PD.TVZ.HR
POLYTECHNIC & DESIGN Vol. 13, No. 1, 2025.
Mrežna inačica: ISSN 2459-6302
Tiskana inačica: ISSN 1849-1995
DOI: 10.19279/TVZ.PD.2024-13-1-01

KLJUČNA OBILJEŽJA OTVORENIH SKUPOVA PODATAKA U ANALIZI SENTIMENTA OBJAVA NA TWITTERU

KEY FEATURES OF OPEN DATA SETS IN SENTIMENT ANALYSIS OF TWITTER POSTS

Gaurish Thakkar

Filozofski fakultet Sveučilišta u Zagrebu, Ivana Lučića 3, 10000, Zagreb

SAŽETAK

Otvoreni skupovi podataka ključni su za usavršavanje modela analize sentimenta, no njihovu praktičnu iskoristivost često otežava nedostatak standardizacije i sveobuhvatne dokumentacije. Ovaj rad pruža kritički pregled otvorenih skupova podataka za analizu sentimenta objava na Twitteru, analizirajući 48 skupova podataka za 30 različitih jezika. Analizirani su ključni elementi, uključujući konvencije o imenovanju, sheme označavanja, metode distribucije podataka i uključivanje bitnih metapodataka poput ID-ova tweetova. Rezultati pokazuju značajne nedosljednosti koje stvaraju izazove za reproducibilnost i komparativnu evaluaciju modela. Identificirana je kritična potreba za standardnim praksama u stvaranju i diseminaciji podatakovnih skupova. Temeljem ove analize ponuđene su konkretne preporuke za unaprjeđenje znanstvene vrijednosti, mogućnosti otkrivanja i dugoročne iskoristivosti otvorenih podatkovnih skupova za znanstvenu zajednicu.

Ključne riječi: *analiza sentimenta, obrada prirodnog jezika, skupovi podataka za analizu sentimenta, multijezičan*

ABSTRACT

Open-source datasets are fundamental to the advancement of sentiment analysis models, yet their practical utility is often hampered by a lack of standardisation and comprehensive documentation. This paper provides a critical review of the open dataset landscape for Twitter sentiment analysis, examining 48 papers that introduce datasets in 30 different languages. We analyse key elements, including naming conventions, labelling schemes, data distribution methods, and the inclusion of essential metadata such as tweet IDs. Our findings reveal significant inconsistencies that create challenges for reproducibility and the comparative evaluation of models. We identify a critical need for standard practices in dataset creation and dissemination. Based on this analysis, we offer concrete recommendations to enhance the scientific value, discoverability, and long-term usability of open datasets for the research community.

Keywords: *sentiment analysis, natural language processing, twitter, sentiment datasets, multilingual*

1. UVOD
1. INTRODUCTION

S obzirom na nedavni napredak u metodologijama dubokog učenja, značaj podatkovnih skupova postao je očigledan. Javno dostupan podatkovni skup koristi se za daljnja istraživanja u određenom području, a njegova vrijednost raste s povećanom upotrebom tijekom faze istraživanja. Takav se ishod može postići samo ako je znanstvena zajednica učinkovita u pronalasku odgovarajućih podataka za svoju istraživačku temu. Podatkovni skup bitna je komponenta algoritama strojnog učenja, a njegova dostupnost ne podrazumijeva nužno i jednostavnost korištenja [1]. Iako ove skupove možemo prenamijeniti, njihova široka dostupnost u javnim ili institucionalnim arhivima može ometati proces pronalaženja

podataka. Objavljene smjernice i okviri za javnu diseminaciju podatkovnih skupova, kao što su podatkovni listovi [2] i podatkovne kartice [3] služe kao sveobuhvatni zapisi, pružajući sustavan pregled vidljivih i nevidljivih aspekata podatkovnih skupova. Iako predložene metodologije prioritet daju pristupačnosti informacijskog konteksta skupova koji utječe na podatke, nedostaju informacije o trenutnom statusu podatkovnih skupova, dostupnim informacijama o njima, njihovom potencijalu za ponovnu upotrebu i drugim povezanim aspektima.

Opseg dostupnih podatkovnih skupova za različite zadatke je ogroman. Stoga je, kako bi se suzilo problematično područje na jedan zadatak, odabran Twitter skup kao primarni izbor, pri čemu napominjemo da je Twitter 2023. godine promijenio naziv u X. Primarni fokus ovog istraživanja je na zadatku analize sentimenta koji se razvio u značajno istraživačko područje zbog analize objava na Twitteru [4, 5].

Istraživanje se prvenstveno fokusira na podatkovne skupove s Twittera iz sljedećih razloga:

- Propisi na Twitteru zabranjuju javno objavljivanje sadržaja tweeta. Jedinstveni identifikatori tweetova (ID-ovi) i oznake mogu se objaviti.
- Ponovna upotreba podatkovnih skupova s Twittera zahtijeva "hidriranje" ID-ova tweetova što je važan aspekt za naknadne studije. Radi se o procesu dohvaćanja svih podataka o tweetu (tekst, korisničke informacije, itd.) uz popis jedinstvenih identifikatora tweetova pomoću Twitter API-ja, gdje se taj popis popunjava stvarnim sadržajem tweetova.
- Tweetovi na Twitteru mogu sadržavati dodatne elemente poput glazbe i slika, te su dosljedno popraćeni implicitnim i vanjskim informacijama kao što su lajkovi, retweetovi i jezik teksta. Korištenjem već dostupne oznake ručno pridružene tweetu proces hidriranja može se ponoviti u svrhu provjere prisutnosti dodatnih modaliteta (slike, videozapise i audio) koje možemo dalje koristiti u modeliranju drugih zadataka [6].

Ova studija bavi se pregledom i analizom 48 istraživačkih radova o otvorenim skupovima podataka za analizu sentimenta i srodne zadatke na 30 jezika. Primarni cilj nije nabrajanje dostupnih resursa, nego istraživanje praktične iskoristivosti ovih skupova ispitivanjem kritičnih aspekata kao što su njihova dokumentacija, metode distribucije i uključivanje vitalnih metapodataka. Ključni su doprinosi ovoga rada sveobuhvatna kompilacija javno dostupnih podatakovnih skupova s Twittera za analizu sentimenta, kategoriziranih prema jeziku, veličini i zadatku. Temeljita studija slučaja detaljno će proučiti trenutno stanje distribucije i ponovne upotrebe podatkovnih skupova, naglašavajući tipične prepreke s kojima se znanstvenici susreću. Ponudit će se niz specifičnih prijedloga za unaprjeđenje znanstvene vrijednosti i trajne korisnosti javnih podatkovnih skupova, s naglaskom na važnosti ovih vrijednih resursa za istraživačku zajednicu.

Rad je strukturiran na sljedeći način: potpoglavlje 2 donosi kratak opis postupka prikupljanja podatakovnih skupova, potpoglavlje 3 daje njihovu analizu, a potpoglavlja 4 i 5 donose preporuke na temelju provedenog istraživanja.

