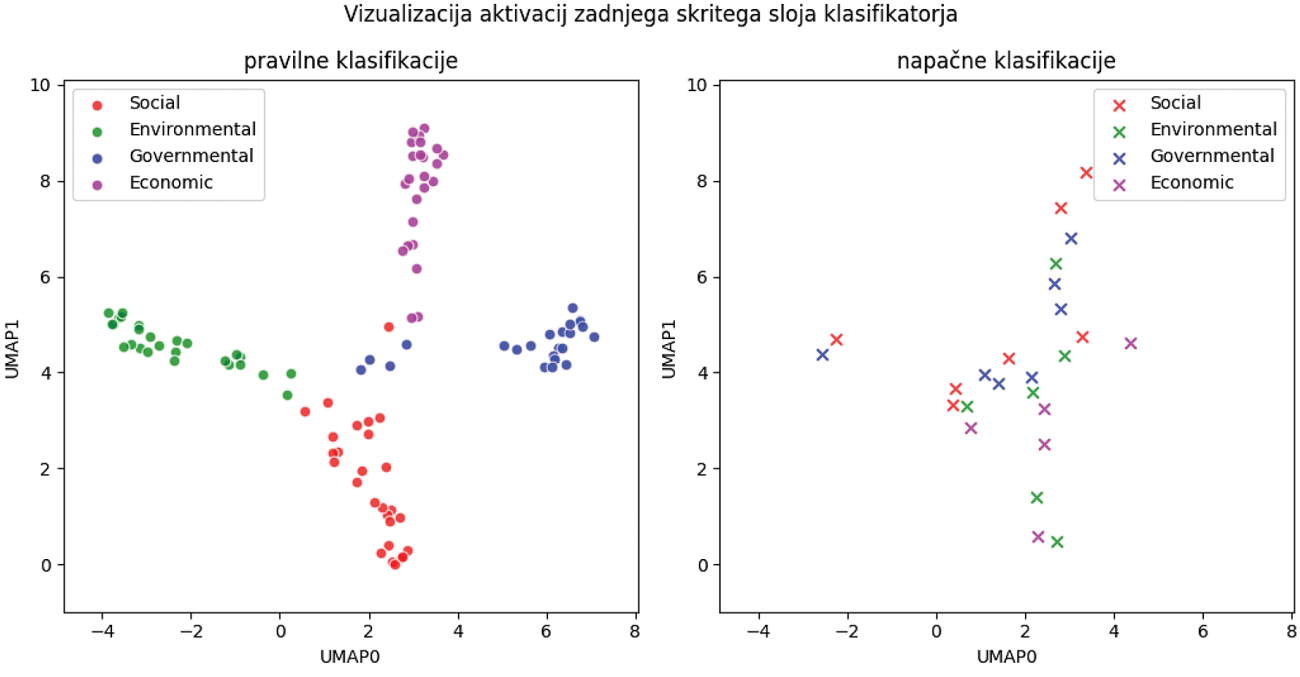
Levi del: Prikazuje aktivacije pravilnih klasifikacij. Pike predstavljajo primere uporabe, ki so bili pravilno razvrščeni v kategorije.

Desni del: Prikazuje aktivacije napačnih klasifikacij. Križci predstavljajo primere uporabe, ki so bili napačno razvrščeni. Barva križca predstavlja pravilno kategorijo.

