Medij. istraž. (god. 28, br. 2) 2022. (63-85)
IZVORNI ZNANSTVENI RAD
DOI: 10.22572/mi.28.2.3
Zaprimljeno: 15. listopada 2021.

Izgradnja jezičnog korpusa govora mržnje na hrvatskom medijskom prostoru društvenih mreža

Slobodan Hadžić*
Artur Šilić**
Tanja Grmuša***

SAŽETAK

Govor mržnje predstavlja neprihvatljiv oblik društveno štetnih komunikacijskih formi čija je raširenost u posljednje vrijeme u porastu. Razvojem digitalnih medija, a posebice jačanjem uloge društvenih mreža u privatnoj i javnoj komunikaciji, otvoren je prostor za brojne javne virtualne forume koji su pasivnu publiku potaknuli na aktivniju participaciju i komunikaciju. Sloboda govora i izražavanja kao temeljna demokratska načela s jedne strane suočavaju se s prostorom za toksičnu komunikaciju s druge strane, okupljajući istomišljenike u virtualnim zajednicama kojima su meta napada nerijetko akteri priloga, novinari, urednici, mediji, pa i drugi korisnici. Uzimajući u obzir neograničenost internetskog prostora čiji je sadržaj teško kontrolirati, postavlja se pitanje kako prepoznati društveno štetne komunikacijske forme u javnom prostoru, može li ih se spriječiti i kako zadržati postojeću publiku.

Mogućnosti upotrebe proširuju se i na područje moderiranja neprimjerenih komentara korisnika oslanjanjem na softvere koji pružaju trenutne odgovore medijskim organizacijama, ali i koji također pokazuju kontinuiranu potrebu za samopobolj-

* Slobodan Hadžić, PRESSCUT d.o.o., Ulica kneza Domagoja 2, 10 000 Zagreb. email: slobodan.hadzic@presscut.hr

** Artur Šilić, EFFICODE SYSTEMS d.o.o., Maksimirska cesta 88, 10 000 Zagreb, email: artur.silic@gmail.com

*** Tanja Grmuša, Poslovno veleučilište Zagreb, Katedra studija Marketinga i komunikacija, Ulica grada Vukovara 68, 10 000 Zagreb, email: tanja.grmusa@pvzg.hr

šanjima. Interes autora u ovom radu vezan je za detekciju govora mržnje prema etnicitetu na društvenim mrežama te istraživanju mogućnosti primjene jezičnih tehnologija u prepoznavanju i sprječavanju širenja govora mrženja. Kvantitativnom i kvalitativnom analizom sadržaja te primjenom niza softverskih rješenja temeljenih na jezičnim tehnologijama omogućena je učinkovita automatska i poluautomatska analiza velike količine korisnički generiranog sadržaja. Korišteni su: program WordFinder za brzi pronalazak riječi u velikim korpusima, alat CRONTIMENT za automatsku dodjelu sentimenta tekstovima na hrvatskom jeziku, aplikacija Text Marker za učinkovito ručno označavanje i izgradnju korpusa. U pojedinim studijama slučaja autori detektiraju promjene u vrstama i frekvencijama pojave govora mržnje u predmetu istraživanja te identificiraju glavne prednosti i nedostatke primjene jezičnih tehnologija, sugerirajući pri tome moguće smjerove razvoja.

Ključne riječi: društveno štetne komunikacijske forme, govor mržnje, digitalni mediji, publika, jezične tehnologije, komentari korisnika

Uvod

Govor mržnje ubrajamo u društveno štetne oblike komunikacije (Labaš, Grmuša, 2011), a ostvaruje se u usmenom i pisanom obliku. Iako jedinstvene i općeprihvaćene definicije ovog pojma nema (usp. McGonagle, 2012; Weber, 2009: 3 prema Klepač Pogrmilović, 2016: 256; Car, 2016; Gardašević, 2016), njegov porast u svakodnevnoj upotrebi i štetne posljedice u javnom prostoru, ali i društvu u cjelini, dovoljan su razlog za nastavak istraživanja. Istodobno, riječ je o temi koja dijeli društva s obzirom na zagovornike slobode govora (Brax, 2016) i izražavanja kao preduvjeta slobodi medija (Svensson, 2016; Kenyon, 2016; Edström, 2016) i zagovornike sankcioniranja komunikacijskih obrazaca koji sadrže neprimjeren govor, stoga je pitanje regulacije i samoregulacije uvijek aktualno (Carlsson, 2016; Đukić, Sić, 2021). Razvoj interneta i novih tehnologija utjecao je na ključne dionike komunikacijskog procesa, stavivši u prvi plan konzumente kojima se nerijetko prilagođava i medijski sadržaj, često podilazeći lakim, spektakularnim (Hromadžić, 2013) i tabloidnim temama (Car, 2016) u kojima etika prva stradava (Vilović, 2003). Uspon društvenih mreža promijenio je komunikacijsku paradigmu u kontekstu identifikacije primarnog izvora (Ostrički, 2017), moderiranja *online* sadržaja (Kalsnes, Ihlebaek, 2021) i korisničke preferencije upotrebe medija, ali i otvorio brojna pitanja o ulozi publike u upravljanju medijima (Hasebrink, 2012). Pandemija bolesti COVID-19 koja je zahvatila svijet početkom 2020. rezultirala je porastom govora mržnje u društvu, pri čemu su društvene mre-

že postale glavni alat za obračun s brojnim neistomišljenicima (Van Dijck, Alinejad, 2020; Haapoja, Laakosonen, Lampinen 2020).

2. Tehnologija u službi otkrivanja i sprječavanja govora mržnje – multidisciplinarni pristup i metodološka ograničenja

Regulacija platformi u središtu je pozornosti javnosti posljednjih nekoliko godina. Skandali poput Cambridge Analytice, uključenost u izborne procese diljem svijeta (SAD, EU, Kina, Indija, Brazil) samo su neki od razloga za inzistiranje na uređivanju digitalne sfere uz pomoć umjetne inteligencije (usp. Bossetta, 2020: 1), a važnost problema prepoznala je i Europska unija donošenjem brojnih regulativa koje obvezuju platforme na odgovorno djelovanje (Elkin-Koren, 2020). Platforme pak reagiraju uvođenjem indikatora povjerenja kako bi osigurale vjerodostojnost u očima publike. Upotreba umjetne inteligencije u praksi ogleda se u nekoliko dimenzija: a) u očuvanju autorskih prava, b) borbi protiv distribucije takvog sadržaja, c) prepoznavanju toksičnoga govora (usp. Gorwa, Binns, Katzenbach 2020: 9). Upotreba algoritama podrazumijeva identifikaciju i/li povezivanje različitih oblika sadržaja (vizualni, auditivni, tekstualni), podsjećajući da njihova upotreba ovisi o svrsi ili tipu podataka koji se promatraju, pri čemu se zanemaruje problem upotrebe osobnog jezika i kontekstualno ovisnog jezika (usp. 2020, 3–10) te pripisivanje pogrešnog sentimenta (Elkin-Koren, 2020).

