PETINPETDESETLET

slovensko sociološko društvo

PANDEMIČNA DRUŽBA

SLOVENSKO SOCIOLOŠKO SREČANJE

Ljubljana, 24.–25. september 2021

Ljubljana, 2021

Izdajatelj:
Slovensko sociološko društvo
Kardeljeva ploščad 5, 1000 Ljubljana

Uredniki:
Miroljub Ignjatović, Aleksandra Kanjuo Mrčela, Roman Kuhar

Tehnični urednik:
Igor Jurekovič

Programski odbor:
Predsedstvo Slovenskega sociološkega društva

Recenzentke:
Anja Zalta, Alenka Švab in Veronika Tašner

Oblikovanje in prelom:
Polonca Mesec Kurdija

Korekture:
avtorji

Elektronska izdaja:
Publikacija je brezplačno dostopna na elektronskem naslovu:
http://www.sociolosko-drustvo.si/.

Ljubljana, 2021

Kataložni zapis o publikaciji (CIP) pripravili
v Narodni in univerzitetni knjižnici v Ljubljani

COBISS.SI-ID 77660931

ISBN 978-961-94302-6-2 (PDF)

SENJA POLLAK
Inštitut Jožef Stefan
MATEJ MARTINC
Inštitut Jožef Stefan
ANDRAŽ PELICON
Inštitut Jožef Stefan
MATEJ ULČAR
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko
ANDREJA VEZOVNIK
Univerza v Ljubljani, Fakulteta za družbene vede

COVID-19 V SLOVENSKIH SPLETNIH MEDIJIH: ANALIZA S POMOČJO RAČUNALNIŠKE OBDELAVE JEZIKA

Povzetek: *V prispevku s pomočjo metod za računalniško obdelavo naravnega jezika analiziramo poročanje slovenskih medijev o epidemiji covid-19. V prispevku najprej identificiramo osrednje teme poročanja na petdesetih slovenskih novičarskih portalih v obdobju 1.1.2020–31.12.2020. Nato z uporabo kontekstualnih besednih vložitev analiziramo razlike v poročanju na manjšem korpusu štirinajstih izbranih novičarskih portalov. Ugotavljamo, da se razlike med portali kažejo predvsem pri besedah povezanih s spornimi temami v javnosti. Na primeru besede kolesar, pokažemo, da se na portalih nova24TV.si, demokracija.si in necenzurirano.si, beseda kolesar intenzivneje povezuje z besedo covid-19, kot na primer na portalu siol.net, ter da beseda na prvih treh portalih nosi politično konotacijo, medtem ko je na siol.net povezana predvsem s športom.*

Ključne besede: *covid-19, novičarski mediji, obdelava naravnega jezika, besedne vložitve, računalniška analiza besedil.*

Uvod

Epidemija covid-19 je v zadnjem letu sprožila širšo družbeno krizo. V času upravljanja s krizo igrajo pomembno vlogo novičarski spletni mediji, saj le-ti služijo javnosti kot primarni vir pridobivanja informacij. Spletni mediji so namreč v Sloveniji v zadnjih letih postali prevladujoči vir za spremljanje novic (iPROM in Valicon 2021). Pomen proučevanja medijskih vsebin ima dolgo zgodovino v komunikologiji. McCombs in Shaw (1972) sta denimo preučevala vpliv televizijske agende na neodločene volivce. Gerbner in dr. so raziskovali, kako televizijske vsebine vzgajajo medijsko občinstvo. Entman (1994) je preučeval, kako mediji na specifične načine okvirjajo teme in kako izbori in poudarki vsebin, vplivajo na dojemanje vsebin s strani občinstva. Cantril (1999) je pokazal, da so ljudje v času družbenih kriz posebej sugestibilni za medijske vsebine. Tradicija ukvarjanja s povezavo med medijskimi vsebinami in občinstvi je teoretsko kompleksna in nam v tem prispevku služi kot iztočnica za utemeljitev, da je preučevanje medijskih vsebin, posebej novičarskih, že več kot stoletje predmet osrednjega komunikološkega zanimanja.

Vrsta novejših študij proučuje medijske vsebine v povezavi s covid-19. S pomočjo modeliranja tematik Liu in dr. (2020) ugotavljajo pomen medijskega poročanja o covid-19 na Kitajskem. Rebello in dr. (2020) preučujejo, kako so se novičarske vsebine povezane s covid-19 manifestirale na spletnih družbenih omrežjih. Mutua in Ong'ong'a (2020) sta s pomočjo analize vsebin in okvirjanja analizirala poročanje mednarodnih tiskovnih agencij o covid-19. Podobno se s pomočjo analize okvirov Hubner (2021) loti novičarskih virov v ZDA. Hart in dr. (2020) s pomočjo računalniško podprte analize ugotavljajo stopnjo politizacije in polarizacije novic o covid-19 v novičarskih medijih v ZDA. Metodološko je najbolj sorodna diahrona analiza poročanja o covid-19 z uporabo gručenja pomenov s kontekstualnimi vložitvami (Montariol in dr. 2021). Zaenkrat so študije novičarskih vsebin v povezavi s covid-19 še vedno maloštevilne. V Sloveniji pa take študije še nimamo.