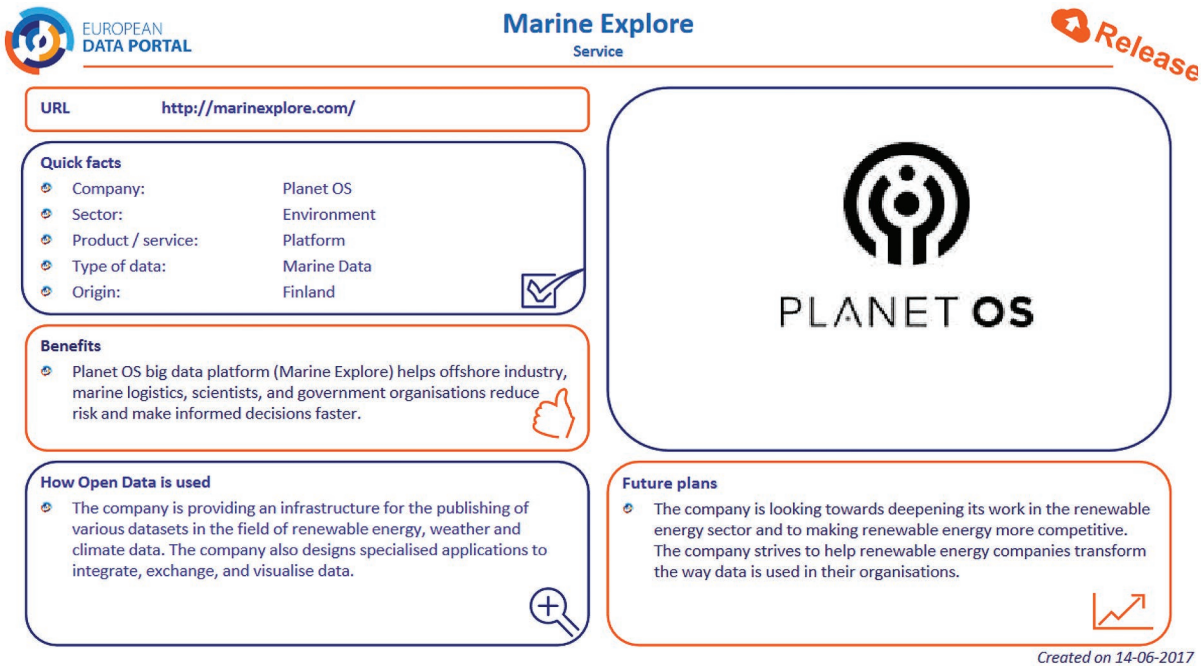
Iz vizualizacije lahko razberemo, da je klasifikator sposoben ločiti med posameznimi kategorijami. Večina napak je pri primerih uporabe, ki so glede na klasifikator povezani z več kategorijami. Napak, kjer predviden vektor spada popolnoma v drugo kategorijo od predvidene s strani domenskega eksperta, ni veliko (dva družbena primera med ekonomskimi (graf zgoraj), dva okoljska in en ekonomski med družbenimi (graf spodaj) in en družbeni in vladni med okolijskimi (graf levo). Te napake bi lahko bile tudi posledica napačnih označb.

Na podlagi te analize lahko sklepamo, da je klasifikator dokaj uspešen pri prepoznavanju kategorij učinkov uporabe odprtih podatkov, vendar obstaja še prostor za izboljšanje, zlasti pri razvrščanju primerov uporabe, ki so povezani z več kategorijami učinka.

Za boljšo predstavitev delovanja klasifikatorja pri primerih učinkov uporabe smo izpisali predvidene verjetnosti kategorij za en primer, kjer se učinek glede na klasifikator kaže v več kategorijah (Slika 5, Tabela 4,5) in primer, pri katerem klasifikator predvideva učinek v eni kategoriji (Slika 6, Tabela 6,7).



Slika 4: **UMAP [28] vizualizacija aktivacij zadnjega skritega sloja klasifikatorja za določanje kategorij učinkov uporabe odprtih podatkov glede na opise primerov uporabe**



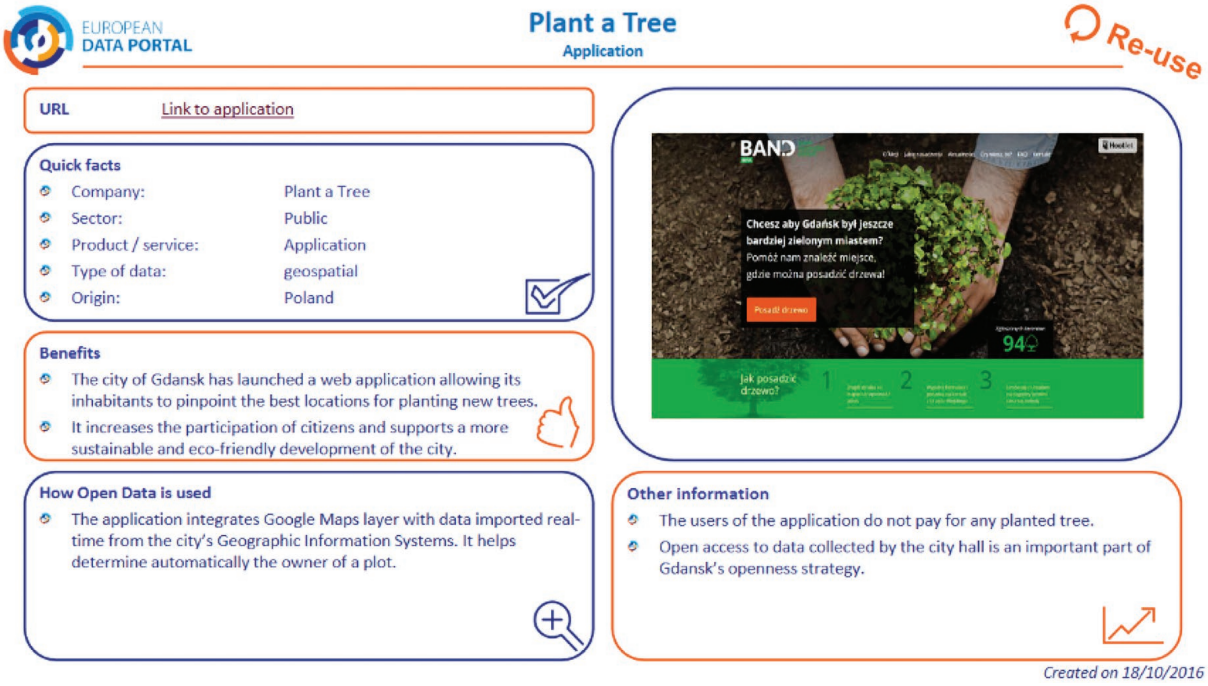
Slika 5: **Zajeta slika PDF dokumenta projekta Marine Explore. [16]**