Tarleton upozorava i na nejasnu klasifikaciju govora mržnje i otvorena pitanja o broju ljudi koji bi trebali pokrivati ovo područje s obzirom na mnoštvo izazova: od kvantitete sadržaja, brzine do raznolikosti, koji nadilaze mogućnosti samih društvenih medija (usp. 2020: 2). Također, postavlja se pitanje koliko precizno ljudski faktor može prepoznati sigurnosne izazove kao društveno neprihvatljiv sadržaj (Ullmann, Tomalin, 2019), što ističe potrebu za zapošljavanjem dodatnih specijalista u medijskim organizacijama kao posljedicu digitalne transformacije organizacije (Grmuša, 2021). Istodobno, ako pitamo hrvatske novinare i urednike to ne izgleda realno (usp. Grmuša, Prelog, 2020: 73–75), a upitno je mogu li ih medijske kuće platiti. Istraživanja govora mržnje u posljednja tri desetljeća pokazala su 12 dominantnih tema ovog komunikacijskog diskursa: odnos govora mržnje i slobode izražavanja, politički aspekti fenomena govora mržnje, govor posvećen ekstremizmu i terorizmu, pitanje društvenih mreža i zajednica, upotreba društvenih mreža i analiza sentimenta, pitanje raširenosti informacija na internetu (usp. Tontodimamma, Nissi, Sarra, Fontanella, 2021: 165). Istodobno u posljednje vrijeme fokus istraživača usmjerava se na istraživanje mogućnosti računalnih tehnologija u prepoznavanju govora mržnje (Ljubešić, Erjavec, Fišer, 2018; Kocijan, Košković, Bajac, 2019). Sustavnu longitudinalnu studiju o prisutnosti

govora mržnje u hrvatskom medijskom prostoru ponudili su Poljak, Hadžić i Martinić (2020) analizirajući razdoblje od 2013. do 2019. godine.

Govor mržnje iz dominantno usko promatrane analize medijskog sadržaja u elektroničkim medijima i na portalima postaje sve zastupljeniji u virtualnoj sferi, posebice na društvenim mrežama (Uyheng i Carley, 2021), što rezultira povećanjem broja radova usmjerenih tom području, ali i promišljanja o multidisciplinarnom pristupu prepoznavanja i praćenja ovog fenomena uz korištenje naprednih računalnih alata. Lingvističke analize komunikacijskih obrazaca i govornih činova (Gibson, 2019) pokazuju promjenjivost jezičnih diskursa korisnika ovisno o grupi kojoj pripadaju. Također, Ullman i Tomalin ističu kako se ne smije zanemariti ni mogućnost igre riječi u engleskom jeziku čime se može poduprijeti govor mržnje (usp. 2019: 6), a posebnu pozornost treba obratiti na mlade korisnike što traži kontinuiranu prilagodbu (usp. Vijavaraghavan, Larochelle i Roy, 2021: 2–3). Yin i Zubiaga ističu kako je otkrivanje govora mržnje otežana zbog ograničenja postojećih NLP (prirodno procesuiranje jezika) metoda, ali i promjenjive prirode *online* govora mržnje (usp. 2021: 19) koji može biti izražen različitim oblicima kao što su sarkazam, stereotip, ironija, humor, metafora (usp. Sap, Gabriel i sur., 2019; Mishra i sur., 2019; Vidgen i sur., 2019 prema Yin i Zubiaga, 2021: 15). Brzi rast i razvoj vokabulara koji se može pripisati govoru mržnje, a koje potiču povezani događaji (usp. Florio i sur., 2020 prema Markov, Ljubešić, Fišer, Daelemans, Walter, 2021: 149) izazov je i za računalne programe. Matamoros-Fernandez i Farkas (2021) opisuju uočene metodološke izazove: skriveni identiteti autora neprimjerena izričaja, pristup tekstnom sadržaju koji su korisnici prethodno uklonili, gubitak konteksta zbog ekstrakcije riječi (usp. Chaudry, 2015; Eddington, 2018; Tulkens i sur., 2016.; Mondal i sur., 2017.; Saleem i sur., 2017 prema 2021: 214). Također, tu su i etički izazovi: mogući napadi na istraživače, emocionalni stres, pitanje poštivanja privatnosti korisnika (usp. 2021: 215). Govor mržnje, kao što pokazuju prethodna istraživanja, ne može se promatrati izvan sociološkog i kulturološkog konteksta koji doprinosi njegovu razumijevanju i kontekstu (usp. Kwok i Wang, 2013; Raisi i Huang, 2016 prema Vijayaraghavan, Lacharelle, Roy, 2021: 1).

3. Metodologija istraživanja

3.1. Ciljevi istraživanja

Ciljevi istraživanja bili su sljedeći:

1) detektirati kategorije neprimjerena govora i govora mržnje prema etnicitetu na društvenim mrežama,

a) identificirati u kojoj mjeri pretraživanje sadržaja društvenih mreža putem negativnih riječi asociranih s nekom skupinom može pridonijeti pronalaženju neprimjerena govora ili u najtežem slučaju govora mržnje,

b) identificirati kojom se strategijom govora mržnje koriste komentatori iznimno neprimjerenih komentara,

c) identificirati u kojoj mjeri iznimno neprimjereni komentari pozivaju na nasilje,

d) identificirati referiraju li se iznimno neprimjereni komentari usmjereni prema skupinama na članak,

e) istražiti teme članaka koji su generirali iznimno neprimjeren govor prema skupinama,

2) ispitati mogućnosti jezičnih tehnologija u identifikaciji govora mržnje

a) utvrditi u kojoj mjeri automatski sentiment može pridonijeti pronalasku neprimjerena govora ili u najtežem slučaju govora mržnje uz primjenu dodatnih softwarskih alata /WordFinder, Crontiment, Text Marker, Metricom – alat za prikupljanje sadržaja s društvenih mreža/.

Polazne hipoteze:

H1: Oslanjanje na negativne izraze povezane s određenom skupinom pridonosi pronalaženju neprimjerena govora ili u najtežem slučaju govora mržnje.

H2: Komentatori iznimno neprimjerenih komentara koriste se dominantno strategijom dehumanizacije.

H3: Većina neprimjerenih komentara poziva na nasilje prema određenim skupinama.

