Spremljevalni korpus Trendi in avtomatska kategorizacija

Iztok KOSEM
Filozofska fakulteta, Univerza v Ljubljani; Institut Jožef Stefan;
Fakulteta za računalništvo in informatiko, Univerza v Ljubljani

Jaka ČIBEJ
Institut Jožef Stefan; Filozofska fakulteta, Univerza v Ljubljani

Kaja DOBROVOLJC
Filozofska fakulteta, Univerza v Ljubljani; Institut Jožef Stefan

Taja KUZMAN
Institut Jožef Stefan

Nikola LJUBEŠIĆ
Institut Jožef Stefan; Fakulteta za računalništvo in informatiko, Univerza v Ljubljani;
Inštitut za novejšo zgodovino

Prispevek predstavlja izdelavo korpusa Trendi, prvega spremljevalnega korpusa za slovenščino. Trenutna različica Trendi 2023-02 pokriva besedila od januarja 2019 do konca februarja 2023, vsebuje pa že več kot 700 milijonov pojavnic oz. več kot 586 milijonov besed. Namen korpusa je, da tako strokovni kot nestrokovni javnosti ponudi podatke o aktualni jezikovni rabi in omogoči spremljanje pojavljanja novih besed ter upadanja ali naraščanja rabe že obstoječih. Poleg same vsebine predstavimo tudi metodologijo in načela izdelave korpusa. Drugi del prispevka opisuje razvoj algoritma za avtomatsko kategorizacijo besedil z novičarskih portalov, ki je bil pripravljen za potrebe korpusa Trendi in tudi drugih korpusov s tovrstnimi besedili. Za namene algoritma je bil izdelan nabor 13 tematskih kategorij, ki so v veliki meri prekrivne z mednarodnimi standardi in

Kosem, I., Čibej, J., Dobrovoljc, K., Kuzman, T., Ljubešić, N.: Spremljevalni korpus Trendi in avtomatska kategorizacija. Slovenščina 2.0, 11(1): 161–188.

1.01 Izvirni znanstveni članek / Original Scientific Article

DOI: https://doi.org/10.4312/slo2.0.2023.1.161-188

kategorijami v primerljivih korpusih drugih jezikov. Na besedilih, označenih s kategorijami, smo naučili več različnih jezikovnih modelov in z najprimernejšim dosegli visoko zanesljivost določevanja tematike besedilom.

Ključne besede: spremljevalni korpus, avtomatska kategorizacija besedil, neologizmi, novičarski portali, slovenščina

1 Uvod

Jezik se nenehno spreminja, pojavljajo se nove besede, obstoječe besede in besedne zveze dobivajo nove pomene, določene besede ali njihovi pomeni se prenehajo uporabljati ipd. V zadnjem času, tudi zaradi epidemije covida-19, ki je prinesla veliko novega izrazoslovja, je še posebej veliko pozornosti deležno področje neologije, tako leksikalne (nove besede) kot semantične (novi pomeni).

Za spremljanje sprememb v jeziku se tipično uporabljajo spremljevalni korpusi, ki vsebujejo najnovejša besedila v jeziku. Spremljevalni korpusi zapolnjujejo manko referenčnih korpusov, katerih izdelava zaradi raznovrstnosti besedil in njihovih formatov ter obsega traja dlje časa. V času tehnološkega napredka in ob dejstvu, da je zdaj zelo veliko besedil dostopnih na spletu, je izdelava spremljevalnih korpusov postala enostavnejša; kar je objavljeno danes, je lahko že jutri vključeno v korpus (seveda pod pogojem, da besedilo ustreza kriterijem za vključitev in ga je tudi z vidika avtorskih pravic mogoče dodati v korpus).

Za slovenščino kljub bogati opremljenosti na področju korpusov do zdaj nismo imeli spremljevalnega korpusa, čeprav se je med različnimi deležniki kazala jasna potreba po njem. Naslavljanja tega manka smo se lotili v okviru projekta *Spremljevalni korpus in spremljajoči podatkovni viri* (SLED),[1] ki je potekal od oktobra 2021 do novembra 2022 in ga je sofinanciralo Ministrstvo za kulturo Republike Slovenije.

Projekt SLED je naslovil še eno s korpusi povezano potrebo jezikoslovne skupnosti, in sicer algoritem za avtomatsko pripisovanje tematike korpusnim besedilom z novičarskih portalov. Tematsko modeliranje je dokaj razvita disciplina, ki je bila v slovenskem prostoru že uporabljena za primerjavo korpusov (Logar idr. 2015, Logar Berginc in

1 https://sled.ijs.si/

Ljubešić 2013), zbirke besedil z metapodatki o tematiki pa še niso bile obogatene s strojnimi metodami, čeprav je za to veliko potenciala. Metapodatek o tematiki, ki jo besedilo naslavlja, je koristen pri semantični analizi besed in besednih zvez, saj lahko takoj pokaže oz. opozori na morebitno omejenost rabe (npr. *dvojni dvojček* je omejen na področje športa). Pri spremljevalnih korpusih se besedila zbirajo dnevno (najpogosteje gre za besedila, ki jih objavljajo novičarski portali) in imajo le redko eksplicitno pripisano tematiko (podobno velja tudi za besedila v drugih korpusih). Ročno pripisovanje tematik bi bilo zaradi velike količine besedil izjemno zamudno in dolgoročno nevzdržno, zato je smiselno izdelati orodje za avtomatsko pripisovanje tematike.

V prispevku najprej ponujamo pregled nekaterih pomembnejših tujih spremljevalnih korpusov, nato pa predstavimo metodologijo in vsebino spremljevalnega korpusa Trendi. Sledi predstavitev izdelave orodja za avtomatsko kategorizacijo tematike besedil novičarskih portalov, pri čemer je bil pomemben del tudi izdelava nabora tematskih kategorij, ki je dovolj podroben, da razdeli besedila v smiselne in karseda celovite kategorije, in hkrati dovolj robusten, da za strojno kategorizacijo ni pretežaven. Na besedilih, označenih s kategorijami, smo naučili več različnih jezikovnih modelov in z najprimernejšim dosegli visoko zanesljivost določevanja tematike besedilom. V zaključku predstavimo načrte za prihodnje delo.

2 Spremljevalni korpusi

V mednarodnem prostoru so spremljevalni korpusi prisotni že od 20. stoletja, saj se je ideja sistematičnega na podatkih temelječega spremljanja sprememb v jeziku pojavila kmalu po nastanku prvih korpusov. Eden prvih spremljevalnih korpusov je bil angleški korpus Bank of English, ki je bil prvič objavljen leta 1991, z namenom rednega posodabljanja z novimi besedili iz pisnih in govorjenih besedil tako britanske kot ameriške angleščine. Korpus je bil kasneje vključen v 4,5-milijardni korpus COBUILD založbe Collins in po zadnjih podatkih obsega več kot 650 milijonov besed, a ima precej omejen dostop, saj ga lahko poleg zaposlenih na založbi Collins prosto uporabljajo zgolj zaposleni in študentje na Univerzi v Birminghamu.

Za angleščino je danes pomemben predvsem korpus NOW (News on the Web; Davies, 2016-), ki se dnevno posodablja z besedili različnih spletnih novičarskih strani (do 200 milijonov besed na mesec) in danes vsebuje več kot 16 milijard besed. Tako kot nekateri drugi v nadaljevanju omenjeni korpusi je korpus NOW prosto dostopen tako za brskanje preko spletnega portala[2] kot prenos podatkov na lokalni računalnik.[3] Med njimi je denimo tudi korpus Coronavirus (Davies 2019-), specializirani podkorpus korpusa NOW, ki se od januarja 2020 dnevno posodablja s 3–4 milijoni novimi besedami iz angleških spletnih novic na temo pandemije covid-19.

Obsežna zbirka korpusov za spremljanje sprememb v jeziku, ki poleg angleščine pokriva še več kot 35 drugih jezikov, so korpusi Timestamped JSI, ki vsebujejo spletne novice, zbrane preko storitve JSI Newsfeed na Institutu »Jožef Stefan« (Trampuš in Novak, 2012). Korpusi za 18 jezikov so na voljo v orodju Sketch Engine (Kilgarriff idr., 2004; Bušta idr., 2017), v katerem imajo poleg ostalih funkcij orodja uporabniki na voljo tudi t. i. Trende (Herman in Kovár, 2013), funkcijo, ki pomaga prepoznavati trende v rabi besed in je na voljo tudi za nekatere obsežnejše diahrone korpuse. Korpusi JSI Newsfeed v Sketch Enginu vsebujejo besedila od 2014 do aprila 2021 (čas zadnje posodobitve) in so različnih velikosti; korpus angleščine na primer vsebuje približno 60 milijard besed. V primerjavi z drugimi spremljevalnimi korpusi je posebnost te večjezične zbirke tudi to, da so poleg običajnih metapodatkov o datumu, viru in jeziku vsebovanih besedil besedila kategorizirana tudi glede na mesec, četrtletje, področje, lokacijo, ključne besede in druge podobne kategorije, relevantne za nadaljnjo analizo.

