Analiza sentimenta u paralelnim korpusima hrvatskog i poljskog jezika

Vrbanec, Dominik

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI

Ak. god. 2022./2023.

Dominik Vrbanec

Analiza sentimenta u paralelnim korpusima hrvatskog i poljskog jezika

Završni rad

Mentor: dr. sc. Nives Mikelić Preradović, red. prof.
Sumentor: dr. sc. Gaurish Pandurang Thakkar

Zagreb, lipanj 2023.

Izjava o akademskoj čestitosti

Izjavljujem i svojim potpisom potvrđujem da je ovaj rad rezultat mog vlastitog rada koji se temelji na istraživanjima te objavljenoj i citiranoj literaturi. Izjavljujem da nijedan dio rada nije napisan na nedozvoljen način, odnosno da je prepisan iz necitiranog rada, te da nijedan dio rada ne krši bilo čija autorska prava. Također izjavljujem da nijedan dio rada nije korišten za bilo koji drugi rad u bilo kojoj drugoj visokoškolskoj, znanstvenoj ili obrazovnoj ustanovi.

_____

(potpis)

Sadržaj

Kazalo tablica

Kazalo slika

1. Uvod

U području obrade prirodnih jezika susrećemo se s pojmovima sentimenta, emocija i njihovom analizom. Kako bi obrada i analiza sentimenata i emocija uopće bila moguća, preduvjet je postojanje podatkovnih skupova - označenih korpusa. S obzirom na to da na hrvatskom jeziku postoji nedostatak korpusa označenog sa sentimentima i emocijama, ovaj rad prikazuje proces stvaranja jednog takvog korpusa. Tijekom tog procesa, stvoren je istovjetni korpus na poljskom jeziku za koji postoji isti nedostatak. Rezultat je, dakle, paralelni hrvatsko-poljski korpus, nastao na temelju postojećih oznaka iz engleskog korpusa - izvornika, koji sadrži oznake sentimenata i emocija.

Sentiment je pojam koji se odnosi na emocionalni stav, mišljenje ili ocjenu prema nečemu ili nekome. Može biti pozitivan, negativan ili neutralan. Prema Hrvatskom jezičnom portalu (Znanje, 2023) sentiment je ideja, mišljenje ili odnos koji je zasnovan više na emocijama negoli na razumu; osjećaj. Prema (Roca, 2022) sentiment najviše odgovara hrvatskoj glagolskoj imenici mnijenje - subjektivni pristup stvarnim doživljajima. Analiza sentimenta, koja se bavi izvlačenjem i razumijevanjem emocionalnog tona (sentimenta) iz teksta, danas se koristi u mnogim djelatnostima kako bi se dobila dublja slika i razumijevanje stavova i mišljenja ljudi te se prema njima usmjerava razvoj budućih proizvoda i usluga (analiza tržišta), čime se smanjuje vjerojatnost njihove loše prihvaćenosti od strane korisnika. Velika količina informacija svakodnevno se generira putem društvenih mreža, web foruma, sustava dopisivanja i drugih online platformi. Te se informacije koriste za analizu sentimenta, a njihova količina prema (Pandurang Thakkar, 2022) eksponencijalno raste. Ipak, prema istom autoru za južnoslavenske jezike broj tekstova označenih sentimentima te drugih resursa i alata za analizu sentimenta prilično je ograničen.

Analiza sentimenta je tehnika koja se koristi za određivanje sentimenta ili tonaliteta teksta, obično na temelju pozitivnih, negativnih ili neutralnih aspekata. Cilj analize sentimenta je razumjeti emocionalnu reakciju autora teksta prema određenoj temi ili entitetu. Za analizu sentimenta potrebno je imati ili moći dohvatiti dovoljno velike relevantne podatkovne skupove. No, iako ih ima puno, za mnoge manje jezike ne postoje ili nisu dostatni, pa se koriste podatkovni skupovi - oznake sentimenata - iz onih jezika gdje ih ima u izobilju, prvenstveno iz engleskog. Slično kao što je u radu (Thakkar i ostali, 2022) korištena međujezična analiza sentimenata između slovenskog i hrvatskog jezika, gdje je prvi bogatiji korišten za stvaranje

drugog, tako je i u ovom radu korišten engleski podatkovni skup sentimenata koji su transferirani na hrvatski i poljski.

Emocije su subjektivna iskustva koja se javljaju u ljudskom umu i tijelu kao odgovor na određene podražaje. Prema Hrvatskom jezičnom portalu (Znanje, 2023) emocije su psihološko stanje duševne pobuđenosti obilježeno skupom subjektivnih osjećaja, obično praćenih fiziološkim promjenama, koje potiču osobu na reakciju (radost, gnjev, strah, ljubav itd.). Emocije se prema Gašparić mogu sagledavati jednodimenzionalno - kroz kategorije, dvodimenzionalno - kategoriji se dodaje intenzitet/valencija i trodimenzionalno - prethodnima se dodaje i dominantnost (Gašparić, 2020). Ista autorica zaključuje da je za tekstne i video zapise dovoljan jednodimenzionalni pristup putem kategorija.

Analiza emocija se bavi identifikacijom i kategorizacijom emocionalnih aspekata teksta. To su aktivnosti koje su primijenjene na prikupljanje, organiziranje, analiziranje i interpretiranje podataka o emocijama, kao i donošenje zaključaka i generalizacija na temelju tih podataka. Ova analiza može prepoznati emocije kao što su sreća, tuga, ljutnja, strah ili iznenađenje. Cilj analize emocija je dobiti dublje razumijevanje njihove prirode i uloge u ljudskim životima te njihov utjecaj na nj. Prema Ekmanovoj teoriji osnovnih emocija (Ekman, 1999), postoje njih šest, po tri iz pozitivnog i negativnog spektra:

- pozitivne
    - sreća - osjećaj zadovoljstva, veselja i radosti,
    - iznenadnost - osjećaj iznenađenja i čuđenja,
    - oduševljenje - snažan osjećaj divljenja i uzbuđenja;
- negativne
    - tuga - osjećaj tuge, jada i potištenosti,
    - strah - osjećaj anksioznosti, strepnje i straha,
    - gnjev - osjećaj bijesa, frustracije i agresije.

Ekmanova klasifikacija osnovnih emocija samo je jedna od mogućih. Primjerice, Plutchik u svom Indeksu profila emocija razlikuje osam prototipova emocija svrstanih redom u parove suprotnih: sreća i tuga, ljutnja i strah, povjerenje i gađenje te iznenađenje i iščekivanje (Plutchik, 1980).

Analiza sentimenta i emocija ne može biti potpuno precizna jer ljudski jezik može biti vrlo složen i često smisao izrečenog ili napisanog ovisi o kontekstu.

2. Izvori podataka

Podatkovni skupovi, korpusi označeni sentimentima i emocijama, relativno su skromni i nedostatni za mnoge manje jezike, pa tako i za hrvatski i poljski. Prema rezultatima jezičnoporedbenih istraživanja u okviru *META-NET White Paper Series* (META-NET, 2020), i hrvatski i poljski jezik jezičnotehnološki su slabije razvijeni jezici kojima manjka naprednih jezičnih resursa i alata za obradu prirodnoga jezika (Štrkalj Despot & Krek, 2023). Stoga je na tim jezicima njihova analiza otežana, ako ne i onemogućena. U nastavku su navedeni korpusi za ta dva jezika, no niti jedan od njih tri nije (više) javno dostupan.