H4: Neprimjereni komentari usmjereni prema određenim skupinama referiraju se na članak (medijsku objavu na društvenim mrežama).

H5: Teme članaka koje su generirale neprimjeren govor prema određenim skupinama odnose se na godišnjice nasilnih događaja i političke komentare.

H6: Jezične tehnologije mogu ubrzati proces prepoznavanja i prikupljanja govora mržnje, ali ne mogu u potpunosti zamijeniti ljudski faktor.

Podatke za analizu sadržaja društvenih mreža prikupila je tvrtka Medianet d.o.o. te ih je ustupila istraživačima.

3.2. Metode i tehnike istraživanja

Osnovna metoda istraživanja je kvalitativna i kvantitativna analiza sadržaja. Riječ je o metodi prikupljanja uglavnom primarnih podataka iz medijskog sadržaja, ali i

masovne komunikacije općenito (usp. Tkalac Verčič, Sinčić Ćorić, Pološki Vokić, 2011: 91). Kvalitativna (nefrekvencijska) analiza sadržaja usmjerena je na otkrivanje obilježja analiziranog sadržaja, dok kvantitativna (frekvencijska) analiza sadržaja mjeri obujam analiziranog sadržaja (frekvenciju) (usp. Tkalac Verčič i sur., 2011: 92; Wimmer i Dominick, 2011: 157–159).

Period istraživanja i obuhvat

Period istraživanja društvenih mreža proveden je na objavama od 6. 5. 2021. do 7. 8. 2021. te podrazumijeva korpus od ukupno 32 848 530 objava hrvatskog medijskog prostora društvenih mreža.

Jedinica analize istraživanja bila je objava na društvenim mrežama Facebook, Twitter, YouTube, Instagram, Forumi i Usenet na prostoru Hrvatske, definirana jezikom korištenom u objavi ili javnom lokacijom domicila korisnika. Objava je bilo koja vrsta javne objave (ne privatne objave i razgovori) na društvenim mrežama (status, komentar, *tweet, retweet*). U ovom istraživanju fokus je stavljen na analizu govora mržnje prema etničkim manjinama u Hrvatskoj. Stoga je za svaku objavu predviđena jedna ili više od sljedećih glavnih kategorija analize: „Albanci“, „Bosanci“, „Hercegovci“, „Mađari“, „Romi“, „Slovenci“, „Srbi“, „Talijani“[1]. Također, za potrebe označavanja dodaju se i kategorije „drugi“, „nitko“, ali i „Hrvati“ jer je u dobrom dijelu slučajeva govor mržnje usmjeren i prema većinskoj etničkoj skupini. Ciljne skupine odabrane su po učestalosti i ne predstavljaju iscrpnu listu svih mogućih etničkih skupina prema kojima se ostvaruje govor mržnje, što je prilika za buduće radove.

3.2.1. Problemi prilikom prikupljanja i analiziranja podataka

Jedan od problema prikupljanja objava koje sadrže govor mržnje je činjenica da mnogi mediji brišu komentare i na taj način slika pravog stanja prisutnoga govora mržnje nije uvijek potpuna.

Kako bi pronalazak neutralnih i negativnih riječi za svaku skupinu bio što sustavniji i iscrpniji, priklonili smo se korijenskoj pretrazi riječi koje se pojavljuju u velikom mrežnom korpusu hrvatskog jezika HrWaC2 – 3,6 milijuna tekstova, 1,4 milijarde pojavnica (Ljubešić, Klubička, 2014). Za tu svrhu kreiran je poseban program WordFinder koji za dani korijen trenutno vraća frekvencije svih pojavnica koje ga sadrže. U suštini, za pronalazak korijena radimo prefiksnu ili infiksnu pretragu svih pojavnica s pažnjom na moguće glasovne promjene. Primjerice, za „Srbi“ to će pretraga prefiksa ili infiksa „srb“, „srp“, „serb“. Nakon pretrage pojedinog korijena,

slijedi njihova pohrana i pregled analitičara. Odabiru se riječi koje se značenjem odnose na ciljnu skupinu i svakoj se riječi dodaje *apriorna* ocjena negativnosti (w0 – neutralna, w1– blago negativna, w2 – jako negativna). Uklanjaju se očite greške i neki *hapax legomena*. Za svaku skupinu odabrano je više početnih korijena koji su se koristili za pretragu.

Tablica 1. Broj pronađenih i korištenih riječi za svaku skupinu

Table 1 Number of words found and used for each group

skupina	korijenski pronađene riječi	korištene ključne riječi za dohvat
Albanci	57	27
Bosanci	294	118
Hercegovci	94	39
Mađari	97	55
Romi	101	49
Slovenci	41	16
Srbi	1143	91
Talijani	52	31

Za slobodan tekst kakav je prisutan na internetu i posebno u objavama na društvenim mrežama svojstvena je sklonost autora kovanju novih riječi, a osim bogatih derivacija, prisutno je i mnogo složenica. Za skupine s najviše pronađenih riječi složenice s niskom frekvencijom velikog su udjela u pronađenim riječima (npr. Bosanci – „bosanskojugoslavenski“, Srbi – „srbojugočetnički“), što ukazuje na moguću dobit u kvaliteti prepoznavanja govora mržnje razdvajanjem složenica ili korištenjem dubokih modela učenja na razini znakova.

3.3. Filtriranje i priprema za označavanje, uzorak za analizu

Iz cijelog korpusa izdvojene su objave koje sadrže u naslovu ili tijelu teksta bilo koju od ključnih riječi. Kod objava foruma, naslov je zadan naslovom dretve, a kod dijeljenja objava na Facebooku naslov je zadan naslovom podijeljenog sadržaja. Na taj se način ostvaruje minimalni kontekst koji često igra ulogu u semantici govora mržnje. Filtriranjem su izdvojene 71 682 objave te je dodatnim uklanjanjem duplikata dobiveno ukupno 64 285 objava. Iskorišten je CRONTIMENT, sustav za automatsku dodjelu sentimenta na hrvatskom jeziku koji je temeljen na strojnom

učenju i velikom ručno konstruiranom rječniku sentimenta (28 000 riječi) te postiže kvalitetu oko 85 % prema F1 mjeri.

Tako je svakoj objavi automatski dodijeljen negativni sentiment u trima razinama:

s0 – manje od 50 % vjerojatnosti negativnog sentimenta,

s1 – 50 % do 65 % vjerojatnosti,

s2 – 65 % i veća vjerojatnost.

Osim toga, svakoj objavi pripisana je i oznaka *apriorne* negativnosti riječi prema ciljnim skupinama:

w0 – nema *apriornih* negativnih riječi,

w1 – pojavnost barem jedne riječi razine 1 /negativna riječ/ i nijedne riječi razine 2 /izrazito negativna/,

w2 – pojavnost barem jedne riječi razine 2, izrazito negativne riječi.