Tabela 4: **Primer surovega in neprocesiranega besedila projekta Marine Explore, ki smo ga uporabili v analizi.**

Marine Explore \n\nService \n\nURL \n\nhttp://marinexplore.com/ \n\n \n\nQuick facts \n\nCompany: \n\nPlanet OS \n\nSector: \n\n \n\nEnvironment \n\n \n\nProduct / service: \n\nPlatform \n\nType of data: \n\nMarine Data \n\nOrigin: \n\n \n\nFinland \n\nBenefits \n\nPlanet OS big data platform (Marine Explore) helps offshore industry, \n\nmarine logistics, scientists, and government organisations reduce \n\nrisk and make informed decisions faster. \n\nHow Open Data is used \n\nFuture plans \n\nThe company is providing an infrastructure for the publishing of \n\nvarious datasets in the field of renewable energy, weather and \n\nclimate data. The company also designs specialised applications to \n\nintegrate, exchange, and visualise data. \n\nThe company is looking towards deepening its work in the renewable \n\nenergy sector and to making renewable energy more competitive. \n\nThe company strives to help renewable energy companies transform \n\nthe way data is used in their organisations. \n\nCreated on 14-06-2017 Release \n

Tabela 5: **Rezultati predvidenih verjetnosti za razrede oziroma kategorije učinkov OP projekta Marine Explore.**

Primer uporabe	**Resnični razred**	**Predvidene verjetnosti za razrede**			
		EKONOMSKI	OKOLJSKI	VLADNI	DRUŽBENI
Marine Explore	okoljski	0,52	0,23	0,03	0,22