2.1. Korpus na hrvatskom jeziku

Na hrvatskom jeziku postoji samo jedan resurs sa oznakom sentimenta: *CroSentiLex* leksikon. Prvi očiti nedostatak ovog leksikona sentimenta je to što je svaka riječ zasebna, tj. izvan konteksta (rečenice ili paragrafa). Drugi nedostatak ogleda se u činjenici da ne sadrži oznake za emocije, nego samo sentimente za riječi, pri čemu se koristi znak + kao oznaka za pozitivni sentiment, - za negativni te 0 za neutralni. Treći nedostatak je mali broj riječi uključen u ovaj leksikon. Cjelokupni leksikon bi trebao sadržavati 37 tisuća hrvatskih lema kojima je algoritamski pripisana numerička vrijednost pozitivnosti i negativnosti (Šikić, 2019), što izgleda kao dostatan broj za ozbiljnije korištenje, no korpus (više) nije javno dostupan *online* (otvaranje poveznice rezultira greškom *500 - Internal Server Error*). Dostupan je samo manji dio leksikona u obliku evaluacijskih podatkovnih skupova koji je podijeljen u tri zasebna leksikona s različitim suglasjem označivača, od malog suglasja (oznaka je rezultat obične većine) koji se sastoji od 2500 riječi (lema), preko umjerenog suglasja (oznaka je rezultat suglasja od 8-9 do ukupno 12 označivača) koji se sastoji od 2189 riječi (lema), do visokog suglasja (oznaka je rezultat suglasja od najmanje 10 od ukupno 12 označivača) koji se sastoji od 1709 riječi (lema). Četvrti nedostatak ovog korpusa ogleda se u činjenici da označivači nisu ljudi nego, kako se može zaključiti iz pratećih tekstova, programski agenti koji se razlikuju u različitim početnim parametrima i koji su ocijenili pojedine riječi koristeći *PageRank* algoritam (Glavaš & Meta-Share, 2013).

2.2. Korpusi na poljskom jeziku

PolEval Sentiment Corpus je jedan od najpoznatijih korpusa na poljskom jeziku s označenim sentimentima. Sadrži veliku kolekciju tekstova iz različitih izvora kao što su društveni mediji, forumi, novinski članci, recenzije i drugo. *PolEval Emotion Corpus* korpus sadrži tekstove na poljskom jeziku koji su označeni emocijama. Oba korpusa nisu (više) dostupna. Podatkovni

skupovi koji se mogu preuzeti sa službenih stranica (PolEval 2017, 2017) ne sadrže oznake sentimenta ni emocija.

2.3. Engleski korpus *MELD*

Izvorni engleski korpus pod nazivom *Multimodal EmotionLines Dataset* (*MELD*) (Poria, 2022; Poria i ostali, 2018) stvoren je i unaprijeđen od drugog korpusa pod nazivom *EmotionLines dataset*, opisanog u radu (Chen i ostali, 2018). *MELD* sadrži iste dijaloge koji se nalaze u EmotionLines, a razlikuju se u tome što je *MELD* bogatiji, s obzirom da sadrži multimedijske (video i zvučne) sekvence te sentimente.

MELD ima više od 1300 dijaloga i 13000 izjava iz TV serije *Friends*. Svaka izjava u dijalogu označena je jednom od sedam emocija (ili njenog nedostatka): radost, tuga, strah, ljutnja, gađenje, iznenađenje i neutralno te jednom od tri oznake sentimenta: pozitivno, neutralno i negativno (Poria, 2022).

MELD je podijeljen u tri standardna podskupa: *dev*, *train* i *test*. Pohranjen je u tri zasebne datoteke formata *csv* (*comma separated values*) koje se mogu otvoriti u bilo kojem uređivaču teksta ili u tabličnom kalkulatoru poput *Excela* te učitati i separirati podatke u nekom programskom jeziku poput *Pythona* (csv problem). Svaki od ta tri podatkovna podskupa načinjen je slučajnim odabirom iz cjelokupnog podatkovnog skupa te sadrži sljedeće stupce: *Sr No.* (redni ili serijski broj), *Utterance* (izjava), *Speaker* (govornik), *Emotion* (emocija), *Sentiment* (sentiment), *Dialogue_ID* (identifikacijski broj dijaloga), *Utterance_ID* (identifikacijski broj izjave), *Season* (sezona), *Episode* (epizoda), *StartTime* (početno vrijeme), *EndTime* (završno vrijeme). Tablica 1 usporedno prikazuje brojčane pokazatelje dijelova izvornog korpusa.

Tablica 1. Brojčani opis *MELD* podatkovnog skupa

Podskup	**Broj zapisa**	
Dev	1109	8.1%
Train	9989	72.9%
Test	2610	19.0%
Ukupno	13708	100%

Cjelokupni podatkovni skup podijeljen je blisko standardnoj podjeli *dev/train/test* na 10/70/20%, preciznije na 8/73/19%. U ovom radu je zadržana istovjetna podjela za rezultirajući hrvatski i poljski podatkovni skup.

2.4. Analiza emocija i sentimenta u *MELD*-u

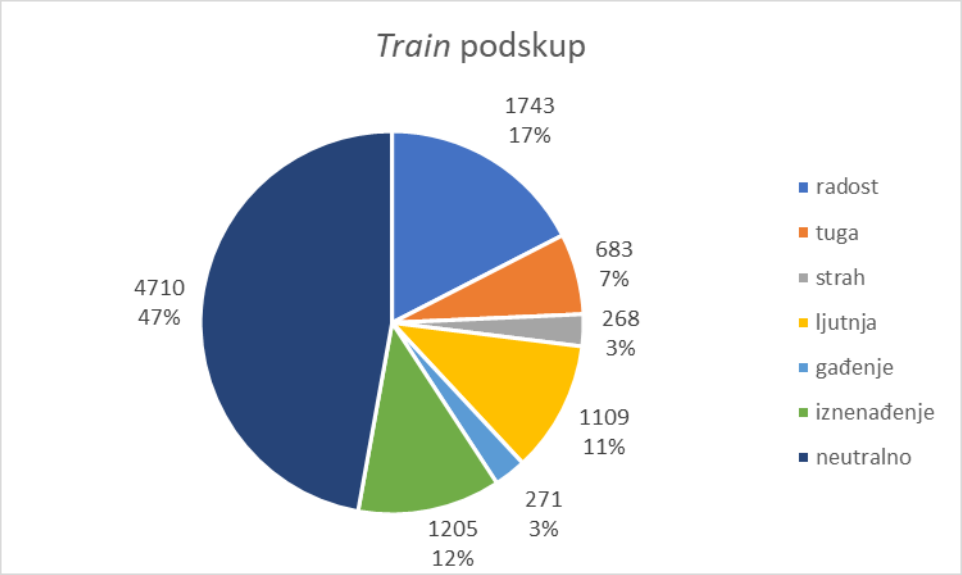
Najčešća emocija i najčešći sentiment u *MELD* podatkovnom skupu je neutralno (Tablica 2). Neutralna emocija, za razliku od svih ostalih emocija, uvijek se veže s neutralnim sentimentom. Druga najučestalija emocija je radost, što je i očekivano u laganom i uglavnom vedrom serijalu poput *Friends*. Na trećem i četvrtom mjestu su iznenađenje i ljutnja, ovisno kako u kojem podskupu. Peta najčešća emocija je tuga, dok su strah i gađenje najrjeđe emocije u sva tri podskupa[1]: *train*, *dev* i *test*. U *train* i *test* podskupovima gađenje je češće nego strah, dok je u *dev* podskupu obrnuto. Također je neočekivano, s obzirom na prirodu serijala, da je negativan sentiment češći od pozitivnog (i to u svakom podskupu).