Naš korpus zajema več deset tisoč člankov. Ročna analiza velikih podatkov je časovno neizvedljiva. V prispevku pokažemo, kako lahko z metodami obdelave naravnega jezika ponudimo nov vpogled v poročanje slovenskih novičarskih portalov o epidemiji. Najprej uporabimo metodo modeliranja tematik s pomočjo latentne Dirichletove alokacije, v nadaljevanju pa metode, ki temeljijo na besednih vektorskih vložitvah. Besedne vektorske vložitve so večstodimenzionalne vektorske predstavitve besed, ki opisujejo besede glede na besedilni kontekst, v katerem se pojavljajo, natrenirane pa so z uporabo nevronskih mrež. Bližina med vektorji v vektorskem prostoru pa odraža semantično povezanost besed. Pri statičnih vložitvah eni besedi ustreza en vektor, pri kontekstualnih pa vektor predstavalja posamezno besedno rabo. Z metodami, ki temeljijo na kontekstualnih besednih vložitvah lahko tako tudi primerjamo rabe besed v različnih medijih.

Korpus

V pričujočem članku obravnavamo korpus o covid-19, ki zajema 89.204 člankov iz obdobja 1. 1.2020–31.12.2020. Z uporabo storitve EventRegistry (Leban in dr. 2014) smo zajeli članke, ki vsebujejo eno izmed besed *covid, koronavirus, sars-cov-2, covid19, covid-19, korona virus, koronavirusna, koronavirusen.* V korpus smo vključili članke tistih portalov, ki so zavedeni v Razvidu medijev, vodenega s strani Ministrstva za kulturo RS. Ta korpus člankov petdesetih portalov[1] (Korpus-50) smo nato uporabili za modeliranje tematik. Za analizo razlik med portali pa smo se omejili le na novičarske portale, ki so v korpusu imeli vsaj 1000 člankov, in izločili portale specializirane za športne in lokalne vsebine. Ta pod-izbor (Korpus-14) zajema članke iz rtvslo.si, siol.net, delo.si, žurnal24.si, vecer.com, 24ur.com, novice.svet24.si, reporter.si, dnevnik.si, demokracija.si, nova24tv.si, politikis.si, mladina.si in necenzurirano.si.

Modeliranje tematik

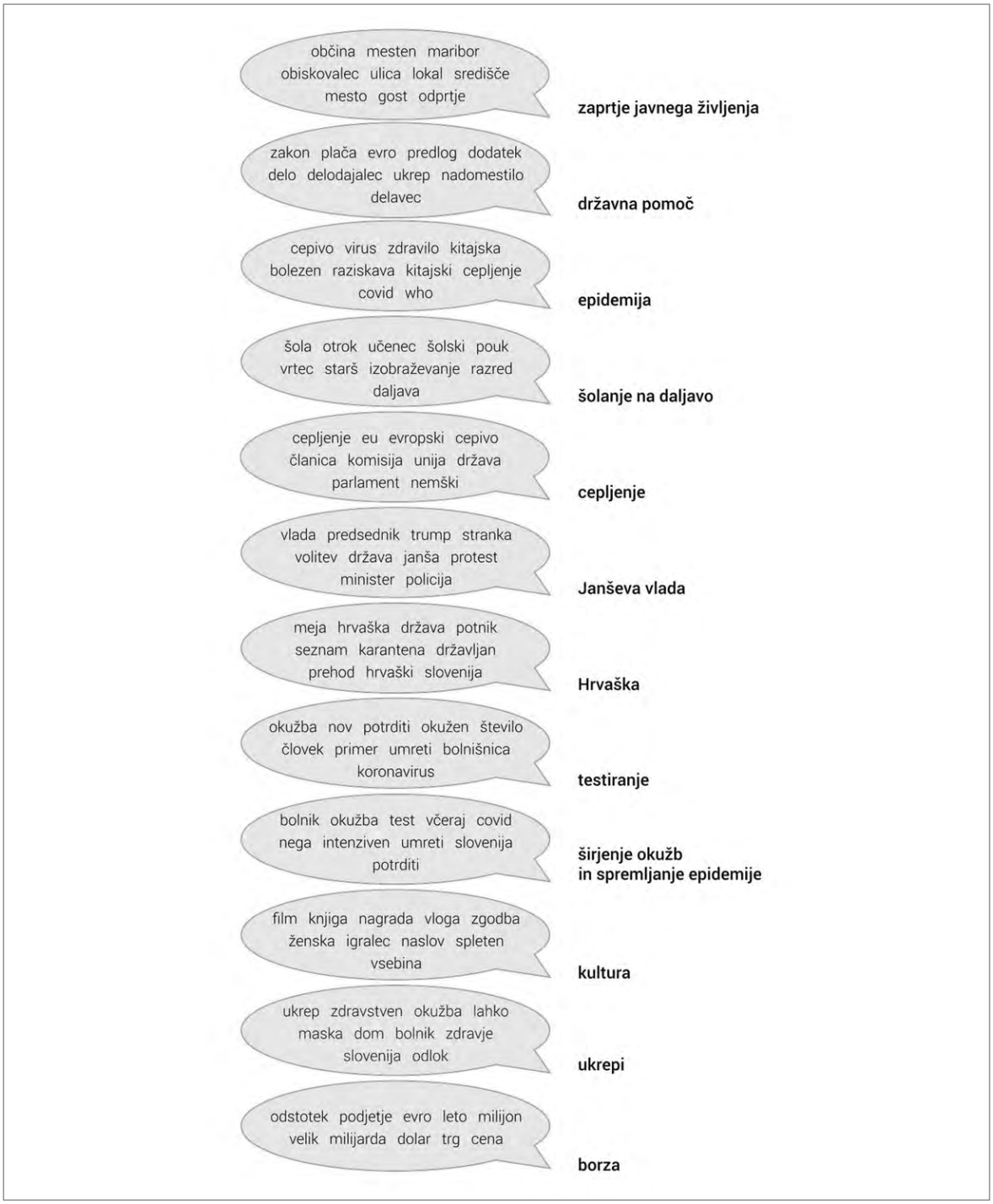
Z uporabo metode latentne Dirichletove alokacije - LDA (Blei in dr. 2003) smo avtomatsko prepoznali tematike v naboru Korpus-50. Metoda predvideva, da so dokumenti iz korpusa sestavljeni iz več tem, pri čemer je večja verjetnost, da vsak dokument obravnava manjše število tem. Podobno velja za besede: vsaka beseda z večjo verjetnostjo pripada manjšemu številu tem. Pred uporabo LDA smo besedila lematizirali, spremenili velike