Obstaja še precej drugih spremljevalnih korpusov, ki so omejeni na določeno vrsto ali področje besedil (Mikko idr., 2018; Mauri idr., 2019; De Smedt, 2021) ali pa so na voljo zgolj za interno rabo. Primer takšnega korpusa je ONLINE, dinamični spremljevalni korpus češkega jezika, ki ga izdeluje Inštitut za češki nacionalni korpus. Velik je približno 6,3 milijarde besed in vsebuje spletne novice in komentarje pod njimi ter besedila s forumov in različnih družabnih omrežij. Korpus ONLINE je razdeljen na dva komplementarna korpusa: ONLINE_NOW

2 https://www.english-corpora.org/
3 https://www.corpusdata.org/

in ONLINE_ARCHIVE. Prvi je posodobljen vsak dan in pokriva obdobje preteklih šestih mesecev. ONLINE_ARCHIVE pokriva obdobje od februarja 2017 do prvega meseca, ki ga vsebuje ONLINE_NOW. Tako se vsebina zadnjega meseca po starosti v korpusu ONLINE_NOW na začetku vsakega meseca preseli v ONLINE_ARCHIVE.

Do določene mere lahko vlogo spremljevalnega korpusa opravljajo tudi diahroni korpusi, seveda pod pogojem, da vsebujejo čim novejša besedila. Kot primer lahko navedemo korpus sodobne ameriške angleščine (Corpus of Contemporary American English; Davies, 2008-), ki vsebuje besedila od leta 1990 do marca 2020 (zadnja posodobitev) in obsega več kot milijardo besed. Prednost korpusa je, da je žanrsko uravnotežen, saj vsebuje besedila iz osmih različnih žanrov (govorjeni jezik, leposlovje, revije, časopise, znanstvena besedila, televizijske in filmske podnapise, bloge in ostale spletne strani). Slovenski ekvivalent bi bil korpus Gigafida 2.0 (Krek idr., 2019), ki obsega 1,13 milijarde besed, vendar pa je v primerjavi s korpusom sodobne ameriške angleščine manj ažuren (vsebuje samo besedila do leta 2018). Kot hibridni pristop med zasnovo spremljevalnega korpusa na eni strani in statičnega na drugi lahko omenimo še češke korpuse serije SYN (Hnátková, 2014), ki se kot neprekrivni sinhroni korpusi izdajajo vsakih pet let (npr. SYN2000, SYN2005, SYN2010 itd.) in skupaj tvorijo referenčni korpus sodobne pisne češčine SYN, ki v zadnji verziji obsega več kot 6 milijard besed.

Za slovenščino pravi spremljevalni korpus do danes še ni obstajal. Obstajajo sicer spletne storitve, kot je Jezikovni sledilnik (Kosem idr., 2021), ki že izkorišča najsodobnejše podatke o jezikovni rabi, v konkretnem primeru podatke že omenjene storitve JSI Newsfeed, za izdelavo neke vrste začasnih korpusov, na katerih se potem izvajajo specifični statistični izračuni.[4] Taka ciljna raba je seveda smiselna, vendar pa je namenjena nestrokovni javnosti; po drugi strani strokovna javnost, kot so leksikografi_ke, jezikoslovci_ke in drugi raziskovalci_ke, potrebujejo tudi dostop do izvirnih besedil in kompleksnejših metapodatkov, če želijo opravljati še druge analize.

4 Podobno je zasnovana priljubljena storitev Google Ngram Viewer (https://books.google.com/ngrams/), ki omogoča analizo rabe besedišča skozi čas. Temelji na podatkovni zbirki Google Books NGram Corpus (Michel idr. 2011), ki vsebuje najpogostejše besede oz. besedne nize iz sicer nedostopne zbirke digitaliziranih besedil od začetka 17. stoletja dalje.

3 Korpus Trendi

Izdelave prvega spremljevalnega korpusa za slovenščino, ki smo ga poimenovali Trendi, smo se lotili v okviru projekta SLED. Poleg izdelave in rednega posodabljanja korpusa Trendi je imel projekt še dva cilja: pripravo na korpusnih podatkih temelječe statistike o različnih vidikih rabe besed in izdelavo orodja, ki besedila avtomatsko opremi s podatkom o tematski kategoriji (glej razdelek 4).

3.1 Metodologija in vsebina korpusa

Z metodološkega vidika smo pri snovanju korpusa Trendi morali sprejeti dve odločitvi: obdobje, ki ga bo korpus pokrival, in kako pogosto bo korpus posodobljen. Pri odločitvi o obdobju smo izhajali iz želje, da bi korpus Trendi vedno pokrival manko najnovejše različice referenčnega (pisnega) korpusa Gigafida, v času pisanja prispevka je bila zadnja različica 2.0. V tem trenutku to pomeni, da Trendi vsebuje besedila od januarja 2019 naprej. Ob objavi nove različice korpusa Gigafida se besedila iz korpusa Trendi dodajo v korpus Gigafida, obdobje, ki ga pokriva nova različica korpusa Trendi, pa se temu ustrezno prilagodi.

Tesna povezanost s korpusom Gigafida tudi pomeni, da bo korpus Trendi predstavljal standardno pisno slovenščino. Odločitev se zdi smiselna tudi zato, ker sta nestandardna oz. govorjena slovenščina pokrita s korpusi, kot sta JANES[5] in Gos,[6] in je torej njun razvoj predmet ločenih projektov.

Pri pripravi seznama virov za vključitev v korpus Trendi smo izhajali iz seznama slovenskih spletnih virov, ki jih najdemo v servisu JSI Newsfeed. Izdelali smo seznam vseh virov od leta 2019 do konca 2021, pridobili smo tudi podatek o skupnem številu besedil na vir. Nato smo pri pripravi seznama za korpus Trendi podrobno analizirali vsakega od 243 virov. 90 virov smo izključili, ker je šlo za tuje ali slovenske spletne strani z vsebino v tujem jeziku. Nato smo s seznama odstranili še več kot 40 virov, nekatere zato, ker niso vsebovali medijskih novic (blogi, spletne strani vladnih uradov in podjetij), druge zato, ker je njihova vsebina preveč specializirana (npr. repozitoriji akademskih publikacij

5 https://www.clarin.si/kontext/query?corpname=janes
6 http://www.korpus-gos.net/

so primernejši za korpuse, kot je Korpus akademske slovenščine – glej Žagar idr., 2022). Ena od strani (preberi.si) je bila s seznama odstranjena zato, ker je agregator novic iz drugih virov. Končni seznam korpusa Trendi tako vsebuje 107 virov, med tistimi, ki so v obdobju 2019–2021 prispevali največ novic, so sta.si (260.080 besedil), rtvslo.si (97.924), siol.net (69.471), delo.si (65.415), 24ur.com (61.623), dnevnik.si (47.749) in vecer.com (45.548).

Seznam virov se bo redno posodabljal, saj lahko pričakujemo pojav novih spletnih strani, pa tudi ukinitev obstoječih. Kot primer lahko navedemo spletno stran necenzurirano.si, ki se je pojavila šele leta 2020 in je že 28. po številu novic (8.494). Dodajanje novih virov v korpus pomeni tudi večje število besed na mesečni ravni in posledično večji korpus Trendi. Trenutni okvirni izračuni kažejo, da se bo Trendi vsak mesec povečal za 10-15 milijonov pojavnic, pri čemer je bil povprečen mesečni obseg leta 2019 12,5 milijona pojavnic, leta 2021 pa že 21 milijonov pojavnic.

Zaradi narave korpusa Trendi bodo potrebne redne posodobitve, ki so zaenkrat predvidene na mesečni ravni, kot je praksa pri podobnih tujih korpusih.[7] To se zdi trenutno realno, upoštevajoč časovno zahtevnost pridobivanja in označevanja besedil, pretvorb v potreben format in vključevanje korpusa v konkordančnike.