2. PRIKUPLJANJE PODATKOVNIH SKUPOVA

2. COLLECTION OF DATASETS

U ovoj studiji korišten je skup specifičnih pojmova na engleskom jeziku, kao što su "Twitter", "sentiment analysis" (analiza sentimenta/mnijenja), "dataset" (podatkovni skup), "social media" (društveni mediji) i "shared task" (dijeljeni zadatak), za pretraživanje nekoliko platformi, uključujući GitHub, Google tražilicu, Clarin [7], podatkovne skupove s Googlea [8], European Language Grid [9], podatkovne skupove s HuggingFacea [10] i Zenodo. Prikupljeni su samo podatkovni skupovi povezani s Twitterom i zadatkom analize sentimenta. Prikupljeni skupovi uređeni su koristeći ID-ove tweetova za polja medij, tekst i datum objave za svaki podatkovni skup. Svi javno objavljeni skupovi bili su propisno citirani, dok su podatkovni skupovi koji su zahtijevali dopuštenje i registraciju prikupljeni slijedeći procedure i kontaktne kanale koje je osigurao tvorac pojedinog podatkovnog skupa.

Dodatno je filtrirana cijela zbirka prikupljenih podatkovnih skupova kako bi se isključilo one koji nisu povezani s Twitterom. Na primjer, isključene su recenzije filmova i skupovi povezani s Redditom.

Prikupljeni su sljedeći podatci za svaki navedeni skup:

- Naziv podatkovnog skupa, izvor i godina objave: naziv je identificiran ako je bio naveden.
- Informacije o jeziku: jezik korišten u podatkovnom skupu.
- Zabilježeni su svi eksplicitno spomenuti zadaci u podatkovnom skupu.
- Izračunat je broj instanci podataka. Ako ukupan broj nije bio izravno prijavljen, agregirane su podjele na skupove za uvježbavanje i testiranje.
- Shema označavanja instanci podataka: dokumentirane su različite klase prisutne u skupu
- Provjera prisutnosti ID-ova tweetova.

Tablica 1 prikazuje sveobuhvatan pregled prethodno spomenutih informacija. Nadalje, ručno je analizirana dokumentacija povezana s pojedinim podatkovnim skupom i znanstvenim radovima koji ga opisuju da bismo stekli sveobuhvatnije razumijevanje sastava skupa. Dodatno su evidentirane distribucijske i pristupne procedure povezane s podatkovnim skupom.

Tablica 1 *Skupovi podataka s Twittera.*

: Dostupni su ID-ovi tweetova. : ID-ovi tweetova nisu dostupni. Brojke u stupcu Veličina su približna procjena. S: sentiment, E: emocija, I: informacija, Sarc: sarkazam, Fig: figurativni jezik (ironija, sarkazam i metafora), ABSA: aspektno-utemeljena analiza sentimenta, LT: naslijeđeni zadaci i skupovi iz prethodnih natjecanja uključeni su u trenutnu verziju. ALI: automatski označene instance - alati i klasifikatori automatski označavaju ove instance podataka. U zadacima TASS klase NEU i NONE tretiraju se kao jedna klasa, ali su označene zasebno.

Jezik	Podatakovni skup	Veličina	Klase	Zapažanja
ar	ASTD [11]	10k	3	S
ar	ArSarcasm-v2 [12]	15.5k	3	Sarc S
ar	SemEval 2017 task 4 [13]	23k	2,5	S
ar	TM-Senti@ar 2021 [14]	16.0M	2	S
bg	Twitter-15@bg [15]	31k	3	S
bn	TR-SLSAMEP 2018 [16]	137k	3	S
bs	Twitter-15@bs	23.4k	3	S
da	AngryTweets 2021 [17]	4.4k	3	S
de	xLiMe@German 2016 [18]	3.4k	3	S
de	Twitter-15@de	76.5k	3	S
de	TM-Senti@de	2.8M	2	S
de	PotTS 2016 [19]	7.9k	3	S
de	SB10k 2017 [20]	10k	3/5	S
de	Germeval Task 2017 [21]	22k	3	S
en	TM-Senti@en	55.7M	2	S
en	SemEval 2013 task 2 [22]	15k	3	S
en	SemEval 2015 task 10 [23]	20k/19.5k/2.9k	3	S
en	SemEval 2015 task 11 [24]	12k	4	Fig
en	SemEval 2016 [25]	132k	2,5	S
en	CB COLING 2014 [26]	12k	3	S
en	CB IJCOL 2015 [27]	80k	3	S ALI
en	RETWEET 2021 [28]	34.9k+1.5k	3	S ALI
en	Twitter-15@en	48.5k	3	S
en	SMILE 2016 [29]	3k	5	E

es	xLiMe@spanish	7.6k	3	S
es	COpres14 2016 [30]	1k	3	S
es	MAVIS 2022 [31]	383.5k+760	2,5	S ALI
es	Twitter-15@es	176.6k	3	S
es	JOSA 2021 [32]	3.4k	3	S
es	TASS 2012 [33]	70k	3,5	S
es	TASS 2013 [34]	68k+2.5k	3,5	S
es	TASS 2014 [35]	68k+2.7k	3,4,6	S TOPIC ABSA
es	TASS 2015 [36]	1.2k	3	LT ABSA
es	TASS 2018 [37]	8.6k	4	LT E
es	TASS 2019 [38]	14k	4	LT
es	TASS 2020 [39]	8.4k	4	LT
es	Corpus Ironic tweets 2016 [40]	690k	2	Ironija
es	HUMOR [41]	27k	5	Humor/funniness
fr	DEFT 2015 task 1 [42]	7.9k	3	S
fr	DEFT 2015 task 2.1	6.7k	4	I S E
fr	DEFT 2015 task 2.2	3.1k	18	I S E
fr	TM-Senti@fr	8.2M	2	S
fr	DEFT 2017 [43]	5.1k	2	FIG
hi	TR-SLSAMEP 2018	396k	3	S
hr	Cro-CoV-Tweets 2021 [44]	10k	3	S
hr	Twitter-15@hr	75.1k	3	S
hu	Twitter-15@hu	48.9k	3	S
id	EDIPO 2022 [45]	7k	6	E
it	CB IJCOL2015 ITA castellucci	55.8k	3	S
it	xLiMe@it	8.6k	3	S
it	sentipolc@Evalita 2016 [46]	7.4k	multidimenzionalan	
it	TM-Senti@it	2.46M	2	S
it	Sardistance 2020 [47]	3.2k	3	Stav
lv	Peisenieks 2014 [48]	1178	3	S
mr	L3CubeMahaSent 2021 [49]	18k	3	S
mt	Malta-Budget 2021 [50]	6.3k	multidimenzionalan	
pl	Twitter-15@pl	144.6k	3	S
pt	Twitter-15@Portuguese	52.8k	3	S
pt	Brazilian tweets 2022 [51]	5.8k	3	S
ru	Twitter-15@ru	48.7k	3	S
sk	Twitter-15@sk	176.6k	3	S
sl	Twitter-15@sl	92k	3	S
sr	doiserbian@tweet 2018 [52]	1.3k	3	S
sr	Twitter-15@sr	38.2k	3	S
su	STDEC 2020 [53]	2.5k	4	E
sq	Twitter-15@sq	42.4k	3	S
sv	Twitter-15@sv	38.3k	3	S
te	TR-SLSAMEP 2018	178k	3	S
tr	BounTi 2022 [54]	7964	3	S
tr	SentimentSet 2022 [55]	2.6k	3	S
ur	Urdu Sentiment Corpus 2020 [56]	10k	3	S
zh	TM-Senti@zh	205.6k	2	S

