Strojna analiza tematik in sentimenta slovenskih novičarskih medijev

Jan Bajt, Marko Robnik-Šikonja
Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Večna pot 113, 1000 Ljubljana
janbajt@gmail.com, marko.robnik@fri.uni-lj.si

Izvleček

V delu primerjamo slovenske novičarske medije s pomočjo analize tematik in sentimenta člankov. Analizirali smo različna stališča sedmih slovenskih medijev do specifičnih dogodkov oziroma tematik v letih 2019 in 2020. Tematike smo modelirali dvofazno z modelom LDA, s katerim smo v množici spletnih člankov poiskali nekaj posameznih tematik. Za nalogo zaznavanja sentimenta smo prilagodili velik vnaprej naučen slovenski maskirni jezikovni model SloBERTa in ga uporabili pri klasifikaciji izbranih člankov v enega izmed treh razredov (pozitivnega, nevtralnega ali negativnega). V množici izbranih tematik smo opazili precejšnje razlike med mediji v pogostosti in sentimentu poročanja.

Ključne besede: analiza sentimenta, latentna Dirichletova alokacija, modeliranje tematik, model BERT, obdelava naravnega jezika, slovenski novičarski mediji

Abstract

We compare topics and sentiment in Slovenian news media. We analysed the sentiment of seven media concerning specific political events or topics in 2019 and 2020. We used two phases of LDA modelling to detect a number of specific topics. For the sentiment analysis task, we fine-tuned large pretrained Slovenian masked language model, SloBERTa, and used it to classify articles in one of three classes (positive, neutral or negative). In the set of selected topics, we observed considerable differences between media in frequency and sentiment of reporting.

Keywords: Sentiment analysis, latent Dirichlet allocation, topic modeling, model BERT, natural language processing, Slovenian news media

1 UVOD

Mediji s pisanjem o dogodkih in z mnenjskimi članki močno vplivajo na družbo in so tudi njen odraz. Ko gre za politične teme, mnogokrat pokažejo tudi ideološka prepričanja, ki jih zastopajo. V delu želimo z metodami procesiranja naravnega jezika poiskati razlike v pisanju slovenskih medijev o nekaj izbranih dogodkih oz. temah s primerjavo sentimenta člankov izbranih tematik. Podobna strojna analiza za slovenske medije še ne obstaja. Naš namen je objektivno analizirati slovenski novičarski medijski prostor in interpretirati razlike med posameznimi tematikami in mediji.

1.1 Obstoječe raziskave na slovenščini

Področje zaznavanja sentimenta je pogosta naloga na področju obdelave naravnega jezika, vendar je bila večina raziskav narejenih za angleški jezik. V zadnjem času so najbolj uveljavljene rešitve z uporabo tehnologij, ki temeljijo na arhitekturi transformer [24]. Med tovrstne modele spada model BERT [7], iz katerega so razvili več različic (npr. RoBERTa [12], ALBERT [10] itd.), ki se uporabljajo za različne naloge predstavitve jezika. Za slovenski jezik so Bučar, Povh in Žnidaršič [4] opravili raziskavo o odkrivanju sentimenta v slovenskih člankih. V tej raziskavi so na množici slovenskih člankov [5] uporabili pet različnih klasifikatorjev (multinomski naivni Bayes, naivni

Bayes, metoda podpornih vektorjev, k-najbližjih sosedov in naključni gozdovi) in jih med seboj primerjali. Pelicon, Pranjić, Miljković, Škrlj in Pollak [18] so analizirali medjezikovne zaznave sentimenta v člankih. Razvili so klasifikacijski sistem, ki s podanim korpusom označenih člankov v slovenskem jeziku določi sentiment ne samo na slovenskih temveč tudi na drugih jezikih brez dodatnih učnih podatkov. Sistem temelji na večjezičnem modelu BERT, v raziskavi pa raziščejo tudi različne pristope za delo z dolgimi besedili. V nasprotju z naštetimi pristopi v našem delu za klasifikacijo sentimenta uporabimo trenutno najuspešnejši slovenski jezikovni model SloBERTa [22].

Za analizo tematik je razvitih kar nekaj modelov. Med najbolj poznanimi so verjetnostni modeli, med katere spadajo probabilistična latentna semantična analiza (pLSA) [9], latentna semantična analiza (LSA) [6] in latentna Dirichletova alokacija (LDA) [2].

Za slovenščino sta Logar Berginc in Ljubešič [14] opravila tematsko primerjavo podatkovnih množic Gigafida in slWaC. V množicah sta poiskala teme z metodo LDA, jih primerjala med seboj in izpostavila nekaj razlik med najdenimi temami obeh množic. Škrajnc in Pollak [21] sta analizirali in primerjali tematike med blogi moških in ženskih avtorjev. Za razliko od Logar Bohinc in Ljubešić sta razvili hierarhične ontologije, kar je omogočilo identifikacijo podtem za vsako izmed tem.

Sistematična analiza tematik v slovenskih medijih še ne obstaja. Razlike med poročanji slovenskih medijev so raziskali Martinc, Perger, Pelicon, Ulčar, Vezovnik in Pollak [17], ki so se osredotočili na temo LGBTIQ+. Raziskali so razlike v zaznanem sentimentu in uporabljenih besedah. V našem delu analiziramo široko množico tematik, vendar zaradi omejenega prostora poročamo o le nekaj ožjih tematikah (politika in epidemija COVID-19).

1.2 Novosti

V zbirki člankov sedmih slovenskih večjih medijev smo najprej opravili analizo tematik in izbrali nekaj širših tem iz preteklih let. Znotraj teh tematik smo določili še podrobnejše teme. Za modeliranje tem smo uporabili statistični model latentne Dirichletove alokacije (LDA) [2], ki zaradi svoje razširjenosti ponuja tudi več vizualizacijskih orodij, ki so nam pomagala pri interpretaciji rezultatov. Za izbrane teme smo analizirali sentiment z modelom SloBERTa, ki smo ga predhodno prilagodili za zaznavanje tri razrednega sentimenta. Analiza je pokazala precejšnje razlike med posameznimi mediji, tako glede pogostosti poročanja o posameznih tematikah kot glede sentimenta pisanja o teh temah. Novosti našega prispevka so naslednje.

- Predlagamo dvonivojsko analizo tematik slovenskih medijev, ki omogoča boljšo interpretacijo in poimenovanje ožjih tem, kot če bi z enofazno analizo določili večjo število tematik, saj bi v tem primeru v veliki množici dobljenih tem težje določili specifične teme, ki jih želimo primerjati med mediji.
- Prvič na slovenskih člankih uporabimo model SloBERTa za analizo sentimenta na nivoju člankov. Model se pokaže kot uspešnejši v primerjavi z dosedanjimi poskusi z modelom SVM.
- Za razliko od dosedanjih večjih analiz sentimenta slovenskih medijev analiziramo trorazredni sentiment, za katerega menimo, da omogoča bolj objektivno analizo v primerjavi z binarnim sentimentom, saj je največ člankov v medijih sentimentno nevtralnih, nevtralni razred pa binarna analiza sentimenta na silo pripoji bodisi pozitivnemu bodisi negativnemu razredu.
- V naši analizi uporabimo doslej največjo podatkovno množico spletnih novic in medijev, kar daje izsledkom večjo težo.
- Opravljena analiza tematik v kombinaciji z analizo sentimenta izbranih tem pokaže precejšnje razlike med mediji tako glede pogostosti pisanja o posameznih temah kot tudi glede sentimenta.

1.3 Struktura prispevka

Delo smo razdelili na pet razdelkov. V drugem razdelku najprej predstavimo uporabljene tehnologije za detekcijo tematik in analizo sentimenta. V tretjem razdelku opišemo analizirano zbirko spletnih člankov in učno množico SentiNews, ki smo jo uporabili za učenje sentimentega klasifikatorja na podlagi modela SloBERTa. Naš pristop k modeliranju tem in analizi sentimenta predstavimo v četrtem razdelku. Peti razdelek vsebuje ovrednotenje razvitih modelov in rezultate modeliranja tematik in sentimenta. Zaključke in ideje za nadaljnjo delo zapišemo v šestem razdelku.

2 TEHNOLOGIJE ZA ANALIZO TEMATIK IN SENTIMENTA

V tem razdelku predstavimo uporabljene tehnologije. Zaradi razpoložljivosti dobrih orodij, s katerimi

lahko interpretiramo rezultate, smo za modeliranje tematik izbrali latentno Dirichletovo alokacijo, ki jo predstavimo v razdelku 4.1.3. Za interpretacijo rezultatov smo izbrali vizualizacijsko orodje LDAvis, katerega zmožnosti opišemo v razdelku 2.2. Za klasifikacijo sentimenta smo uporabili slovenski model SloBERTa, ki ga opišemo v razdelku 2.3.

2.1 Analiza tem z latentno Dirichletovo alokacijo

Model latentne Dirichletove alokacije (LDA) [2] je verjetno najpogosteje uporabljen pristop za modeliranje tem v besedilih. Modeliranje tem je metoda nenadzorovanega učenja, kjer v korpusu besedil poiščemo potencialno skrite ali abstraktne teme [1]. Poleg modela LDA sta za modeliranje tem pogosto uporabljena še pristopa LSA (angl. Latent Semantic Analysis) [8] in pLSA (angl. Probabilistic Latent Semantic Analysis) [9].

Model LDA, katerega delovanje prikazuje slika 1, predpostavlja, da je v korpusu besedil določeno število tem (na sliki so štiri teme: Tema 1, Tema 2, Tema 3, Tema 4). Vsaka izmed tem vsebuje verjetnostno porazdelitev besed, ki se v temi nahajajo (npr. za Temo 1: 5% Beseda 1, 4 % Beseda 2, 2,5 % Beseda 3 itn.). Vsak dokument je zgrajen iz naključne mešanice tem v korpusu (npr. 60% Tema 1, 20 % Tema 2, 15% Tema 3 in 5 % Tema 4). Za tvorjenje novega dokumenta verjetnostni model predpostavi, da iz vsake teme naključno izberemo določeno število besed (iz Teme 1 60% vseh besed v dokumentu, iz Teme 2 20 % itd.).

V realnosti je proces ravno obraten. Na sliki 2 je grafični prikaz delovanja modela LDA. Model na vhodu prejme število tem K in korpus besedil z M dokumenti in N besedami v posameznem besedilu. Na nivoju dokumenta (notranji okvir na sliki 2) model vsaki besedi w naključno določi temo z, s tem pa pridobimo porazdelitev tem θ v posameznem dokumentu.

Model ima še dva hiperparametra, α in β, ki sta parametra Dirichletove porazdelitve. LDA namreč predvideva, da sta porazdelitvi tem v besedilu in besed v temah Dirichletovi. Hiperparameter α vpliva na porazdelitev tem v posameznih dokumentih, β pa na porazdelitev besed v posameznih temah. Večja vrednost parametra α pomeni, da bodo dokumenti mešanica večjega števila tem, večja vrednost parametra β pa pomeni, da bodo teme mešanica večjega števila besed. Obratno velja za majhne vrednosti obeh parametrov.