3.2 Priprava besedil

Za pripravo besedil smo pripravili cevovod, ki vključuje pridobivanje besedil, označevanje na različnih ravneh, združevanje po virih in obdobjih ter pretvorbo v različne formate. Pridobivanje besedil je zaenkrat vezano na servis JSI Newsfeed, ki uporablja protokol RSS novic. Nekateri viri, kot so sta.si, delo.si itd. imajo določene vsebine zaklenjene oziroma so dostopne samo naročnikom, zato so v teh primerih pri pridobivanju prek protokola RSS prosto dostopni samo povzetki ali prvih nekaj odstavkov, včasih celo samo naslov in podnaslov. Pri reševanju tega problema smo združili moči z ekipo, ki v okviru projekta *Razvoj slovenščine v digitalnem okolju* (RSDO) sklepa pogodbe z besedilodajalci.

7 Pri začetnih verzijah smo se še srečevali z določenimi hrošči pri pridobivanju in pretvarjanju besedil, zato je prihajalo do nekajmesečnih zamikov pri objavi korpusa.

Dogovor z besedilodajalci vključuje redno dostavljanje celotnih besedil. Posledično bo končna oblika cevovoda za korpus Trendi kombinacija priprave besedil, pridobljenih s spleta, in besedil, ki jih bodo v digitalni obliki poslali tisti besedilodajalci, ki preko protokola RSS ne omogočajo dostopa do celotnih besedil.

Del postopka pridobivanja besedil je tudi deduplikacija, ki je trenutno omejena zgolj na raven vira besedila; del cevovoda je namreč preverjanje, da se besedilo z istim URL-jem ne ponovi. Zavedamo se, da zaradi pokrivanja istih dogodkov obstaja velika prekrivnost med viri. Še več, mnogi viri osnujejo številne novice na podlagi vsebin sta.si, kar pripelje do podvajanja besedila na ravni stavkov, odstavkov ali tudi celotne vsebine. Kljub temu za namene korpusa Trendi deduplikacija na ravni vsebine ni predvidena, saj želimo uporabnikom omogočiti analizo vsebin posameznih virov ter primerjalne analize med viri. Deduplikacija pa bo najbrž opravljena pri pripravi besedil za novo različico korpusa Gigafida, kot je bila praksa v preteklih različicah (Krek idr., 2019).

Sledi postopek strojnega označevanja besedil, za kar uporabljamo označevalni cevovod CLASSLA-Stanza (Ljubešić in Dobrovoljc, 2019),[8] ki se kot referenčno orodje za slovnično označevanje besedil v slovenščini aktivno razvija v okviru projekta RSDO. Orodje je nadgradnja odprtokodnega orodja Stanza (Qi idr., 2020), ki v primerjavi z izvorno programsko opremo podrobneje naslavlja specifike slovenščine, zlasti na ravni stavčne segmentacije, tokenizacije, oblikoskladenjskega označevanja in lematizacije po sistemu JOS (Erjavec idr., 2010). Poleg navedenih ravni orodje besedila tudi skladenjsko razčleni po sistemu Universal Dependencies (Dobrovoljc idr. 2017) in v njih označi imenske entitete (Zupan idr., 2017), kot so imena oseb, krajev, organizacij ipd.

Po končanem označevanju se v cevovodu opravi še pretvorba besedil iz privzetega formata označevalnega orodja (CONNL-U) v TEI XML, ki ga med drugim potrebujemo za statistične izračune s programom LIST (Krsnik idr., 2019). V ta proces sta vključena še dva povezana postopka združevanja besedil: združevanje besedil po viru na dan (vsakodneven postopek) in združevanje besedil istega vira za cel mesec (enkrat na mesec, na začetku novega meseca za nazaj). V zadnjem koraku, ki ga

8 https://pypi.org/project/classla/

izvajamo enkrat mesečno in ga moramo pognati ločeno zaradi kombinacije XSLT in skripte Perl, je opravljena še pretvorba mesečnih datotek (razdeljenih po viru) v format VERT, ki ga uporabljata konkordančnika KonText (Machálek, 2020)[9] in NoSketch Engine (Rychlý, 2007)[10].

3.3 Zadnja različica in dostopnost korpusa Trendi

Prva različica korpusa Trendi, imenovana Trendi 2022-05, je bila objavljena junija 2022 in je vsebovala 565.308.991 pojavnic oz. malo več kot 473 milijonov besed. Trenutna zadnja različica Trendi 2023-02 pokriva besedila do konca februarja 2023, vsebuje pa že več kot 700 milijonov pojavnic oz. več kot 586 milijonov besed, kar je v skladu z našimi ocenami, da se bo korpus mesečno povečeval za 10–15 milijonov besed. V korpusu je 1.786.645 besedil 71 izdajateljev, pri čemer imajo največje deleže Slovenska tiskovna agencija (417.718 besedil; 23,4 %), Delo d.o.o. (164.300; 9,2 %), Radiotelevizija Slovenija (150.039; 8,4 %), Media24 d.o.o. (118.651; 6,6 %), PRO PLUS d.o.o. (108.736; 6,1 %) in TSMedia d.o.o. (101.846; 5,7 %).[11]

Korpus Trendi je za brskanje prosto dostopen v treh konkordančnikih CLARIN.SI – konkordančniku KonText in dveh različicah konkordančnika NoSketch Engine; tako KonText kot NoSketch Engine imata več enakih funkcionalnosti (enostavno in napredno iskanje ipd.), vendar pa KonText ponuja možnost registracije in shranjevanje iskanj in priljubljenih korpusov, NoSketch Engine pa dodatne funkcionalnosti, kot je luščenje ključnih besed (angl. *keywords*) iz korpusov, za uporabo katerih ni potrebna registracija. Konkordančnik NoSketch Engine je na CLARIN.SI poleg starejše različice (Bonito) po novem na voljo tudi v novejši različici uporabniškega vmesnika (Crystal),[12] ki zagotavlja izboljšano uporabniško izkušnjo in dolgoročnejše vzdrževanje.

Korpus Trendi kot podatkovna množica trenutno ni na voljo, saj avtorskopravna razmerja z izdajatelji besedil še niso urejena. Pogodbe z

9 https://www.clarin.si/kontext/query?corpname=trendi
10 https://www.clarin.si/noske/run.cgi/corp_info?corpname=trendi
11 Analiza pokaže, da so deleži izdajateljev iz meseca v mesec dokaj nespremenljivi. Najnovejši podatki so na voljo na https://www.clarin.si/ske/#text-type-analysis?corpname=trendi&wlminfreq=1&wlicase=1&include_nonwords=1&showresults=1&wlnums=frq&wlattr=text.publisher.
12 https://www.clarin.si/ske/#dashboard?corpname=trendi

besedilodajalci se sicer hkrati urejajo tudi za referenčni korpus Gigafida. Ko bo to urejeno, bo korpus Trendi na voljo pod odprto licenco z vzorčenimi odstavki (glej Logar idr. 2013[13] kot primer takšnega vira), v celoti pa za posamezne raziskovalce, ki bodo podpisali posebno pogodbo za uporabo korpusa v znanstvenoraziskovalne namene.

4 Avtomatska kategorizacija besedil

Sorodni spremljevalni korpusi, predstavljeni v 2. razdelku, poleg podatkov o letu nastanka besedil pogosto vsebujejo tudi podatek o področni kategorizaciji besedil, ki omogoča opazovanje jezikovnih trendov znotraj posameznih tematskih področij ali identifikacijo trendov, ki se pojavljajo zgolj na določenem področju. Korpusi Timestamped JSI tako vsebujejo informacijo o ključnih besedah in področjih glede na kategorizacijo ontologije DMOZ (Grobelnik idr., 2006), npr. družba, posel, šport, ki je besedilom avtomatsko pripisana z orodjem Enrycher (Štajner 2009).

Podobno tudi češki korpusi serije SYN in ONLINE vsebujejo podrobno večnivojsko kategorizacijo besedil (Cvrček idr., 2020) glede na skupino besedil (npr. leposlovje, stvarna besedila, publicistika), vrsto besedil (npr. znanstvena, strokovna, poljudna besedila), žanrsko skupino (npr. družboslovje) in žanrsko področje (npr. ekonomija, politika, pravo, psihologija), pri čemer sama metoda klasifikacije ni podrobneje dokumentirana.