Slika 6: **Zajeta slika PDF dokumenta projektaPlant a Tree. [16]**

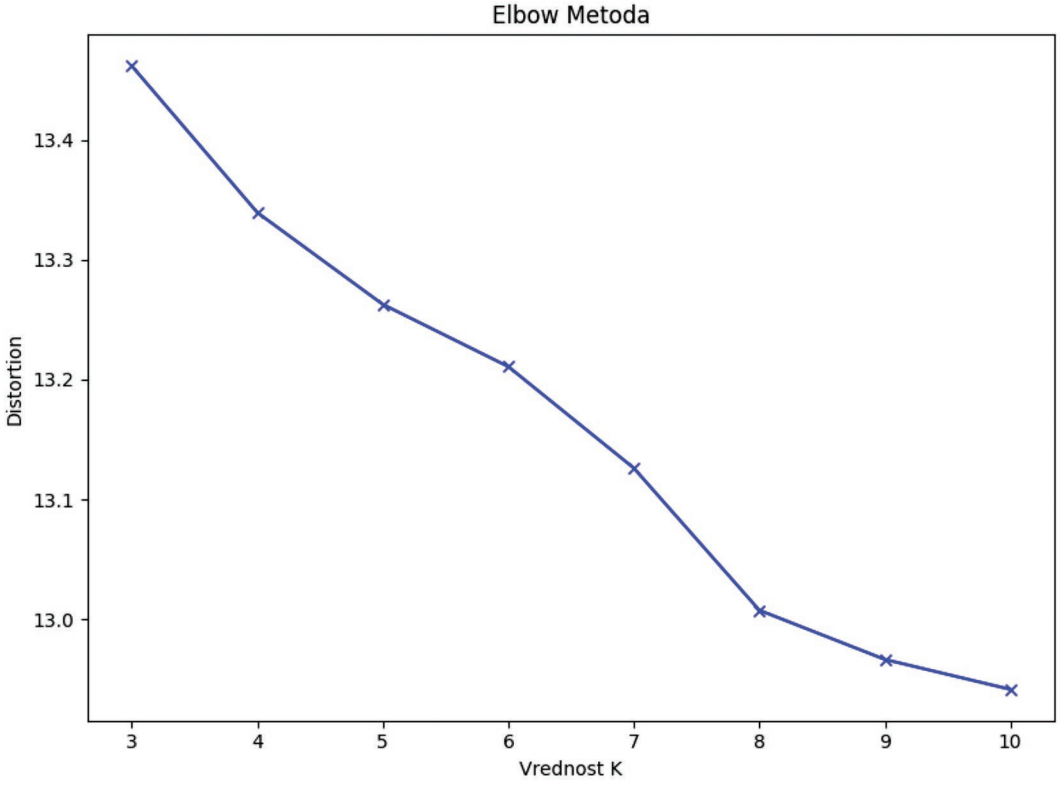
Tabela 6: **Primer surovega in neprocesiranega besedila projekta Plant a Tree, ki smo ga uporabili v analizi.**

Plant a Tree\nApplication\nURL\nLink to application\n\nQuick facts\nCompany:\nPlant a Tree\n\n\nSector:\n\nPublic\n\nProduct / service:\nApplication\nType of data:\ngeospatial\nOrigin:\n\nPoland\nBenefits\nThe city of Gdansk has launched a web application allowing its\ninhabitants to pinpoint the best locations for planting new trees.\nIt increases the participation of citizens and supports a more\nsustainable and eco-friendly development of the city.\nHow Open Data is used\nOther information\nThe users of the application do not pay for any planted tree.\nOpen access to data collected by the city hall is an important part of\nGdansk's openness strategy.\nThe application integrates Google Maps layer with data imported real-\ntime from the city's Geographic Information Systems. It helps\ndetermine automatically the owner of a plot.\n\nCreated on 18/10/2016 Re-use\nCreated on 14-06-2017 Release \n

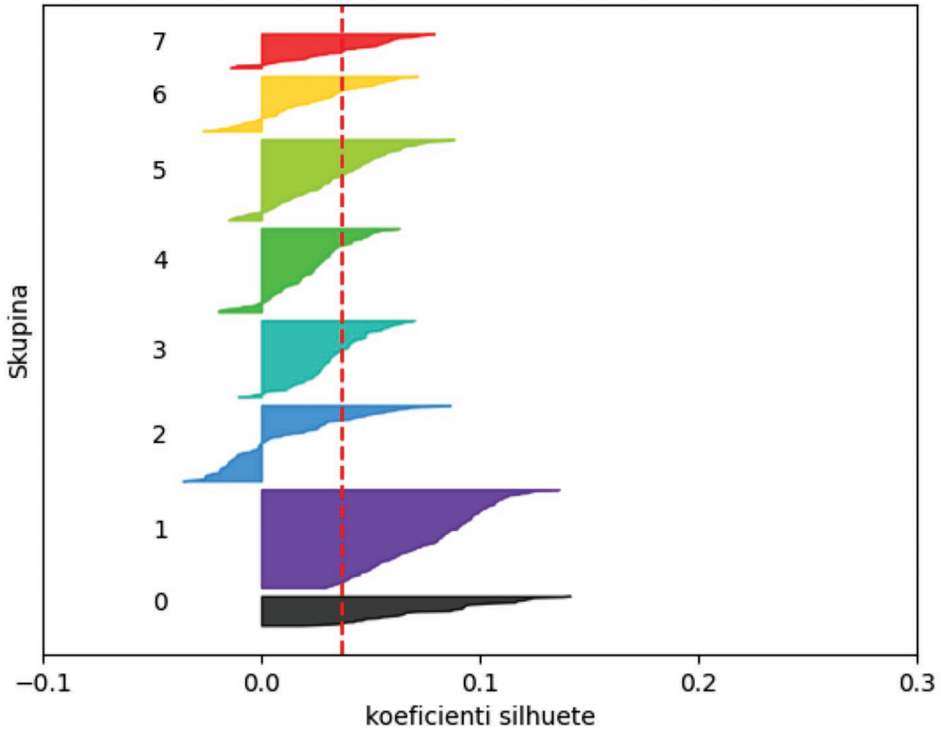
Tabela 7: **Rezultati predvidenih verjetnosti za razrede oziroma kategorije učinkov OP projekta Plant a Tree.**