3. ANALIZA OTVORENIH PODATKOVNIH SKUPOVA
3. ANALYSIS OF OPEN DATASETS

Prikupljeno je ukupno 48 znanstvenih radova koji se bave podatkovnim skupovima na 30 različitih jezika. Te publikacije uključuju nekoliko podjela za različite zadatke i jezike. Svaki zadatak i jezik evidentirani su kao zasebni unosi budući da su imali različite sustave označavanja i distribucije. Slijedi pregled opažanja proizašlih iz ručne analize podatkovnih skupova.

3.1. JEZIK PODATKOVNOG SKUPOVA
3.1. LANGUAGE OF THE DATASETS

Skupovi analizirani u studiji obuhvaćaju znatan broj jezika, od široko rasprostranjenih do jezika s malo digitalnih resursa. Odabrani skupovi uključuju jezike poput arapskog, bengalskog, bugarskog, bosanskog, hrvatskog, danskog, nizozemskog, engleskog, francuskog, njemačkog, hindskog, mađarskog, indonezijskog, talijanskog, latvijskog, marathi, malteškog, poljskog, portugalskog, ruskog, slovačkog, slovenskog, srpskog, sundskog, albanskog, švedskog, telugu, turskog, urdu i kineskog. Svi skupovi jasno su naveli jezik tweetova. U nekim slučajevima podatkovni skupovi sadržavali su podjele na različite jezike ili dijalekte.

3.2. NAZIV PODATKOVNOG SKUPOVA
3.2. NAME OF THE DATASETS

Značajan problem je što mnogim skupovima nedostaje jedinstveno ime ili identifikator. Znanstvenici često objavljuju skupove kao dopunski materijal svojih publikacija, usredotočujući se prvenstveno na aspekte modeliranja, a ne na sam podatkovni skup. Stoga skupovi ne posjeduju naziv koji može funkcionirati kao jedinstveni identifikator, a to za naknadne studije otežava identificiranje, referenciranje i uspostavu veze s podacima.

3.3. ZADACI I SHEMA OZNAČAVANJA
3.3. TASKS AND LABELING SCHEME

Zadaci povezani s podatkovnim skupovima uključuju klasifikaciju sentimenta, klasifikaciju emocija, klasifikaciju subjektivnosti, aspektno-utemeljenu analizu sentimenta, klasifikaciju objektivnosti naspram subjektivnosti, humor, figurativni jezik, ironiju, sarkazam, detekciju stava, negacije, detekciju jezika i klasifikaciju tema. Iako je primarni zadatak većine analiziranih skupova analiza sentimenta, specifična implementacija shema označavanja značajno varira, što predstavlja izazov za standardizaciju i unakrsnu usporedbu modela. Najčešći pristup je sustav s tri oznake: "pozitivan", "negativan" i "neutralan". Međutim, čak ni ova osnovna struktura nije univerzalna.

Naša analiza identificirala je brojne varijacije koje otežavaju izravne usporedbe:

- **Prošireni polaritet:** Neki skupovi koriste pet klasa za opis detaljnijeg sentimenta, dodajući kategorije "vrlo pozitivan" (pos+) i "vrlo negativan" (neg+) standardnim trima klasama.
- **Uključivanje klasa koje nisu obilježene sentimentom:** Mnogi skupovi uvode klase poput "neutralno", "objektivno", "nepoznato" ili "mješovito" za obilježavanje tweetova koji ne izražavaju jasan sentiment. Iako su vrijedne za izgradnju robusnih klasifikatora, definicija ovih klasa može se razlikovati između skupova. Na primjer, skupovi podataka TASS tretiraju 'NEU' (neutralno) i 'NONE' (nijedno) kao zasebne oznake, iako se ponekad spajaju.
- **Višedimenzionalne i višenamjenske oznake:** Sve veći broj skupova koristi višedimenzionalnu klasifikaciju [46, 50], istovremeno analizirajući jedan tweet prema više atributa. Na primjer, podatkovni skup može klasificirati tweet s obzirom na polaritet (pozitivan/negativan/neutralan) i ironiju (da/ne). Ovi višedimenzionalni klasifikacijski skupovi omogućuju nijansiranije i lakše razumijevanje sentimenta jer imaju više oznaka klasa za jednu instancu podataka. Skup podataka sentipolc@Evalita [46] analizira tweetove u više dimenzija, uključujući subjektivnost, polaritet, ironiju i emociju. Ovaj višestruki pristup pruža bogatije podatke, ali otežava izravnu usporedbu s jednostavnijim, jednodimenzionalnim skupovima podataka.
- **Oznake specifične za zadatak:** Skupovi podataka stvoreni za srodne, ali različite

zadatke, poput detekcije emocija, detekcije stava ili analize humora, koriste potpuno različite skupove oznaka poput Plutchikovih [57] osam osnovnih emocija ili kategorija stava specifičnih za projekt [58].

Ova fragmentacija u metodologijama označavanja znači da primjena velikih jezičnih modela na jednom skupu podataka nije izravno usporediva s primjenom na drugom. To ukazuje na nužnost da tvorci skupova podataka osiguraju dosljednu i opsežnu dokumentaciju o svojim kriterijima označavanja i, gdje je moguće, preslikaju svoju shemu u ustaljenije klasifikacijske sustave.

Tablica 2 *Nazivi zadataka, broj oznaka povezanih sa svakim zadatkom i definicije tih oznaka: "pos" za pozitivno, "neg" za negativno, "neu" za neutralno,"+" za vrlo, "obj" za objektivno i "sub" za subjektivno.*

Zadaci	Broj oznaka	Oznake klasa
figurativni jezik	2	da/ne
sentiment	2	pos/neg
sentiment	3	pos/neg/neu
stance	3	za/protiv/neutralno/ nijedno
sentiment	4	pos/neg/neu/nijedno
sentiment	4	pos/neg/neu/obj
figurativni jezik	4	sarkazam/ironija/ metafora/ostalo
emocija	4	radost/ljutnja/strah/ tuga
sentiment	5	pos/neg/neu/ mješovito/nepoznato
sentiment	5	pos+/pos/neu/neg/ neg+
sentiment	6	pos+/pos/neu/neg/ neg+/none
multidimenzionalan	2+3+2	[sub/obj],[pos/ neg/neu or mješovito],[ironija]
multidimenzionalan	2+3+8+2+ 2+2+2+3	[subjektivnost] [polaritet][emocija] [sarkazam][ironija] [negacija][izvan teme][jezik]
emocija	8	radost/tuga/strah/ ljutnja/iščekivanje/ iznenađenje/ gađenje/povjerenje
emocija	18	emocije

Tablica 2 pruža sveobuhvatan sažetak zadataka, zajedno s odgovarajućom metodom označavanja i oznakama za podatkovne skupove.