Iz korpusa dokumentov in števila tem model LDA izračuna dva tipa distribucij:

- distribucije besed za vsako izmed zaznanih tem in
- distribucije tem, ki se pojavljajo v posameznem dokumentu iz korpusa dokumentov.

Izračunan model lahko uporabimo za analizo izbranega korpusa ali za klasifikacijo novih dokumentov, ki niso del izbranega korpusa. Pri analizi izbranega korpusa dobimo pregled tem, ki jih vsebuje, interpretiramo jih lahko z besedami, ki jih posamezne teme vsebujejo. Pri klasifikaciji novih dokumentov model uporabi zgolj besede iz učnega korpusa, ne upošteva pa besed iz dokumentov, ki jih ni v naučenem modelu. Za vsak nov dokument dobimo verjetnostno porazdelitev tem tega dokumenta.

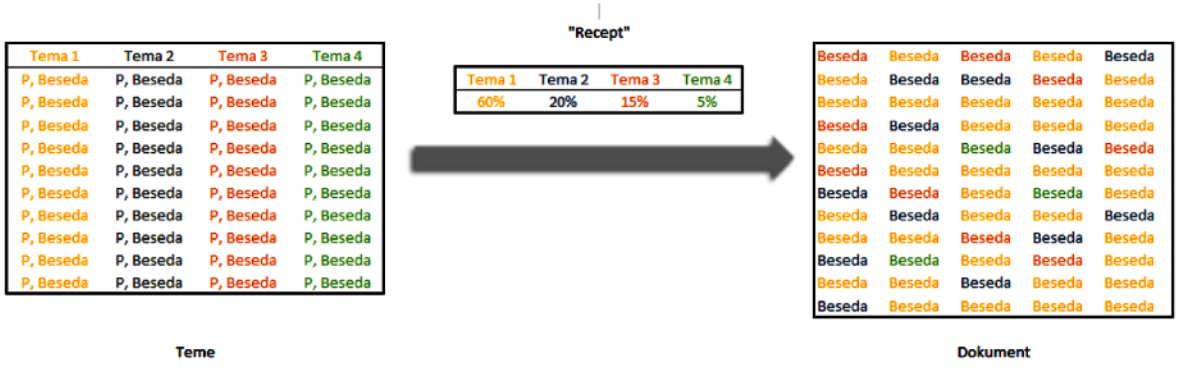
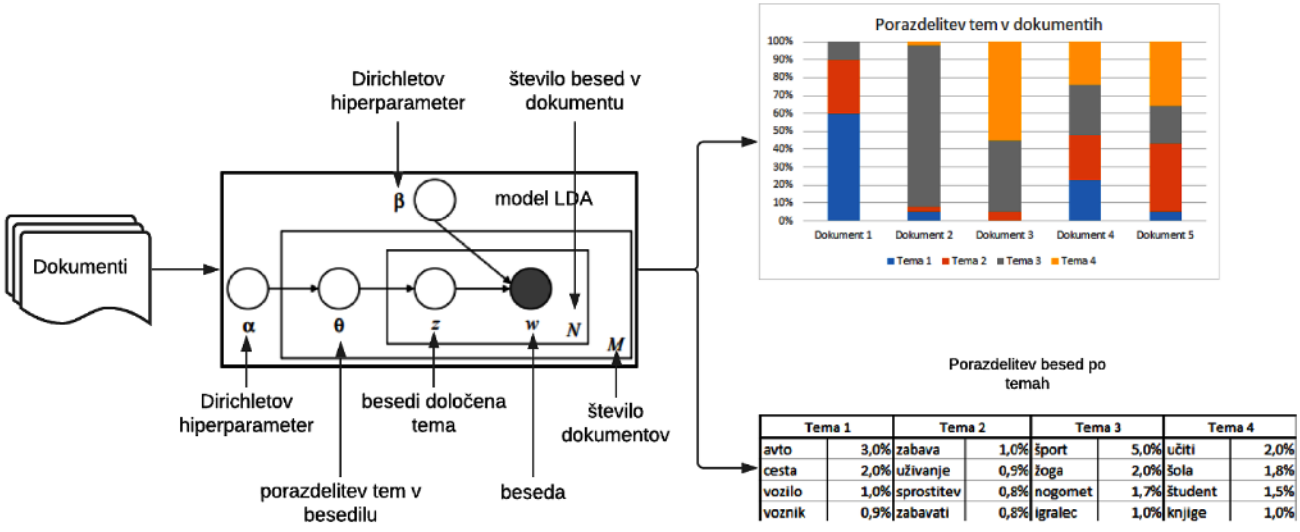


Figure 1: **Primer generiranja dokumenta z verjetnostnim modelom, ki ga predpostavlja LDA.**



Porazdelitev besed po temah

Tema 1		Tema 2		Tema 3		Tema 4	
avto	3,0%	zabava	1,0%	šport	5,0%	učiti	2,0%
cesta	2,0%	uživanje	0,9%	žoga	2,0%	šola	1,8%
vozilo	1,0%	sprostitev	0,8%	nogomet	1,7%	študent	1,5%
voznik	0,9%	zabavati	0,8%	igralec	1,0%	knjige	1,0%

Figure 2: **Shema delovanja modela LDA.**

2.2 Vizualizacija tem z orodjem LDAvis

LDAvis [20] je orodje, ki omogoča interaktivno vizualizacijo tem, pridobljenih z modelom LDA. Z orodjem imamo pregled nad vsemi temami in razlikami med njimi, kot tudi pregled besed povezanih z izbrano temo, kot ilustrira slika 3. Poleg vizualizacije tem orodje vpelje še mero primernosti besed (angl. relevance) za vsako besedo znotraj teme.

Vizualizacija modela je razdeljena na dva dela:

- globalen pregled vseh tematik (leva polovica slike 3) in
- pregled najpogostejših besed znotraj izbrane teme (desna polovica slike 3).

Z analizo globalnega prikaza vseh tem lahko ugotovimo, kako pogosto se tema v besedilu nahaja in kako so teme med seboj povezane. Posamezne teme so predstavljene s krogi v dvodimenzionalnem prostoru. Večji krogi pomenijo, da je tema bolj razširjena med opazovanimi besedili, razdalja med krogi pa pove, kako podobne so si teme. Dobro interpretabilen model LDA je predstavljen z velikimi krogi, ki se med sabo ne prekrivajo in so razpršeni po celotnem prostoru.

V desni polovici vizualizacije na sliki 3 so v obliki stolpčnega diagrama predstavljene najprimernejše besede za interpretacijo izbrane teme. Modri del posameznega stolpca predstavlja pogostost besede v celotnem korpusu besedil, rdeči del pa predstavlja pogostost besede v izbrani temi.

Pomemben del orodja je mera primernosti besede λ za določeno temo, ki lahko zavzame vrednosti med 0 in 1. Vrednost $\lambda = 1$ pomeni, da so besede v desni polovici vizualizacije razvrščene po pogostosti besed znotraj izbrane teme (po velikosti rdečega dela stolpca). Nižja kot je vrednost λ, večjo pozornost dajemo besedam, ki se bolj izključno pojavljajo v izbrani temi. Sievert in Shirley [20] predlagata nastavitev λ na vrednost 0.6. V interaktivni vizualizaciji lahko vrednost λ prilagajamo in s tem spreminjamo vrstni red besed v desni polovici vizualizacije, kar nam pomaga pri interpretaciji modela.

2.3 Klasifikacija sentimenta s prilagajanjem modela SloBERTa

Model BERT (angl. Bidirectional Encoder Representations from Transformers) [7] je jezikovni model, ki pri učenju predstavitve besed upošteva kontekst vseh besed v stavku (tako pred kot za opazovano besedo). S prilagajanjem naučenega modela (angl. fine tuning) dobimo trenutno najuspešnejše modele za različne naloge na področju obdelave naravnega jezika, tudi za klasifikacijo sentimenta, ki jo uporabimo v našem delu. Model BERT sestavljajo kodirniki nevronske arhitekture transformer [24], ki uporablja mehanizem samopozornosti (angl. self-attention). Osnovna inačica modela BERT (BERT base) vsebuje 12 slojev kodirnikov, kjer ima vsaka plast 768 skritih nevronov. Model ima 12 glav pozornosti (angl. attention head) in skupno 110 milijonov parametrov.

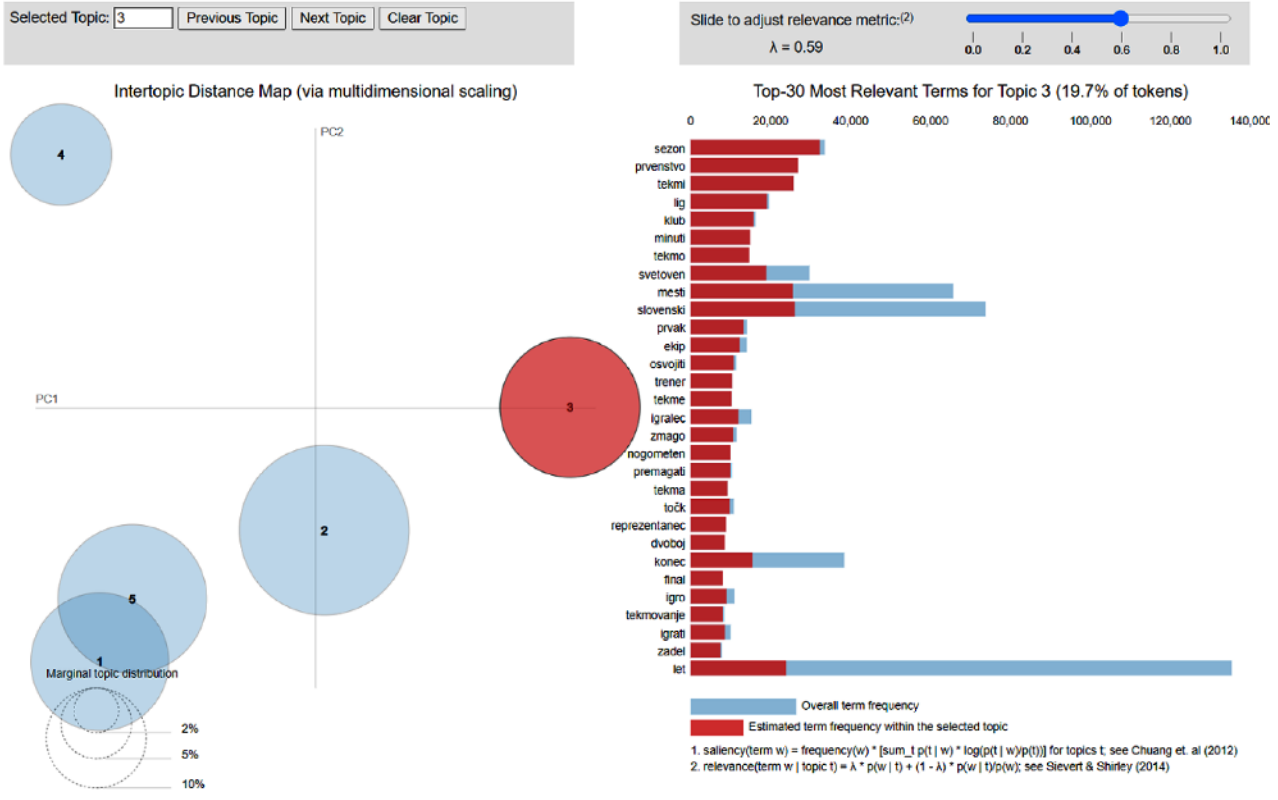


Figure 3: **Vizualizacija tem z orodjem LDAvis. Na desni je prikazana interpretacija teme 3 (rdeč krog na levi strani), ki zadeva temo o nogometu.**