1. https://docs.google.com/document/d/1gVpYwjCcmjuwVXDuNFKZp4FefXXZqFa7pACnsZnskg/edit?usp=sharing

začetnice v male, odstranili nepolnopomenske besede in utežili besede z mero TF-IDF, ki daje poudarek bolj specifičnim besedam za dokument.

Kot rezultat dobimo skupine besed, ki predstavljajo 20 najpogostejših tematik. Imena tematik smo določili ročno. Iz rezultatov smo odstranili teme, ki so vsebovale veliko šumnih podatkov ali se nam niso zdele zanimive za analizo (npr. športni dogodki). Končni seznam 12 tematik je prikazan na Sliki 1.

Slika 1: Najpogostejše tematike (metoda LDA).



Ekstrakcija besednih vložitev

Besedne vektorske vložitve, ki so natrenirane z uporabo nevronskih mrež, so predstavitve besed v prostoru, kjer vsako besedo opisuje vektor z več sto dimenzijami. Besede, ki so si blizu v vektorskem prostoru (kar lahko merimo s kosinusno razdaljo), so si tudi semantično podobne.

Pri statičnih vložitvah je posamezna beseda v korpusu predstavljena z enim vektorjem. Če reprezentacijo Korpusa-50 generiramo z modelom fastText[2] (Bojanowski in dr. 2017), so besedi koronavirus najbližje besede *nov, virus, enterovirus, pozitiven, testiranje, test, razširiti, izvid, okužba*. Za besedo *Janša* pa med 10 najbližjimi besedami najdemo tudi besedo *Šarec* in politike Višegrajske skupine ter desne evropske politike (Morawieck, Orban, Kurz).

Za razliko od statičnih vložitev, kjer vsako besedo predstavlja en vektor, pri kontekstualnih vložitvah vsako pojavitev besede opisuje svoj vektor. To je pomembno predvsem z vidika večpomenskih besed ali kjer analiziramo razlike med besedami v različnih kontekstih. Za eksperimente v nadaljevanju smo Korpus-14 lematizirali, kontekstualne vložitve pa smo zgradili z uporabo modela SloBERTA[3] (Ulčar in Robnik-Šikonja 2020). Povprečenje kontekstualnih reprezentacij na nivoju posamezne leme (osnovne oblike besede) (cf. Martinc in dr. 2020) v podkorpusu specifičnega medija nam omogoča primerjavo različnih medijskih vsebin.