3.4. ID TWEETA
3.4. TWEET ID

Ključni element za ponovnu iskoristivost i proširivost podatkovnog skupa s Twittera je uključivanje ID-ova tweeta. Oni su poveznica s metapodacima i drugim kontekstualnim informacijama. Od analiziranih primjera, 10 ih ne nudi ID tweeta. Umjesto toga distribuira ga kao kombinaciju tekstova i oznaka.

Prema [59], jedan od zadataka povezanih s analizom sentimenta uključuje izdvajanje informacije o vremenu objave sadržaja.

Naša analiza pokazala je da, iako većina analiziranih skupova pruža ID-ove tweetova, značajna manjina (10 instanci) distribuira podatke kao predobrađeni tekst i oznake, čime se prekida veza s izvorom. Ovaj propust ozbiljno ograničava potencijal za buduća istraživanja. Pomoću ID-a tweeta, moguće je programatski dohvatiti:

- **Vremenske informacije:** Točan datum i vrijeme objave tweeta, što je kritičan čimbenik u analizi sentimenta budući da se javno mišljenje može brzo mijenjati.
- **Metrike korisničke interakcije:** Podatke o lajkovima i retweetovima koji mogu služiti kao implicitni signali sentimenta ili utjecaja.
- **Multimodalni sadržaj:** Mogućnost provjere i dohvaćanja povezanih medija poput slika ili videozapisa koji su sve važniji za multimodalnu analizu sentimenta.
- **Geoprostorne podatke:** Informacije o lokaciji, kada ih korisnik pruži, omogućuju regionalne ili lokalizirane studije sentimenta.

Bez ID-a tweeta skup podataka postaje statičan artefakt, ograničen samo na tekst. Ne može se lako ažurirati, provjeriti ili obogatiti novim slojevima informacija kako se pojavljuju nova istraživačka pitanja. Stoga uključivanje ID-ova tweetova treba smatrati obveznom praksom za svaki javno objavljeni podatkovni skup s Twittera.

3.5. ALI: AUTOMATSKI OZNAČENE INSTANCE
3.5. ALI: AUTOMATICALLY LABELLED INSTANCES

Tri skupa podataka [27, 28, 31] u zbirci bila su popraćena instancama koje su označene vanjskim metodama, a ne ručno. Podatkovni skup MAVIS [31] autori su ručno označili binarnim oznakama, ali također sadrži automatiziranu kategorizaciju. Za provođenje automatizirane kategorizacije korištene su tri različite tehnologije, IBM Watson, Google Cloud Natural Language i Meaning Cloud.

Udaljena supervizija uključuje stvaranje kompilacije ključnih riječi (kao što su emotikoni, smajlići i hashtagovi) koje se mogu koristiti kao neovisne oznake ili kao kombinirani prikazi za naprednije oznake. Ove se ključne riječi zatim koriste za automatsko anotiranje određenog skupa. Podatkovni skup TM-Senti [14] stvoren je pomoću tehnike udaljene supervizije, no autori priznaju da ovaj pristup ima ograničenja i da nije potpuno učinkovit.

Značajan trend u stvaranju velikih podatkovnih skupova je korištenje automatiziranih ili poluautomatiziranih tehnika označavanja za dopunu ili zamjenu ručnih anotacija.

Ovo je istraživanje identificiralo nekoliko skupova koji uključuju automatski označene instance (ALI), odnosno instance označene vanjskim alatima ili klasifikatorima, a ne ručno. To su, na primjer, skup MAVIS koji je ručno označen binarnim oznakama, ali distribuiran s automatiziranim kategorijama. Drugi primjer je TM-Senti koji je nastao udaljenom supervizijom, gdje se heuristike poput emotikona, smajlića ili specifičnih hashtagova koriste kao oznake za automatsko anotiranje velikih količina tweetova s milijunima instanci na nekoliko jezika.

Iako ove tehnike omogućuju stvaranje podatkovnih skupova u razmjeru koji bi bio previše skup za ručnu anotaciju, one imaju značajna ograničenja. Prisutnost strojno generiranih i potencijalno nečistih podataka može imati dubok utjecaj na obuku i evaluaciju modela. Stoga je bitno da znanstvenici mogu jasno razlikovati ručno anotirane od automatski označenih instanci. Ovo razlikovanje temelj je preporuke koju donosi ovaj rad, a to je usvajanje jasnog standarda imenovanja kao što je prefiks "auto-" za označavanje skupova podataka generiranih bez izravne ljudske anotacije kako bi znanstvenici bili svjesni podrijetla podataka.

3.6. DISTRIBUCIJA PODATKOVNOG SKUPA
3.6. DATASET DISTRIBUTION

Pristupačnost i metode distribucije skupa izravno utječu na njegovo korištenje i ponovnu upotrebu unutar istraživačke zajednice. Ovo je istraživanje identificiralo trodijelnu kategorizaciju dostupnosti podataka koja predstavlja niz izazova za istraživače. Prva kategorija, koja nije uključena u istraživanje, zatvoreni su podatkovni skupovi koji nisu javno dostupni i stoga ne doprinose širem napretku zajednice. Druga i najpristupačnija kategorija otvoreni su skupovi koji su slobodno i javno distribuirani. Često se nalaze na sveučilišnim mrežnim stranicama, u repozitorijima zajedničkih zadataka ili na platformama poput GitHuba i Zenoda, olakšavajući neposredan pristup za istraživanje i razvoj.

Treća kategorija, međutim, uvodi značajnu prepreku pristupu. Ovi skupovi nisu javno dostupni za preuzimanje te istraživači moraju formalno zatražiti pristup od izvornih autora. Ovaj proces obično uključuje slanje e-pošte ili ispunjavanje obrasca za registraciju i pristanak na specifične uvjete, odredbe i ograničenja autorskih prava. Ovo ograničavanje pristupa često je nužno zbog uvjeta licenciranja povezanih s izvornim podacima. Međutim, ono može odgoditi istraživanje i rezultirati neodgovorenim zahtjevima ako su se autori preselili u druge institucije ili više nisu aktivni u tom području. U konačnici, to može utjecati na potencijal skupa za ponovnu upotrebu i provjeru od strane šire zajednice.

4. PREPORUKE
4. RECOMMENDATIONS

Naša analiza otkrila je da mnogim podatkovnim skupovima nedostaju jedinstveni identifikatori što otežava njihovo otkrivanje i citiranje. Stoga

preporučujemo davanje jedinstvenog naziva svim novim podatkovnim skupovima i, gdje je to prikladno, akronima kako bi se olakšalo referenciranje i dohvaćanje u akademskim pretraživanjima. Ova procedura olakšava identifikaciju i korelaciju skupa podataka u naknadnim studijama. Osim toga, olakšava pretraživanje ključnim riječima na internetu, na primjer, naziv može olakšati pristup skupu koji sadrži recenzije proizvoda s Amazona.