Manj podrobna je kategorizacija besedil v korpusih NOW, COCA in Coronavirus, saj vsebujejo zgolj delitev na žanre (npr. govorjeni jezik, leposlovje, revije) in za določene med njimi (Sharoff 2018) tudi funkcijske tipe (npr. pravna besedila, navodila, recenzije, promocije), a to pomanjkljivost delno naslavlja njihov skupni konkordančnik, ki za dani konkordančni niz izpiše tudi topike oz. seznam ključnih besed glede na avtomatsko detekcijo pogosto ponavljajočih se besed v prikazanih besedilih (npr. simptom, alergija, pljuča).

4.1 Priprava tematskih kategorij

Ena od aktivnosti projekta SLED je bila tudi izdelava orodja za avtomatsko kategorizacijo besedil glede na tematiko. Za izdelavo takšnega orodja oz. modela zanj potrebujemo dvoje: nabor kategorij in učne množice.

13 https://www.clarin.si/repository/xmlui/handle/11356/1035

Pri izdelavi nabora kategorij smo se opirali na podatke iz treh skupin virov:

- slovenskih novičarskih portalov, izbrali smo jih šest, tj. rtvslo.si, delo.si, sta.si, dnevnik.si, 24ur.com in vecer.com;
- nabora tematskih kod oz. kategorij Mednarodnega tiskovnega telekomunikacijskega sveta (IPTC).[14] S tem smo tudi želeli zagotoviti čim boljšo usklajenost naših kategorij z mednarodnim standardom;
- kategorij v sodobnih sinhronih in spremljevalnih korpusih, pri čemer sta bila relevantna predvsem češki korpus SYN_2015 (Křen idr., 2016) in estonski nacionalni korpus (Koppel in Kallas, v tisku).

Glavno vodilo pri pripravi klasifikacije je bilo pripraviti relativno majhen nabor kategorij, v katere lahko uvrstimo vse novice na različnih portalih. S tem bi zagotovili tudi boljše delovanje modela. Posledično smo pri analizi uporabljenih virov več pozornosti posvečali krovnim kategorijam, kar je bilo sploh potrebno pri naboru IPTC, ki ima približno 1.400 kategorij, razdeljenih v tri nivoje (s tem da krovni nivo sestavlja le 17 kategorij). Za ponazoritev smiselnosti uporabe zgolj krovnih kategorij lahko vzamemo kategorijo *šport*, ki ima na večini novičarskih portalov nadaljnje kategorije, od katerih se povsod pojavita samo *nogomet* in *košarka*, ostale pa le na nekaterih portalih, npr. dnevnik.si nima *zimskih športov*, ima pa ločeno podstran za novice o *Luki Dončiću*; rtvslo.si je edini, ki ima podstran za novice o *Formuli 1*, 24ur.si ima ločene podstrani za *Ligo prvakov* in *Ligo Evropa* (nogomet) ter *borilne športe*.

Končna klasifikacija vsebuje 13 kategorij:

- **Umetnost in kultura**. Vključuje besedila o kulturi, umetnosti, filmih, knjigah, gledališču, pa tudi recenzije ipd.
- **Črna kronika**. Naravne in ostale nesreče, človeški delikti, kriminal.
- **Gospodarstvo**. Vključuje besedila s področja ekonomije, trgov, financ, zaposlitev ipd.
- **Okolje**. Zajema okoljevarstvo, planet, energente, tudi kmetijske teme.
- **Zdravje**. Fizično in mentalno zdravje ljudi, medicina, farmacija, zdravstvena infrastruktura.
- **Prosti čas**. Hobiji, rekreacija, potovanja, turizem, ljubljenčki, dom in družina, bivanje.

14 https://cv.iptc.org/newscodes/subjectcode

- **Politika in pravo**. Mednarodne in nacionalne novice s področna državne uprave, pravnih postopkov in družbenih razmerij, konfliktov, vojn.
- **Znanost in tehnologija**. Znanstvena odkritja, zanimivosti, tehnološke inovacije, informacijska tehnologija, računalništvo.
- **Družba**. Družbena vprašanja in razmerja, enakost, diskriminacija, religija, etika ipd.
- **Šport**. Športni rezultati in zanimivosti z različnih športnih področij.
- **Vreme**. Meteorološke napovedi, opisi vremenskih posebnosti, stanj, procesov.
- **Zabava**. Estrada, moda, slog.
- **Izobraževanje**. Procesi posredovanja in pridobivanja znanja ter veščin. Vse stopnje izobraževanja, od vrtca do univerzitetnega izobraževanja, pa tudi vseživljenjsko učenje.

Kot prikazuje primerjalna Tabela 1, obstaja precejšnja prekrivnost tako s kategorijami novičarskih portalov kot s kategorijami IPTC in tujih korpusov. V nekaterih primerih, npr. *gospodarstvo, prosti čas, politika in družba*, naša kategorija zajema več kategorij ostalih virov. Tako ima za prosti čas estonski korpus kar sedem ločenih kategorij. Edini primer, ko se eno od kategorij tujih virov lahko uvrsti v dve naši, sta *umetnost in kultura* ter *zabava*. Kategoriji smo namreč ločili po eni strani zato, ker ima veliko slovenskih novičarskih portalov ločene podstrani zanju, po drugi strani pa zaradi samega jezika – kulturno-umetniške vsebine so za razliko od zabavnih pogosto precej bolj strokovne.

Medtem ko v naše kategorije lahko umestimo vseh 17 kategorij IPTC, pa češki oz. estonski korpus določenih kategorij nimata, npr. estonski nima *črne kronike*, češki pa ne *okolja*, *zdravja*, *znanosti in tehnologije* ter *zabave*. Oba tudi nimata ločene kategorije za *vreme*, ki pa jo ima IPTC in smo jo dodali zato, ker jo ima večina slovenskih novičarskih portalov.

Tabela 1: *Primerjava tematskih kategorij projekta SLED z domačimi novičarskimi portali in tujimi viri*

kategorija	zastopanost na šestih slovenskih portalih	češki korpus	estonski korpus	IPTC
umetnost in kultura	5	culture	culture & entertainment	arts, culture and entertainment
črna kronika	6	crime	/	disaster and accident
gospodarstvo	6	economy	economy, finance & business; agriculture; construction & real estate	economy, business and finance; labour
okolje	2	/	nature & environment	environmental issue
zdravje	3	/	health	health
prosti čas	4	leisure	beauty; cars; food & drinks; gambling & casinos; home, family & children; pets and animals; travel & tourism; video games	lifestyle and leisure
politika in pravo	1	politics	politics & government	politics; crime, law and justice; unrest, conflicts and war
znanost in tehnologija	5	/	science, technology & IT	science and technology
družba	1	social life	society; religion; sex; women	social issue; religion and belief; human interest
šport	6	sports	sports	sport
vreme	4	/	/	weather
zabava	4	/	culture & entertainment*	arts, culture and entertainment*
izobraževanje	1	/	education	education

Če pogledamo še prekrivnost kategorij s stranmi oz. podstranmi šestih slovenskih novičarskih portalov, vidimo, da so problematične kategorije predvsem *politika*, *družba* in *izobraževanje*. Gre za sicer legitimne kategorije, ki pa na novičarskih portalih nimajo svojih podstrani,

temveč so novice razpršene po drugih podstraneh, ki so večinoma opredeljene glede na geografski izvor novice, npr. Slovenija, Svet, Lokalno. Medtem ko so se avtorji češkega korpusa odločili slediti takšni delitvi tudi pri kategorijah (*current events, foreign news, domestic news, regional news*), smo se mi raje držali tematike. To za izdelavo učnih množic pomeni nekoliko več ročnega dela oz. iskanje drugih kazalcev, s katerimi lahko odkrijemo tematiko prispevka na posameznem portalu. Izjema je portal sta.si, ki že ima ustrezne kategorije, in sicer *Šolstvo* in *Družba*, za politiko pa *Državni zbor, Evropska unija, Mednarodna politika, Slovenska notranja politika in Slovenska zunanja politika*.

4.2 Priprava učnih množic

Na podlagi pripisa kategorije s pomočjo URL-naslovov smo preverili, koliko besedil je na voljo v posamezni kategoriji. Izračunali smo njihovo minimalno, maksimalno in povprečno dolžino (v pojavnicah, tj. besedah, ločenih s presledki) ter dolžino besedil glede na kvartile. Kot kaže Tabela 2, je npr. v kategoriji šport najkrajše besedilo dolgo le 24 pojavnic, najdaljše pa 6028 pojavnic. 25 % vseh besedil v kategoriji šport je krajših od 150 pojavnic (q_1), 25 % pa je daljših od 374 pojavnic (q_3). Pri sestavi učne množice smo želeli vzorčiti predvsem besedila povprečne dolžine in se izogniti ekstremom, zato smo upoštevali le besedila z dolžino nad q_1 in pod q_3 (besedila srednje dolžine).