Primer uporabe	**Resnični razred**	**Predvidene verjetnosti za razrede**			
		EKONOMSKI	OKOLJSKI	VLADNI	DRUŽBENI
Plant a Tree	okoljski	0	0,98	0	0,02

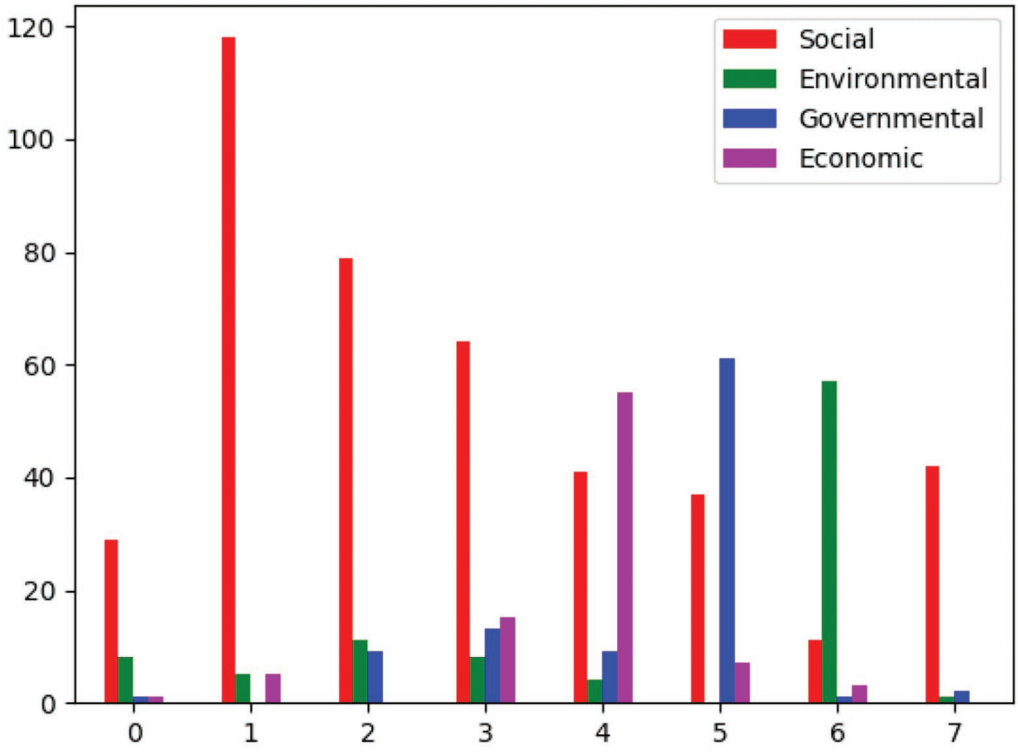
Rezultati, pridobljeni z razvojem klasifikatorja, nakazujejo, da od uporabljenih metod vektorji pridobljeni s pomočjo GTE v1.5 [25], [26] modela brez dodatnega preprocesiranja najbolj natančno zajemajo vsebino opisov uporabe za namen ugotavljanja kategorije učinka. Ker so bili vektorji dovolj deskriptivni za klasifikacijo, smo nadaljevali z analizo teh vektorjev v iskanju morebitnih novih kategorij učinkov, ki bi omogočile bolj natančno in uporabno klasifikacijo učinkov uporabe odprtih podatkov. Z uporabo elbow metode [38] (Slika 7) in Silhouette metode [39] (Slika 8) smo ugotovili, da bi bilo možno učinke razdeliti na 8 skupin z uporabo metode gručenja K-means [36], [37].