Za učenje modela BERT se uporabita dve različni nalogi in velika množica neoznačenih podatkov. Prva naloga je maskirni jezikovni model (angl. masked language model), kjer je 'zamaskiranih' naključnih 15% besed oz. vhodnih žetonov (angl. tokens). Model napoveduje zamaskirane besede, s tem pa se nauči konteksta zamaskiranih besed. Druga naloga je predvidevanje naslednje povedi, kjer model prav tako učimo na neoznačenih podatkih. Na vhod model dobi zaporedje dveh povedi, ugotoviti pa mora ali druga poved sledi prvi ali ne. S tem se model uči smiselnih povezav na nivoju povedi. Naučen model lahko prilagodimo za uporabo na specifičnih nalogah obdelave naravnega jezika, kot so odkrivanje sentimenta, povzemanje besedil, odgovarjanje na vprašanja ipd. Pri prilagajanju modela za določeno nalogo uporabimo označene podatke, primerne za nalogo, ki jo izvajamo, modelu pa dodamo nov izhodni sloj.

Model RoBERTa [12] (angl. A Robustly Optimized BERT Pretraining Approach) je izboljšava modela BERT, ki je uporabila veliko večjo podatkovno množico in daljši čas učenja. RoBERTa uporablja večje velikosti učnih paketov (angl. batch size), v fazi učenja pa ne uporablja naloge predvidevanja naslednjih povedi. Spremenjen je tudi način maskiranja besed. Pri modelu BERT so besede maskirane statično (samo enkrat, na začetku učenja), model RoBERTa pa besede maskira dinamično, kar pomeni, da se maskiranje izvede posebej za vsako iteracijo učenja.

Model SloBERTa [23] je enojezičen slovenski model RoBERTa, ki sledi arhitekturi in učenju francoskega modela CamemBERT [16]. Naučili so ga na korpusu besedil, ki vsebuje 3.47 milijard besed, njegov slovar pa vsebuje 32.000 žetonov.

3 UČNI PODATKI

Za opravljeno analizo smo potrebovali dve zbirki podatkov. Zbirko spletnih novic, ki smo jo analizirali glede tem in sentimenta, opišemo v razdelku 3.1, učno množico, s katero smo naučili klasifikator sentimenta, pa v razdelku 3.2.

3.1 Zbirka spletnih člankov iz storitve Event Registry

Dostop do velike množice člankov s slovenskih novičarskih medijev nam je omogočil spletni servis Event Registry [11]. Pridobljena množica vsebuje 2.2 mili-

jona člankov v formatu JSON, objavljenih med 1. 1. 2014 in 31. 12. 2020. Posamezen članek poleg besedila vsebuje še meta podatke, kot so naslov, datum, čas objave, spletni naslov in informacije o viru članka.

Za potrebe naše analize smo za vsak članek izluščili njegovo vsebino, naslov in medij. Zaradi omejenega prostora smo uporabili članke iz let 2019 in 2020 za nekaj največjih slovenskih medijev (Dnevnik, 24ur.com, RTV Slovenija in Siol.net) in nekaj medijev iz desnega političnega pola, za katere smo pričakovali večje razlike v poročanju (Nova24TV, Tednik Demokracija in Portal Politikis).

V tabeli 1 so prikazane povprečne dolžine člankov posameznih medijev v posameznem letu. Iz člankov izbranih medijev smo odstranili tiste z manj kot 25 besedami in duplikate, ostale članke pa smo uporabili pri zaznavanju tematik. Končno število uporabljenih člankov izbranih medijev v posameznem letu je prikazano v tabeli 2.

Table 1: **Povprečna dolžina člankov v številu besed za posamezne medije v analiziranih letih 2019 in 2020.**

Medij	2019	2020
RTV Slovenija	403	452
Siol.net	410	405
24ur.com	202	214
Svet24	340	346
Tednik Demokracija	424	463
Nova24TV	523	546
Portal Politikis	353	356
Dnevnik	253	253
Povprečje	351	373

Table 2: **Število uporabljenih člankov izbranih medijev v letih 2019 in 2020.**

Medij	2019	2020
RTV Slovenija	28.948	33.466
Siol.net	25.871	23.863
24ur.com	18.831	21.281
Tednik Demokracija	8.869	8.213
Nova24TV	6.524	7.170
Portal Politikis	6.142	5.321
Dnevnik	20.990	15.304
Skupaj	116.175	114.618

3.2 Učna množica za analizo sentimenta SentiNews

Za učenje sentimenta smo uporabili podatkovno množico slovenskih člankov SentiNews [3, 5]. Celotna podatkovna zbirka je sestavljena iz več kot 250.000 člankov s politično, poslovno, ekonomsko in finančno vsebino petih slovenskih spletnih medijskih virov (24ur, Dnevnik, Finance, RTV Slovenija in Žurnal24). V tej zbirki je bilo 10.427 dokumentov ročno označenih s sentimentom, merjenim s petstopenjsko Likertovo lestvico (1 – zelo negativno, 2 – negativno, 3 – nevtralno 4 – pozitivno in 5 – zalo pozitivno). Dokumente je anotiralo od 2 do 6 anotatorjev, povprečne vrednosti ocen pa so pretvorjene v eno izmed končnih oznak po naslednjih merilih [5]:

- razred »negative«, če je povprečje ocen manjše ali enako 2.4,
- razred »neutral«, če je povprečje ocen med 2.4 in 3.6,
- razred »positive«, če je povprečje ocen večje ali enako 3.6.

Članki so označeni s sentimentom na treh nivojih: nivo stavka, odstavka in dokumenta. V naši raziskavi smo uporabili članke, ki so s sentimentom označeni na nivoju celotnega dokumenta. Izmed skupno 10.427 anotiranih člankov je 5.425 člankov označenih s nevtralnim sentimentom, 3.337 z negativnim in 1.665 z pozitivnim sentimentom. Povprečno število besed v teh člankih je 309.

4 METODOLOGIJA ANALIZE TEMATIK IN SENTIMENTA

V tem razdelku opišemo uporabo tehnologij, opisanih v razdelku 2, za namen analize tem (razdelek 4.1) in določitve sentimenta novic posameznih medijev v izbranih temah (razdelek 4.2). Shema celotnega postopka je ilustrirana na sliki 4.

4.1 Modeliranje tem

S postopki modeliranjem tem želimo v množici slovenskih člankov odkriti različne teme, o katerih pišejo slovenski mediji. Za te teme želimo, da so dovolj podrobne, da bomo lahko na njih odkrivali razlike v sentimentu in opravili primerjavo različnih medijev. Primer tem, ki bi jih lahko odkrili in analizirali je npr. cepljenje proti COVID-19 ali menjava vlade v letu 2020. Za pridobivanje tako podrobnih tem bi lahko zgradili model LDA za veliko število tem (100 in več), vendar bi si s tem otežili interpretacijo pridobljenih tem. Namesto tega smo se modeliranja tem lotili dvofazno. Najprej smo zgradili model LDA na celotnem korpusu besedil za majhno število širših tem (okrog 10), ki jih je hitreje in lažje interpretirati.

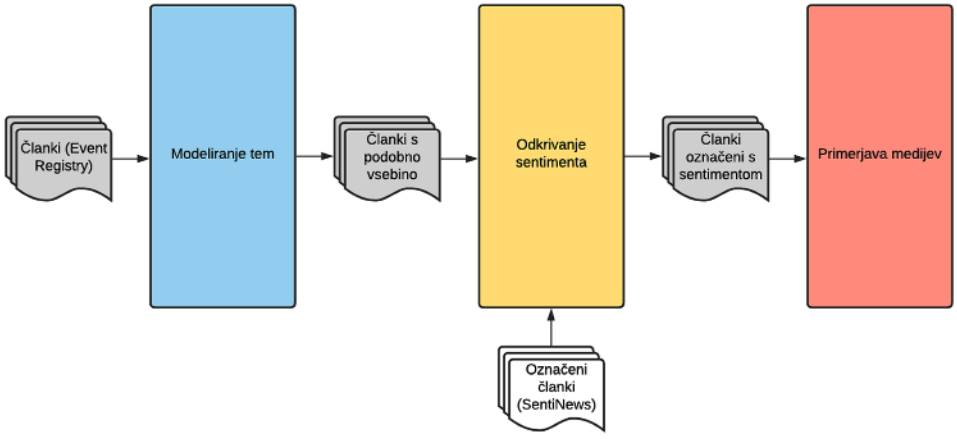


Figure 4: **Shema uporabljene metodologije za analizo člankov.**

Po opravljeni interpretaciji prvostopenjskega modela smo za nadaljnjo obravnavo izbrali le dve širši temi o politiki in epidemiji COVID-19. Iz množice dokumentov smo tako izbrali le tiste, ki imajo verjetnost, da pripadajo širši temi višjo od določene meje (vsaj 0.55). Na tako pridobljeni novi množici člankov smo zgradili nov podrobnejši model LDA. Na koncu smo iz nekaj drugostopenjskih tem znotraj širših tem o politiki in epidemiji COVID-19 izluščili najbolj tipične članke in jih uporabili za analizo sentimenta in primerjavo medijev. Celoten postopek modeliranje tem sestavljajo naslednji koraki, ki so prikazani tudi na sliki 5:

1. Predobdelava člankov (razdelek 4.1.1).
2. Priprava podatkov za računanje prvostopenjskega modela LDA (razdelek 4.1.2).
3. Konstrukcija modela LDA (razdelek 4.1.3).
4. Interpretacija prvostopenjskega modela LDA (razdelek 4.1.4).
5. Izbor teme in člankov za računanje podrobnejšega drugostopenjskega modela.
6. Ponovimo 2., 3. in 4. korak za modeliranje podtem izbranih širših tem.
7. Izbor podtem in člankov za nadaljnjo analizo sentimenta (razdelek 4.1.5).

3.0.1 Predobdelava člankov

V predobdelavi podatkov za odkrivanju tematik iz besedil izluščimo za nas pomembne dele. Postopek na besedilih člankov izvede naslednje korake, ki jih shematsko prikazuje slika 6.