Prema tih šest oznaka korpus je razdijeljen u devet segmenata s razdiobom navedenom u Tablici 2.

Tablica 2. Broj objava filtriranog korpusa prema razinama negativnosti sentimenta i ključnih riječi

Table 2 Number of posts from filtered corpus based on sentiment negativity and key words

	w0	w1	w2
s0	15777	1985	2229
s1	11440	1494	2126
s2	20978	4161	4095

3.4. Ručno označavanje

Radi ekonomičnosti označavanja, za objave kojima je dodijeljena oznaka w0 (tj. ne sadrže *apriorne* negativne ključne riječi) izdvojen je slučajan podskup. Tako je stvoren konačan korpus od 33 704 objave s razdiobom prema razini negativnosti predočenom u Tablici 3.

Tablica 3. Ukupan broj označenih objava prema razinama negativnosti
Table 3 Total of used posts based on negativity

	w0	w1	w2
s0	5190	1985	2229
s1	5049	1494	2126
s2	7375	4161	4095

Cilj je bio pronaći neprimjeren govor u komentaru uz moguće minimalni kontekst naslova.

Primjeri objava (komentara) uključenih u ručno označavanje nalaze su u nastavku. Zvjezdicama su izostavljena osobna imena privatnih korisnika društvenih mreža.

---

vrijeme: 08/06/2021 10:02:10

izvor: facebook.com

naslov: Rijeka Danas [SHARE] Skejo u Kninu održao govor pa se derao 'Za dom spremni' (VIDEO)

komentar: `**** **** kaže srbenda koja bi veliku Srbiju.`

---

vrijeme: 11.5.2021. 13:24:06

izvor: forum.index.hr

naslov: Hasanbegović: Tomašević nije mirni dečko iz susjedstva, on je strana franšiza

komentar: `**** **** Balija idi odakle si došao pa tamo dijeli lekcije i pametuj, a nas ostavi na miru.`

Označavanje je izvršilo ukupno pet osoba. Na kontrolnim dijelovima skupa za označavanje (oko 10 % ukupnog broja objava) dobivena je prosječna podudarnost od 79 %. Za jednu objavu, par označavanja dvaju označivača smatra se podudarnim ako se radi o potpuno istim skupovima. Preostalih 90 % oznaka dodijeljeno je od strane samo jedne osobe radi obuhvata što veće količine materijala.

4. Interpretacija rezultata

4.1. Udjeli neprimjerena govora na označenom korpusu

Nakon završetka označavanja, objave s više označivača unificirane su unijom skupova oznaka (pazeći na semantiku oznake „nitko") i dobiveni su rezultati prikazani u Tablici 4.

Tablica 4. Udio neprimjerena govora ovisno o prediktorima {s0, s1, s2} i {w0, w1, w2}

Table 4 Proportion of inappropriate speech depending on predictors {s0, s1, s2} i {w0, w1, w2}

	w0	w1	w2
s0	5,7 %	22,3 %	65,8 %
s1	3,0 %	12,7 %	60,4 %
s2	30,5 %	84,5 %	68,9 %

U Tablici 5. dani su brojevi objava s neprimjerenim govorom (očekivani za stupac w0 i stvarni za stupce w1 i w2, vidi poglavlje 5.).

Tablica 5. Količina neprimjerena govora ovisno o prediktorima {s0, s1, s2} i {w0, w1, w2}, za stupac w0 dana je procjena s obzirom na inicijalne količine skupova u filtriranom korpusu

Table 5 The amount of inappropriate speech depending on the predictors {s0, s1, s2} and {w0, w1, w2}, for the column w0 an estimate is given with regard to the initial amount of sets in the filtered corpus

	w0*	w1	w2
s0	*894*	442	1467
s1	*342*	190	1285
s2	*6397*	3516	2822

Iz Tablica 4 i 5 čitamo da su prediktori {w0, w1, w2} i {s0, s1, s2} korisni za izdvajanje neprimjerena govora. Na konkretnom slučaju danog korpusa društvenih mreža

i manjinskih etničkih grupa možemo očekivati da ako imamo popis ciljnih riječi s razinama negativnosti {w0, w1, w2} i koristimo tehnologiju dodjele automatskog sentimenta u više negativnih razina {s0, s1, s2}, moguće je znatno smanjenje količine ljudskog označavanja za pronalazak primjera neprimjerena govora. Primjerice, uzevši samo objave sa s2 ili uključivo w2 prediktorima i ručnim označavanjem tek 52 % korpusa možemo obuhvatiti čak 89 % neprimjerena govora.

U Tablici 6 vidi se količina neprimjerena govora na označenom korpusu prema pojedinim skupinama. Iz nje možemo pročitati kako opisanom metodom dohvaćamo objave koje imaju oko 52 % vjerojatnosti da sadrže neprimjeren govor i kako je neprimjeren govor doista zastupljen prema svim ciljnim skupinama u različitim udjelima.

Tablica 6. Količina neprimjerena govora prema pojedinim skupinama

Table 6 Amount of inappropriate speech targeted at specific groups

podskup	br. objava	Albanci	Bosanci	Hercegovci	Hrvati	Mađari	Romi	Slovenci	Srbi	Talijani	drugi	nitko
w0s0	5190	8	6		62		4		204		25	4541
w0s1	5049	1	3		47				89		12	4898
w0s2	7375	3	35	1	432	3	2	8	1419	2	521	1600
w1s0	1985	87	30	10	43	11	7	118	187	4	8	1578
w1s1	1494	35	4	2	12		12	11	130	1	5	1288
w1s2	4161	1033	160	24	401	8	20	475	1927	8	207	699
w2s0	2229	93	233	26	42	67	420	1	360	305	123	765
w2s1	2126	61	208	17	38	60	454		277	239	103	818
w2s2	4095	157	642	33	162	77	596	6	1215	287	325	1324
ukupno	33704	1478	1321	113	1239	226	1515	619	5808	846	1329	17511
	100,0 %	4,4 %	3,9 %	0,3 %	3,7 %	0,7 %	4,5 %	1,8 %	17,2 %	2,5 %	3,9 %	52,0 %

Za daljnje zaključke potrebno je više eksperimenata. Udio neprimjerena govora prema pojedinim skupinama iz ove tablice nije moguće generalizirati na cijeli medijski prostor društvenih mreža zbog provedene metodologije. Naime, nisu obuhvaćene sve objave s neprimjerenim govorom (npr. one koje ne sadrže odabrane ključne riječi) i potencijalno su omjeri drugačiji (primjerice za skupinu „Srbi“ korištene su neutralne i frekventne riječi, dok je za druge skupine dohvat proveden samo s negativnim riječima). Također, riječi većinske etničke skupine („Hrvati“) nisu bile uvjet

filtriranja originalnog korpusa pa je očekivano mnogo neprimjerena govora prema toj skupini koji ovdje nije obuhvaćen.