Tabela 2: *Dolžina in število besedil po kategorijah*

Kategorija	Min. št. pojavnic	Maks. št. pojavnic	q_1	q_2	q_3	Vsa besedila	Besedila srednje dolžine
Šport	24	6.028	150,0	226,0	374,0	70.617	35.099
Okolje	30	4.705	212,75	336,0	519,0	2.896	1.445
Gospodarstvo	26	7.309	98,0	222,0	401,0	11.489	5.725
Umetnost in kultura	26	7.264	236,0	362,0	535,0	17.207	8.596
Znanost in tehnologija	27	6.920	177,0	308,0	608,0	3.465	1.725
Zabava	50	5.944	117,0	156,0	222,0	16.210	7.975
Zdravje	7	31.221	129,0	189,0	503,0	25.134	12.510

Kategorija	Min. št. pojavnic	Maks. št. pojavnic	q_1	q_2	q_3	Vsa besedila	Besedila srednje dolžine
Črna kronika	50	2.301	117,0	154,0	236,0	7.342	3.656
Prosti čas	37	5.835	194,0	284,0	456,0	4.478	2.227
Vreme	56	1.934	128,0	146,0	1573,0	3.804	1.881
Izobraževanje	10	1.844	129,0	141,0	272,25	2.736	1.351
Politika in pravo	7	3.304	127,0	141,0	285,0	38.126	19.013
Družba	9	2.241	126,0	136,0	217,0	10.941	5.426

Najbolje zastopana kategorija je *Šport*, sledita pa ji *Politika* in *Zdravje*. Po drugi strani so slabše zastopane kategorije *Izobraževanje*, *Okolje* ter *Znanost in tehnologija*. Povprečne dolžine besedil (q_2) so med kategorijami podobne – daljša so predvsem besedila v kategorijah *Znanost in tehnologija*, *Okolje* in *Umetnost, kultura*. Na tej točki je treba omeniti, da lahko besedila po tem sistemu spadajo v natanko eno kategorijo – šlo je za pragmatično odločitev, saj tovrstni sistem odseva delitev pri večini virov, kjer je po eno besedilo uvrščeno le v eno kategorijo, obenem pa je implementacija enodimenzionalnih oznak pri klasifikaciji manj kompleksna, manj težav pa to povzroča tudi pri zapisu metapodatkov v končni format VERT. Poleg tega bi bilo treba opraviti tudi eksperiment ročnega označevanja z več kategorijami, da bi videli, ali je večplastna kategorizacija res potrebna in v kolikšni meri. V prvem koraku smo zato eksperimente omejili na enostavnejši scenarij z eno kategorijo.

Za razvoj modela za kategorizacijo tematike novičarskih besedil smo pripravili dve učni množici. Prva, manjša, je vsebovala približno 10.000 besedil (oz. po največ 800 besedil srednje dolžine za vsako od 13 kategorij). Druga, večja, pa približno 36.000 besedil – po 2.800 na kategorijo; pri podreprezentiranih kategorijah, v katerih ni bilo na razpolago vsaj 2.800 besedil srednje dolžine, smo v tem primeru deficit zapolnili tudi s krajšimi in daljšimi besedili. Večja množica je bila zato v primerjavi s prvo nekoliko slabše vzorčena, a smo želeli preveriti, ali lahko z večjo učno množico pri razvoju modela dosežemo višjo napovedno točnost.

Vzorčenje besedil (sestava množic je prikazana v Tabeli 3) je potekalo na naslednji način:

(1) Vzeli smo vsa besedila srednje dolžine znotraj posamezne kategorije.
(2) Vsaki kategoriji smo naključno vzorčili 3.000 besedil (750 besedil na leto za leta 2019, 2020, 2021 in 2022); če je bilo besedil manj kot 3.000, smo v vzorec zajeli vsa.
(3) Iz vzorca 3.000 besedil (iz točke (2)) smo naključno vzorčili 1.000 besedil (250 na leto).
(4) Iz teh 1.000 besedil smo naključno vzorčili 100 besedil za razvojno množico in še ločenih 100 besedil za testno množico.
(5) V manjšo učno množico smo dodali približno 800 besedil (iz vzorca 1.000 besedil iz točke (3) brez 200 besedil za razvojno in testno množico).
(6) V večjo učno množico smo dodali približno 2.800 besedil (iz točke (2)), a brez besedil, ki so bila v točki (4) dodana v razvojno in testno množico. Če ni bilo dovolj besedil srednje dolžine, smo naključno dodali nekaj krajših in daljših besedil, da smo dosegli približno 2.800 besedil (v primeru nekaterih slabše zastopanih kategorij te številke kljub temu ni bilo mogoče doseči).

Tabela 3: *Število besedil v učni, razvojni in testni množici*

Kategorija	Testna množica	Razvojna množica	Manjša učna množica	Večja učna množica
Črna kronika	100	100	800	2.800
Družba	100	100	800	2.800
Gospodarstvo	100	100	800	2.800
Izobraževanje	100	100	695	2.536
Umetnost in kultura	100	100	800	2.800
Okolje	100	100	711	2.696
Politika in pravo	100	100	800	2.800
Prosti čas	100	100	718	2.834
Šport	100	100	800	2.800
Vreme	100	100	726	2.800
Zabava	100	100	800	2.800
Zdravje	100	100	800	2.800
Znanost in tehnologija	100	100	753	2.814
Skupaj	**1.300**	**1.300**	**10.003**	**36.080**

Poskrbeli smo, da je vzorčenje besedil v največji možni meri upoštevalo enakomerno distribucijo tako med kategorijami (npr. črna kronika, politika) kot med viri (npr. rtvslo.si, delo.si). Ker smo besedila za učno množico črpali iz korpusa Trendi, ki v času pisanja tega članka zajema besedila med letoma 2019 in 2022, smo poskrbeli tudi, da so bila besedila karseda enakomerno vzorčena iz vseh štirih let – besedila iz let 2020 in 2021 so bila namreč zaradi pandemije novega koronavirusa v mnogo tematskih kategorijah zaznamovana s pandemsko vsebino; preveč besedil iz tega obdobja bi lahko v model vneslo neželene pristranskosti oz. znižalo njegovo robustnost.

4.3 Razvoj kategorizacijskih modelov in klasifikacija

Preizkusili smo več modelov za kategorizacijo, razvili pa smo jih tako z orodjem fastText (Joulin idr., 2016) kot z orodjem Simple Transformers (Rajapakse, 2019). Rezultate opišemo v nadaljevanju v ločenih razdelkih.

4.3.1 Modeli fastText

Za potrebe učenja modela z orodjem fastText smo učno, testno in razvojno množico predobdelali, tako da smo v besedilih vse znake za odstavke zamenjali s presledki, odstranili ločila in spremenili vse besede v male črke.

Glede na uporabljeno učno množico in glede na to, ali smo pri razvoju upoštevali vnaprej naučene vektorske vložitve za slovenščino na nivoju pojavnic po modelu fastText (Ljubešić in Erjavec, 2018), smo razvili štiri modele:

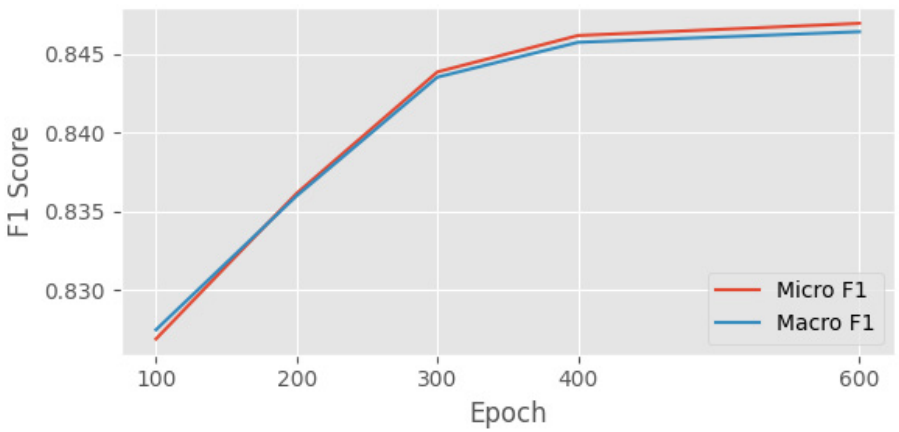
(1) z manjšo učno množico in brez vnaprej naučenih vektorskih vložitev;
(2) z večjo učno množico in brez vnaprej naučenih vektorskih vložitev;
(3) z manjšo učno množico in z uporabo vnaprej naučenih vektorskih vložitev;
(4) z večjo učno množico in z uporabo vnaprej naučenih vektorskih vložitev.