1. Filtriranje člankov.
2. Tokenizacija.
3. Pretvorba besed v male črke.

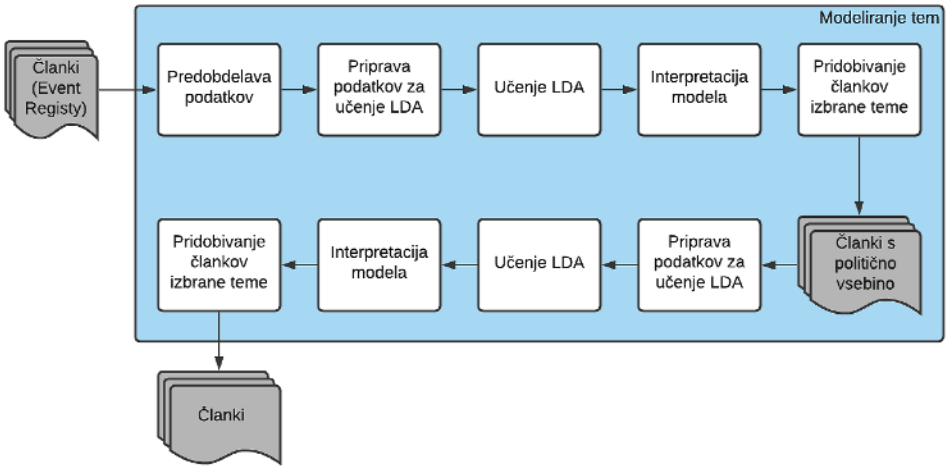


Figure 5: **Dvostopenjski postopek modeliranja tem za pridobivanja člankov določene ožje teme, ki jo analiziramo glede sentimenta.**

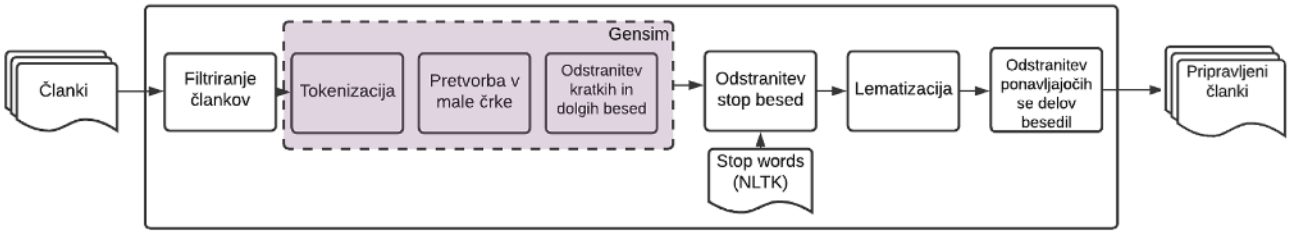


Figure 6: **Proces predobdelave podatkov za analizo tematik.**

4. Odstranitev besed krajših oz. daljših od določene dolžine.
5. Odstranitev nepomembnih besed (angl. stop--words).
6. Lematizacija.
7. Odstranitev ponavljajočih se delov besedil.

V prvem koraku smo članke filtrirali, kar pomeni, da smo odstranili vse članke, ki so krajši od določenega števila besed, in vse podvojene članke, t.j. članke z identičnimi naslovi. Pri odstranjevanju duplikatov smo v tem koraku odstranili zgolj članke s popolnoma identičnimi naslovi, ne pa tudi člankov, kjer je v naslovu spremenjena zgolj posamezna beseda, je pa očitno, da gre za enako vsebino članka.

Naslednje tri korake predobdelave smo opravili s pomočjo metode *simple_preprocess* iz knjižnice *Gensim* [19]. Metoda besedilo razdeli na posamezne besede (tokenizacija), jih pretvori v male črke in odstrani vse besede, ki so krajše oz. daljše od določene dolžine.

Iz preostalih besed odstranimo še t.i. *stop besede* (angl. stop words), ki nimajo posebnega pomena v povedih (npr. vezniki, zaimki, imena mesecev itn.). Seznam stop besed za slovenski jezik smo pridobili iz knjižnice *NLTK* [15]. Vse preostale besede smo pretvorili v njihove leme (osnovne oblike). Za lematizacijo slovenskih besed smo uporabili orodje *Classla* [13].

V zadnjem koraku smo pregledali dobljene članke posameznih medijev in pri določenih medijih opazili ponavljajoče se dele besedil, ki smo jih odstranili. Pri več medijih se je v člankih velikokrat pojavljala beseda »foto«, ki je bila v spletni obliki članka del naslova priloženih fotografij. Odstranili smo tudi oznake virov, npr. 'reuters', 'getty images', 'urbanec' in 'sportid', ki predstavljajo vir fotografij. Pri mediju 24ur.com smo odstranili ponavljajoč se začetni del besedila, ki od uporabnika zahteva omogočenje piškotkov spletne strani. Iz člankov Siol.net Novice smo odstranili ponavljajoč se začetni del besedila, ki se je nanašal na t.i. *termometer*, ki bralcu članka razloži vlogo le-tega pri poročanju o popularnosti članka. Pri ostalih medijih večjih ponavljajočih se delov na začetku člankov nismo opazili.

3.0.2 Priprava n-gramov za model LDA

S postopkom predobdelave podatkov smo iz člankov izluščili posamezne besede, ki nam lahko nekaj povedo o temah člankov. V člankih se določene besede večkrat pojavljajo skupaj (npr. Marjan Šarec, Janez Janša, državni zbor itd.), kar lahko pomaga pri interpretaciji tem. V naslednjem koraku smo zato v člankih zaznali pogoste dvojice besed (bigrame) in pogoste trojice besed (trigrame). Knjižnica *Gensim* [19] ponuja model za avtomatsko zaznavanje pogostih besednih zvez imenovan *Phrases*. Zaznane bigrame in trigrame smo pretvorili v en sam niz besed ločenih s podčrtajem (npr. državni_zbor) in jih dodali v seznam besed predprocesiranih člankov. Pri tem nismo upoštevali tistih bigramov in trigramov, ki se pojavijo v manj kot 15 člankih in tistih, ki imajo vrednost *threshold*[1] nižjo od 100.

Iz predobledanih člankov smo pridobili podatke, ki jih potrebujemo za učenje modela LDA s knjižnico *Gensim* [19]. Ta na vhodu sprejme za vsak članek t.i. vrečo besed (angl. bag of words) in slovar besed (angl. dictionary) za celoten korpus besedil. Slovar besed vsebuje vse unikatne besede iz korpusa predobdelanih besedil, za vsako od besed pa določi unikatno identifikacijsko število (id). S pomočjo slovarja knjižnice *Gensim* tvorimo vrečo besed za vsak članek v korpusu z uporabo metode *doc2bow*.

S tem imamo pripravljen slovar besed in korpus člankov predstavljenih z vrečami besed in lahko začnemo z učenjem modela LDA.

[1] Podrobnosti o parametru *threshold* so predstavljene v dokumentaciji orodja Gensim na https://radimrehurek. com/gensim/models/phrases.html#gensim.models.phrases.Phrases.

3.0.3 Izgradnja modela LDA

Delovanje modela LDA smo razložili v razdelku 2.1. Implementacija modela LDA v knjižnici *Gensim* na vhodu sprejme število tem, ki jih želimo odkriti v besedilu, slovar besed in korpus člankov v formatu vreče besed, katerih pripravo smo opisali v razdelkih 4.1.1 in 4.1.2. Glavno merilo evalvacije modela LDA je smiselnost in interpretabilnost tem, zato smo izračunali več modelov z različnim številom tematik. Izračunane modele smo poizkusili interpretirati in za nadaljevanje postopka izbrali subjektivno najbolj interpretabilen in smiseln model.

3.0.4 Interpretacija tem modela LDA

Pri interpretacijah modelov LDA smo si pomagali z več pripomočki. Uporabili smo najpogostejše besede posameznih tem, poleg tega pa še vizualizacijo modela LDA z orodjem pyLDAvis [20]. Pri interpretaciji je pomembna mera primernosti besed, katere večje vrednosti dajejo prednost besedam, ki pripadajo opazovani temi v večji meri kot drugim temam. Na ta način tematik ne interpretiramo le na podlagi najpogostejših besed ampak uporabimo tudi besede, ki so najbolj primerne za opazovano temo.

Za razumevanje širšega konteksta smo si pri interpretaciji modela LDA pomagali tudi z naslovi člankov. Vsakemu članku smo najprej določili temo, ki ji pripada v največji meri (največja verjetnost). Članke smo nato združili po temah in za vsako temo izbrali 20-30 člankov, ki najbolje predstavljajo posamezno temo (imajo najvišjo verjetnost, da pripadajo temi). Iz teh člankov smo izluščili naslove in jih uporabili pri interpretaciji.

V postopku interpretacije teme poimenujemo s pomočjo treh elementov, ki jih prikazuje slika 7:

- najpogostejše oz. najprimernejše besede določene teme,
- vizualizacija z orodjem pyLDAvis,
- naslovi člankov.

3.0.5 Izbor člankov za nadaljnjo analizo

Ko smo v analizi drugostopenjskega modela LDA določili in izbrali podrobnejše teme, moramo iz njih izbrati članke, ki jih bomo uporabili pri nadaljnji analizi sentimenta. Pri tem želimo izbor tematsko homogene množice člankov, ki bo omogočala smiselno primerjavo med mediji. To dosežemo tako, da iz vsake izbrane teme, izberemo članke z dovolj visoko verjetnostjo pripadnosti temi. Prag verjetnosti je lahko med posameznimi temami različen, zato smo preizkusili različne vrednosti verjetnosti in pri vsaki primerjali še naslove člankov, ki dosegajo prag.

4.2 Analiza sentimenta

Analiza sentimenta je sestavljena iz dveh faz. Najprej na označenih podatkih naučimo napovedni model za sentiment dokumentov, čemur sledi klasifikacija člankov, ki smo jih izbrali v razdelku 4.1.5.

Kot napovedni model smo uporabili vnaprej naučen maskirni jezikovni model SloBERTa [22], ki smo ga podrobneje opisali v razdelku 2.3. Model smo prilagodili za napovedovanje sentimenta z uporabo učne množice SentiNews [5], ki smo jo opisali v razdelku 3.2. Model SloBERTa smo prilagajali s pomočjo knjižnice *transformers* [25]. Osnovni model je bil naučen za nalogo napovedovanja maskiranih besed, zato smo iz modela odstranili zadnji sloj, ki je namenjen tej nalogi, in mu dodali dva nova sloja. Prvi dodani sloj je linearen s 768 nevroni, ki smo mu dodali še opuščanje nevronov (angl. dropout). Kot zadnji sloj smo dodali tri nevrone, vsakega za eno izmed oznak sentimenta (pozitivno, negativno in nevtralno).