Slika 7: **Črtni grafikon distorzije po vrednosti K metode k-means gručenja [36], [37], [38].**



Slika 8: **Vizualizacija koeficientov silhuete [39] za vrednost K=8 za k-means metodo gručenja [36], [37]**



Slika 9: **Stolpčni grafikon prikazuje ujemanje prej določenih kategorij s kategorijami določenimi z k-means metodo gručenja [36], [37] (na x osi so označene k-means kategorije, na y osi pa število primerov uporabe, barve ločujejo prej določene kategorije).**

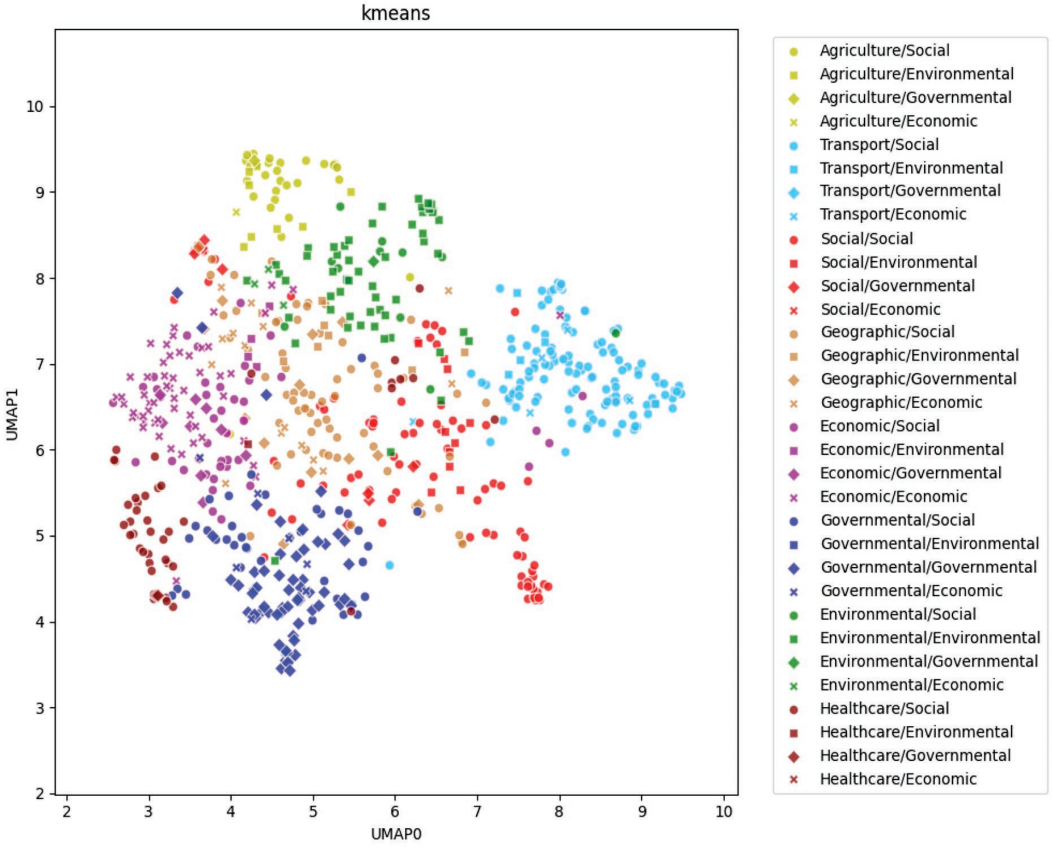
Po uporabi k-means metode gručenja [36], [37] na vektorjih smo novo pridobljene skupine primerjali s prej določenimi kategorijami učinkov uporabe. Opazili smo jasno prekrivanje treh novih skupin s tremi od štirih prej določenih kategorijah učinkov (vladni, okoljski, ekonomski) , kot je razvidno na Sliki 9. V preostale nove skupine so bili večinoma razdeljeni primeri uporabe z učinkom v družbeni kategoriji.

Ob pregledu novih kategorij smo določili nove kategorije učinkov, ki so predstavljene v Tabeli 8.

Slika 10 prikazuje UMAP [28] projekcijo vektoriziranih besedil primerov uporabe odprtih podatkov, ki so bili vgrajeni z modelom GTE v1.5 [25], [26]. Barve ločujejo z kategorije pridobljene z k-means metodo [36], [37]. Oblike oznak pa ločujejo med 4 prej določenimi kategorijami.