Tablica 2. Distribucija emocija i sentimenta u *MELD* podatkovnom skupu

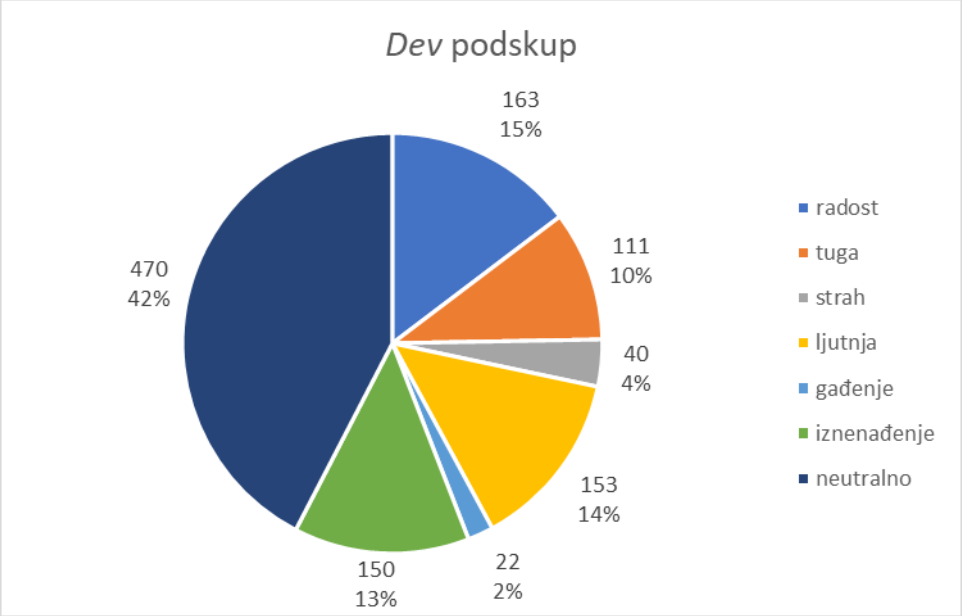
	Kategorije	Train		Dev		Test	
Emocija	radost	1743	17%	163	15%	402	15%
	tuga	683	7%	111	10%	208	8%
	strah	268	3%	40	4%	50	2%
	ljutnja	1109	11%	153	14%	345	13%
	gađenje	271	3%	22	2%	68	3%
	iznenađenje	1205	12%	150	13%	281	11%
	neutralno	4710	47%	470	42%	1256	48%
Sentiment	pozitivno	2334	23%	233	21%	521	20%
	neutralno	4710	47%	470	42%	1256	48%
	negativno	2945	30%	406	37%	833	32%

Na temelju: https://arxiv.org/abs/1810.02508

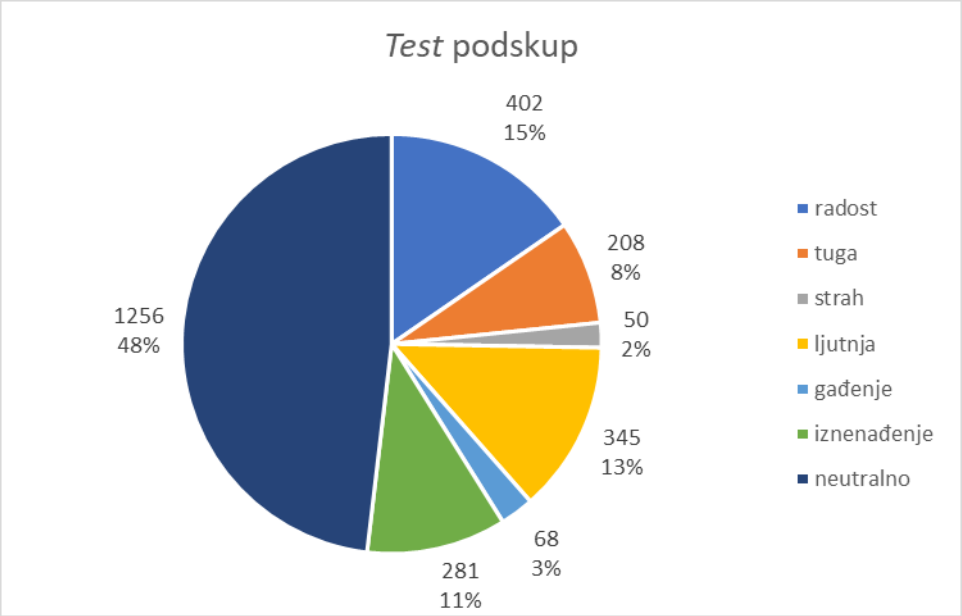
[1] Za uvježbavanje modela kojima se obrađuju podatkovni skupovi koristi se *train* podatkovni podskup, za procjenu različitih verzija modela tijekom razvoja koristi se *dev* podatkovni podskup, a za potvrdu odgovora na istraživačka pitanja koristi se *test* podatkovni podskup (van der Goot, 2021). Dakle, cjelokupni podatkovni skupovi se najčešće dijele na *train*, *dev* i *test* podskupove.



Slika 1. Zastupljenost emocija u *train* podatkovnom podskupu



Slika 2. Zastupljenost emocija u *dev* podatkovnom podskupu



Slika 3. Zastupljenost emocija u *test* podatkovnom podskupu

Rezultati ovog završnog rada, tj. generirani podatkovni skupovi na hrvatskom i poljskom jeziku u cijelosti strukturom odgovaraju engleskom izvorniku.

3. Korištene *Python* biblioteke

Svaka se biblioteka u *Python* skriptu uvozi naredbom `import` iza koje slijedi naziv biblioteke. Neke funkcije/moduli mogu se iz biblioteka zasebno uvesti naredbom `from`, a prema potrebi i želji programera preimenovati ključnom riječju `as` (poput `from deep_translator import GoogleTranslator as GT`).

Python* biblioteka *os (Python Software Foundation, 2023b) omogućuje rad s operacijskim sustavom, posebno njegovim datotečnim podsustavom na jednostavan i učinkovit način. Biblioteka omogućuje manipulaciju s datotekama, mapama, procesima i drugim resursima računalnog i operacijskog sustava. Pruža mnoge funkcije za provjeru postojanja datoteka ili mapa, kreiranje, premještanje, brisanje, promjenu imena i pristupanje informacijama o datotekama kao što su veličina, vremenske oznake i dozvole. U programu su korištene sljedeće funkcije za manipulaciju datotekama i mapama:

- os.path.exists(path): provjerava postoji li datoteka ili mapa na zadanoj putanji,
- os.path.isdir(path): provjerava je li zadana putanja mapa,
- os.makedirs(path): stvara novu mapu i sve potrebne mape na putu zadane putanje,
- os.listdir(path): vraća popis svih datoteka i mapa na zadanoj putanji.

Python* biblioteka *time (Python Software Foundation, 2023c) omogućava funkcionalnosti poput upravljanja vremenom, izračunavanje vremenskih razlika, oblikovanje datuma i sl. U radu je korištena funkcija:

- time.sleep(sekunde): pauzira izvršavanje programa na određeno vrijeme.

Python* biblioteka *codecs pruža funkcionalnosti za rad s različitim oblikovanjima, zapisima odnosno kodiranjima teksta, omogućujući čitanje, pisanje i pretvaranje teksta između različitih kodova. U Hrvatskoj se najčešće koriste kodovi windows-1250, ISO-8859-2 i UTF-8 (Moj web dizajn, 2023), ali u svijetu se najviše koristi UTF-8, tako da su u radu tekstovi otvarani (i prema potrebi konvertirani) u UTF-8, na način:

```
with codecs.open(naziv_dokument, "r", encoding="utf-8", errors="ignore")
```

Za pisanje je umjesto „r“ (*read*) korištena opcija „w“ (*write*). U slučaju da se na ulazu dogodi znak koji nije prepoznatljiv, ukoliko ga ne uspije pretvoriti u UTF-8 kodiranje, umjesto da program stane i javi grešku, on taj znak ignorira (opcija errors="ignore") i nastavlja dalje.

Python* biblioteka *csv2tsv omogućuje konverziju csv (*Comma-Separated Values*) datoteka u tsv (*Tab-Separated Values*) format (Python Software Foundation, 2023a), a detaljnije je opisano u poglavlju 4.2. Formati datoteka.

Python* biblioteka *deep_translator pruža sučelje za prevođenje teksta koristeći različite online prevoditelje kao što su *Google Translate, Microsoft Azure Translator* i *Yandex.Translate*. U radu je korišten samo *Google Prevoditelj* kao pretpostavljeno najbolji besplatni online prevoditelj, a što je detaljnije prikazano u poglavlju 4.3.2. Posebni slučajevi.