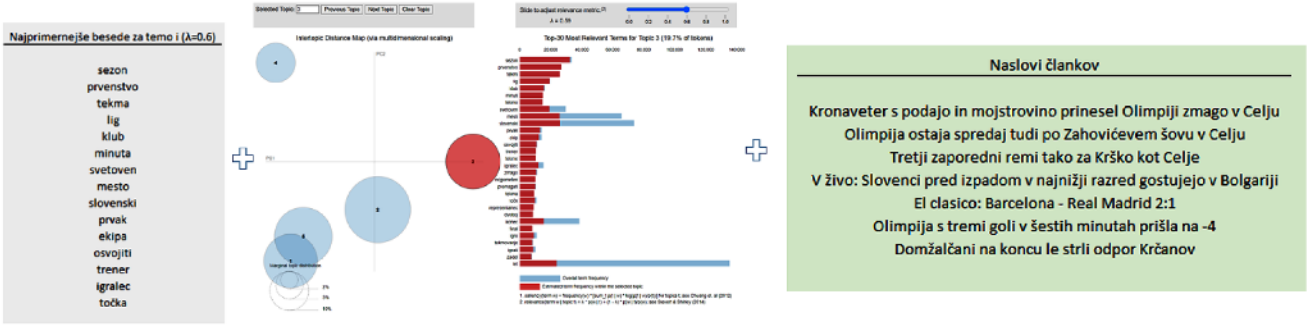


Figure 7: **Informacije, ki jih uporabimo pri interpretacije in poimenovanju posamezne teme: najprimernejše besede (levo), vizualizacija z orodjem LDAvis (sredina)in naslovi člankov (desno).**

Pred prilagajanjem modela smo pripravili vhodne podatke. Model lahko na vhodu sprejme največ 512 žetonov, zato smo predolge članke skrajšali, krajše pa podaljšali s posebnimi žetoni *pad_token*. S tem smo dobili enako dolge predstavitve vseh člankov. Postopek smo izvedli s tokenizatorjem modela SloBERTa, ki pripravi podatke v primernem formatu za učenje:

1. 'input_ids': seznam unikatnih id-jev za vsako izmed besed v članku
2. 'attention_mask': seznam ničel in enic, kjer mesta z ničlami predstavljajo *pad_token*.

Za učenje modela smo uporabili okolje Google Colaboratory, ki ponuja brezplačno uporabo GPU z omejeno rabo spomina. Ta nam je dovoljevala velikosti paketov podatkov (angl. batch size) največ 8 za zaporedja žetonov dolžine 512. Podatke smo razdelili na učno, validacijsko in testno množico, kjer 80% podatkov predstavljajo učni podatki, po 10% pa validacijski in testni podatki. Izpopolnjevanje modela je potekalo 6 epoh; takrat se klasifikacijska točnost na validacijski množici ni več izboljševala in je dosegla najvišjo vrednost 70%.

Za zanesljivo interpretacijo rezultatov naše analize je nujno uporabiti kakovosten napovedni model za analizo sentimenta, zato smo prilagojen model SloBERTa pred dejansko uporabo ovrednotili. Poleg klasifikacijske točnosti, ki je dosegla 70% na testni množici, smo za ocenjevanje modela uporabili še mere točnost, priklic in F_1. Klasifikacijska točnost predstavlja delež pravilno napovedanih vseh primerov. Točnost, priklic in F_1 so namenjene za ocenjevanje klasifikatorjev v dvorazrednih problemih. V večrazrednih problemih enega od razredov izberemo kot pozitivnega, ostali razredi skupaj pa predstavljajo negativni razred. Točnost predstavlja delež pravilno klasificiranih primerov med napovedanimi pozitivnimi primeri, priklic pa nam pove delež pravilno klasificiranih primerov dejanskega pozitivnega razreda. Meri točnost in priklic uporabimo za izračun mere F_1:

$$F_1 = \frac{2 \cdot \text{točnost} \cdot \text{priklic}}{\text{točnost} + \text{priklic}}$$

Podrobnejši rezultati napovednega modela so prikazani v tabeli 3.

Table 3: **Rezultati napovednega modela za zaznavanje sentimenta SloBERTa na testni množici SentiNews.**

razred	točnost	priklic	F_1	št. primerov
negativni	0,67	0,72	0,70	331
nevtralni	0,75	0,69	0,72	558
pozitivni	0,64	0,72	0,68	154
povprečje	0,69	0,71	0,70	1043
uteženo povprečje	0,71	0,70	0,70	1043

Model ima v primerjavi s podobnimi modeli BERT, učenimi na binarnem sentimentu angleških člankov [7, 10], nižjo točnost, kar je posledica več razlogov. Ker smo bili omejeni z delovnim spominom v okolju Google Colaboratory, smo model učili z dokaj majhnimi velikostmi paketov. Slabost modela SloBERTa pri napovedovanju sentimenta celotnih člankov je tudi v omejeni dolžini vhoda v model (512 žetonov), s krajšanjem članka pa izgubljamo informacije. Tretji razlog za nižjo uspešnost modela v primerjavi z angleškimi je tudi v trirazredni klasifikaciji sentimenta, ki je težji problem kot binarna klasifikacija. Menimo, da je uporaba trorazrednega sentimenta zaradi objektivnosti analize bolj smiselna od binarne, saj večina člankov spada prav v tretji, nevtralni razred.

Tako kot Bučar, Povh in Žnidaršič [4] smo tudi mi preizkusili klasifikator SVM (angl. support vector machine), ki se je pri zaznavanju binarnega sentimenta izkazal za dokaj uspešnega (85%). Za učenje sentimenta s tremi možnimi oznakami smo z modelom SVM dobili klasifikacijsko točnost le okrog 60%.

Naučeni napovedni model smo uporabili na člankih iz razdelka 4.1.5, ki smo jih pridobili s procesom modeliranja tematik, in jim določili sentiment. Rezultati modeliranja tem in napovedan sentiment predstavljajo podatke, s pomočjo katerih smo ovrednotili celoten postopek in interpretirali rezultate, kar opišemo v 5. poglavju.

5 INTERPRETACIJA REZULTATOV

V tem razdelku predstavimo rezultate analize tematik, njihovo interpretacijo in primerjamo različne medije. Analizo smo izvedli za leti 2019 in 2020, kjer smo za vsako leto posebej odkrili teme, o katerih pišejo članki, in primerjali odnos medijev do zaznanih tem.

Ker smo želeli ugotoviti, kakšne razlike je zmožna zaznati predstavljena metodologija, smo izbrali štiri osrednje medije (MMC RTV Slovenija, 24ur.com, Siol.

net Novice in Dnevnik) in tri desno usmerjene medije (Nova24TV, Tednik Demokracija in Portal Politikis). Z metodami opisanimi v 4. poglavju želimo poiskati razlike v opredelitvah posameznih medijev do nekaterih tem med tema dvema skupinama medijev.

V razdelku 5.1 predstavimo primerjavo pokritosti posameznih širših tem med različnimi mediji, v razdelku 5.2 pa ožjih podtem. V razdelku 5.3 predstavimo še primerjavo med mediji glede uporabljenega sentimenta pri pisanju o posameznih ožjih temah.

5.1 Rezultati modeliranja splošnih tem

Nekaj statistik o podatkih uporabljenih pri modeliranju splošnih tem z modelom LDA smo za obe analizirani leti zbrali v tabeli 4.

Table 4: **Podatki o učni množici za pripravo modela LDA za splošne teme.**

	2019	2020
število splošnih tem	8	10
št. člankov za učenje modela	94.640	93.914
št. besed v slovarju	66.653	64.051

Razlike med mediji smo opazili že pri modeliranju splošnih tem. Za vsako leto smo naučili model LDA na člankih vseh vključenih medijev in opazili, da model LDA odkrije eno ali več tem, ki jo sestavljajo večinoma članki iz desnih medijev. Na sliki 8 je prikazana porazdelitev tematik, ki jih pokrivajo različni mediji. Opazimo, da je večina člankov desnih medijev dodeljenih ločeni temi o politiki, medtem ko članki s politično vsebino ostalih medijev uporabljajo toliko drugačen jezik, da so dodeljeni drugi temi. Dokaj presenetljiva je tudi nizka frekvenca poročanja treh desnih medijev o športu in svetu slavnih.

Ker želimo v naši analizi preveriti opredelitve medijev do enakih tem, smo iz množice člankov izločili članke desnih medijev in ponovno izračunali model LDA, kot to prikazuje slika 9. Pri novo izračunanem modelu smo dobili bolj enakomerno zastopane teme. Članke desnih medijev smo nato klasificirali z dobljenim modelom LDA. Problem precej drugačnega pisanja desnih medijev in prevlada pisanja o politiki pri njih se je pojavil tako pri splošnem modelu LDA (za pridobitev splošnih tem) kot tudi pri podrobnejših modelih. V obeh primerih smo postopali na zgoraj opisan način.

Za članke iz leta 2020 smo naučili model za 10 tem (slika 9). Večino tem smo lahko poimenovali že samo s pregledom besed z najvišjimi vrednostmi mere primernosti besed (tabela 5), pri nekaterih temah pa smo si pomagali še s preverjanjem naslovov člankov (tabela 5). Temo epidemije virusa COVID-19 smo

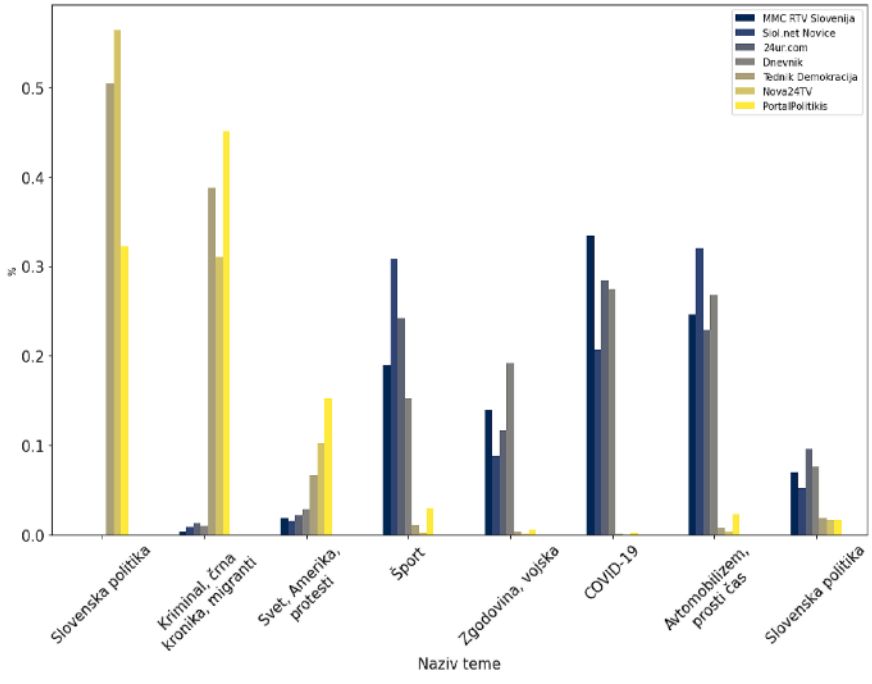


Figure 8: **Distribucija člankov izbranih medijev po temah pridobljenih z modelom LDA, ki je bil naučen na celotni množici podatkov za leto 2020.**