Ovo istraživanje također je otkriloo da izostavljanje ID-ova tweetova u nekim skupovima ozbiljno ograničava njihovu znanstvenu korisnost. Kako bi otvoreni podatkovni skupovi ostali vrijedni i prilagodljivi za buduća istraživanja, preporučujemo da se uključivanje ID-ova tweetova usvoji kao standardna i obvezna praksa za svaki skup izveden iz Twittera. ID tweeta jedina je pouzdana veza s izvorom podataka, što je bitno za provjeru i replikaciju. Ono također omogućuje istraživačima hidriranje skupova omogućujući pristup bogatom nizu metapodataka, uključujući informacije o vremenu, metrici angažmana korisnika i multimodalnom sadržaju. Sve to uklonjeno je u tekstualnim formatima. Paralelno s razvojem modela analize sentimenta u smjeru uključivanja više kontekstualnih i multimodalnih signala, pristup ovim metapodacima putem ID-ova tweetova nije samo koristan već je i ključan za napredak područja. Uvođenje ove prakse kao standardne značajno bi poboljšalo dugoročnu vrijednost i ponovnu iskoristivost javno financiranih i stvorenih podatkovnih skupova.

Ako je skup javno dostupan i potječe s Twittera, obvezno je uključiti ID-ove tweetova prilikom distribucije informacija. ID-ovi tweetova obogaćuju skup podataka metapodatcima potencijalno dobivenima tijekom faze hidracije. Pristup ID-u tweeta omogućuje pristup svim multimodalnim i višedimenzionalnim informacijama koje Twitter pohranjuje u vezi s tweetom.

Preporučujemo grupiranje instanci koje ne sadrže sentiment ili rezidualne klasificirane instance s onima označenim sentimentom. Prilikom primjene velikih jezičnih modela u stvarnim situacijama važno je izbjeći semantički pomak, odnosno pogrešnu klasifikaciju unosa bez sentimenta kao pripadajućih sentimentnim kategorijama. Taj cilj možemo postići korištenjem nesentimentnih klasa i implementacijom metode obuke koja točno detektira i klasificira takve pojave kao zasebnu kategoriju.

Također preporučujemo implementaciju standarda imenovanja za skupove generirane automatiziranim tehnologijama gdje nema ljudske uključenosti. Za razlikovanje podatkovnih skupova koji sadrže automatski označene instance od onih koji su ručno anotirani ili ispravljeni, može se dodati prefiks kao što je "auto-". Skupovi proizvedeni heurističkim metodama mogu se dodatno razlikovati korištenjem različitih afiksa. Takav pristup pružit će jasniju indikaciju istraživačima da skup podataka uključuje instance koje nisu označili ljudi.

Konačno, podatkovne skupove nastale iz zajedničkih zadataka bilo bi korisno objaviti na platformama kao što je GitHub, ukoliko to licenca dopušta. Također je važno uključiti relevantne ključne riječi za poboljšanje vidljivosti i ponovne upotrebe skupova. Osim toga, preporučujemo korištenje alata za obradu prirodnog jezika (NLP) za automatizaciju procesa generiranja podatkovnih kartica i podatkovnih listova za već postojeće skupove. Zadaci uključuju identificiranje glavnog jezika i dodatnih jezika prisutnih u tweetovima, kao i određivanje njihove distribucije. Dodatne aktivnosti uključuju izdvajanje subjekata i podtema iz skupa podataka, određivanje domene i distribucije skupa, identificiranje relevantnih oznaka za olakšavanje otkrivanja podatakovnog skupa i povezivanje vrsta i naziva zadataka (kao što su klasifikacija teksta i analiza sentimenta) sa skupom.

5. ZAKLJUČAK
5. CONCLUSION

Ovaj rad predstavlja sveobuhvatan pregled i analizu otvorenih podatkovnih skupova s Twittera za analizu sentimenta. Detaljnom analizom 48 različitih skupova identificirano je nekoliko kritičnih područja koja zahtijevaju pažnju istraživačke zajednice. Iako je dostupno mnogo skupova, njihova ukupna korisnost zbog nepostojanja jedinstvenog načina imenovanja, označavanja ili metode distribucije, a često im

nedostaju važni ID-ovi tweetova. Nedostatak standardizacije u konvencijama imenovanja, nedosljedne sheme označavanja, rascjepkane metodama distribucije i izostavljanje ključnih ID-ova tweetova stvaraju značajne poteškoće za istraživače jer ometaju otkrivanje skupova, ograničavaju potencijal za obogaćivanje podataka i kompliciraju usporedbu performansi velikih jezičnih modela u različitim studijama.

Kako bi se prevladalo ove nedostatke, u radu se donosi skup djelotvornih preporuka osmišljenih za poboljšanje korisnosti i dugotrajnosti otvorenih podataka. Glavni prijedlozi su: obvezno uključivanje ID-ova tweetova, standardizacija imenovanja podatkovnih skupova, pružanje jasne dokumentacije za sheme označavanja i poboljšanje pristupačnosti putem sustava kontrole verzija. Prijedlozi su usmjereni na poticanje kohezivnijeg i učinkovitijeg istraživačkog ekosustava. Primjenom ovih praksi otvoreni skupovi podataka mogu postati čvrst, pouzdan i uistinu ponovno iskoristiv resurs za napredak u analizi sentimenta.

Na temelju nalaza i ograničenja prepoznatih u ovom radu, predlažemo nekoliko smjerova za buduća istraživanja:

- **Kvalitativna analiza i analiza temeljena na performansama:** Ovaj se rad usredotočio na metapodatke i karakteristike distribucije podatkovnih skupova. Koristan sljedeći korak bila bi analiza kvalitete podataka, promatranje stupnja slaganja između različitih anotatora kada je to moguće te provođenje velike studije koja ispituje funkcioniranje standardnih modela na ovim različitim skupovima. Na taj način otkrilo bi se kako njihove različite značajke utječu na rezultate u stvarnom svijetu.
- **Istraživanje degradacije podatkovnih skupova:** Problem obrisanih ili nedostupnih tweetova ("smrt tweeta") s vremenom dovodi do degradacije skupa. Buduća istraživanja mogla bi longitudinalno proučavati zbirku ovih skupova kako bi se kvantificirala stopa degradacije i njezin utjecaj na ponovljivost istraživanja. To bi moglo dovesti do najboljih praksi za održavanje i osvježavanje starijih podatkovnih skupova.
- **Automatizacija dokumentacije podatkovnih skupova:** Kao što smo naglasili, proces stvaranja sveobuhvatne dokumentacije poput podatkovnih kartica i listova podataka iznimno je važan, ali može biti opterećujući. Istraživanje korištenja modernih alata obrade prirodnog jezika za automatizaciju generiranja ovih dokumenata za postojeće podatkovne skupove bilo bi vrlo korisno. Takvi projekti mogli bi uključivati zadatke poput automatizirane identifikacije jezika, izdvajanja tema i klasifikacije domena s ciljem obogaćivanja metapodataka povezanih sa skupom.
- **Proširenje istraživanja:** Kontinuirani, zajednički napor za održavanje živog repozitorija ovih podatkovnih skupova nužan je radi osiguravanja pristupa najnovijem popisu dostupnih resursa svim istraživačima.