Tabela 4: *Delovanje modelov fastText*

Št. modela	Velikost učne množice	Vložitve	Mikro F1	Makro F1
1	manjša	ne	0,83	0,83
2	večja	ne	0,85	0,85
3	manjša	da	0,85	0,85
4	večja	da	0,85	0,85

Kot prikazuje Tabela 4, pri učenju na manjši učni množici uporaba vnaprej naučenih vektorskih vložitev nekoliko izboljša delovanje modela (2 odstotni točki pri merah makro F1 in mikro F1), pri učenju na večji učni množici pa ne pride do razlik. Učenje po drugi strani pri uporabi manjše učne množice traja le 15 minut, pri uporabi večje pa približno eno uro.

Pri učenju modelov fastText nismo avtomatsko optimizirali hiperparametrov, temveč smo pri večini upoštevali privzete nastavitve. Izjema je bilo število epoh, pri katerem smo model naučili večkrat na različnem številu epoh in preverili, kdaj sta meri F1 optimalni. Za optimiziranje hiperparametrov smo modele učili na učni množici, testirali pa na razvojni. Slika 1 npr. prikazuje meri F1 v odvisnosti od števila epoh za model 3 (manjša učna množica, vektorske vložitve). Meri F1 opazno naraščata do epohe 400, zatem se rast umiri.



Slika 1: Meri F1 v odvisnosti od števila epoh za model fastText z manjšo učno množico in vnaprej naučenimi vložitvami.

Za končni model smo izbrali model 4 (večja učna množica, vektorske vložitve), ki smo ga naučili na 1000 epohah. Model je pod imenom *fastText-Trendi-Topics 1.0* na voljo na repozitoriju CLARIN.SI (Kuzman idr., 2022). Tabela 5 prikazuje, kako se model odreže pri klasifikaciji besedil različnih kategorij na testni množici. Največjo točnost dosega pri kategorijah *Vreme*, *Šport* in *Črna kronika* (F1 > 0,90), najnižjo pa pri kategorijah *Znanost in tehnologija* ter *Prosti čas*. Razlika je pričakovana, saj kategorije z najvišjo točnostjo vsebujejo besedila, ki so si med seboj zelo podobna ne glede na vir ali čas objave (npr. opisi vremena, poročila o prometnih nesrečah, opisi športnih tekem); po drugi strani je kategorija *Prosti čas* precej bolj raznolika, enako pa velja tudi za *Znanost in tehnologijo*, ki pokriva različne discipline in opisuje vedno nova odkritja, zato so si besedila manj podobna.

Tabela 5: *Natančnost, priklic in F1 za različne kategorije besedil pri modelu fastText z večjo učno množico in vektorskimi vložitvami*

Kategorija	Natančnost	Priklic	F1
Vreme	0,98	0,96	0,97
Šport	0,94	0,97	0,96
Črna kronika	0,94	0,92	0,93
Umetnost in kultura	0,85	0,91	0,88
Zabava	0,91	0,86	0,88
Izobraževanje	0,90	0,83	0,86
Zdravje	0,86	0,83	0,85
Politika in pravo	0,83	0,83	0,84
Družba	0,87	0,79	0,83
Gospodarstvo	0,81	0,83	0,82
Okolje	0,77	0,88	0,82
Znanost in tehnologija	0,81	0,77	0,79
Prosti čas	0,65	0,70	0,67
Uravnoteženo povprečje	**0,86**	**0,86**	**0,85**

Matrika zamenjav je prikazana v Tabeli 6. Večina najpogostejših zamenjav je do neke mere smiselnih, npr. *Zabava - Umetnost* (6) in *Zabava - Prosti čas* (4) ter *Izobraževanje - Družba* (8) in *Družba - Politika* (7), *Družba - Okolje* (5), *Gospodarstvo - Okolje* (4), *Gospodarstvo - Politika* (7), *Okolje - Politika* (4), *Znanost - Okolje* (6), *Znanost*

- *Gospodarstvo* (4), *Prosti čas - Umetnost* (4). V kontekstu covidne krize so smiselne tudi zamenjave v kategoriji *Zdravje*, npr. *Zdravje - Politika* (4) in *Zdravje - Družba* (5). Manj smiselne so zamenjave iz kategorije *Prosti čas*, ki je bila tudi najslabše klasificirana: *Gospodarstvo - Prosti čas (5)*, *Okolje - Prosti čas* (4), *Znanost - Prosti čas* (7), *Prosti čas - Gospodarstvo* (5).

Tabela 6: *Matrika zamenjav pri modelu fastText z večjo učno množico in vektorskimi vložitvami*

Nap.→ Res.↓	**Vre.**	**Šp.**	**Črn.**	**Ume.**	**Zab.**	**Izo.**	**Zdr.**	**Pol.**	**Dru.**	**Gos.**	**Oko.**	**Zna.**	**Pro.**
Vre.	**97**	-	-	-	-	-	-	-	-	-	3	-	-
Šp.	-	**97**	-	-	1	-	-	-	-	-	1	-	1
Črn.	-	2	**91**	-	1	1	1	-	1	1	-	-	2
Ume.	-	1	-	**92**	2	-	-	-	1	-	-	2	2
Zab.	-	2	-	6	**85**	-	-	1	1	-	-	1	4
Izo.	-	-	-	1	2	**82**	1	1	8	1	1	-	3
Zdr.	-	-	1	1	-	-	**85**	4	5	-	-	2	2
Pol.	-	-	3	3	1	1	-	**91**	-	1	-	-	-
Dru.	-	1	-	1	1	3	2	7	**76**	-	5	-	-
Gos.	-	-	-	-	-	1	1	7	1	**80**	4	1	5
Oko.	1	-	-	1	-	-	-	4	-	1	**87**	2	4
Zna.	-	-	-	3	-	-	3	1	1	4	6	**75**	7
Pro.	1	5	2	4	2	1	4	-	1	5	4	5	**66**

4.3.2 Modeli tipa Bert

Jezikovni modeli tipa BERT temeljijo na nevronskih mrežah in se jih z učenjem na več milijardah besed pripravi na učinkovito reševanje različnih nalog obdelave naravnega jezika. Pripravljeni modeli, ki smo jih uporabili, so na voljo na repozitoriju Hugging Face, za nalogo kategorizacije tematike novičarskih besedil pa smo jih doučili z učenjem na učni množici, za kar smo uporabili orodje Simple Transformers.

Tudi pri učenju modelov tipa BERT smo uporabili privzete nastavitve hiperparametrov in preverili le optimalno število epoh pri učenju na učni množici in testiranju na razvojni. Za učenje smo uporabili slovenske vektorske vložitve SloBERTa 2.0 (Ulčar in Robnik-Šikonja,

2021), preizkusili pa smo tudi večjezične vložitve XLM-RoBERTa (Conneau idr., 2020), a ker so se odrezale slabše od enojezičnih, smo jih uporabili le v enem eksperimentu. Razvili smo tri modele (rezultati so prikazani v Tabeli 7):

(1) model z vložitvami XLM-RoBERTa in manjšo učno množico,
(2) model z vložitvami SloBERTa 2.0 in manjšo učno množico,
(3) model z vložitvami SloBERTa 2.0 in večjo učno množico.

Tabela 7: *Delovanje modelov tipa BERT*

Št. modela	Velikost učne množice	Vložitve	Mikro F1	Makro F1
1	manjša	XLM-RoBERTa	0,91	0,91
2	manjša	SloBERTa 2.0	0,93	0,93
3	večja	SloBERTa 2.0	0,94	0,94

Za model (1) se je kot najustreznejša nastavitev izkazalo 6 epoh, za model (2) 8 epoh in za model (3) 2 epohi. Učenje modelov (1) in (2) je trajalo približno 2 uri, modela (3) pa 3 ure.