Raznolikost poročanja tematik

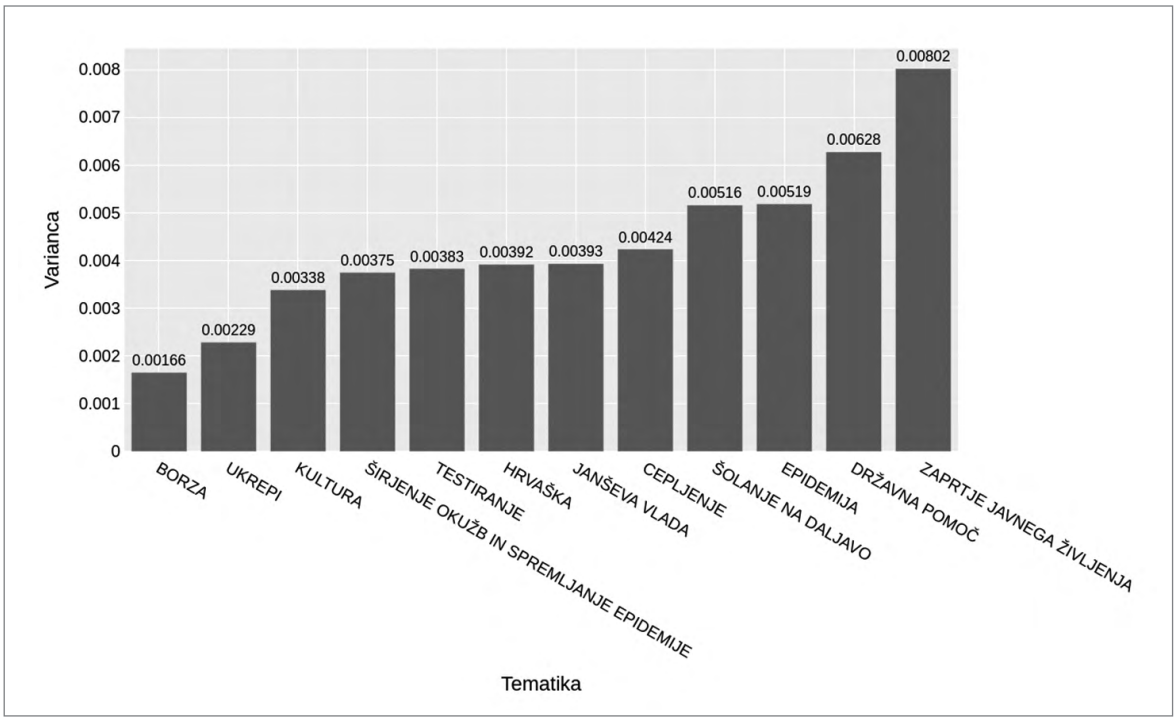
Za vektorsko reprezentacijo tematike za posamezen medij smo povprečili vse vložitve lem, ki tematiko opisujejo (besede v tematiki pripadajočem oblaku na Sliki 1). Z izračunom variance na množici reprezentacij na nivoju posameznega medija za vsako specifično tematiko lahko pridobimo oceno, kako se razlikuje poročanje o posamezni temi. Bolj kot se konteksti, v katerih se skupina besed iz tematike pojavlja, razlikujejo med mediji, večja bo varianca.

Iz Slike 2 vidimo, da se med štirinajstimi mediji najbolj raznoliko poroča o zaprtju in državni pomoči, najbolj enovito pa je poročanje o borzni tematiki. Raznolikost poročanja bi lahko bila povezana z različnimi zornimi koti in poudarki, ki jih imajo mediji na različne teme, vendar bi bilo za natančnejše razumevanje povezave med variancami in vsebinami potrebno nadaljne raziskovanje.

2. Pri metodi fastText je vsaka beseda predstavljena kot vsota vektorskih vložitev znakovnih n-gramov, ki jih beseda vsebuje. V praksi to pomeni, da metoda pri modeliranju semantične bližine upošteva tudi morfološko podobnost besed, zaradi česar je ta metoda še posebej uporabna za generiranje besednih vložitev v morfološko bogatih jezikih, kot je slovenščina.

3. Ta metoda za izdelavo kontekstualnih vložitev temelji na nevronski arhitekturi Transformer (Vaswani in dr. 2017), ki uporablja mehanizem pozornosti za določanje semantičnih relacij med besedami v kontekstu. Model, ki smo ga uporabili, je bil naučen na nenadzorovan način, na nalogi napovedovanja maskiranih žetonov v slovenskem korpusu, ki vsebuje 3,5 milijarde besed. Pri tej nalogi se 15% žetonov v korpusu zamenja z maskiranimi žetoni, model pa se nauči napovedovanja teh maskiranih žetonov s pomočjo nezamaskiranega konteksta.

Slika 2: Varianca tematik v medijih Korpusa-14.



Povezanost konceptov s covid-19

Z računanjem povezanosti med besedami (s pomočjo kosinusne razdalje med njihovimi vektorji) primerjamo povezanost izbranih konceptov s covid-19 v različnih portalih Korpusa-14.