6. OGRANIČENJA
6. LIMITATIONS

Nalazi i opažanja u ovom radu trebaju se razmatrati u kontekstu nekoliko ograničenja.

Prvo, veličine skupova podataka navedene u Tablici 1 i Tablici 2 temelje se na brojkama koje su dali izvorni autori u trenutku objave. Zbog dinamične prirode Twittera, stvarni broj dostupnih tweetova može biti manji. Tweetove korisnici mogu obrisati ili učiniti nedostupnima u određenim zemljopisnim regijama što dovodi do potencijalnog odstupanja između dokumentirane veličine i hidrirane veličine skupa.

Drugo, ovaj popis nije iscrpan i možda ne uključuje neke najnovije skupove.

Iako sve to može utjecati na studiju, temeljna opažanja u vezi sa standardima imenovanja, distribucije i dokumentacije podatkovnih skupova ostaju valjana za istražene podatke. Konačno, ova se analiza temelji na dokumentaciji koja prati svaki skup. Nismo neovisno ponovno anotirali ili potvrđivali kvalitetu oznaka, niti smo testirali performanse velikih jezičnih modela na tim podatkovnim skupovima. Fokus ove studije je na meta-informacijama i čimbenicima ponovne iskoristivosti samih skupova, a ne na kvalitativnoj procjeni njihovih anotacija.

6. RIZICI
6. RISKS

U ovom radu identificirani su neki objavljeni podatkovni skupovi koji sadrže tekst s platforme Twitter. Tekst može uključivati osobne identifikacijske podatke (PII) koji bi se mogli koristiti u zlonamjerne svrhe. Prikupljanje i distribucija podataka s platformi društvenih medija poput Twittera nose neizbježne rizike, prvenstveno u vezi s privatnošću korisnika. Naše je istraživanje identificiralo nekoliko skupova koji su objavljeni s neobrađenim tekstom tweetova. Ova je praksa problematična jer tekst može sadržavati osobne identifikacijske podatke (PII) koji bi se mogli zloupotrijebiti. PII u ovom kontekstu može uključivati ne samo imena i lokacije, već i mišljenja, zdravstvene informacije ili druge osjetljive podatke koji se, kada se agregiraju, mogu koristiti za deanonimizaciju pojedinaca. Objavljivanje neobrađenog teksta zaobilazi korisnikovu mogućnost kontrole svojih podataka. Čak i ako korisnik izbriše tweet, njegov sadržaj može neograničeno ostati u distribuiranom podatkovnom skupu.

Stoga, na kraju, naglašavamo iznimnu važnost pridržavanja Twitterovih pravila o distribuciji podataka koja dopuštaju dijeljenje ID-ova tweetova, ali ne i punog teksta tweetova. Distribucijom ID-ova istraživači omogućuju dohvaćanje podataka iz izvora, što poštuje korisnikovu odluku o brisanju sadržaja i ublažava rizik da se PII pohranjuju u nekontroliranim podatkovnim skupovima trećih strana. Ovaj pristup temeljno je etičko razmatranje za odgovorno istraživanje u analizi društvenih medija.

7. REFERENCE
7. REFERENCES

[1.] KOESTEN L.; Vougiouklis P.; Simperl E.; Groth P.; Dataset Reuse: Toward Translating Principles to Practice; Patterns, Vol. 1, No. 8, 2020.; str.100136; DOI: https://doi.org/10.1016/j.patter.2020.100136

[2.] Gebru T.; Morgenstern J.; Vecchione B.; Vaughan J. W.; Wallach H.; Iii H. D.; Crawford K.; Datasheets for datasets; Communications of the ACM, Vol. 64, No. 12, 2021.; str.86—92; DOI: https://doi.org/10.1145/3458723

[3.] Pushkarna M.; Zaldivar A.; Kjartansson O.; Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI; Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency; str.1776–1826; ISBN 9781450393522; Seoul, Republic of Korea, 2022; DOI: https:// doi.org /10.1145/3531146.3533231

[4.] Severyn A.; Moschitti A.; Twitter sentiment analysis with deep convolutional neural networks; Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval; str.959--962; 2015; DOI: https:// doi.org/10.1145/2766462.2767830.

[5.] Giachanou A.; Crestani F.; Like it or not: A survey of twitter sentiment analysis methods; ACM Computing Surveys (CSUR), Vol. 49, No. 2, 2016.; str.1—41; DOI: https://doi.org/10.1145/2938640

[6.] Thakkar G.; Hakimov S.; Tadić M.; M2SA: Multimodal and Multilingual Model for Sentiment Analysis of Tweets; Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), ed. Calzolari N., Kan M., Hoste V., Lenci A., Sakti S., Xue N., str.10833--10845; Torino, Italia 2024; DOI: https://doi.org/10.48550/arXiv.2404.01753

[7.] CLARIN.SI repository; https://www.clarin.si/repository/xmlui/; [01.06.2022.]

[8.] Google Dataset Search; https:// datasetsearch.research.google.com/; [01.06.2022]

[9.] European Language Grid; https://live.european-language-grid.eu/catalogue/; [01.06.2022];

[10.] Hugging Face Datasets repository; https:// huggingface.co/datasets; [01.06.2022]

[11.] Nabil M.; Aly M.; Atiya A.; ASTD: Arabic Sentiment Tweets Dataset; Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, ed. Màrquez L., Callison-Burch C., Su J., str.2515--2519; Lisbon, Portugal 2015.; DOI: https://doi.org/10.18653/v1/D15-

1299

[12.] Abu Farha I.; Zaghouani W.; Magdy W.; Overview of the WANLP 2021 Shared Task on Sarcasm and Sentiment Detection in Arabic; Proceedings of the Sixth Arabic Natural Language Processing Workshop; Ukraine (Virtual) 2021; https://aclanthology.org/2021.wanlp-1.36

[13.] Rosenthal S.; Farra N.; Nakov P.; SemEval-2017 Task 4: Sentiment Analysis in Twitter; Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017); str.502--518; Vancouver, Canada, 2017; DOI: https://doi.org/10.18653/v1/S17-2088

[14.] Yin W.; Alkhalifa R.; Zubiaga A.; The emojification of sentiment on social media: Collection and analysis of a longitudinal Twitter sentiment dataset; 2021.; DOI: 10.48550/arXiv.2108.138981; https://figshare.com/articles/dataset/TM-Senti/16438281

[15.] Mozetič I.; Grčar M.; Smailović J.; Multilingual Twitter Sentiment Classification: The Role of Human Annotators; PLOS ONE, Vol. 11., No.5, 2016.; str.1-26; DOI: 10.1371/journal.pone.0155036; http://hdl.handle.net/11356/1054