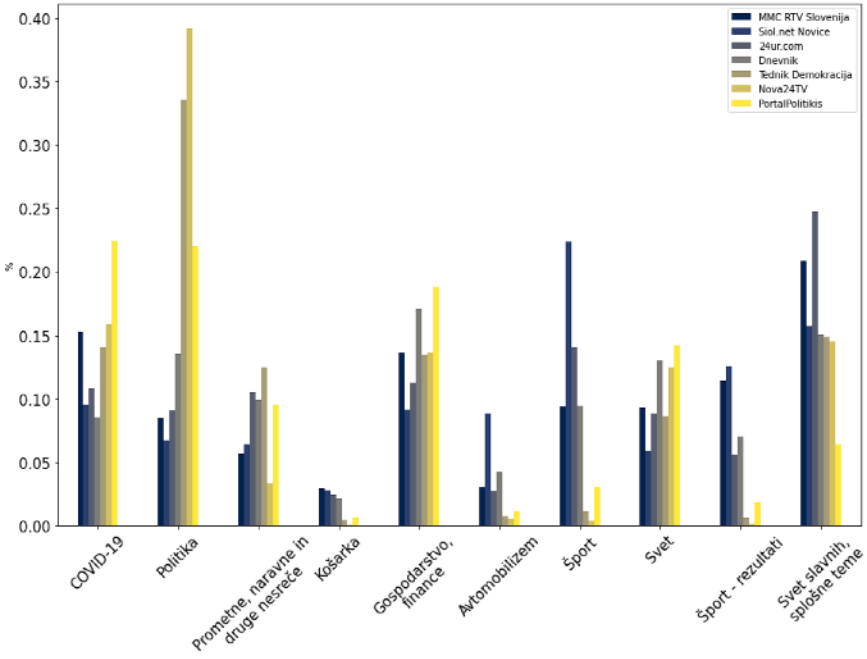


Figure 9: **Distribucija tem člankov po medijih za leto 2020 z modelom LDA, izračunanim na osrednjih medijih.**

lahko prepoznali s pomočjo besed *korona virus, okužb, covid, okužen* itd. Pri športnih tematikah samo s pregledom besed ne moremo dovolj dobro interpretirati tematik. Temi poimenovani Šport in Šport – rezultati sta si po najprimernejših besedah zelo podobni, saj obe v večji meri predstavljata temo nogometa. Po

Table 5: **Pregled 20 lem, pridobljenih z orodjem Classla [13], z največjo vrednostjo mere primernosti za posamezne splošne teme za leto 2020.**

	COVID-19	Politika	Prometne, naravne in ostale nesreče	Košarka	Gospodarstvo, finance	Avtomobilizem	Šport (splošno)	Svet	Šport (rezultati)	Film, glasba svet slavnih
1	koronavirus	stranki	dom	košarkar	podjetji	lahek	sezon	ameriški	minuti	let
2	nov	predsednik	policist	lig	evrov	avtomobiti	klub	trump	tekmi	film
3	okužb	policij	ljubljan	dallas	odstotkov	motor	prvenstvo	let	zmago	življenje
4	covid	vlade	otrok	točk	odstotek	model	nogometen	britanski	mesti	lahek
5	ljud	političen	občin	košarkarski	lahek	električen	lig	kitajski	dvoboj	knjig
6	okužen	minister	območji	končnic	javen	voziti	tekmovanje	dejati	dirko	čas
7	števiti	poslanec	voziti	dončić	let	hitrost	športen	volitev	premagati	imeti
8	potrditi	zakon	center	luk_dončić	zaposlen	avtomobilov	evropski	biden	polčas	svet
9	ukrep	vlad	bolnišnic	igre	ministrstvo	kilometrov	šport	predsednik	zadetek	glasben
10	zdravstven	desus	zdravnik	miami	dejavnost	kolo	let	sporočiti	gol	fotografij
11	držav	sodišče	voznik	parket	deloven	meter	nogomet	držav	kolesar	otrok
12	virus	javen	gasilec	četrtin	gospodarski	let	prvak	poroč	zadel	videti
13	umrl	predlog	občine	tekmi	gospodarstvo	avto	igralec	vojen	zmagovalec	družin
14	okužbe	politik	župan	dosegel	vrednost	zrak	trener	zvezen	priložnost	ljudje
15	karanten	dejati	šolski	boston	pomoč	vozil	slovenski	ruski	prednost	slovenski
16	hrvaški	odločitev	zaposlen	dragić	sredstev	polet	svetoven	trumpov	točki	nagrad
17	bolezen	odstop	pomoč	točka	epidemij	zato	lig_prvakov	napad	niz	zato
18	teden	slovenski	oddelek	miamij	deti	znamki	reprezentanec	dolarjev	tour	filmski
19	bolnikov	opozicij	policij	zvezdnik	družbe	avtomobilski	lahek	vojaški	etap	ljubezen
20	zdravje	postopek	star	tekmo	finančen	hitro	tek	oblast	izgubiti	misel

pregledu naslovov člankov obeh tem smo ugotovili, da ena tema predstavlja predvsem športne rezultate oz. izide tekem, druga pa govori o športnikih in športnih tekmovanjih. Za interpretacijo nekaterih tem je bilo potrebno pregledati več besed in naslovov.

Za članke iz leta 2019 smo določili 8 različnih tem. Postopek interpretacije tem je bil enak prej opisanemu za leto 2020. V obeh letih smo zaznali podobne teme, nekaj pa je tudi razlik. Tako smo v letu 2020 zaznali temo, ki je v celoti pripadala politiki, v letu 2019 pa smo politično tematiko zaznali skupaj z gospodarsko in finančno tematiko. Prav tako smo opazili nekaj razlik pri zaznanih športnih tematikah.

V obeh letih je opaziti razliko v distribuciji tematik med posameznimi mediji. Desni mediji imajo visok delež člankov s temo politike, nizko pa pri tematikah športa. Osrednji mediji imajo bolj enakomerno porazdelitev tem, vseeno pa so tudi med njimi opa-

Table 6: **Primeri naslovov člankov, ki pripadajo posameznim splošnim temam za leto 2020.**

COVID 19	Rekordne številke okužb v Italiji, Nemčiji in Avstriji, v Franciji je umrlo že več kot 40.000 ljudi Rekordno število novookuženih v Avstriji in Nemčiji, na Hrvaškem največ smrti v enem dnevu doslej Na Poljskem število potrjenih okužb preseglo milijon, na Švedskem dnevni rekord okuženih Na Hrvaškem 2399 novih okužb in največ smrti v enem dnevu, 35 V Grčiji ustavljajo javno življenje, v Belgiji prvi znaki izboljšanja
Politika	Polnar: Iz poslanske skupine DeSUS lahko izstopim le sam Pivčeva: Pred sejo sveta stranke ne bom odstopila Direktor Ukoma za opozarjanje na manipulativne novinarske prispevke Predčasne volitve ostajajo prva izbira SD-ja Urbanija za opozarjanje na »manipulativne prispevke«; oster odziv v Levici in SD-ju
Prometne, naravne in druge nesreče	Ceste na Dolenjskem v le nekaj urah vzele dve življenji V Piranu zagorela hiša, sedem oseb prepeljali v bolnišnico Še bo vetrovno, občasno bo rahlo snežilo Voznik pri Kranju zapeljal s ceste in po prevračanju vozila umrl č rni petek: V čelnem trčenju štirje mrtvi
Košarka	LeBronu Jamesu v Houstonu vzklikali »MVP« Naveza Jokić-Porter novo udarno orožje Denverja Dragić in Butler poskrbela za Miamijev zmagovalni 'ogenj' Zmagi za vodilno moštvo obeh konferenc Dallas zmagal brez poškodovanega Dončića
Gospodarstvo, finance	Prispevki delavcev državo doslej stali več kot 90 milijonov evrov Delodajalci v treh dneh oddali 1154 vlog za čakanje zaposlenih Finančni minister: Leta 2021 ne bo novih davkov, nekateri se bodo znižali 2021 ne bo prineslo višjih davkov, tudi davka na nepremičnine (še) ne bo Za 6,6 odstotka BDP-ja primanjkljaja in 16 odstotkov manj proračunskih prihodkov
Avtomobilizem	Clio E-tech Uradno: prenova in novost enega najbolj priljubljenih križancev Električni leaf z večjo baterijo na preizkusu uporabnosti Volkswagnov ofenziva ob koncu leta Kako je voziti male SUV? Ti se izkažejo najbolje.
Šport	Final four v ligi prvakov? Presednika NZS Mijatovića čaka nujni sestanek s Kekom As: V ligi prvakov možen tudi final four Kevin Kampl izvedel, kdaj bo začel sezono Prvi seznam po letu 2005, pod katerim ni podpisa Dobovičnika
Svet	Novi predsednik ZDA je Joe Biden Florida in Teksas Trumpu, ki napoveduje »veliko zmago«; Biden nagovoril javnost: Bodite potrpežljivi Lukašenko na skrivaj prisegel za šesti mandat Macron na skeptičnem vzhodu brani svoj strateški dialog z Moskvo Iran grozi Združenim arabskim emiratom s posledicami zaradi dogovora z Izraelom
Šport – rezultati	Celje – Triglav 1:1 Tekma zapravljenih priložnosti v Kidričevem #video Prva zmaga Celjanov, nocoj na TV SLO 2/MMC Olimpija – Domžale Razigrani Vekić paral živce Domžalam #video Nov poraz Maribora, Aluminij slavil v Domžalah
Film, glasba, svet slavnih	Sestra Miley Cyrus se je večkrat počutila pozabljeno: Živela sem v njeni senci Poslovila se je plesalka Lojzka Žerdin Ko izza znamenitega oranžnega kavča skoči kar sama Rachel Bowrain – 2020 sekund v živo Lana Del Rey svojo glamurozno obleko kupila v nakupovalnem središču

zne razlike v pokrivanju tematik. Tako izstopa večje pokrivanje športa in avtomobilizma portala Siol.net in večje pokrivanje sveta slavnih portalov 24ur.com in MMC RTV.

5.2 Analiza rezultatov modeliranja podrobnejših tem

Naslednji korak analize posameznega leta je interpretacija rezultatov pridobljenih z modeliranjem podrobnejših podtem izbrane splošne teme. V tem koraku pridobimo končne teme in njihove članke, ki jih nato uporabimo za primerjavo sentimenta medijev. Postopek določanja podtem je bil enak kot pri modeliranju splošnih tem.

Za nadaljnjo analizo smo iz leta 2020 vzeli članke s splošnimi temami o epidemiji koronavirusa in slovenski ter svetovni politiki. Za leto 2019 smo poiskali podobne tematike in sicer: politika, gospodarstvo, finance in svetovna politika. Za izbrane teme obeh let smo ponovno izračunali model LDA in s tem pridobili podrobnejše podteme. Za analizo sentimenta smo na koncu izbrali tiste podteme, ki so vsebovale najbolj podobne članke.

Za leto 2020 smo v člankih o slovenski politiki z nadaljnjim modeliranjem našli 12 podtem (slika 10). Naslovi posameznih tem povzemajo dogajanje na slovenski politični sceni v letu 2020. Zaznali smo temo menjave vlade, ki zajema odstop premiera Marjana Šarca, kot tudi sestavo nove vlade. Poleg tega smo zaznali tudi dogajanje znotraj stranke DeSUS in afero Aleksandre Pivec, afero o nabavi zaščitne medicinske opreme, proteste, ukinitev financiranja STA itd. Opaziti je, da imajo desni mediji največji del svojih člankov s temo, ki smo jo označili kot temo razmišljanj o slovenski politiki in njeni zgodovini. Znotraj te teme članki omenjajo predvsem 30 letnico plebiscita o samostojnosti Slovenije, kar je verjetno vzrok, da imajo desni mediji tako visok delež člankov s to temo.