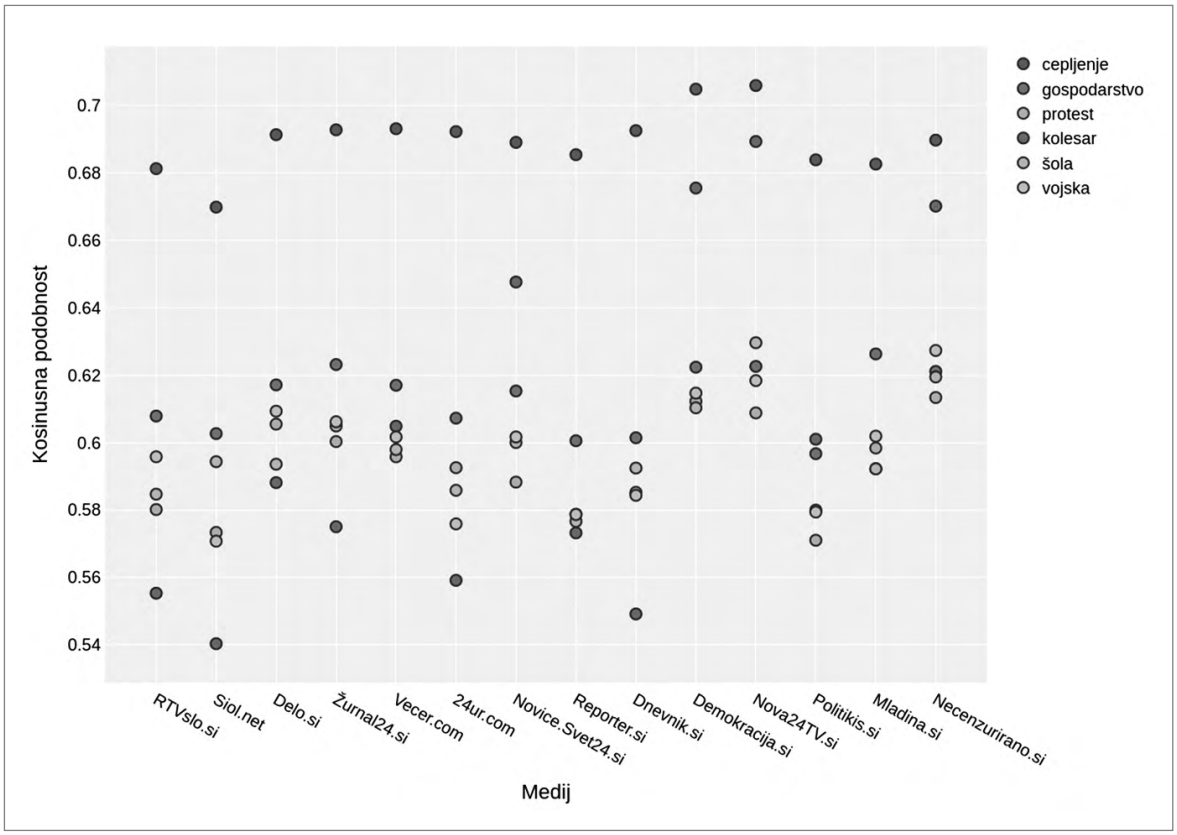
Besede so bile izbrane ročno glede na kriterij nudenja vpogleda v več aspektov epidemije in poročanja s strani različnih medijev. Kot zelo očitno povezano besedo smo izbrali *cepljenje*. Ker nas je zanimalo, ali se je več poročalo o gospodarskih ali izobraževalnih posledicah epidemije, smo vključili besede *gospodarstvo* in *šola*, dodali smo besedo *vojska*. Prav tako smo izbrali besede, za katere smo predpostavljali večje razlike med mediji (*kolesar* in *protest*).

Iz Slike 3 je razvidno, da je na vseh portalih najmočnejša povezava med covid-19 in cepljenjem, kar je pričakovano. Zanimiva je beseda *gospodarstvo*, ki se pri večini portalov bolj povezuje s covid-19 kot na primer beseda šola, kar morda priča o tem, da mediji v kontekstu epidemije covid-19 večji poudarek dajejo gospodarskim temam kot šolstvu, četudi so se zdele javne razprave o izvajanju šolanja v času epidemije prav tako v ospredju kot teme vezane na gospodarstvo.