Tabela 8: **Predlagane kategorije učinkov.**

K-means skupina	Predlagane kategorije	Ang
0	Kmetijski	Agriculture
1	Transportni	Transport
2	Družbeni	Social
3	Geografski	Geographic
4	Ekonomski	Economic
5	Vladni	Governmental
6	Okoljski	Environmental
7	Zdrvstveni	Healthcare



Slika 10: **UMAP [28] projekcija vektoriziranih besedil primerov uporabe odprtih podatkov, ki so bili vgrajeni z modelom GTE v1.5 [25], [26] z označenimi novimi kategorijami.**

4 ZAKLJUČEK

V tem prispevku smo predstavili način avtomatizacije kategoriziranja učinkov uporabe odprtih podatkov glede na opise primerov uporabe. Pokazali smo, da je z uporabo modelov umetne inteligence možno uspešno kategorizirati primere uporabe v trenutno prepoznane in določene kategorije učinkov s strani Evropske komisije. Pokazali smo, da se učinki posameznih primerov uporabe pogosto kažejo v več kategorijah ter da meje med posameznimi kategorijami niso jasne. Po potrditvi možnosti klasifikacije v trenutno poznane kategorije učinkov z uporabo modelov umetne inteligence smo poskusili identificirati morebitne nove kategorije, ki bi ponudile bolj podroben in uporaben pregled nad kategorijami učinkov uporabe odprtih podatkov. Bolj podrobna klasifikacija kategorij učinkov bi lahko prispevala k kasnejšemu bolj natančnem prepoznavanju učinka, saj bi bilo verjetno potrebno prepoznavati učinek v kategoriji zdravstva drugače kot v kategoriji transporta.

5 LITERATURA

[1] Open Government Data. (b. d.). Organisation for Economic Co-operation and Development. https://www.oecd.org/gov/digital-government/open-government-data.htm. (Dostopano dne: 28. Julij 2024)

[2] Attard, J., Orlandi, F. in Auer, S. (2016). Value Creation on Open Government Data. 2016 *49th Hawaii International Conference on System Sciences (HICSS)*.

[3] Safarov, I., Meijer, A. in Grimmelikhuijsen, S. (2017). Utilization of open government data: A systematic literature review of types, conditions, effects and users. *Information Polity*, 22(1), pp. 1-24.

[4] Ubaldi, B. (2013), Open Government Data: Towards Empirical Analysis of Open Government Data Initiatives. OECD *Working Papers on Public Governance*, No. 22, OECD Publishing, Paris.

[5] Jaeger, P. T. in Bertot, J. C. (2010). Transparency and technological change: Ensuring equal and sustained public access to government information. *Government Information Quarterly*, 27(4), pp. 371–376.

[6] Leviäkangas, P. in Molarius, R. (2020). Open government data policy and value added – Evidence on transport safety agency case. *Technology in Society*, 63(2).

[7] Kalampokis, E., Tambouris, E., in Tarabanis, K. (2013). Linked Open Government Data Analytics. *Electronic Government,* pp. 99–110.

[8] Keserű, J. in James Kin-sing C. (2015). The Social Impact of Open Data. *3rd International Open Data Conference 2015 (IODC).*

[9] Huyer, E. in van Knippenberg, L. (2020). The Economic Impact of Open Data: Opportunities for value creation in Europe, Capgemini Invent. *European, Data Portal*.

[10] OECD. (2018). Open Government Data Report: Enhancing Policy Maturity for Sustainable Impact. *OECD Digital Government Studies*. OECD Publishing, Paris.

[11] Zuiderwijk, A. in Janssen, M. (2014). Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, 31(1), pp. 17–29.

[12] Ferencek, Aljaž, Kljajić Borštnar, Mirjana, Pretnar Žagar, Ajda. Categorisation of open government data literature. *Business systems research*. 2022, vol. 13, no. 1, str. 66–83

[13] Ferencek, Aljaž, Kljajić Borštnar, Mirjana. Topic modelling of open government data impact areas using GPT 3.5 model. V: Drobne, Samo (ur.), et al. *SOR ,23 : proceedings of the 17th International Symposium on Operational Research in Slovenia : Bled, Slovenia, September 20–22, 2023*. 1st electronic version. Ljubljana: Slovenian Society Informatika – Section for Operational Research, 2023. Str. 71–76

[14] Ferencek, Aljaž, Kljajić Borštnar, Mirjana. Open government data impact areas identification with data mining techniques. V: Drobne, Samo (ur.), et al. *SOR ,21 proceedings : the 16th International Symposium on Operational Research in Slovenia : September 22–24, 2021*, online. Ljubljana: Slovenian Society Informatika, Section for Operational Research, 2021. Str. 101-106.