Za leto 2019 med splošnimi temami nismo dobili teme z zgolj politično vsebino, ampak smo dobili politično tematiko skupaj s financami in gospodarstvom. Ker se te tri teme pogosto prepletajo, se to zdi upravičeno. Za temo s političnimi, finančnimi in gospodarskimi članki smo izračunali nov model LDA za 10 podtem. Ena izmed podtem predstavlja celotno politično dogajanje, ostale pa govorijo o podrobnejših gospodarskih in finančnih temah. Primeri takih tem so stečaj letalske družbe Adrie Airways, prevzemi in prodaje podjetij (npr. Mercator), različne sodne obravnave itd.

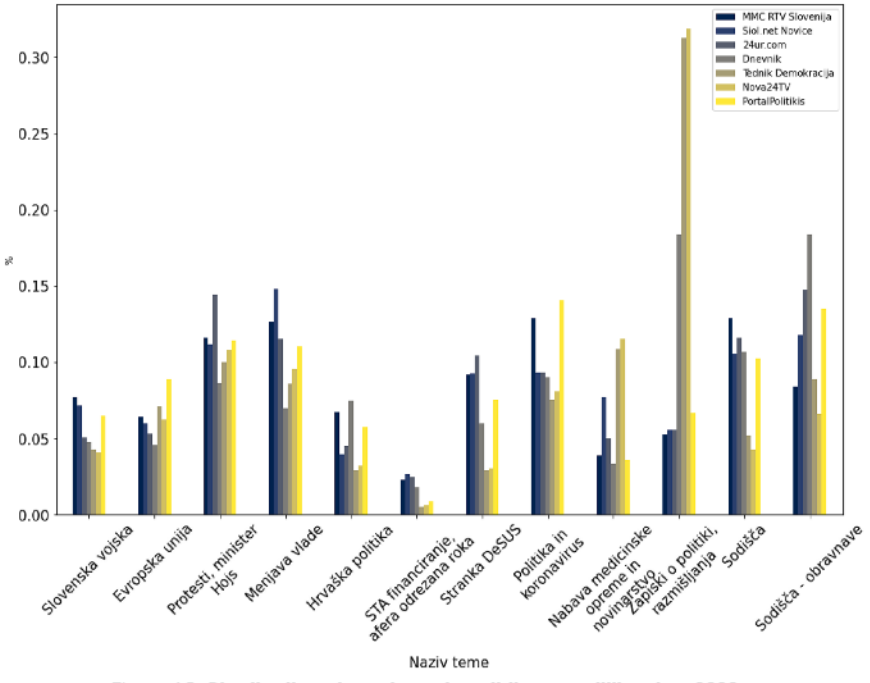


Figure 10: **Distribucija podtem slovenske politike po medijih za leto 2020.**

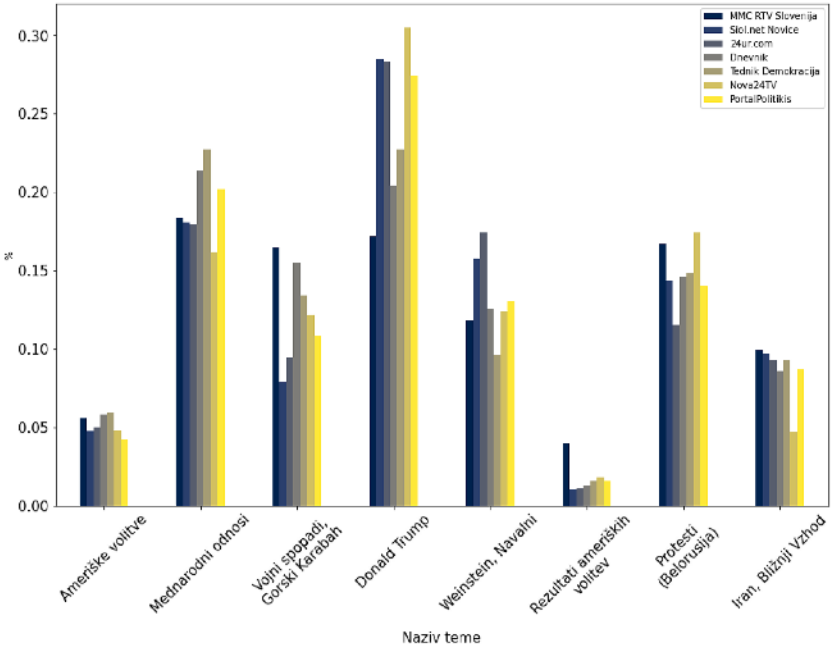


Figure 11: **Distribucija podtem svetovne politike po medijih za leto 2020.**

Pri člankih, ki smo jih določili temi poimenovani 'svet', smo za leto 2020 zaznali 8 različnih podtem, za leto 2019 pa 10. V obeh letih smo zaznali teme o ameriški politiki in predsedniku Donaldu Trumpu, vojnih spopadih na Bližnjem vzhodu in protestih. V letu 2020 je posebna podtema ameriške volitve, v letu 2019 pa evropska politika. Zaradi ameriških volitev v letu 2020 je bilo več govora o predsedovanju Donalda Trumpa, ki je najbolj zastopana svetovna politična tema v letu 2020. V letu 2019 je veliko člankov na temo evropske politike, saj so to leto potekale evropske parlamentarne volitve. Deleži tem se v letu 2020 med posameznimi mediji ne razlikujejo prav veliko, za leto 2019 pa smo opazili, da desni mediji v večji

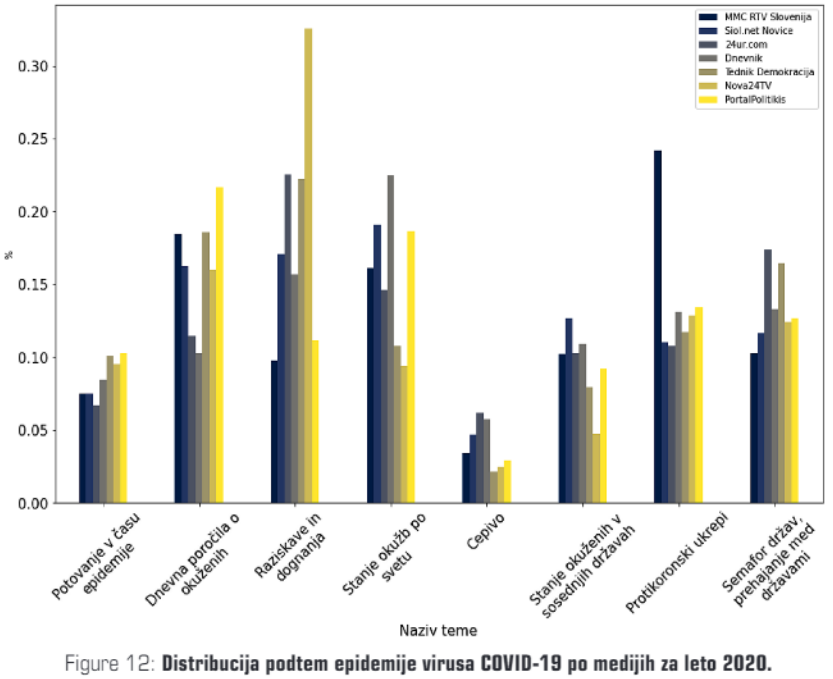


Figure 12: **Distribucija podtem epidemije virusa COVID-19 po medijih za leto 2020.**

meri pišejo o migrantski krizi in terorističnih napadih ter o evropski politiki.

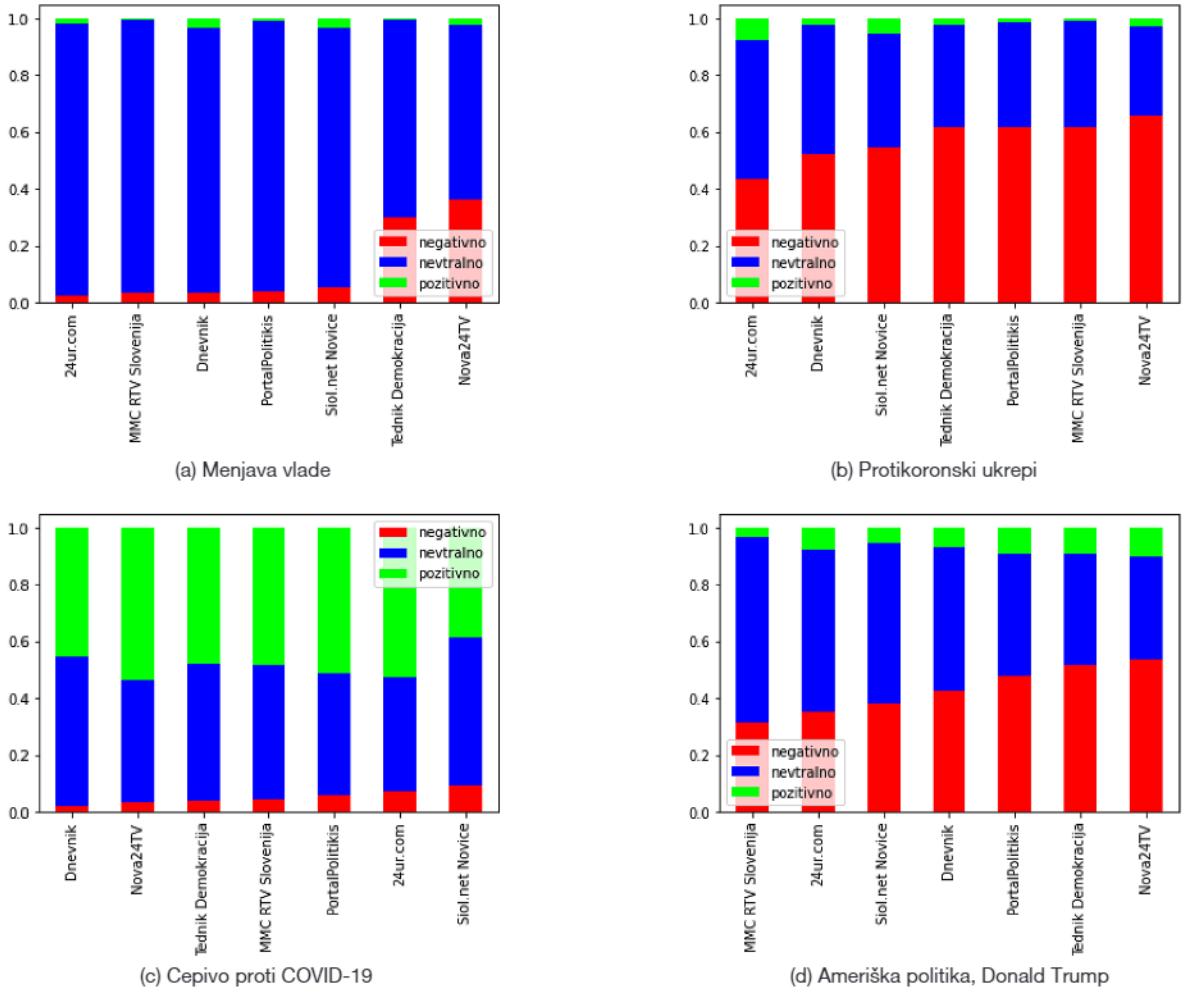
V letu 2020 je posebna tema epidemije virusa COVID-19, ki je pri nas izbruhnila v začetku meseca marca. Znotraj te tematike smo zaznali podteme, kot so poročanje o številu okuženih, o cepivu, evakuacijah in potovanjih v času epidemije ter o protikoronskih ukrepih (slika 12).