Z vidika primerjave med portali se največja odstopanja pojavijo pri besedah, ki se manj samoumevno pojavljajo v povezavi s covid-19. To so hkrati tudi besede, za katere bi lahko rekli, da imajo večjo "ideološko obteženost", ker v javnih razpravah pogosto nastopajo kot označevalci s polariziranimi ideološkimi pomeni. To ponazarja primer besede *kolesar*, ki se močneje povezuje s covid-19 na portalih nova24tv.si, demokracija.si in necenzurirano.si. Veliko manj izrazita pa je povezava med besedo *kolesar* in covid-19 na siol.net. Povezanost besede kolesar s covid-19 je mogoče razložiti s politizacijo besede *kolesar*, tako pri medijih, ki so družbeni iniciativi kolesarjev naklonjeni (necenzurirano.si) kot tistimi, ki skušajo iniciativo diskreditirati (nova24tv.si, demokracija.si), manj pa tam, kjer je beseda rabljena izrazito v športnem kontekstu. Beseda *kolesar* je namreč v času epidemije, ko se je kolesarjenje

vzpostavilo kot protivladno protestniško gibanje, dobila nove konotativne pomene (konotira junaški upor proti vladni represiji na eni strani, na drugi pomeni razdiralno gibanje, ki škoduje aktualnemu političnemu establišmentu).

Slika 3: Povezave med izbranimi koncepti in covid-19 po različnih medijih.

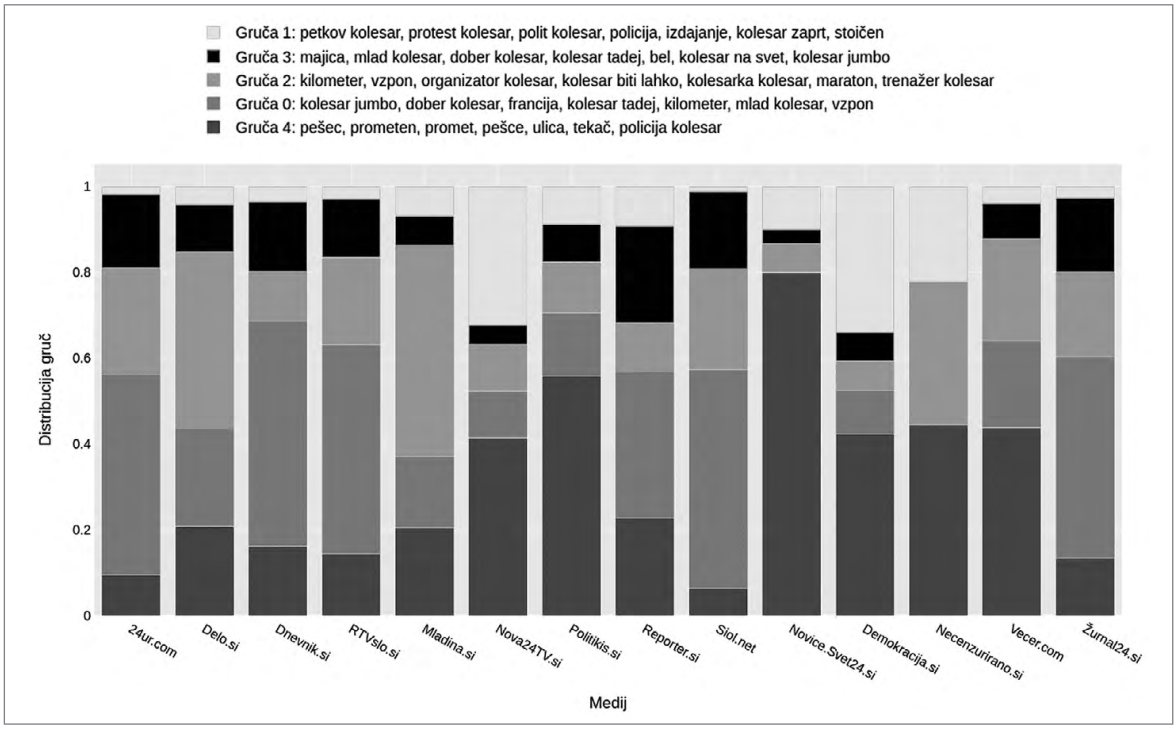


Razlikovanje besednih rab

Podobno kot Martinc in dr. (2021) smo za analizo razlik med mediji uporabili metodo gručenja pomenov in primerjavo distribucije gruč. Kontekstualne vektorje besednih pojavitev gručimo s pomočjo algoritma k-means (Steinley 2006). Za vsako besedo v Korpusu-14 njene rabe razdelimo na 5 gruč. Vsako gručo opišemo s skupkom ključnih besed oz. besednih nizov glede na TF-IDF. Razlike med portali lahko nato preučujemo z vidika razlik med distribucijami gruč.