[16.] Choudhary N.; Singh R.; Rao V. A.; Shrivastava M.; Twitter corpus of Resource-Scarce Languages for Sentiment Analysis and Multilingual Emoji Prediction; Proceedings of the 27th International Conference on Computational Linguistics; str.1570--1577; Santa Fe, New Mexico, USA, 2018; https://aclanthology.org/C18-1133/

[17.] Pauli A. B.; Barrett M.; Lacroix O.; Hvingelby R.; DaNLP: An open-source toolkit for Danish Natural Language Processing; Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa); str.460--466; 2021; https://aclanthology.org/2021.nodalida-main.53/

[18.] Rei L.; Krek S.; Mladenić D.; xLiMe Twitter Corpus XTC 1.0.1, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1078

[19.] Sidarenka U.; PotTS: The Potsdam Twitter Sentiment Corpus; Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16); str.1133--1141; Portorož, Slovenia, 2016; https://aclanthology.org/L16-1181/

[20.] Cieliebak M.; Deriu J. M.; Egger D.; Uzdilli F.; A Twitter Corpus and Benchmark Resources for German Sentiment Analysis; Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, ed. Ku L., Li C., str.45--51; Valencia, Spain 2017.; DOI: https://doi.org/10.18653/v1/W17-1106

[21.] Wojatzki M.; Ruppert E.; Holschneider S.; Zesch T.; Biemann C.; Germeval 2017: Shared task on aspect-based sentiment in social media customer feedback; Proceedings of the GermEval, 2017.; str.1—12; DOI: https://doi.org/10.17185/duepublico/72074

[22.] Nakov P.; Rosenthal S.; Kozareva Z.; Stoyanov V.; Ritter A.; Wilson T.; SemEval-2013 Task 2: Sentiment Analysis in Twitter; Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013); str.312--320; Atlanta, Georgia, USA, 2013; https://aclanthology.org/S13-2052/

[23.] Rosenthal S.; Nakov P.; Kiritchenko S.; Mohammad S.; Ritter A.; Stoyanov V.; SemEval-2015 Task 10: Sentiment Analysis in Twitter; Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015); str.451--463; Denver, Colorado, 2015; DOI: https://doi.org/10.18653/v1/S15-2078

[24.] Ghosh A.; Li G.; Veale T.; Rosso P.; Shutova E.; Barnden J.; Reyes A.; SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter; Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015); str.470--478; Denver, Colorado, 2015; DOI: https://doi.org/10.18653/v1/S15-2080

[25.] Nakov P.; Ritter A.; Rosenthal S.; Sebastiani F.; Stoyanov V.; SemEval-2016

Task 4: Sentiment Analysis in Twitter; Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016); str.1--18; San Diego, California, 2016; DOI: https://doi.org/10.18653/v1/S16-1001

[26.] Vanzo A.; Croce D.; Basili R.; A context-based model for Sentiment Analysis in Twitter; COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers; str.2345--2354; Dublin, Ireland, 2014; https://aclanthology.org/C14-1221/

[27.] Castellucci G.; Vanzo A.; Croce D.; Basili R.; Context-aware Models for Twitter Sentiment Analysis; IJCoL vol. 1, n. 1 December 2015: Emerging Topics at the First Italian Conference on Computational Linguistics; str.69; 2015; DOI: https://doi.org/10.4000/ijcol.322

[28.] Tayebi Arasteh S.; Monajem M.; Christlein V.; Heinrich P.; Nicolaou A.; Naderi Boldaji H.; Lotfinia M.; Evert S.; How Will Your Tweet Be Received? Predicting the Sentiment Polarity of Tweet Replies; Proceedings of the 2021 IEEE 15th International Conference on Semantic Computing (ICSC); str.370-373; DOI: https://doi.org/10.1109/ICSC50631.2021.00068; Laguna Hills, CA, USA, 2021.

[29.] Wang B.; Liakata M.; Zubiaga A.; Procter R.; Jensen E.; Smile: Twitter emotion classification using domain adaptation; CEUR Workshop Proceedings, Vol. 1619, 2016.; str.15—21; https://doi.org/10.48550/arXiv.2307.00135

[30.] Unal C. J. A.; A Sentiment Analysis Model of Spanish Tweets. Case Study: Colombia 2014 Presidential Election/diplomski rad; Universidad Nacional de Colombia; 2016.; https://repositorio.unal.edu.co/handle/unal/56482

[31.] Prieto Santamaría L.; Tuñas J. M.; Fernández Peces-Barba D.; Jaramillo A.; Cotarelo M.; Menasalvas E.; Conejo Fernández A.; Arce A.; Gil de Miguel A.; Rodríguez González A.; Influenza and Measles-MMR: two case study of the trend and impact of vaccine-related Twitter posts in Spanish during 2015-2018; Human Vaccines & Immunotherapeutics, Vol. 18, No. 1, 2022.; str.1-16; DOI: https://doi.org/10.1080/21645515.2021.1877597

[32.] Agüero-Torales M.; Vilares D.; López-Herrera A.; On the logistical difficulties and findings of Jopara Sentiment Analysis; Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching; str.95--102; Online, 2021; DOI: https://doi.org/10.18653/v1/2021.calcs-1.12

[33.] Román J. V.; Serrano S. L.; Cámara E. M.; González-Cristóbal J. C.; TASS - Workshop on Sentiment Analysis at SEPLN; Procesamiento del Lenguaje Natural, Vol. 50, No. 0, 2013.; str.37—44; http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4657

[34.] Román J. V.; Morera J. G.; Serrano S. L.; González-Cristóbal J. C.; TASS 2013 - A Second Step in Reputation Analysis in Spanish; Procesamiento del Lenguaje Natural, Vol. 52, No. 0, 2014.; str.37—44; http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/4901

[35.] Román J. V.; Cámara E. M.; Morera J. G.; Zafra S. M. J.; TASS 2014 - The Challenge of Aspect-based Sentiment Analysis; Procesamiento del Lenguaje Natural, Vol. 54, No. 0, 2015.; str.61—68; http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5095

[36.] Cámara E. M.; Cumbreras M. Á. G.; Román J. V.; Morera J. G.; TASS 2015 – The Evolution of the Spanish Opinion Mining Systems; Procesamiento del Lenguaje Natural, Vol. 56, No. 0, 2016.; str.33—40; http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5284

[37.] Martínez Cámara E.; Almeida-Cruz Y.; Díaz Galiano M. C.; Estévez-Velarde S.; García Cumbreras M. Á.; García Vega M.; Gutiérrez Y.; Montejo Ráez A.; Montoyo A.; Munoz R.; Overview of TASS 2018: Opinions, health and emotions; TASS@SEPLN; 2018; https://ceur-ws.org/Vol-2172/p0_overview_tass2018.pdf

[38.] Díaz-Galiano M. C.; Vega M. G.; Casasola E.; Chiruzzo L.; Cumbreras M. Á. G.; Cámara E. M.; Moctezuma D.; Montejo-Ráez A.; Cabezudo M. A. S.; Tellez E.