[15] Ferencek, Aljaž, Kljajić Borštnar, Mirjana, Pretnar Žagar, Ajda. Text mining approach to research gap definition in open government data. V: Čeh Časni, Anita (ur.), Arnerić, Josip (ur.). *Book of abstracts*. 18th International Conference on Operational Research, KOI 2020, Šibenik, Croatia, 23–25 September, 2020. Zagreb: Croatian Operational Research Society: University, Faculty of Economics and Business, 2020. Str. 56.

[16] Data.europa.eu. (b. d.). Publications Office of the European Union. https://data.europa.eu/en/impact-studies/use-cases. (Dostopano dne: 24. Julij 2024)

[17] OECD. (2018). Open Government Data Report: Enhancing Policy Maturity for Sustainable Impact. *OECD Digital Government Studies*, OECD Publishing, Paris.

[18] European Data Portal. (2023). Open Data Maturity Report 2023. *Publications Office of the European Union*. https://data.europa.eu/sites/default/files/odm2023_report.pdf (Dostopano dne: 28. Julij 2024)

[19] Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, 60(5), pp. 503-520.

[20] Campos, R., Mangaravite, V., Pasquali, A., Jorge, A., Nunes, C., & Jatowt, A. (2020). YAKE! Keyword extraction from single documents using multiple local features. *Information Sciences*, 509, 257–289. https://doi.org/10.1016/j.ins.2019.09.013

[21] Hevner, A., March, S., Park, J. in Ram, S. (2004). Design Science in Information Systems Research. *MIS Quarterly*, 28(1), pp. 75-105.

[22] Wirth, R. in Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. *Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining.*

[23] McCulloch, W. S. in Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, 5(4), pp. 115–133.

[24] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., in Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, pp. 5999–6009.

[25] Alibaba-NLP/gte-large-en-v1.5 Hugging Face. (b.d.). https://huggingface.co/Alibaba-NLP/gte-large-en-v1.5 (Dostopano dne: 2. Julij 2024)

[26] Li, Z., Zhang, X., Zhang, Y., Long, D., Xie, P., Zhang, M. in Group, A. (2023). Towards General Text Embeddings with Multi-stage Contrastive Learning.

[27] Devlin, J., Chang, M. W., Lee, K. in Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 – 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies – Proceedings of the Conference*, pp. 4171–4186.

[28] McInnes, L., Healy, J. In Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

[29] Gudivada, V., Apon, A. in Ding, J. (2017). Data quality considerations for big data and machine learning: Going beyond data cleaning and transformations. *International Journal on Advances in Software*, 10(1), pp. 1–20.ž

[30] Ho, T. K. (1995). Random decision forests. *Proceedings of the International Conference on Document Analysis and Recognition*, pp. 278–282.

[31] Cortes, C. in Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), pp. 273–297.

[32] Hendrycks, D. in Gimpel, K. (2016). Gaussian Error Linear Units (GELUs).

[33] Bridle, J. S. (1990). Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition. *Neurocomputing*, pp. 227–236.

[34] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. in Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(56), pp. 1929–1958.

[35] Loshchilov, I. in Hutter, F. (2016). SGDR: Stochastic Gradient Descent with Warm Restarts. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*.

[36] Lloyd, S. P. (1957). Least squares quantization in PCM. *Technical Report RR-5497*, Bell Lab, September 1957.

[37] MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, Vol. 1, pp. 281–297.

[38] Robert Tibshirani, Guenther Walther, Trevor Hastie, Estimating the Number of Clusters in a Data Set Via the Gap Statistic, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 63(2), pp. 411–423.

[39] Rousseeuw, Peter. (1987). Rousseeuw, P.J.: Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, Vol. 20, pp. 53–65.

[40] Hossin, Mohammad in M.N, Sulaiman. (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, Vol. 5, pp. 01–11.

[41] Bradley, A.P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, Vol. 30, pp. 1145-1159.ss

[42] Goutte, C. in Gaussier, E. (2005). A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation. *Advances in Information Retrieval, ECIR 2005 Lecture Notes in Computer Science*, Vol. 3408.

▪

Nejc Čelik je asistent za področje Informacijski sistemi na Fakulteti za organizacijske vede na Univerzi v Mariboru. Njegovi raziskovalni interesi so vezani na uporabo umetne inteligence v organizacijah.

▪

Aljaž Ferencek je doktorski študent na Fakulteti za organizacijske vede na Univerzi v Mariboru. Magisterij je pridobil na isti fakulteti. Njegovi raziskovalni interesi vključujejo podatkovno znanost in odprte podatke, o čemer je že objavil raziskave.