Če primerjamo model SloBERTa, ki je bil učen samo na slovenskih podatkih (3,47 milijardah pojavnic), in večjezični model XLM-RoBERTa, ki je bil učen na 100 jezikih, od tega na 1,7 milijarde slovenskih pojavnic, rezultati iz tabele potrjujejo, da je tudi pri tej nalogi enojezični model primernejši od večjezičnega. To se sklada z ugotovitvami Ulčar idr. (2021), ki so primerjali modela na drugih pogostih nalogah, povezanih s procesiranjem naravnega jezika.

Čeprav sta si bila glede na rezultate modela (2) in (3) zelo podobna in je model (2) zahteval nekoliko manj časa za učenje, smo za končni model izbrali model (3), ki je bil naučen na večji učni množici in je morda kljub vsemu nekoliko bolj robusten. Na voljo je na repozitoriju CLARIN.SI pod imenom *SloBERTa-Trendi-Topics 1.0* (Čibej idr., 2022) ter na repozitoriju HuggingFace.[15] Rezultati klasifikacije po posameznih kategorijah so prikazani v Tabeli 8.

15 https://huggingface.co/cjvt/sloberta-trendi-topics

Tabela 8: *Natančnost, priklic in F1 za različne kategorije besedil pri modelu SloBERTa 2.0 z večjo učno množico in vektorskimi vložitvami*

Kategorija	Natančnost	Priklic	F1
Vreme	0,98	0,98	0,98
Šport	0,97	0,99	0,98
Črna kronika	0,99	0,98	0,98
Zabava	0,97	0,99	0,98
Umetnost, kultura	0,99	0,96	0,97
Politika	0,93	0,99	0,96
Družba	0,94	0,96	0,95
Zdravje	0,93	0,95	0,94
Znanost in tehnologija	0,89	0,94	0,92
Izobraževanje	0,93	0,88	0,90
Gospodarstvo, posel, finance	0,92	0,87	0,89
Okolje	0,89	0,89	0,89
Prosti čas	0,88	0,83	0,85
Uravnoteženo povprečje	**0,94**	**0,94**	**0,94**

Podobno kot modeli fastText se tudi izbrani model s SloBERTo najbolje odreže pri klasifikaciji besedil iz kategorij *Vreme*, *Šport* in *Črna kronika*. Najslabše se odreže pri kategorijah *Gospodarstvo, posel, finance* ter *Okolje* in *Prosti čas*, precej bolje pa klasificira besedila iz kategorije *Znanost in tehnologija* (F1 znaša 0,92, pri izbranem modelu s fastTextom pa je ta dosegla le 0,79). Tudi nasploh pri vseh kategorijah SloBERTa dosega občutno boljše rezultate – mera F1 je namreč v primerjavi s fastTextom pri posameznih kategorijah višja tudi za 9 odstotnih točk.

Matrika zamenjav je prikazana v Tabeli 9. Do zamenjav pri tem modelu za razliko od modela fastText (Tabela 6) prihaja le sporadično. Tudi tukaj je najpogostejša zamenjava smiselna: *Izobraževanje – Družba* (5), še vedno pa je najbolj problematična kategorija *Prosti čas*, npr. *Gospodarstvo – Prosti čas* (4), *Okolje – Prosti čas* (4).

Tabela 9: *Matrika zamenjav pri modelu fastText z večjo učno množico in vektorskimi vložitvami*

Nap.→ Res.↓	Vre.	Šp.	Črn.	Ume.	Zab.	Izo.	Zdr.	Pol.	Dru.	Gos.	Oko.	Zna.	Pro.
Vre.	**98**	-	-	-	-	-	-	-	-	-	2	-	-
Šp.	-	**99**	-	-	-	-	-	-	-	-	-	-	1
Črn.	-	-	**98**	-	-	-	-	-	-	-	1	1	-
Ume.	-	-	-	**94**	1	1	-	-	-	-	1	1	-
Zab.	-	-	1	-	**99**	-	-	-	-	-	-	-	-
Izo.	-	1	-	-	-	**88**	2	-	5	3	-	-	1
Zdr.	-	-	-	-	-	-	**95**	3	-	-	-	2	-
Pol.	-	-	-	-	-	-	-	**99**	-	1	-	-	-
Dru.	-	-	-	-	-	-	2	2	**95**	-	-	-	-
Gos.	-	-	-	-	-	3	1	1	1	**87**	2	1	4
Oko.	2	-	-	-	-	-	1	1	-	2	**87**	1	4
Zna.	-	-	-	-	-	1	1	-	-	-	3	**93**	1
Pro.	-	2	-	1	2	2	-	1	-	2	2	5	**82**

4.3.3 Klasifikacija korpusa Trendi

Za klasifikacijo besedil v korpusu Trendi (Kosem idr., 2022) smo uporabili model *SloBERTa-Trendi-Topics 1.0*, ki je sicer računsko požrešnejši in počasnejši, a v primerjavi z modelom *fastText-Trendi-Topics 1.0* tudi natančnejši. Na primer, za klasifikacijo verzije 2022-10 (1.635.614 besedil od 1. januarja 2019 do vključno 31. oktobra 2022) je bilo potrebnih približno 152 ur procesiranja z grafično procesno enoto (GPU) na spletni strani Kaggle.[16] Za klasifikacijo besedil enega meseca (npr. junij 2022; 41.262 zajetih besedil) smo potrebovali povprečno 3,5 ure (približno 3,3 sekunde na besedilo). Rezultati klasifikacije so prikazani v Tabeli 10.

Tabela 10: *Število klasificiranih besedil v korpusu Trendi 2022-10 po kategorijah*

Kategorija	Število besedil
Vreme	11.262
Šport	311.117
Črna kronika	116.208
Zabava	108.721
Umetnost in kultura	98.076
Politika in pravo	176.608

16 Kaggle: https://www.kaggle.com/ (Dostop: 16. december 2022)

Kategorija	Število besedil
Družba	25.681
Zdravje	131.001
Znanost in tehnologija	109.593
Izobraževanje	44.471
Gospodarstvo	186.385
Okolje	97.852
Prosti čas	218.634
Brez kategorije	5
Skupaj	1.635.614

Ob prvem pregledu korpusa se rezultati zdijo obetavni. Na podlagi že analiziranih leksikografskih podatkov smo opravili poizvedbe v korpusu Trendi in večinoma je kategorizacija potrdila jezikoslovne ugotovitve. Tako se na primer izraz *dvojni dvojček*, ki prihaja iz košarkarske terminologije, pojavlja skoraj izključno v besedilih, katerim je bila pripisana kategorija Šport (99,8 % zadetkov). Podobno *aforizem,* izraz s področja književnosti, pričakovano prevladuje v besedilih, ki jim je bila pripisana kategorija Kultura (67,3 %). Besedna zveza *državni zbor* kaže večjo razpršenost po kategorijah (pojavlja se v 12 od 13 kategorij), vseeno pa po pričakovanjih prevladuje v Politiki (49,4 %); a tudi v ostalih kategorijah najdemo primesi politike, npr. zadetki iz športnih besedil pogosto omenjajo politično potrjevanje zakonov in finančnih sredstev, relevantnih za določeno športno panogo.

5 Sklep

S spremljevalnim korpusom Trendi je jezikovna infrastruktura za slovenščino bogatejša za še en pomemben vir, ki omogoča spremljanje povsem sodobne rabe jezika. Najnovejša različica Trendi 2023-02 pokriva besedila od začetka 2019 do konca februarja 2023, vsebuje pa že več kot 700 milijonov pojavnic iz 107 virov. Izdelani cevovod omogoča dnevno zbiranje besedil, njihovo avtomatsko označevanje in pretvorbo, mesečno pa pretvorbo v format VERT, ki ga zahtevajo konkordančniki repozitorija CLARIN.SI, v katerih je korpus Trendi prosto dostopen.