Python* biblioteke *Stanza* i *stanfordnlp

Stanza je novija biblioteka za obradu prirodnog jezika (Stanford NLP Group, 2020) koja je razvijena kao nasljednik i unaprijeđena verzija biblioteke *Stanford NLP* (Qi i ostali, 2018; Stanford NLP Group, 2020). Obje biblioteke sadrže razne module za obradu teksta kao što su tokenizacija, anotacija, prepoznavanje imenovanih entiteta, analiza ovisnosti i drugi. Ti moduli su bazirani na raznim modelima strojnog učenja i statističkim tehnikama. *Stanza* je donijela bolje performanse i skalabilnost. *Stanza* je optimiziranija u odnosu na Stanford NLP, pa brže obrađuje tekst. Obje biblioteke pružaju podršku za razne jezike (*Stanza* za nešto veći broj), s predtreniranim modelima za nekoliko jezika. Podrška za jezike korištene u ovom radu prilikom preuzimanja u obliku `Stanza.download("en"/"hr"/"pl")` su sljedećih veličina: 545MB za engleski (235MB u slučaju *stanfordnlp*), 207MB za hrvatski (227MB za *stanfordnlp*) i 414MB za poljski (228MB za *stanfordnlp*), no samo engleski, njemački i kineski su podržani za analizu sentimenta, a svi jezici su bez mogućnosti analize emocija. Pored toga, za *standfornlp* je potrebno preuzeti i pokrenuti *stanfordnlp* poslužitelj na lokalnom računalu.

Python biblioteka *nrclex*

NRCLex je *Python* biblioteka koju je izradio i javno objavio za slobodno korištenje i modifikaciju na pypi.org Mark M. Bailey (Bailey, 2019), a nastao je na temeljima leksikona kojeg je prethodno objavio *National Research Council Canada* (S. M. Mohammad, 2016; S. Mohammad & Turney, 2013) te *WordNet*-ovim skupovima sinonima (*WordNet synonym sets*) koji se koristi unutar *NLTK* biblioteke (Bird i ostali, 2009; NLTK Project, 2023).

NRCLex je u stanju razlučiti, ocijeniti rezultatom između 0 i 1 sljedeće vrste emocija i sentimenta, tj. emocionalnim afektima: strah, ljutnja, iščekivanje, povjerenje, iznenađenje, pozitivno, negativno, tuga, gađenje i sreća (u originalu: *fear, anger, anticipation, trust, surprise, positive, negative, sadness, disgust, joy*). U radu su za analizu sentimenta korišteni rezultati za emocionalne afekte pozitivnost i negativnosti na sljedeći način: Ukoliko je rezultat

za pozitivan afekt veći od negativnog, ukupni afekt se proglašava pozitivnim, u suprotnom, ako je manji, ukupni afekt se proglašava negativnim, a u slučaju jednakog rezultata, ukupni afekt se proglašava neutralnim.

4. Tijek rada i problemi

U nastavku su opisani koraci rada, uočeni problemi i načini njihova rješavanja.

4.1. Izvori podatkovnih skupova

U prvom koraku rada preuzeti su titlovi za tri jezika: engleski, hrvatski i poljski. Početna je ideja bila da se iskoriste titlovi koji su nastali ljudskim prijevodom te su provjereni i ocijenjeni od korisnika web sjedišta s kojih su preuzeti: www.opensubtitles.org, www.titlovi.com i www.tvsubtitles.net. Prvo web sjedište nije dozvoljavalo preuzimanje titlova u cijelosti ili barem po sezoni, što je opcija samo za VIP korisnike s pretplatom, pa u konačnici to web sjedište odnosno titlovi s njega nisu bili korišteni (za preuzimanje nisu od koristi bili ni programski alati poput *wget*-a). S drugog su web sjedišta preuzeti titlovi za engleski i hrvatski jezik, a s trećeg za poljski. Analiza preuzetih titlova, rezultirala je uočavanjem dva problema.

Prvi je bio očekivani: titlovi nisu poravnati. Titlovi preuzeti s interneta na različitim jezicima, od različitih autora i za različite oblike video zapisa, nisu međusobno kompatibilni, tj. ne sadrže nužno iste tekstove niti odgovarajuća vremenska poravnanja. Dodatno, svaki prevoditelj ima svoj stil prevođenja.

Drugi, neočekivani problem, je taj što se dijalozi nisu preslikavali 1:1, što je prikazano u nastavku na primjeru prve epizode prve sezone (Tablica 3).

Tablica 3. Usporedba prvih sekvenci titlova prve epizode prve sezone serije *Friends*

#	**Engleski**	**Hrvatski**	**Poljski**
1	00:00:47,881 --> 00:00:49,757 [CAR HORNS HONKING]	00:00:01,320 --> 00:00:04,839 PRIJATELJI	00:00:00,042 --> 00:00:04,171 www.napiprojekt.pl - nowa jakość napisów. Napisy zostały specjalnie dopasowane do Twojej wersji filmu.
2	00:00:49,966 --> 00:00:52,760 There's nothing to tell. It's just some guy I work with.	00:00:46,840 --> 00:00:49,159 Nemam vam što reći. Radim s njime.	00:00:48,966 --> 00:00:51,218 "Początek"
3	00:00:52,969 --> 00:00:55,137 Come on. You're going out with a guy.	00:00:49,799 --> 00:00:53,560 Daj! Izlaziš s njim, očito s njim nešto nije u redu.	00:00:55,639 --> 00:00:57,891 Nie ma o czym mówić!
4	00:00:55,305 --> 00:00:57,848 There's gotta be something wrong with him.	00:00:54,799 --> 00:00:57,159 Ima li grbu? Grbu i tupe?	00:00:57,975 --> 00:01:00,269 To tylko facet z, którym pracuję!
5	00:00:58,058 --> 00:00:59,933 So does he have a hump and a hair piece?	00:00:58,840 --> 00:01:00,920 Čekaj! Jede li kredu?	00:01:00,352 --> 00:01:03,063 Daj spokój... Chodzisz z nim!
6	00:01:02,395 --> 00:01:03,771 Wait. Does he eat chalk?	00:01:01,359 --> 00:01:05,200 Ne bih htjela da proživi ono što sam ja prošla s Carlom.	00:01:03,188 --> 00:01:05,274 Coś musi być z nim nie tak!

7	00:01:05,023 --> 00:01:08,233 I don't want her to go through what I went through with Carl. Oh.	00:01:05,280 --> 00:01:10,319 Smirite se. To nije čak ni spoj. Izlazak dvoje ljudi, bez seksa.	00:01:05,274 --> 00:01:07,359 Dobra Joey, bądź miły.
8	00:01:08,443 --> 00:01:11,236 Okay, everybody relax. This is not even a date.	00:01:10,680 --> 00:01:12,799 Za mene je to spoj.	00:01:07,943 --> 00:01:09,069 Co, ma garba?
9	00:01:11,446 --> 00:01:14,406 It's just two people going out to dinner and not having sex.		00:01:09,111 --> 00:01:10,153 Garba i perukę?
10	00:01:14,616 --> 00:01:16,158 Sounds like a date to me.		00:01:12,489 --> 00:01:14,575 Chwila, może je kredę?
11			00:01:14,825 --> 00:01:20,038 Po prostu nie chce, żeby przechodziła to samo co ja z Carlem!
12			00:01:20,038 --> 00:01:22,124 Dobra! Wszyscy spokój.
13			00:01:22,124 --> 00:01:23,167 To nawet nie randka.
14			00:01:23,167 --> 00:01:25,252 Po prostu dwoje ludzi
15			00:01:25,252 --> 00:01:27,337 idzie razem na kolację bez sexu.
16			00:01:27,337 --> 00:01:29,423 Dla mnie to brzmi jak randka.

Iz tablice je vidljivo da je deset početnih dijaloga na engleskom jeziku prevedeno na osam dijaloga na hrvatskom te 16 na poljskom. Ta činjenica sasvim onemogućava programsko rješenje koje bi poravnavalo titlove koristeći već provjeren gotov prijevod.