S.; Overview of TASS 2019: One More Further for the Global Spanish Sentiment Analysis Corpus.; IberLEF@ SEPLN; str.550--560; 2019; https://ceur-ws.org/Vol-2421/TASS_overview.pdf

[39.] García-Vegaa M.; Díaz-Galianoa M. C.; García-Cumbrerasa M. Á.; del Arcoa F. M. P.; Montejo-Ráeza A.; Jiménez-Zafraa S. M.; Cámarab E. M.; Aguilarc C. A.; Antonio M.; Cabezudod S.; Overview of TASS 2020: Introducing emotion detection; Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020) co-located with 36th Conference of the Spanish Society for Natural Language Processing (SEPLN 2020); str.163--170; 2020; https://ceur-ws.org/Vol-2664/tass_overview.pdf

[40.] López G. J.; Ruiz I. M.; Character and Word Baselines Systems for Irony Detection in Spanish Short Texts; Procesamiento del Lenguaje Natural, Vol. 56, No. 0, 2016.; str.41—48; http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/5285

[41.] Castro S.; Chiruzzo L.; Rosá A.; Garat D.; Moncecchi G.; A Crowd-Annotated Spanish Corpus for Humor Analysis; Proceedings of SocialNLP 2018, The 6th International Workshop on Natural Language Processing for Social Media; 2018; DOI: https://doi.org/10.18653/v1/w18-3502

[42.] Vukotić V.; Claveau V.; Raymond C.; IRISA at DeFT 2015: supervised and unsupervised methods in sentiment analysis; DeFT, Défi Fouille de Texte, joint à la conférence TALN 2015; 2015; https://www.irisa.fr/texmex/people/raymond/pub/DEFT2015.pdf

[43.] Benamara F.; Grouin C.; Karoui J.; Moriceau V.; Robba I.; Analyse d'opinion et langage figuratif dans des tweets: présentation et résultats du Défi Fouille de Textes DEFT2017; Atelier TALN 2017: Défi Fouille de Textes (DEFT 2017); str.1--12; 2017; https://hal.science/hal-01912785/document

[44.] Babić K.; Petrović M.; Beliga S.; Martinčić-Ipšić S.; Matešić M.; Meštrović A.; Characterisation of COVID-19-Related Tweets in the Croatian Language: Framework Based on the Cro-CoV-cseBERT Model; Applied Sciences, Vol. 11, No. 21, 2021.; ARTICLE-NUMBER: 10442; DOI: https://doi.org/10.3390/app112110442

[45.] Riccosan; Saputra K. E.; Pratama G. D.; Chowanda A.; Emotion dataset from Indonesian public opinion; Data in Brief, Vol. 43, 2022.; str.108465; DOI: https://doi.org/10.1016/j.dib.2022.108465

[46.] Moctezuma D.; No C. T. N.; Tellez E. S.; Graff M.; Miranda-Jiménez S.; Sur C. T.; On the performance of B4MSA on SENTIPOLC'16; of the Final Workshop 7 December 2016, Naples; str.200; 2016; DOI: http://dx.doi.org/10.4000/books.aaccademia.2017

[47.] Cignarella A. T.; Lai M.; Bosco C.; Patti V.; Rosso P.; SardiStance@EVALITA2020: Overview of the Task on Stance Detection in Italian Tweets; Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020); 2020; https://ceur-ws.org/Vol-2765/paper159.pdf

[48.] Peisenieks J.; Skadins R.; Uses of Machine Translation in the Sentiment Analysis of Tweets; Baltic HLT; 2014; DOI: https://doi.org/10.3233/978-1-61499-442-8-126

[49.] Kulkarni A.; Mandhane M.; Likhitkar M.; Kshirsagar G.; Joshi R.; L3CubeMahaSent: A Marathi Tweet-based Sentiment Analysis Dataset; Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis; str.213--220; 2021; https://aclanthology.org/2021.wassa-1.23/

[50.] Cortis K.; Davis B.; A Dataset of Multidimensional and Multilingual Social Opinions for Malta's Annual Government Budget, https://zenodo.org/record/4650232

[51.] Patrick H.; Belfo F. P.; Trigo A.; Brazilian tweets classified for sentiment analysis, https://zenodo.org/record/6557683

[52.] Ljajić A.; Marovac U.; Improving sentiment analysis for twitter data by handling negation rules in the Serbian language; Computer Science and Information Systems, Vol. 16, 2018.;

str.13-13; DOI: https://doi.org/10.2298/CSIS180122013L

[53.] Putra O. V.; Wasmanson F. M.; Harmini T.; Utama S. N.; Sundanese Twitter Dataset for Emotion Classification; 2020 International Conference on Computer Engineering, Network, and Intelligent Multimedia (CENIM) (CENIM 2020); 2020; DOI: https://doi.org/10.1109/CENIM51130.2020.9297929

[54.] Mutlu M. M.; Özgür A.; A Dataset and BERT-based Models for Targeted Sentiment Analysis on Turkish Texts; Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop; str.467--472; Dublin, Ireland, 2022; DOI: https://doi.org/10.18653/v1/2022.acl-srw.39

[55.] Balli C.; Guzel M. S.; Bostanci E.; Mishra A.; Sentimental analysis of Twitter users from Turkish content with natural language processing; Computational Intelligence and Neuroscience, Vol. 2022, 2022; DOI: https://doi.org/10.1155/2022/2455160

[56.] Khan M. Y.; Nizami M. S.; Urdu Sentiment Corpus (v1.0): Linguistic Exploration and Visualization of Labeled Datasetfor Urdu Sentiment Analysis.; 2020 IEEE 2nd International Conference On Information Science and Communication Technology (ICISCT); 2020; DOI: 10.1109/ICISCT49550.2020.9080043

[57.] Plutchik R.; The nature of emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice; American scientist, Vol. 89, No. 4, 2001.; str.344—350; http://www.jstor.org/stable/27857503.

[58.] Fraisse A.; Paroubek P.; Toward a unifying model for Opinion, Sentiment and Emotion information extraction; Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), ed. Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J., Piperidis S., str.3881--3886; Reykjavik, Iceland 2014; http://www.lrec-conf.org/proceedings/lrec2014/pdf/1010_Paper.pdf

[59.] Liu B.; Sentiment analysis and subjectivity.; Handbook of natural language processing, Vol. 2, No. 2010, 2010.; str.627—666; https://dl.acm.org/doi/book/10.5555/1738958

AUTORI · *AUTHORS*

• **Gaurish Pandurang Thakkar** - rođen je u Indiji 1989. godine. Doktorirao je 2022. godine na Filozofskom fakultetu Sveučilišta u Zagrebu u polju informacijskih znanosti. Nakon doktorata zaposlen je na Filozofskom fakultetu Sveučilišta u Zagrebu kao viši asistent u Zavodu za lingvistiku. Aktivno sudjeluje na međunarodnim projektima, objavljuje radove u znanstvenim i stručnim časopisima te izlaže na relevantnim domaćim i međunarodnim skupovima. Na diplomskom studiju Odsjeku za informacijske i komunikacijske znanosti održava kolegije iz obrade prirodnog jezika.

Korespondencija · *Correspondence*
gthakkar@m.ffzg.hr