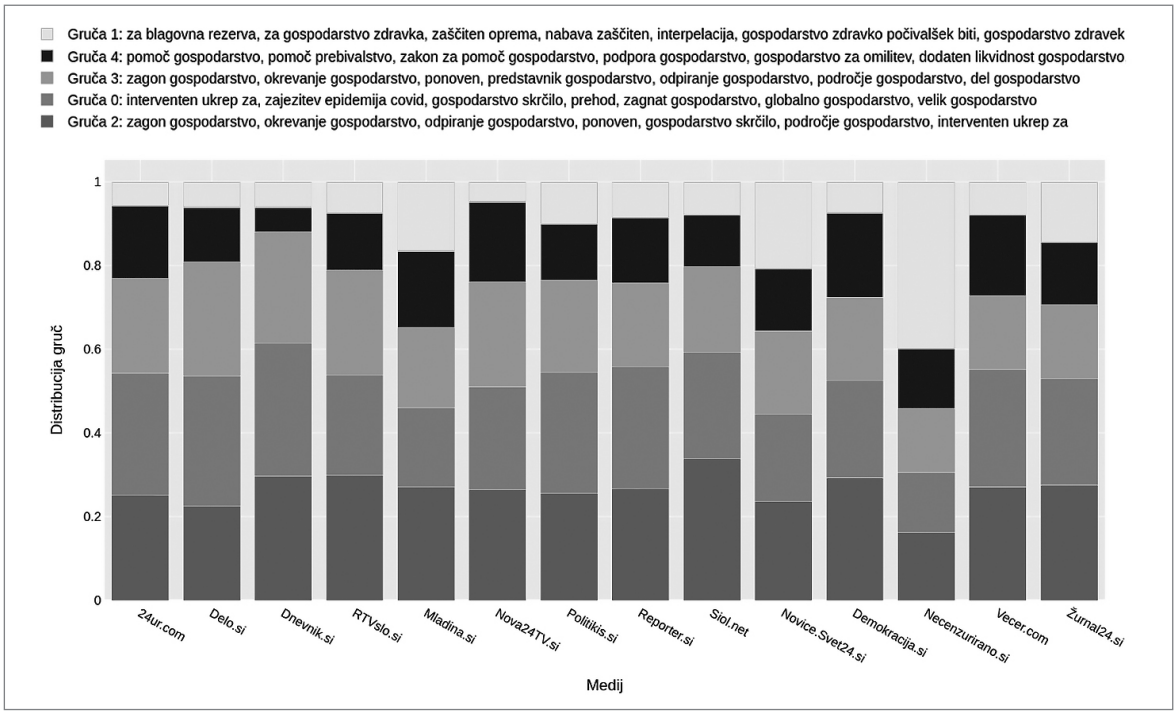
Metodo ponazorimo na primeru besede *kolesar* (Slika 4). Različne rabe besede *kolesar* v različnih gručah opisujejo pripadajoče besede. Zanimiva je predvsem gruča 1, ki se nanaša na politično konotacijo besede, saj jo označujejo pojmi kot *petkov kolesar, protest kolesarjev, policija*. Ostale štiri gruče pa se nanašajo na rabe besede kolesar v drugih, predvsem športnih kontekstih. Iz razlik med distribucijami lahko vidimo, da je gruča 1 izrazitejše zastopana na portalih demokracija.si, nova24tv.si in v nekoliko manjši meri v necenzurirano.si.

Slika 4: Prikaz gruč za besedo kolesar.



Tudi pri analizi gruč za besedo *gospodarstvo* (Slika 5) je zanimiva gruča 1 (z besedami blagovna rezerva, zaščitna oprema in interpelacija), ki zaznamuje ključne sporne teme v javnosti. Gruča 1 je najbolj zastopana v necenzurirano.si, ki je v času epidemije te teme tudi najbolj izpostavljal v kontekstu kritike delovanja vlade. Ta gruča je izrazitejša tudi na portalu mladina.si, ki se prav tako izrazito postavlja kot kritik vladnega delovanja, ter pri novičarskem tabloidu novice.svet24.si.

Slika 5: Prikaz gruč za besedo gospodarstvo.



Zaključki

V prispevku s pomočjo računalniških metod analiziramo poročanje slovenskih novičarskih portalov o epidemiji covid-19. Z metodo LDA identificiramo osrednje teme poročanja, kot so *epidemija, državna pomoč, šolanje na daljavo, cepljenje, gospodarstvo*, idr. Nato z uporabo kontekstualnih besednih vložitev analiziramo razlike v poročanju izbranih portalov. Zanimive razlike so predvsem pri bolj „ideološko obteženih" besedah, kot je *kolesar*, kjer je povezava s covid-19 močnejša tako na portalih, ki so protestom izraziteje naklonjeni (necenzurirano.si) kot med tistimi, ki skušajo iniciativo diskreditirati (nova24tv.si, demokracija.si). Prav tako z analizo različnih kontekstualnih pomenov pokažemo, da je na teh portalih politična raba besede *kolesar* bistveno bolj zastopana. V nadaljevanju bi bilo zanimivo pogledati tudi druge besede, ki polarizirajo javno razpravo (npr. *cepljenje, migrant, meja*).

Zahvala

Prispevek je rezultat raziskovalnega projekta Računalniško podprta večjezična analiza novičarskega diskurza s kontekstualnimi besednimi vložitvami (št. J6-2581) in programa Tehnologije znanja (št. P2-0103), ki ju financira ARRS, ter evropskega projekta EMBEDDIA (No. 825153), ki ga v okviru okvirnega programa za raziskave in inovacije Obzorje 2020 financira EU. Predstavljeni izsledki ne predstavljajo mnenja Evropske komisije in predstavlja izključno mnenja avtorjev.