Avtomatska kategorizacija besedil z novičarskih portalov glede na tematiko se je z uporabljenim naborom kategorij in z naučenimi orodji

izkazala za zelo uspešno – najboljši model pri večini kategorij dosega visoko točnost (nad 90 % oz. celo nad 98 %), do zamenjav med kategorijami pa prihaja redko, kar omogoča zanesljivo pripisovanje metapodatkov in gradnjo visoko zanesljivih tematskih podkorpusov, ki jih je mogoče uporabiti za nadaljnje analize distribucije jezikovne rabe besed. V okviru nadaljnjega dela na projektu bomo rezultate strojne klasifikacije tematik primerjali tudi z ročno pripisanimi oznakami, predvsem zato, da ugotovimo, ali tudi pri človeških označevalcih pogosto prihaja do določenih napak v kategorizaciji, zlasti med podobnimi ali delno prekrivnimi kategorijami (npr. *Zabava* in *Prosti čas* ter *Znanost in tehnologija* in *Okolje*). Upoštevati je treba tudi, da so bili modeli evalvirani na množici, ki je bila uravnotežena po kategorijah, zato je klasifikacija na realnih podatkih morda pogosteje lažno pozitivna za redkejše kategorije, kot je npr. *Okolje*. Ker dejanske porazdelitve besedil v kategorije v celotnem slovenskem medijskem prostoru ne poznamo, je treba uspešnost modela oceniti z ročno evalvacijo. Poleg tega je bil klasifikator razvit le za uporabo na člankih z novičarskih portalov, v prihodnje pa bi podobno metodo veljalo preizkusiti tudi za pripisovanje metapodatkov drugim besedilnim zvrstem, ki jih najdemo npr. v korpusu Gigafida 2.0 (npr. leposlovje, časopisni članki, revije).

Načrti za prihodnost vključujejo predvsem optimizacijo cevovoda, odpravo določenih hroščev (npr. nepričakovani simboli kot posledica napačnega kodiranja besedilnih datotek) in pa redno povečevanje korpusa. Del izboljšav korpusa je tudi zamenjava besedil, ki so bila pridobljena z virov s plačljivimi vsebinami, saj je pogosto brezplačno na voljo samo prvih nekaj odstavkov. Za tovrstna besedila je potrebno urediti pogodbo z besedilodajalci in besedila pridobiti izven obstoječega cevovoda. Korpus je lahko uporaben tudi za kompleksnejše naloge, kot so npr. analiza sloga, identifikacija avtorjev in ocenjevanje kompleksnosti besedila.

Zahvala

Projekt SLED (*Spremljevalni korpus in spremljajoči podatkovni viri*) je financiralo Ministrstvo za kulturo Republike Slovenije kot del *Javnega razpisa za (so)financiranje projektov, namenjenih gradnji in posodabljanju infrastrukture za slovenski jezik v digitalnem okolju 2021–2022*. Raziskovalna programa št. P6-0411 (*Jezikovni viri in tehnologije za*

slovenski jezik) in št. P6-0215 (*Slovenski jezik – bazične, kontrastivne in aplikativne raziskave*) je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije iz državnega proračuna.

Literatura

Bušta, J., Herman, O., Jakubíček, M., Krek, S., & Novak, B. (2017). JSI Newsfeed corpus. *The 9th International Corpus Linguistics Conference*. University of Birmingham.

Caterina, M., Silvia, B., Eugenio, G., Massimo, C., & Francesco, S. (2019). KIParla corpus: a new resource for spoken Italian. *CEUR WORKSHOP PROCEEDINGS*. SunSITE Central Europe.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 8440–8451).

Cvrček, V., Křen, M., Čermáková, A., Chlumská, L., Škrabal, M., in Kováříková, D. (2020). Overview of text classification in SYN2015. Pridobljeno s https://wiki.korpus.cz/doku.php/en:cnk:klasifikace_textu_syn2015

Čibej, J., Kuzman, T., Ljubešić, N., Kosem, I., Ponikvar, P., Dobrovoljc, K., & Krek, S. (2022). *Text classification model SloBERTa-Trendi-Topics 1.0*. Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1709.

Davies, M. (2008-). *The Corpus of Contemporary American English (COCA)*. Retrieved from https://www.english-corpora.org/coca/

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and linguistic computing*, *25*(4), 447–464.

Davies, M. (2016-). *Corpus of News on the Web (NOW)*. Pridobljeno s https://www.english-corpora.org/now/

Davies, M. (2019-). *The Coronavirus Corpus*. Pridobljeno s https://www.english-corpora.org/corona/

De Smedt, K. (2020). Contagious “Corona” Compounding by Journalists in a CLARIN Newspaper Monitor Corpus. *CLARIN Annual Conference*.

Grobelnik, M., Brank, J., Mladenić, D., Novak, B., & Fortuna, B. (2006). Using DMoz for constructing ontology from data stream. *28th International Conference on Information Technology Interfaces* (pp. 439–444).

Herman, O., & Kovár, V. (2013). Methods for Detection of Word Usage over Time. *RASLAN*.

Hnátková, M., Křen, M., Procházka, P., & Skoumalová, H. (2014). The SYN-series corpora of written Czech. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*.

Joulin, A., Grave, É., Bojanowski, P., & Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*: *Volume 2, Short Papers* (pp. 427–431).

Kilgarriff, A., Rychlý, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. V G. Williams in S. Vessier (ur.): *Proceedings of the Eleventh EURALEX International Congress* (pp. 105–116). Lorient, France.

Kosem, I., Čibej, J., Dobrovoljc, K., Erjavec, T., Ljubešić, N., Ponikvar, P., Šinkec, M., & Krek, S. (2022). *Monitor corpus of Slovene Trendi 2022-10*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1681

Krek, S., Arhar Holdt, Š., Erjavec, T., Čibej, J., Repar, A., Gantar, P., Ljubešić, N., Kosem, I., & Dobrovoljc, K. (2020). Gigafida 2.0: the reference corpus of written standard Slovene. *Proceedings of the 12th Language Resources and Evaluation Conference*.

Kuzman, T., Čibej, J., Ljubešić, N., Kosem, I., Ponikvar, P., Dobrovoljc, K., & Krek, S. (2022). *Text classification model fastText-Trendi-Topics 1.0*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1710

Laitinen, M., Lundberg, J., Levin, M., & Martins, R. M. (2018). The Nordic Tweet Stream: A dynamic real-time monitor corpus of big and rich language data. *Digital Humanities in the Nordic Countries 3rd Conference*.

Ljubešić, N., & Erjavec, T. (2018). *Word embeddings CLARIN.SI-embed.sl 1.0*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1204

Logar, N., Erjavec, T., Krek, S., Grčar, M. in Holozan, P. (2013). Written corpus ccGigafida 1.0, Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1035

Logar Berginc, N., & Ljubešić, N. (2013). Gigafida in slWaC: tematska primerjava. *Slovenščina 2.0*, *1*(1), 78–110.

Logar, N., Ljubešić, N., & Erjavec, T. (2015). Kres in Gigafida kot korpusna osnova za slovar: razlike in podobnosti. In M. Smolej (ur.), *Slovnica in slovar – aktualni jezikovni opis* (str. 479–486). Ljubljana: Znanstvena založba Filozofske fakultete.

Michel, J. B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Google Books Team, Pickett, J. P., ..., & Orwant, J. (2011). Quantitative analysis of culture using millions of digitized books. *science*, *331*(6014), 176–182.

Rajapakse, T. C. (2019). *Simple Transformers*. Pridobljeno s https://github.com/ThilinaRajapakse/simpletransformers

Sharoff, S. (2018). Functional text dimensions for the annotation of web corpora. *Corpora*, *13*(1), 65–95.

Štajner, T., Rusu, D., Dali, L., Fortuna, B., Mladenić D., & Grobelnik, M. (2010). A service oriented framework for natural language text enrichment. *Informatica, 34*(3), 307–313.

Trampuš, M., & Novak, B. (2012). Internals of an aggregated web news feed. *Proceedings of 15th Multiconference on Information Society*.

Ulčar, M., Žagar, A., Armendariz, C. S., Repar, A., Pollak, S., Purver, M., in Robnik-Šikonja, M. (2021). Evaluation of contextual embeddings on less-resourced languages. arXiv preprint arXiv:2107.10614. Pridobljeno s https://arxiv.org/pdf/2107.10614.pdf

Monitor Corpus Trendi and Automatic Text Categorization

The paper presents the compilation of the Trendi corpus, the first monitor corpus of Slovene. The current version (Trendi 2023-02) contains texts published between January 2019 and October 2023, with a total of over 700 million tokens (more than 586 million words). The purpose of the corpus is to provide linguists and non-linguists with data on current language use and to enable the monitoring of new words as well as the increase and decline in the use of existing words. In the paper, we present the contents of the corpus and the methods and criteria used in its compilation. The second part of the paper is focused on the development of a tool for categorizing text topics in news articles. The tool was developed specifically for the Trendi corpus but can be used for other corpora containing similar texts. A set of 13 thematic categories was developed for the tool. The set generally follows international standards and categories used in comparable corpora for other languages. Using texts annotated with these categories, we trained multiple language models and achieved a high classification accuracy when categorizing text topics.

Keywords: monitor corpus, automatic text categorization, neologisms, news sites, Slovene