Neophodno je napomenuti da je u promatranom periodu bilo mnogo nacionalno i etnički značajnih obljetnica te da bi u nekom drugom periodu rezultati istraživanja mogli biti značajno različiti. U tom kontekstu u idućem bi se istraživanju trebale uzeti u obzir objave s društvenih mreža i to cijelu godinu dana kako bi se izbjegli specifični periodi u godini, a posebno važni za pojedinu ciljnu skupinu govora mržnje. Za točnu kvantifikaciju količine govora mržnje prema pojedinim skupinama bilo bi nužno obraditi veći dio izvornog korpusa koji sadrži i mnoge objave izostavljene ovim eksperimentom. To bi se moglo učiniti povećanjem skupa ključnih riječi i fraza te uključivanjem *apriori* neutralnih ključnih riječi za svaku skupinu. S druge strane, pomoću metoda strojnog učenja bilo bi moguće izgraditi model za prepoznavanje neprimjerena govora koji bi dvojbene objave za ljudsku odluku učinkovito automatski dohvaćao iz većeg skupa.

4.2. Količina iznimno neprimjerenih objava

Prilikom označavanja neprimjerena govora, analitičari su dodjeljivali oznake koje su imale značenje „neprimjeren govor“ bez obzira na njegovu gradaciju. Korisno je uvođenje više razina govora mržnje, primjerice Evkovski (2021) navodi mogućnost gradacije između primjerena (0), neprihvatljiva (1), uvredljiva (2) i nasilna (3) govora u kontekstu govora mržnje. Dodatan zadatak prilikom označavanja 33 704 objave bio je izdvojiti i one zapise koji su bili iznimno neprimjereni prema subjektivnom sudu označivača. Takvih jako uvredljivih do nasilnih objava pronađeno je ukupno 131, a prema prethodnoj skali na razini su 2 i 3.

Tablica 7. Količina objava iznimno neprimjerena govora ovisno o prediktorima {s0, s1, s2} i {w0, w1, w2}

Table 7 Amount of posts of extremely inappropriate speech depending on predictors {s0, s1, s2} and {w0, w1, w2}

	w0	w1	w2	P(x(s))
s0	4	1	2	5,3%
s1	9	0	1	7,6%
s2	54	27	33	87,0%
P(x(w))	51,1 %	21,4 %	27,5 %	100,0 %

Objave s iznimno neprimjerenim govorom svrstane su u Tablicu 7 prema prediktorima negativnosti riječi prema ciljnim skupinama {w0, w1, w2} i prema prediktorima automatskog negativnog sentimenta {s0, s1, s2}. Iz marginalne distribucije P(x(w)) čitamo kako prediktor prema negativnim riječima prema ciljnim skupinama nije dobar izbor za lakši pronalazak iznimno neprimjerena govora. To je i intuitivno jasno jer se iznimno neprimjeren govor ne ostvaruje samo kroz ciljane ključne riječi prema skupinama (bilo da su neutralne ili uvredljive). S druge strane, prediktor automatskog sentimenta značajno je bolji i može nam pomoći da lakše pronađemo iznimno neprimjeren govor ili govor mržnje.

Osim izdvajanja 131 objave iznimno neprimjerena govora prema ciljnim etničkim skupinama, za svaku pronađenu objavu napravljena je i sadržajna analiza prema varijablama koje se čitaju iz teksta objave, a ispisane su u Tablicama 8 do 11.

Tablica 8. Korištena strategija komentatora govora mržnje prema Čolović (2020) u izdvojenim objavama; jedan komentar može ostvariti jednu ili više navedenih kategorija

Table 8 The used strategy of hate speech commenters according to Čolović (2020) in separate posts; one comment can achieve one or more listed categories

Korištena strategija komentatora govora mržnje		
dehumanizacija	109	83 %
hiperseksualizacija	18	14 %
ilegalizacija	14	11 %
poziv na nasilje	10	8 %
dekulturalizacija	5	4 %

Prema Čolović (2020) postoje četiri strategije govora mržnje. Prvo, to je dehumanizacija koja se može odnositi na opredmećivanje, demonizaciju, animalizaciju i druge podstrategije koje nastoje maknuti ljudskost iz omražene skupine. Hiperseksualizacija je s druge strane kategorija koja se koristi kada se govori o seksualnom nasilju nad nekom skupinom ili o njihovoj reprodukciji i drugim seksualnim tabuima. Ilegalizacija je strategija koja se koristi kada je tvrdnja da su postupci i/ili postojanje neke skupine u nekom kontekstu ilegalno, ono može biti administrativnog i/ili kriminalno/terorističkog tipa. Dekulturalizacija je strategija koja se koristi kada se neku skupinu želi prikazati nekulturnom ili necivilizirano. Poziv na nasilje nije

jedna od strategija koju definira Čolović, no bila je neophodna za one slučajeve kada nije bilo drugih indikatora osim poziva na nasilje.

Tablica 9. Količina iznimno neprimjerenih objava s pozivom na nasilje

Table 9 The amount of extremely inappropriate posts calling for violence

Poziv na nasilje		
ne	71	54 %
da	60	46 %

Sve objave iznimno neprimjerena govora ostvarene su kao komentari na društvenim mrežama koji se u dretvi (engl. *thread*) vežu s drugim komentarima na seminalnu objavu, tj. podjelu sadržaja (engl. *share*). Sve seminalne objave načinjene su od strane profila medijskih kuća (tj. internetskih portala) i u manjoj mjeri od strane poznatijih profila nezavisnih medija (osobne stranice ili *podcasti*) i vežu se (engl. *link*) na originalni sadržaj. Analiza sadržaja obavljena je na tekstu članaka internetskih portala, dok su druge vrste objava na društvenim mrežama zbog raznolikosti formata i količine informacija kodirane kao „neprimjenjivo". Upravo analiza tog originalnog sadržaja daje nam uvid u vrstu sadržaja na koje komentatori najčešće reagiraju s iznimno neprimjerenim govorom. U nastavku se „originalni sadržaj" navodi kao „članak".

Tablica 10. Distribucija objava koje se sadržajno referiraju na sadržaj originalnog članka

Table 10 Distribution of publications that refer to the content of the original article

Referira li se komentar na članak?		
da	78	60 %
ne	35	27 %
neprimjenjivo	18	14 %

Tablica 11. Distribucija objava prema gruboj temi originalnog članka (svaki članak je u točno jednoj kategoriji).