Literatura

Blei, David M., Ng, Andrew. Y., in Jordan, Michael. I. (2003): Latent Dirichlet Allocation. Journal of Machine Learning Research, 3: 993–1022.

Bojanowski, Piotr, Grave, Edouard, Joulin, Armand, in Mikolov, Tomas (2017): Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5: 135–146.

Cantril, Hadley (1999): Invazija z Marsa. V: S. Splichal (ur.): Komunikološka hrestomatija: 137–150. Ljubljana: Fakulteta za družbene vede.

Entman, Robert M. (1994): Representation and Reality in the Portrayal of Blacks on Network Television News. Journalism Quarterly, 71(3): 509–520.

Hart, Sol P., Chinn, Sedona, Soroka, Stuart (2020): Politicization and Polarization in COVID-19 News Coverage. Science communication, 42(5): 679–697.

Leban, Gregor, Fortuna, Blaž, Brank, Janez, in Grobelnik, Marko (2014): Event Registry: Learning about World Events from News. V: Proceedings of the 23rd International Conference on World Wide Web (WWW ,14 Companion): 107–110. New York: Association for Computing Machinery.

McCombs, Maxwell E., in Shaw, Donald L. (1972): The Agenda-Setting Function of Mass Media. The Public Opinion Quarterly, 36(2): 176–187. Dostopno prek: http://www.jstor.org/stable/2747787 (14. 6. 2021).

Hubner, Austin (2021): How did we get here? A framing and source analysis of early COVID-19 media coverage. Communication Research Reports, 38(2): 112–120

iPROM in Valicon (2021): Medijska potrošnja 2021. Dostopno prek: https://iprom.si/files/2021/05/iPROM-in-Valicon-raziskava-Medijska-potrosnja-2021-Porocilo-iPROM-Press.pdf (15. 6. 2021).

Liu, Qian, Zheng, Zequan, Zheng, Jiabin, Chen, Qiuyi, Liu, Guan, Chen, Sihan, Chu, Bojia, Zhu, Hongyu, Akinwunmi, Babatunde, Huang, Jian, Zhang, Casper J. P., in Ming, Wai-Kit (2020):

Health Communication Through News Media During the Early Stage of the COVID-19 Outbreak in China: Digital Topic Modeling Approach. Journal of Medical Internet Research, 22(4): e19118

Martinc, Matej, Kralj Novak, Petra, in Pollak, Senja (2020): Leveraging Contextual Embeddings for Detecting Diachronic Semantic Shift. V: Proceedings of the 12th Conference on Language Resources and Evaluation: 4811–4819.

Martinc, Matej, Perger, Nina, Pelicon, Andraž, Ulčar, Matej, Vezovnik, Andreja, in Pollak, Senja (2021): EMBEDDIA Hackathon Report: Automatic Sentiment and Viewpoint Analysis of Slovenian News Corpus on the Topic of LGBTIQ+. V: Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation: 121–126.

Montariol, Syrielle, Martinc, Matej, in Pivovarova Lidia (2021): Scalable and Interpretable Semantic Change Detection. V: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: 4642–4652.

Mutua, Sylvia Ndanu, in Ong'ong'a, Daniel Oloo (2020): Online News Media Framing of COVID-19 Pandemic: Probing the Initial Phases of the Disease Outbreak in International Media. European Journal of Interactive Multimedia and Education, 1(2): e02006.

Rebello, Katarina, Schwieter, Christian, Schliebs, Marcel, Joynes-Burgess Kate, Elswah, Mona, Bright, Jonathan, in Howard, N. Philip. (2020): Covid-19 News and Information from State-Backed Outlets Targeting French, German and Spanish-Speaking Social Media Users. Understanding Chinese, Iranian, Russian and Turkish Outlets. Data memo. Dostopno prek: https://demtech.oii.ox.ac.uk/wp-content/uploads/sites/93/2020/06/Covid-19-Misinfo-Targeting-French-German-and-Spanish-Social-Media-Users.pdf (3. 6. 2021).

Steinley, Douglas (2006): K-Means Clustering: a Half-Century Synthesis. British Journal of Mathematical and Statistical Psychology, 59(1): 1–34.

Ulčar, Matej, in Robnik-Šikonja, Marko (2020): Slovenian RoBERTa contextual embeddings model: SloBERTa 1.0, Slovenian language resource repository CLARIN.SI. Dostopno prek: http://hdl.handle.net/11356/1387 (1. 5. 2021).

Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion in Gomez, Aidan N., Kaiser, Lukasz in Polosukhin, Illia (2017): Attention is all you need. V: Proceedings of the 31st International Conference on Neural Information Processing Systems: 6000–6010.