Rješenje gore opisanih problema je neposredno korištenje *csv* datoteke s engleskim oznakama dobivene s *MELD* podatkovnim skupom, tj. konačni proizvod autora koji su načinili taj podatkovni skup.

4.2. Formati datoteka

Potencijalna neobična kodiranja teksta i greške koje pri tome mogu nastati prilikom učitavanja datoteke u *csv* formatu savladane su korištenjem *Python* modula *codecs* i opcijom ignoriranja grešaka, na sljedeći način:

```
import codecs
with codecs.open(naziv_dokumenta, "r", encoding="utf-8", errors="ignore") as f:
       tekst=f.read().strip()
    tekst=tekst.split("\n")
```

Izvršenje gornjeg koda rezultira redcima - stringovima koji sadrže 11 polja, no izdvajanje tih polja iz stringa metodom *split(",")* nije bilo uspješno iz razloga što se unutar drugog polja (Izjava) nalaze zarezi, a ista je situacija i s posljednja dva polja (Pocetno_vrijeme,

Zavrsno_vrijeme). Alternativni pokušaj razdvajanja po navodnicima također nije bio uspješan, jer neke izjave sadrže, a druge ne sadrže navodnike. Nameće se zaključak da autori izvornog podatkovnog skupa na engleskom jeziku koji se u ovom radu koristi nisu bili svjesni mogućih problema pri korištenju tako strukturiranih podataka.

Pronađeno rješenje je konverzija *csv* datoteka u *tsv* (*Tab Separated Values*), koristeći dodatni *Python* modul *csv2tsv*:

```
from csv2tsv import to_tsv
for doc in documents:
	to_tsv(doc, doc.replace(".csv",".tsv"), encoding="utf-8")
```

Rad s datotekama u *tsv* formatu je potom bio moguć jer su zapisi polja razdvojeni tabulatorom koji se ne nalazi u izvornim podacima.

4.3. Strojno prevođenje

Za strojno prevođenje korišten je Google Prevoditelj, preko *Python* modula *deep_translator*. Iako strojni prevoditelji, a na to nije imun ni Google Prevoditelj, tijekom strojnog prevođenja rade greške, učestalije kod flektivnih jezika poput hrvatskog i poljskog, prijevod je i više nego zadovoljavajući za potrebe ovog rada. Pri korištenju Google Prevoditelja, kao i u prethodnim koracima rada, nužno je bilo riješiti nekoliko problema.

4.3.1. Opterećenje prevoditelja

Google Prevoditelj ima ograničenje glede broja zahtjeva koje je spreman izvršiti s jedne IP adrese u nekom vremenu, pa je za najveći podatkovni podskup (*train*) javljao greške poput:

```
requests.exceptions.ConnectTimeout:
HTTPSConnectionPool(host='translate.google.com', port=443): Max retries exceeded
with url: /m?tl=hr&sl=en&q=Because+uh%C2%85we-
we%C2%85we+split+up.+Monica+and+I+split+up.+Hold+me. (Caused by
ConnectTimeoutError(<urllib3.connection.HTTPSConnection object at
0x000002502F301150>, 'Connection to translate.google.com timed out. (connect
timeout=None)'))
```

Taj je problem riješen preventivnim dodavanjem pauze nakon svakih tisuću zahtjeva za prijevodom:

```
for num,red in enumerate(tekst[1:]):
	if num%1000==0 and not num==0:
		print("Google Translator pauza :))")
		time.sleep(120)
```

U gornjem programskom kodu `tekst` je lista sastavljena od cijelih redova zapisa. Indeksiranje `tekst [1:]` je potrebno jer je prvi red podatkovnog skupa opis polja, odnosno zaglavlja stupaca gledano iz perspektive tablice otvorene u *Excelu*. Radi brojanja redaka, potrebno je koristiti *Python* ključnu riječ `enumerate`, zajedno s varijablom `num` koja broji retke. Ostatak pri dijeljenju

u *Pythonu* dobije se operandom `%`, a dodatni uvjet `not num==0` koristi se radi izbjegavanja pauze odmah na početku petlje. Pauza se dobije korištenjem modula `time` te metodom `sleep()`.

4.3.2. Posebni slučajevi

Google Prevoditelj nije htio prevesti:

- specijalne znakove ukoliko su oni bili s vrlo malo teksta,
- male količine teksta u žargonu,
- naglašavanje izraza ponavljanjem slova (npr. *maaaaad* umjesto *mad*).

U takvim slučajevima Google Prevoditelj je ponekad znao vratiti prijevod u jednom od dva oblika: *nontype* ili prazan red. Takav se prijevod morao ignorirati te ostaviti neprevedeni tekst.

```
from deep_translator import GoogleTranslator as GT
try:
    prijevod = GT(source="en", target="hr").translate(izjava)
except:
    print("Google Prevoditelj nije uspio prevesti tekst:", izjava)
    prijevod = izjava
    pass
```

5. Rezultati

Pored korištenja oznaka emocija i sentimenta dobivenih iz ***MELD*** podatkovnog skupa, a s obzirom na nepostojanje drugih označenih korpusa, u radu su korištene biblioteke odnosno programski alati za *Python* razvijeni od strane Sveučilišta u Stanfordu (*Stanford NLP*, tj. njegovu novu i napredniju inačicu slobodnog otvorenog koda pod nazivom ***Stanza***) i ***NRCLex***.

S obzirom da ne postoje odgovarajući korpusi na hrvatskom i poljskom jeziku označeni sentimentom, te je za njihovo stvaranje bilo nužno koristiti engleski *MELD* korpus, *Stanza* i *NCRLex* koji su u stanju riječi i rečenice teksta ocjenjivati sentimentima, korišteni su radi usporedbe s rezultatima sentimenta *MELD* korpusa, s ciljem odgovora na pitanje može li se neki od njih koristiti za automatizirano označavanje sentimenta. Tablica 4 prikazuje primjere ocjenjivanja *Stanze* i *NCRLexa* u odnosu na *MELD*. Već u prvih pet izjava *dev* podskupa podataka možemo pronaći četiri tipična slučajeva: potpuno slaganje u pozitivnoj, negativnoj i neutralnoj ocjeni te potpuno neslaganje u njima za istu izjavu.

Tablica 4. Primjeri rezultata za iste izjave

Izjava	***MELD***	***Stanza***	***NCRLex***
Oh my God, he's lost it. He's totally lost it.	negativno	negativno	negativno
Or! Or, we could go to the bank, close our accounts and cut them off at the source.	neutralno	neutralno	neutralno
You're a genius!	pozitivno	pozitivno	pozitivno
Aww, man, now we won't be bank buddies!	negativno	pozitivno	neutralno

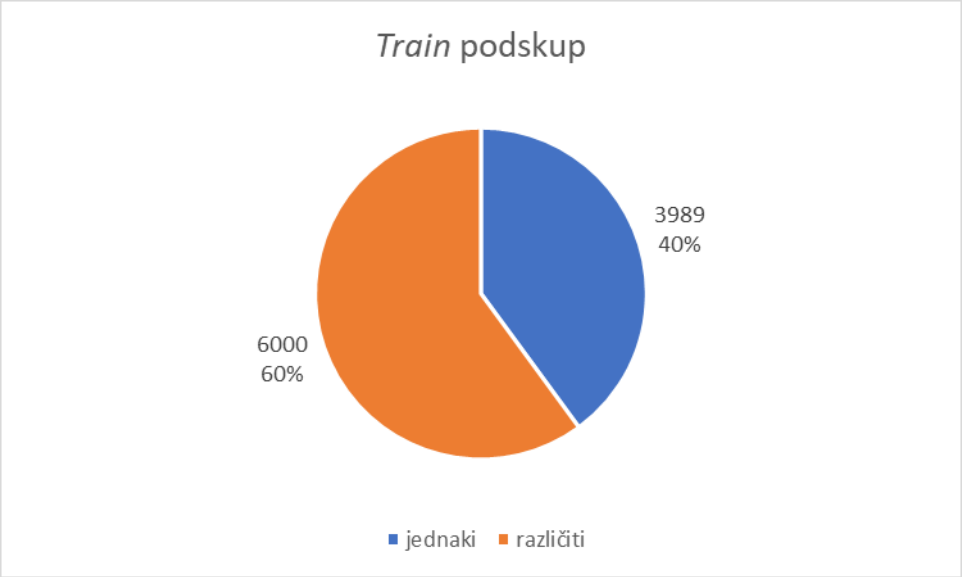
S obzirom na to da rezultati nisu identični, korisno je sagledati i usporediti ukupnost njihova (ne)slaganja i brojčane pokazatelje za sva tri načina ocjenjivanja sentimenta (*MELD*, *Stanza* i *NCRLex*) te za sva tri podskupa podataka (*train*, *dev* i *test*), što je prikazano u tablicama i grafičkim prikazima u nastavku. Nažalost, slična usporedba na razini emocija nije moguća iz sljedećih razloga:

- *Stanza* nema mogućnosti analize emocija,
- *MELD* i *NCRLex* nemaju jednaki skup emocija niti njihov broj (*MELD* ima radost, tugu, strah, ljutnju, gađenje, iznenađenje i neutralno, a *NCRLex* ima strah, ljutnju, iščekivanje, povjerenje, iznenađenje, pozitivno, negativno, tugu, gađenje i sreću).

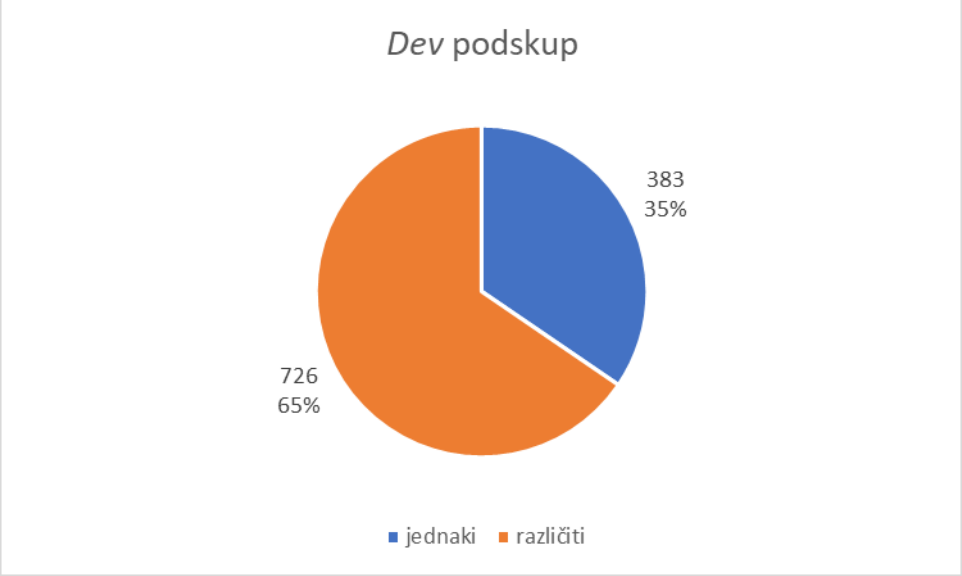
Ukupni podatkovni skup opisan je tablicom 5 i grafičkim prikazima na slikama 4-6 koji prikazuju rezultate podskupova, a u suštini govori o tome da je suglasje ocjena prisutno u 35-40% slučajeva, ovisno o podskupovima.

Tablica 5. Usporedba brojnosti ocjena sentimenta na cjelokupnom podatkovnom skupu

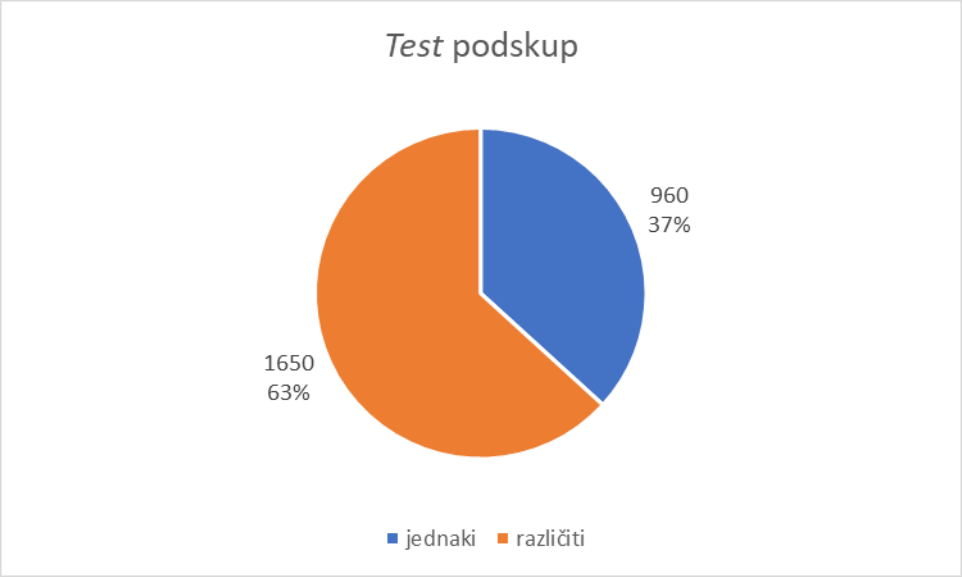
	MELD	*Stanza*	*NRCLex*
negativni	2945	2242	1110
neutralni	4710	5799	7212
pozitivni	2334	1948	1667



Slika 4. Odnos različitih i jednakih ocjena sentimenta na *train* podatkovnom skupu

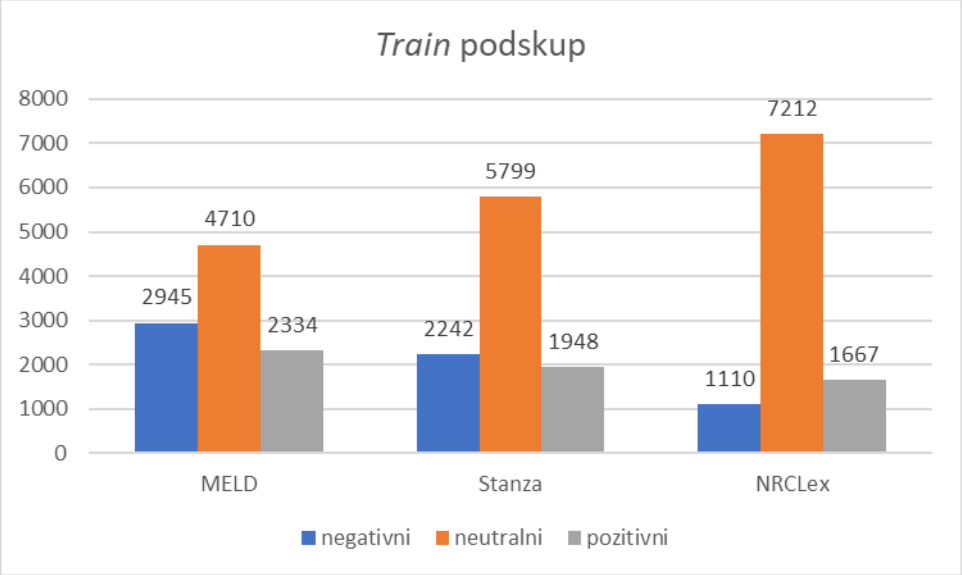


Slika 5. Odnos različitih i jednakih ocjena sentimenta na *dev* podatkovnom skupu