Table 11 Distribution of posts according to the broad topic of the original article (each article is in exactly one category)

Gruba tema članka		
godišnjica nasilnog čina i/ili državne obljetnice	72	55 %
Sport	13	10 %
kriminalne radnje	10	8 %
politički komentari	7	5 %
turistički incident	4	3 %
Celebrity	2	2 %
rasprava o govoru mržnje	2	2 %
smrtni slučaj	2	2 %
Neprimjenjivo	19	15 %

Kategorija „godišnjica nasilnog čina i/ili državne obljetnice“ odnosi se velikim dijelom na članke koji govore o obljetnicama poput Oluje, Jasenovca i sličnih događaja koji imaju velik emotivan odjek u Hrvatskoj i regiji. Kategorije poput „sporta“, „kriminalnih radnji“, „političkih komentara“, „celebrity“, „rasprava o govoru mržnje“ i „smrtni slučajevi“ nećemo dublje objašnjavati. Kategorija „turistički incidenti“ odnosi se na članke koji govore o problemu dugotrajnog ostavljanja ručnika na plaži ili tjeranja ljudi s plaže. Kategorija „neprimjenjivo“ odnosi se na objave koje nisu članci s internetskih portala i stoga nisu analizirane.

Primjeri najgorih objava u promatranom periodu:

Naslov: Stipo Mlinarić Ćipe uništio SDSS, četnike i sve protivnike hrvatskih branitelja u 10 minuta (objava na društvenim mrežama – nije članak)

Komentar: SAMO BRIŠITE MOJ KOMENTAR ALI UZALUD !!! OPET ĆU PONOVITI : *Pa dajte jednom izvedite preostale Srbe pred s/t/r/e/l/j/a/č/k/i vod i u/n/i/š/t/i/t/e četnike jednom zauvek . I onako vam ne bi bilo prvi put da se na taj način obračunavate sa r/e/m/e/t/i/

l/a/č/k/i/m/ faktorom u Ante - Starčevićevskom stilu. I nakon toga, kada niti JEDAN Srbin ne bi ostao, vi bi i dalje kmečali o četnicima i uništavanju , o velikosrpskoj agresiji ....ZMS !!!*"

Naslov: PARTIZANI MU NIKAD NEĆE OPROSTITI: Potez Janeza Janše šokirao antifašiste

Komentar: ********, čišćenje Hrvatske od orjunista, srbo-komunista, marksista, lenjinista, titoista, atesita i sotonista, nije etničko čišćenje. Nego dezinfekcija hrvatskoga prostora od insekticida i zagađenosti hrvatskoga zraka od njihova jugo-vazduha, bre!

Naslov: Vučić danas glumi mirotvorca. Vučić 1995.: Nikad Glina i Banija neće biti Hrvatska

Komentar: Privremeni stanovnici privremene ndh2, konjojebci, pokatoliceni Srbi, odvratne ustase, olos ljudski.

Naslov: Vučić: U Oluji je ubijeno više tisuća Srba, a stotine tisuća su protjerane

Komentar: Malo,malo.

Iz navedenih primjera komentara vidljivo je da se razvijaju strategije izbjegavanja brisanja komentara kreativnom upotrebom znakova, kao i da postoje komentari koji samostalno ne izražavaju ikakvu emociju, no kada se stave u kontekst objave, dobivaju emotivan naboj (posljednji komentar odnosi se na naslov članka koji je podijeljen). U svrhu generalizacije ovih analiza, u sljedećim će istraživanjima biti potrebno prikupiti korpus duljeg razdoblja i s većim obuhvatom potencijalno neprimjerena govora.

Zaključak

Temeljem podataka iznesenih u prethodnom poglavlju možemo zaključiti sljedeće:

- Prva hipoteza isticala je kako oslanjanje na negativne izraze povezane s određenom skupinom pridonosi pronalaženju neprimjerena govora ili u najtežem slučaju govora mržnje što je i potvrđeno u ovome istraživanju. Treba istaknuti kako udio neprimjerenih objava ovisi i o jasnom određivanju i označavanju indikatora (u ovom slučaju: w0, w1, w2). Iako je udio neprimjerena govora zabilježen prema određenim skupinama, rezultati pokazuju da je najviše neprimjerena govora uočeno za kategoriju „nitko" (52 %), slijede „Srbi" (17,2 %), „Romi" (4,5 %), „Albanci" (4,4 %) te „Bosanci" (3,9 %). No, zbog metodoloških razloga (izostavljanje svih objava s neprimjerenim govorom i izostavljanje kategorije većinskog naroda) potrebna su daljnja istraživanja i daljnje provjere.
- Rezultati istraživanja pokazali su da je u čak 83 % analiziranih slučajeva korištena strategija dehumanizacije (usp. Čolović, 2020) **čime je potvrđena druga hipoteza. Slijede komentari koji koriste strategije hiperseksualizacije (14 %) i ilegalizacije (11 %).**
- Nadalje, rezultati istraživanja pokazali su da većina analiziranih komentara (njih 54 %) ipak ne poziva na nasilje prema određenim društvenim skupinama, čime treća hipoteza nije potvrđena. Istodobno, uočeno je da je u 46 % analiziranih komentara poziv na nasilje prema određenim društvenim skupinama ipak prisutan, što otvara neka nova istraživačka pitanja za buduće radove.
- Neprimjereni komentari usmjereni prema određenim skupinama referiraju se na članak u 60 % analiziranih slučajeva, čime je potvrđena i **četvrta** hipoteza istraživanja. U svega 27 % analiziranih komentara to nije bio slučaj.
- Istraživanje je djelomično potvrdilo petu hipotezu da se teme članaka koje su generirale neprimjeren govor prema određenim skupinama odnose na godišnjice nasilnih događaja (poput Oluje i Jasenovca) koje imaju snažan emocionalan aspekt. Navedeno je potvrđeno u 55 % slučajeva u kojima je analizirana tema članka, dok su politički komentari, koji su problematizirali spomenutu tematiku, zastupljeni u svega 5 % analiziranih slučajeva, dakle manje od očekivanog.
- Posljednja, šesta hipoteza, potvrđena je budući da je istraživanje pokazalo da su jezične tehnologije korisne u pretrazi opsežnijih korpusa medijskih objava. Istodobno, treba istaknuti kako govor mržnje nije jednoznačan pojam prisutan u određenim oblicima verbalnih iskaza, već je riječ o višeslojnom fenomenu koji traži gradaciju (usp. Evkovski, 2021) u kojoj veliku ulogu i dalje ima čovjek. Analiza kategorija neprimjerena govora i njegovo razlikovanje prema prediktorima temelje se još uvijek na subjektivnoj procjeni označivača (analitičara).