5.3 Analiza sentimenta

V prejšnjem razdelku smo analizirali zaznane podteme, v tem razdelku pa izluščimo članke izbranih podtem in primerjamo razlike sentimenta med mediji. Ker nekaj tem govori o enem samem dogodku oz. temi (menjava vlade, stranka DeSUS), nekaj pa jih pokriva več dogodkov (npr. financiranje STA in afera odrezana roka sta del iste teme), smo za analizo sentimenta izbrali tiste, ki predstavljajo en sam dogodek oz. temo.

Za leto 2020 smo izbrali teme menjave vlade, ameriške politike (predsednik Donald Trump), cepivu proti COVID-19 in protikoronskih ukrepih. Porazdelitve sentimentov po posameznih temah so prikazane na sliki 13.

Pri temi menjave slovenske vlade opazimo, da pri vseh medijih prevladuje nevtralen sentiment. Nova24TV in Tednik Demokracija imata nekoliko višji delež člankov z negativnim sentimentom, ostali mediji pa so si zelo podobni po porazdelitvi sentimenta. Pri temah o protikoronskih ukrepih in o ameriški politiki opazimo, da v člankih prevladuje negativni sentiment. Opaziti je nekoliko višji delež pozitivnega sentimenta medijev Nova24TV in Portal Politikis pri temi o ameriškem predsedniku Donaldu Trumpu. Pri vseh treh omenjenih temah je delež negativnega sentimenta najvišji pri desnih medijih, predvsem pri Nova24TV. Pri temi o cepivih je porazdelitev senti-



(a) Menjava vlade

(b) Protikoronski ukrepi

(c) Cepivo proti COVID-19

(d) Ameriška politika, Donald Trump

Figure 13: **Distribucije sentimenta po medijih za izbrane podteme v letu 2020.**

menta precej drugačna. Večina člankov o temi piše s pozitivnim sentimentom, saj večinoma pišejo o razvoju in dobavi cepiv.

Za analizo leta 2019 smo izbrali podobne teme kot za leto 2020: slovensko politiko, stečaj Adrie Airways in ameriško politiko. Temi slovenske in ameriške politike sta dokaj široki, a razbitje na še več podtem bi razdelilo članke na premajhne skupine za zanesljivo analizo sentimenta. Pri temi slovenske politike smo opazili podoben delež sentimenta kot v letu 2020 pri menjavi vlade. Obe temi namreč veliko omenjata koalicijo, opozicijo, politične stranke itd. Pri temi ameriške politike v letu 2019 smo opazili, da imajo mediji zelo podobne porazdelitve sentimenta, pri temah menjave vlade in protikoronskih ukrepov pa imata Tednik Demokracija in Nova24TV opazno višji delež negativnega sentimenta. Zanimivo je, da v letu 2019 Portal Politikis nekoliko odstopal od ostalih dveh desnih medijev.

Z analizo sentimenta v obeh letih smo opazili, da imajo članki desnih medijev pogosto višji delež negativnega sentimenta kot uveljavljeni mediji. Večina negativnih desnih člankov opisuje teme, ljudi, stranke in medije z nasprotnega političnega pola, kar je razlog za višji delež negativnega sentimenta. Prav tako smo v obeh letih opazili nizek delež pozitivnih člankov. Razlogov za to je lahko več. Eden glavnih je zagotovo ta, da v večini obravnavanih tematik težko pričakujemo članke s pozitivnim sentimentom. Drug razlog bi lahko bil v našem modelu za napovedovanje sentimenta. Model smo naučili na relativno majhni količini podatkov označenih s pozitivnim sentimentom, ki večinoma izhaja iz finančne oz. gospodarske tematike (zaslužki, delnice, prodaja itd.).

Rezultate moramo torej jemati z nekaj rezerve, ki izhajajo iz omejitev strojne analize. Omeniti je potrebno tudi, da model zaznava sentiment zgolj za prvih 512 žetonov vsakega članka, kar pomeni izgubo dela informacij. Ločiti je potrebno tudi med zaznavo sentimenta (naklonjenosti) in izraženimi stališči. Sentiment namreč detektiramo v člankih s podobnimi temami, kar pa ne pomeni, da mediji o istih temah govorijo iz istih stališč. V politični temi se velikokrat pokaže, da desni mediji z negativnim sentimentom pišejo o levo opredeljenih strankah ali osebah, med tem ko levo opredeljeni mediji počnejo obratno. Tako članki medijev obeh opredelitev pišejo o isti politični temi, a oboji izpostavljajo negativne aspekte.

6 ZAKLJUČEK

V delu smo s pristopi obdelave naravnega jezika poskusili objektivno analizirati dosedaj največjo zbirko člankov izbranih slovenskih medijev. Z analizo smo želeli primerjati razlike med uveljavljenimi in manj uvel- javljenimi desno opredeljenimi mediji, do katerih smo imeli dostop preko servisa Event Registry. Predstavili smo postopek predobdelave besedil in tehniko modeliranja tematik, kjer smo z dvonivojskim modelom LDA iz člankov zaznali podrobne tematike in z naučenim modelom SloBERTa analizirali sentiment.

Z modeliranjem tematik smo pridobili nekaj smiselnih in interpretabilnih tem, ki opisujejo posamezne dogodke oz. teme. Ugotovili smo, da desni mediji v veliki večini pišejo le o političnih temah. Nekaj tem smo izbrali in njihove članke klasificirali glede sentimenta z naučenim modelom SloBERTa. Ugotovili smo, da je večina člankov političnih tem z negativnim in nevtralnim sentimentom, zelo malo pa s pozitivnim. Opazili smo razliko med desnimi mediji in uveljavljenimi mediji, kjer so za več tem desni mediji imeli višji delež negativnega sentimenta.

Rezultati so pokazali nekaj razlik med mediji tako na nivoju zaznanih tematik kot na nivoju sentimenta. Rezultati bi bili lahko natančnejši, če bi se uspeli izogniti omejitvam našega pristopa. Pri modeliranju tem bi lahko namesto iskanja podobnih člankov glede na najvišjo verjetnost pojavljanja v dani temi uporabili metodo gručenja. Članke bi predstavili s porazdelitvami verjetnosti vektorskih vložitev za posamezne teme in poskusili poiskati smiselne skupine člankov, ki bi vsebovale podobne članke. Rezultate zaznavanja sentimenta bi lahko izboljšali z uporabo natančnejšega modela, ki bi ga naučili z več raznovrstnimi učnimi množicami. Namesto krajšanja člankov na začetnih 512 žetonov bi lahko uporabili kakšen drug pristop, kot na primer kombiniranje začetka in konca članka.

ZAHVALE

Avtorja se zahvaljujeva Gregorju Lebanu iz podjetja Event Registry, ki je omogočil dostop do slovenskih člankov. Raziskavo je sofinancirala Javna agencija za raziskovalno dejavnost Republike Slovenije skozi projekt J6-2581 (Računalniško podprta večjezična analiza novičarskega diskurza s kontekstualnimi besednimi vložitvami) in raziskovalni program P6-0411 (Jezikovni viri in tehnologije za slovenski jezik).

LITERATURA

[1] David Blei, Lawrence Carin, and David Dunson. Probabilistic topic models: A focus on graphical model design and applications to document and image analysis. *IEEE signal processing magazine*, 27:55–65, 11 2010.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of machine learning research*, 3:993–1022, 2003.

[3] Jože Bučar. Manually sentiment annotated Slovenian news corpus SentiNews 1.0, 2017. Slovenian language resource repository CLARIN.SI.

[4] Jože Bučar, Janez Povh, and Martin Žnidaršič. Sentiment classification of the Slovenian news texts. In *Proceedings of the 9th International Conference on Computer Recognition Systems CORES 2015*, pages 777–787, 2016.

[5] Jože Bučar, Martin Žnidaršič, and Janez Povh. Annotated news corpora and a lexicon for sentiment analysis in Slovene. *Language Resources and Evaluation*, 52(3):895–919, 2018.

[6] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391– 407, 1990.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[8] Susan T. Dumais. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38(1):188–230, 2004.

[9] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual interna- tional ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

[10] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. *CoRR*, abs/1909.11942, 2019.

[11] Gregor Leban, Blaz Fortuna, Janez Brank, and Marko Grobelnik. Event Registry: Learning about world events from news. In *Proceedings of the 23rd International Conference on World Wide Web*, page 107–110, 2014.

[12] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *ArXiv*, abs/1907.11692, 2019.

[13] Nikola Ljubešić and Kaja Dobrovoljc. What does neural bring? Analysing improvements in morpho- syntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 29–34, 2019.

[14] Nataša Logar and Nikola Ljubešić. Gigafida in slWaC: tematska primerjava. *Slovenščina 2.0*, 1(1):78–110, 2013.

[15] Edward Loper and Steven Bird. NLTK: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics.*, 2002.

[16] Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric Vil- lemonte de la Clergerie, Djamé Seddah, and Benoit Sagot. CamemBERT: A tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[17] Matej Martinc, Nina Perger, Andraž Pelicon, Matej Ulčar, Andreja Vezovnik, and Senja Pollak. EM- BEDDIA hackathon report: Automatic sentiment and viewpoint analysis of Slovenian news corpus on the topic of LGBTIQ+. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 121–126, 2021.

[18] Andraž Pelicon, Marko Pranjić, Dragana Miljković, Blaž Škrlj, and Senja Pollak. Zero-shot learning for cross-lingual news sentiment classification. *Applied Sciences*, 10(17):5993, 2020.

[19] Radim Řehůřek and Petr Sojka. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, 2010.

[20] Carson Sievert and Kenneth Shirley. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70, 2014.

[21] Iza Skrjanec and Senja Pollak. Topic ontologies of the Slovene blogosphere: A gender perspective. 2016.

[22] Matej Ulčar and Marko Robnik-Šikonja. Slovenian RoBERTa contextual embeddings model: SloBERTa 1.0, 2020. Slovenian language resource repository CLARIN.SI.

[23] Matej Ulčar and Marko Robnik-Šikonja. SloBERTa: Slovene monolingual large pretrained masked language model. In *Proceedings of Data Mining and Data Warehousing, SiKDD*, 2021.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 6000–6010, 2017.

[25] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, 2020.