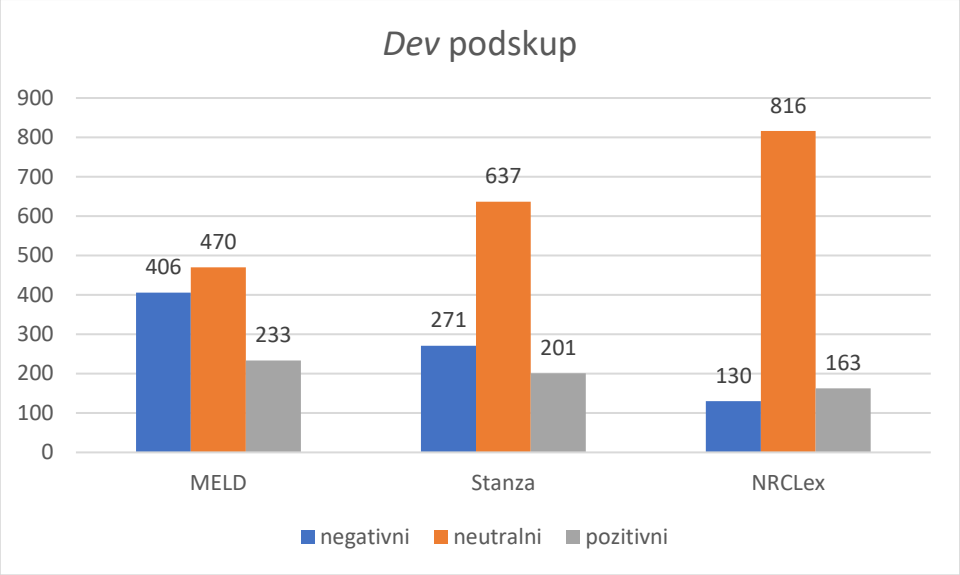


Slika 6. Odnos različitih i jednakih ocjena sentimenta na *test* podatkovnom skupu

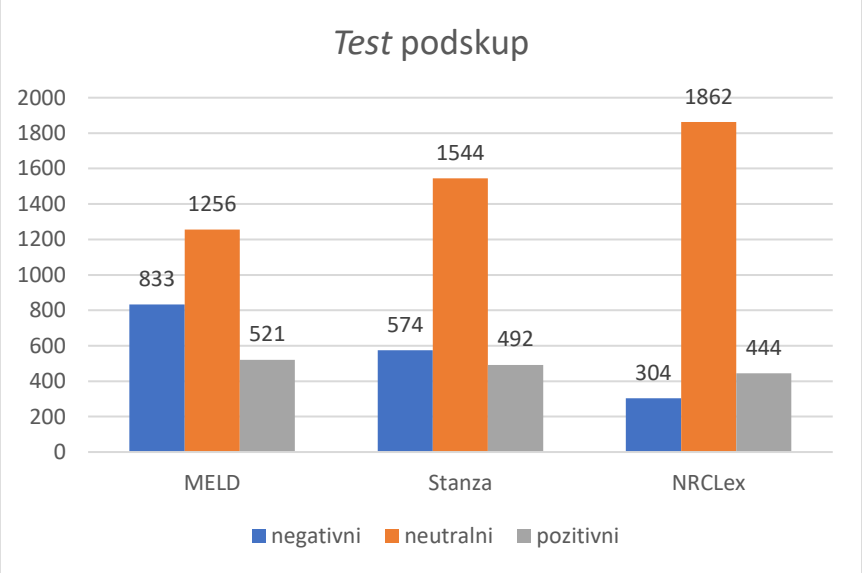
Ako promatramo pojedine podatkovne podskupove, u svima je najzastupljeniji neutralni sentiment, no uspoređujući *MELD*, *Stanza* i *NRCLex* pristupe, rezultati se razlikuju relativno: *MELD* ima od svih njih najmanje neutralnih, *Stanza* više, a *NCRLex* najviše. To znači da se *Stanza* „teže“, a *NCRLex* „najteže“ „odlučuju“ na neki polaritet sentimenta, a onda bismo mogli zaključiti da su i manje korisni.



Slika 7. Usporedba brojnosti ocjena sentimenta na *train* podakovnom podskupu



Slika 8. Usporedba brojnosti ocjena sentimenta na *dev* podakovnom podskupu



Slika 9. Usporedba brojnosti ocjena sentimenta na *test* podakovnom podskupu

S obzirom na prikazane brojke i njihovu usporedbu, vrijednost korpusa stvorenog strojnim prevođenjem Google Prevoditeljem na temelju kvalitetnog *MELD* korpusa, a u nedostatku hrvatskih i poljskih korpusa vrjednija je nego li bi bila da su načinjeni pomoću označavanja korištenjem *Stanze* ili *NCRLexa*.

6. Zaključak

U području obrade jezika nedostaje hrvatski tekstualni korpus s oznakama sentimenta koji bi omogućio njihovu analizu. U ovom radu se opisuje proces stvaranja dva takva korpusa, tj. jedan paralelni hrvatsko-poljski korpus koji su paralelni s engleskom korpusom - izvornikom, koji je nastao na temelju televizijske serije *Friends*. Tekstovi iz engleskog korpusa su strojno prevedeni na hrvatski i poljski pomoću *Python* programa i Google Prevoditelja, te je provedena analiza sentimenta sva tri korpusa. Taj novonastali korpus dostupan je online (https://tinyurl.com/frenhrpl) za slobodno korištenje.

Strojno prevođenje korištenjem jezika visoke razine poput Pythona daje u današnje vrijeme vrlo dobre rezultate, posebno u stvaranju jezičnih resursa za male (ne-svjetske) jezike. Pri tome se, naravno, mora riješiti mnoštvo problema (neki su u ovom prikazu spomenuti), no na taj način se u razumnom vremenu mogu stvoriti vrijedni resursi. Rad je uspješno riješio problem koji povezuje dva područja: informacijske znanosti te poljski jezik i književnost. Ovo iskustvo može se koristiti za stvaranje korpusa i za druge manje jezike.

7. Literatura

Bailey, M. M. (2019). *NRCLex: An affect generator based on TextBlob and the NRC affect lexicon* (4.0) [Python; OS Independent]. https://github.com/metalcorebear/NRCLex

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: Analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Chen, S.-Y., Hsu, C.-C., Kuo, C.-C., Ting-Hao, Huang, & Ku, L.-W. (2018). *EmotionLines: An Emotion Corpus of Multi-Party Conversations* (arXiv:1802.08379). arXiv. https://doi.org/10.48550/arXiv.1802.08379

Ekman, P. (1999). Basic Emotions. U *Handbook of Cognition and Emotion* (str. 45–60). John Wiley & Sons, Ltd. https://doi.org/10.1002/0470013494.ch3

Gašparić, B. (2020). *Leksikon emocija hrvatskog jezika* [Info:eu-repo/semantics/bachelorThesis, University of Zagreb. University of Zagreb, Faculty of Humanities and Social Sciences. Department of information and Communication sciences]. https://urn.nsk.hr/urn:nbn:hr:131:577940

Glavaš, G. & Meta-Share. (2013, travanj 2). *Croatian Sentiment Lexicon*. http://metashare.ilsp.gr:8080/repository/browse/croatian-sentiment-lexicon/940fe19e6c6d11e28a985ef2e4e6c59eff8b12d75f284d58aacfa8d732467509/

META-NET. (2020). *META-NET White Paper Series: Key Results and Cross-Language Comparison—META Multilingual Europe Technology Alliance*. http://www.meta-net.eu/whitepapers/key-results-and-cross-language-comparison

Mohammad, S. M. (2016). *NRC Emotion Lexicon*. http://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm

Mohammad, S., & Turney, P. (2013). Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, *29*. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Moj web dizajn. (2023). *HTML hrvatske kodne stranice*. https://www.mojwebdizajn.net/skriptni-jezici/vodic/html/html-hrvatske-kodne-stranice.php

NLTK Project. (2023). *NLTK: Sample usage for wordnet*. https://www.nltk.org/howto/wordnet.html

Pandurang Thakkar, G. (2022). *Cross-lingual sentiment analysis of official EU Slavic languages* [PhD Thesis, University of Zagreb. Faculty of Humanities and Social Sciences]. https://repozitorij.ffzg.unizg.hr/islandora/object/ffzg:7344

Plutchik, R. (1980). A general psychoevolutionary theory of emotion. U *Theories of emotion* (str. 3–33). Elsevier.