Osim istaknutih nalaza koji se mogu izravno povezati s početnim hipotezama, rezultati istraživanja pokazali su nam i još neke indikatore:

1. Pomoću opisane metodologije filtriranja pomoću ključnih riječi može se obuhvatiti značajan dio neprimjerena govora, od kojih je manji broj objava iznimno neprimjerena govora.
2. Jezične tehnologije, poput automatske analize sentimenta, mogu značajno pomoći u prikupljanju korpusa za sadržajnu analizu ili za treniranje automatskog modela za detekciju govora mržnje.
3. Za prikupljanje sveobuhvatnog korpusa neprimjerena govora nužno je u više iteracija obogatiti rječnik negativnih riječi prema pojedinim skupinama koje su ciljevi govora mržnje.
4. Iznimno neprimjeren govor u pravilu se ostvaruje kao komentar na dijeljene članke (vijesti) od strane portala ili drugih mrežnih izvora pa je za izradu smjernica oko smanjenja količine govora mržnje i ublažavanja njegovih negativnih društvenih posljedica nužno analizirati i originalne medijske sadržaje koji na određeni način uzrokuju govor mržnje. U raspravu o govoru mržnje neophodno je uključiti sve dionike medijske okoline – od korisnika društvenih mreža, preko medijskih kuća do zakonodavaca.

Za uspjeh izrade i korištenja modela za automatsko prepoznavanje govora mržnje nužno je okupiti interdisciplinaran tim – medijske stručnjake, jezikoslovce (lingviste), podatkovne znanstvenike i sociologe. Buduća istraživanja trebaju uključiti povećanje ključnih riječi analizom postojećega korpusa, kao i usmjerenost na pronalazak specifičnih fraza. Također, potrebno je treniranje modela za automatsko prepoznavanje govora mržnje i označavanje na širem korpusu s obzirom na ciljne skupine.

BILJEŠKE

[1] Ne uzima se u obzir složenost etničkog/nacionalnog identiteta, već se promatraju kategorije realno ostvarene kroz pisani tekst.

LITERATURA I IZVORI

Čolović N. (2020) "Strategije govora mržnje", rad u postupku objave.

Bossetta, M. (2020) "Scandalous Design: How Social Media Platform's Responses to Scandal Impacts Campaigns and Election", *Social Media & Society*, 1–4.

Brax, D. (2016) "Hate Speech and the Distribution of the Costs and Benefits of Freedom of Speech", 185–193. U: M. Edström, A. T. Kenyon i E.-M. Svensson (ur.): *Blurring the Lines Market-Driven and Democracy-Driven Freedom of Expression.* Gothenburg: Nordicom.

Car, V. (2016) "Moć medija: između slobode izražavanja i govora mržnje?", 187–217. U: E. Kulenović (ur.): *Govor mržnje u Hrvatskoj*. Zagreb: Biblioteka Političke analize.

Carlsson, U. (2016) "Opening speech: Freedom of Expression in Transition. A Media Perspective", 19–29. U: M. Edström, A. T. Kenyon i E.-M. Svensson (ur.): *Blurring the Lines Market-Driven and Democracy-Driven Freedom of Expression*. Gothenburg: Nordicom.

Đukić, M. i D. Sić (2021) "Regulation of fake news and hate speech on social networks", 51–73. U: M. Đukić (ur.): *4th International Science Conference „European Realities – Movements", Conference Proceedings 2019*. Osijek: Akademija za umjetnost i kulturu u Osijeku, Sveučilište Josipa Jurja Strossmayera u Osijeku.

Edström, M. (2016) "Audience Advertising Fatigue and New Alliances to Finance Content in Broadcasting", 131–141. U: M. Edström, A. T. Kenyon i E.-M. Svensson (ur.): *Blurring the Lines Market-Driven and Democracy-Driven Freedom of Expression*. Gothenburg: Nordicom.

Elkin-Koren, N. (2020) "Contesting Algorithms: Restoring the public interest in content filtering by artificial intelligence", *Big Data & Society*, 1–13.

Evkoski, B., Pelicon, A., Možetić, I., Ljubešić, N. i P. Kralj Novak (2021) "Retweet communities reveal the main sources of hate speech", *Computer Science > Social and Information Networks,* arXiv:2105.14898 [cs.SI], pristupljeno 3. kolovoza 2021.

Gardašević, Đ. (2016) "Govor mržnje i hrvatski ustavnopravni okvir", 151–187. U: E. Kulenović (ur.): *Govor mržnje u Hrvatskoj*. Zagreb: Biblioteka Političke analize.

Gibson, A. (2019) "Free Speech and Safe Spaces: How Moderation Policies Shape Online Discussion Spaces", *Social Media + Society*, 1–15.

Gillespie, T. (2020) "Content moderation, AI, and the question of scale", *Big Data & Society,* (2000), 1–5.

Gorwa, R., Binns R. i K. Katzenbach (2020) "Algorithmic content moderation: Techical and political challenges in the automation of platform governance", *Big Data & Society*, 1–15.

Grmuša, T. (2021) "Učinci digitalne transformacije na inovativnost medijskih organizacija i novi poslovni modeli", 71–87. U: J. Jurišić i Z. Hrnjić Kuduzović (ur.): *Zbornik radova 10. regionalne znanstvene konferencije Vjerodostojnost medija „Vjerodostojnost medija: medijska agenda 2020. – 2030."*. Zagreb: Fakultet političkih znanosti Sveučilišta u Zagrebu, Hanns-Seidel-Stiftung,

Grmuša, T. i L. Prelog (2020) "Uloga novih tehnologija u borbi protiv lažnih vijesti – iskustva i izazovi hrvatskih medijskih organizacija", *Medijske studije*, 11 (22), 62–80.

Haapoja, J., Laakosonen, S. M. i A. Lampinen (2020) "Gaming Algorithmic Hate-Speech Detection: Stakes, Parties and Moves", *Social Media Society*, 1–10.

Hasebrink, U. (2012) "Uloga publika u upravljanju medijima: zapostavljena dimenzija medijske pismenosti", *Medijske studije*, 3 (6), 58–72.

Hromadžić, H. (2013) "Politika, društvo spektakla i medijska konstrukcija realnosti", *Politička misao: časopis za politologiju*, 50 (2), 60–74.

Kalsnes, B. i K. A. Ihlebaek (2021) "Hiding hate speech: political moderation on Facebook", *Media, Culture & Society*, 43 (2), 326–342.