PolEval 2017. (2017). *PolEval 2017 Tasks*. http://2017.poleval.pl/index.php/tasks/

Poria, S. (2022). *MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversation* [Data set]. Deep Cognition and Language Research (DeCLaRe) Lab. https://github.com/declare-lab/*MELD*

Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., & Mihalcea, R. (2018). *MELD*: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*.

Python Software Foundation. (2023a). *csv2tsv: Convert csv file to tsv file* (1.1.0). https://github.com/not-dev/csv-to-tsv

Python Software Foundation. (2023b). *os—Miscellaneous operating system interfaces*. Python documentation. https://docs.python.org/3/library/os.html

Python Software Foundation. (2023c). *time—Time access and conversions*. Python documentation. https://docs.python.org/3/library/time.html

Qi, P., Dozat, T., Zhang, Y., & Manning, C. D. (2018). Universal Dependency Parsing from Scratch. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 160–170. https://nlp.stanford.edu/pubs/qi2018universal.pdf

Roca, M. (2022). *Označavanje i analiza sentimenta IMDB recenzija u englesko-hrvatskom podatkovnom skupu* [Info:eu-repo/semantics/masterThesis, University of Zagreb. Faculty of Humanities and Social Sciences. Department of information and Communication sciences]. https://urn.nsk.hr/urn:nbn:hr:131:089040

Stanford NLP Group. (2020). *Stanza – A Python NLP Package for Many Human Languages*. *Stanza*. https://stanfordnlp.github.io/*Stanza*/

Šikić, L. (2019, lipanj 3). *Analiza tekstova u Poslovnom dnevniku*. CroEcon. http://croecon.contentio.biz/post/analiza-teksta-poslovni-dnevnik/

Štrkalj Despot, K., & Krek, S. (2023). *Semantic role labeling in Slovenian and Croatian*. Institut za hrvatski jezik i jezikoslovlje. http://ihjj.hr/projekt/semantic-role-labeling-in-slovenian-and-croatian/66/

Thakkar, G., Preradovic, N. M., & Tadic, M. (2022). Multi-task learning for cross-lingual sentiment analysis. *arXiv preprint arXiv:2212.07160*.

van der Goot, R. (2021). We Need to Talk About train-dev-test Splits. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 4485–4494. https://doi.org/10.18653/v1/2021.emnlp-main.368

Znanje. (2023). *Hrvatski jezični portal*. https://hjp.znanje.hr/index.php?show=search

Analiza sentimenta u paralelnim korpusima hrvatskog i poljskog jezika

Sažetak

U području obrade jezika postoji nedostatak hrvatskog korpusa teksta označenog sa sentimentima koji omogućuje njihovu analizu. U radu je opisan proces stvaranja dva takva korpusa, tj. paralelnog hrvatsko-poljskog korpusa, a na temelju postojećih oznaka iz engleskog korpusa. Engleski izvornik je nastao na temelju televizijske serije *Friends*. Strojnom obradom u vlastitom *Python* programu, prevedeni su na hrvatski i poljski jezik tekstovi iz engleskog korpusa te izvršena njihova analiza. U radu su prikazani i dokumentirani problemi s kojima se je pri tome trebalo suočiti i koje je trebalo riješiti.

Ključne riječi: paralelni korpus, sentiment, strojna obrada jezika

Sentiment analysis in parallel corpora of the Croatian and Polish languages

Summary

In the field of natural language processing, there is a lack of a Croatian text corpus marked with sentiments, which enables their analysis. This paper describes the process of creating two such corpora, i.e., a parallel Croatian-Polish corpus, based on existing tags from the English corpus. The English original was created based on the TV Show *Friends*. Using machine processing, i.e., translation via the original Python program, texts from the English corpus were translated into Croatian and Polish equivalents and their analysis was carried out. The paper presents and documents the problems that had to be faced and solved during that process.

Key words: parallel corpus, sentiment, machine language processing

Dodatak: Programski kod za strojno prevođenje

```
# -*- coding: utf-8 -*-

import os
import time
import codecs
from csv2tsv import to_tsv
from deep_translator import GoogleTranslator as GT

izvorna_mapa="English"
odredisna_mapa={"hr":"Croatian", "pl":"Polish"}
for mapa in odredisna_mapa.values():
    if not os.path.isdir(mapa):
        os.makedirs(mapa)

izvorni_jezik="en"
odredisni_jezik=["hr", "pl"]

rjecnik={"anger_hr":"ljutnja",
    "disgust_hr":"gađenje",
    "fear_hr":"strah",
    "joy_hr":"radost",
    "neutral_hr":"neutralno",
    "sadness_hr":"tuga",
    "surprise_hr":"iznenađenje",
    "positive_hr":"pozitivno",
    "negative_hr":"negativno",
    "anger_pl":"złość",
    "disgust_pl":"obrzydzenie",
    "fear_pl":"strach",
    "joy_pl":"radość",
    "neutral_pl":"neutralne",
```

```
    "sadness_pl":"smutek",
    "surprise_pl":"zaskoczenie",
    "positive_pl":"pozytywne",
    "negative_pl":"negatywne",}

prijevod={"hr":
"Broj\tIzjava\tGovornik\tEmocija\tSentiment\tDijalog_ID\tIzjava_ID\tSezona\tEpizoda\tPocetno_vrijeme\tZavrsno_vrijeme",
         "pl": "Numer\tWyraz\tMówca\tEmocja\tSentyment\tDialog_ID\tWyraz_ID\tSezon\tOdcinek\tCzas_startu\tCzas_końca"}

svi_dokumenti=sorted(os.listdir(izvorna_mapa))
svi_dokumenti=[dokument for dokument in svi_dokumenti if ".csv" in dokument]
for dokument in svi_dokumenti:
    #print(dokument)
    if not os.path.exists(izvorna_mapa+"\\"+dokument.replace(".csv",".tsv")):
        to_tsv(izvorna_mapa+"\\"+dokument, izvorna_mapa+"\\"+dokument.replace(".csv",".tsv"), encoding="utf-8")
svi_dokumenti=[dokument.replace(".csv",".tsv") for dokument in svi_dokumenti]
for dokument in svi_dokumenti:
    for jezik in odredisni_jezik:
        if not os.path.exists(odredisna_mapa[jezik]+"\\"+dokument):
            print("Pocinje obrada", dokument, "za", jezik, "jezik.")
            with codecs.open(izvorna_mapa+"\\"+dokument, "r", encoding="utf-8", errors="ignore") as f:
                tekst=f.read().strip()
                tekst=tekst.split("\n")
                with codecs.open(odredisna_mapa[jezik]+"\\"+dokument,"w", encoding="utf-8", errors="ignore") as odrediste:
                    odrediste.write(prijevod[jezik]+"\n")
                    for num,red in enumerate(tekst[1:]):
                        if num%1000==0 and not num==0:
                            print("Google Translator pauza :))")
                            time.sleep(120)
                        podijeljeni_red=red.split('\t')
                        if len(podijeljeni_red)==11:
                            novi_red = podijeljeni_red
                            for brojac in range(len(novi_red)):
```

```
                            if brojac==1:
                                novi_red[brojac].replace('"','').replace('"','')
                                try:
                                    google_prijevod = GT(source=izvorni_jezik, target=jezik).translate(novi_red[brojac])
                                    if google_prijevod:
                                        novi_red[brojac] = google_prijevod
                                except:
                                    print("Google Prevoditelj nije uspio prevesti tekst:", novi_red[brojac])
                                    pass
                            elif brojac in [3,4]:
                                novi_red[brojac] = rjecnik[novi_red[brojac]+"_"+jezik]
                        novi_tekst="\t".join(novi_red)
                        #print(novi_tekst)
                        odrediste.write(novi_tekst+"\n")
                    else:
                        print("Nešto nije u redu sa zapisom:", podijeljeni_red)
            print("\tKraj obrade dateoteke", dokument, "za", jezik, "jezik")
        else:
            print(dokument, "za jezik", jezik, "već je obrađen")

print("Svi zadaci su izvršeni")
```