Kenyon, A. T. (2016) "Who, What, Why and How. Questions for Positive Free Speech and Media Systems", 29–41. U: M. Edström, A. T. Kenyon i E.-M. Svensson (ur.): *Blurring the Lines Market-Driven and Democracy-Driven Freedom of Expression.* Gothenburg: Nordicom.

Klepač Pogrmilović, B. (2016) "Govor mržnje i politička korektnost – Hrvatski sabor kao "slika društva"", 251–307. U: E. Kulenović (ur.): *Govor mržnje u Hrvatskoj*. Zagreb: Biblioteka Političke analize.

Kocijan, K., Košković, L. i P. Bajac (2019) "Detecting hate speech online: A case of Croatian", 185–197. U: *International Conference on Automatic Processing of Natural-Language Electronic Texts with NooJ*.

Labaš, D. i T. Grmuša (2011) "Istinitost i objektivnost u informaciji i društveno štetne komunikacijske forme", *Kroatologija: časopis za hrvatsku kulturu*, 2 (2), 87–121.

Ljubešić, N., Erjavec, T. i D. Fišer (2018) "Datasets of Slovene and Croatian moderated news comments", 124–131. U: *Proceedings of the 2nd Workshop on Abusive Language Online* (ALW2).

Ljubešić, N. i F. Klubička (2014) "bs,hr,srWaC - Web Corpora of Bosnian, Croatian and Serbian", 29–35. U: *Proceedings of the 9th Web as Corpus Workshop* (WaC-9). Association for Computational Linguistics.

Markov, I., Ljubešić, N., Fišer, D. M. i W. Daelemans (2021) "Exploring stylometric and Emotion-Based Features for Multilingual Cross-Domain Hate Speech", *Sentiment and Social Media Analysis,* WASSAP, 149–159.

Matamoros-Fernandez, F. (2021) "Racism, Hate Speech and Social Media: A Systematic Review and Critique", *Televsion & New Media*, 22 (2), 205–224.

McGonagle, T. (2012) "A Survey and Critical Analysis of Council of Europe Strategies of Countering "Hate Speech"", 456–498. U: M. Herz i P. Molnar (ur.): *The Content and Context of Hate Speech – Rethinking Regulation and Responses.* New York: Cambridge University Press.

Newman, N. (2018) *Journalism, Media and Technology Trends and Predictions 2018*. Oxford Institute: Reuters.

Ostrički, I. (2017) "Medijski tekst kao pokretač digitalnih gomila", *In medias res: časopis filozofije medija*, 6 (10), 1601–1627.

Peruško, Z. (2021) "Croatia", 70–71. U: N. Newman s R. Fletcher, A. Schulz, S. Andı, C. T. Robertson i R. K. Nielsen (ur.): *Digital News Report, 10th Edition.* Reuters Institute: Univeristy of Oxford.

Poljak, M., Hadžić, J. i M. Martinić (2020) "Govor mržnje u hrvatskom medijskom prostoru", *In medias res: časopis filozofije medija*, 9 (17), 2709–2444.

Svensson, E. M. (2016) "Upholding the Division Between Editorial and Commercial Content in Legislation and Self-Regulation", 109–121. U: M. Edström, A. T. Kenyon i E.-M. Svensson (ur.): *Blurring the Lines Market-Driven and Democracy-Driven Freedom of Expression*. Gothenburg: Nordicom.

Tontodimamma, A., Nissi, E., Sarra, A. i L. Fontanella (2021) "Thirty years of research into hate speech - topics of interest and their evolution", *Scientometrics*, 126, 157–179.

Tkalac Verčič, A., Sinčić Ćorić, D. i N. Pološki Vokić (2011) *Priručnik za metodologiju istraživačkog rada; Kako osmisliti, provesti i opisati znanstveno istraživanje*, 2. izdanje. Zagreb: M.E.P.

Ullmann, S. i M. Tomalin (2019) „Quarantining online hate speech - technical and ethical perspectives", *Ethics and Information Technology*, 22, 69–80.

Uyheng J., Carley, K. M. (2021) "Characterizing network dynamics of online hate communities around the Covid 19 pandemic", *Applied Network Science,* 6 (20), 1–21.

Van Dijck, J. i D. Alinejad (2020) "Social Media and Trust in Scientific Expertise: Debating the Covid-19 Pandemic in The Netherlands", *Social Media + Society,* 1–11.

Vijavaraghavan, P., Larochelle, H. i D. K. Roy (2021) "Interpretable Multi-Modal Hate Speech Detection". U: H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan i H. Lin (ur.): *Advances in Neural Information Processing Systems* 33 (NeurIPS 2020).

Vilović, G. (2003) "Istraživačko novinarstvo, tabloidizacija i etika", *Društvena istraživanja* 12 (2003), 6 (68), 957–974.

Yin, W. i A. Zubiaga (2021) "Towards generalisable hate speech detection: a review on obstacles and solutions", *PeerJ Comput Sci*. 7, e598.

Wimmer, Roger D. i J. R. Dominick (2011) *Mass Media Research: An Introduction*, 9. izdanje. WADSWORTH CENGAGE Learning.

Building a Language Corpus of Hate Speech in the Croatian Media Space of Social Networks

Slobodan Hadžić
Artur Šilić
Tanja Grmuša

ABSTRACT

Hate speech is an unacceptable form of socially harmful communication that has recently increased in prevalence. The development of digital media, and in particular the strengthening of the role of social networks in private and public communication, has opened the space for numerous public virtual forums that have encouraged passive audiences to participate and communicate more actively. Freedom of opinion and expression as fundamental democratic principles on the one hand are juxtaposed with the space for toxic communication on the other, but also with the potential for radicalization of certain social groups of like-minded individuals who congregate in virtual communities where they often select contributors, journalists, editors, media and other users as targets for future attacks. Given the boundlessness of Internet space, whose content is difficult to control, the question is how to detect socially harmful forms of communication in the public sphere, whether they can be prevented, and how to retain existing audiences. The potential uses extend into the area of moderating inappropriate user comments by relying on software that provides instant responses to media organizations, but also has a constant need for self-improvement. The author's interest in this work relates to detecting hate speech against ethnic groups on social networks and exploring the possibility of using speech technologies to detect and prevent its spread. Using quantitative and qualitative content analysis in conjunction with a set of software solutions built on speech technologies, we enable efficient automatic and semi-automatic analysis of user-generated content. We use: the WordFinder program for fast word retrieval in large corpora, the Crontiment tool for automatic sentiment analysis of texts in Croatian, and the Text Marker application for efficient manual albeling and building of corpora. The authors use a longitudinal study to identify

changes in the types and frequencies of hate speech in this topic and identify the main advantages and disadvantages of language technologies, suggesting possible directions for development.

Keywords: socially harmful communication forms, hate speech, digital media, audience, natural language processing, user comments.