SVEUČILIŠTE U ZAGREBU

FILOZOFSKI FAKULTET

ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI

Ak. god. 2022./2023.

Mila Bikić, Valerija Bočkaj, Tea Krolo, Lorena Šubić

**Završni izvještaj analize sentimenta recenzija filmova**

Kolegij: Obrada prirodnog jezika

Zagreb, lipanj 2023.

Sadržaj

[1. Podaci o projektnom timu 1](#_Toc136774495)

[2. Detaljni podaci o projektu 1](#_Toc136774496)

[2.1. Prikupljanje i čišćenje korpusa 1](#_Toc136774497)

[2.2. Priprema neanotiranog korpusa 2](#_Toc136774498)

[3. Rezultati projekta 3](#_Toc136774499)

[3.1. Korpus 3](#_Toc136774500)

[3.2. Analiza sentimenta 4](#_Toc136774501)

[3.3. Slaganje među anotatorima 6](#_Toc136774502)

[4. Aktivnosti projekta 10](#_Toc136774503)

[5. Aktivnosti svakog člana projektnog tima 11](#_Toc136774504)

[6. Evidencija radnih sati svakog člana projektnog tima 13](#_Toc136774505)

[7. Završna evaluacija 16](#_Toc136774506)

[Literatura 18](#_Toc136774507)

[Popis slika i tablica 19](#_Toc136774508)

[Prilozi 20](#_Toc136774509)

[Prilog 1: Kod za *scrapeanje* mrežne stranice 20](#_Toc136774510)

[Prilog 2: Kod za segmentiranje rečenica i izradu Excel tablice 22](#_Toc136774511)

[Prilog 3: Brojanje pojavnica i različnica 23](#_Toc136774512)

# Podaci o projektnom timu

Projektni tim sastoji se od 4 članova. Jedan član pokriva tehnički aspekt projekta, odnosno prikupljanje podataka i sastavljanje korpusa, te koordinaciju grupe, tri člana su anotatori, od kojih dvoje analizira rezultate, a treća će pripremiti prezentaciju i izlagati. Svi članovi bavit će se čišćenjem podataka i kontrolom njihove kvalitete. Rad svih članova tima će biti praćen putem Excel tablica koje su prikazane u šestom poglavlju.

# Detaljni podaci o projektu

Projekt se sastoji od 3 glavne faze unutar kojih se nalazi nekolicina aktivnosti. Prva faza se sastoji od planiranja projekta, organizacije i podijele projektnih aktivnosti te definiranja ključnih rokova. U drugoj se fazi prikupljaju i čiste podatci, izrađuje se neanotirani korpus te se provodi njegova anotacija kao i kontrola kvalitete samog korpusa i rezultata anotacije. U trećoj fazi će se analizirati rezultati projekta te će se izraditi završni izvještaj s detaljima o tijeku projekta i rezultatima kao i prezentacija istih. Cilj ovog projekta je odrediti sentiment rečenicama koje čine recenzije filmova te izračunati razinu slaganja tri različita anotatora koja će označavati bez međusobnog dogovora. Rezultati tog postupka bit će anotirani korpus i same oznake rečenica koje su pridodali anotatori. Dobiveni rezultati će se analizirati s ciljem dobivanja uvida u poteškoće koje nastaju pri izradi korpusa i analizi sentimenta. Također, opisat će se završni format i veličina korpusa.

## Prikupljanje i čišćenje korpusa

Neanotirani korpus se prikupio u .tsv formatu (*tab separated values*) s domene https://www.recenzijefilmova.com/komedije-filmovi/ koristeći biblioteku *BeautifulSoup* za takozvano *scrapeanje* mrežnih stranica. Kod je pisan u *Python* distribuciji *PyCharm* primjenom Python jezika verzije 3.9. Prilikom prikupljanja korpusa prvotno je bilo potrebno istražiti kako su označeni potrebni elementi u HTML kodu. U tu svrhu se koristio alat *Inspect element* u *Mozilla Firefox* pretraživaču. Kod kojim se *scrapeala* mrežna stranica nalazi se u Prilogu 1. Zbog nekonzistentnosti u pisanju recenzija, prvu verziju korpusa je potrebno čistiti. Prilikom detaljnog pregleda neočišćenog korpusa, napravljen je popis nedostataka koji se najčešće pojavljuju te koje je najlakše riješiti programski.

Prilikom preuzimanja korpusa, na nekim mjestima se izgubio razmak iza interpunkcije. Problem je riješen primjenom regularnog izraza r"(?<=[.,:\-!?])(?=[^\s])" u Pythonu, koji razmak dodaje nakon znakova interpunkcije “.”, “,”, “:”, “-”, “!” te “?”, ali samo u slučaju interpunkciju ne slijedi razmak. Regularni izraz je primijenjen samo na *review* stupac .csv datoteke. Pritom je ispred točke u riječi “itd.” se dodao nepotreban razmak koji je riješen alatom *Find and replace* u programu *Notepad++*.

U sljedećem koraku započelo je ručno čišćenje korpusa u programu *Notepad++*. Svaki od tri anotatora je ručno očistio 25 recenzija tako da je sveukupno očišćeno 75 recenzija. Prvotno su se očistile generalne informacije o filmovima, npr. IMDb ocjena, godina izdavanja i slično, kojih se trebalo riješiti, u svrhu dobivanja isključivo rečenica recenzije.

Na kraju, provedena je kontrola kvalitete očišćenog korpusa. Primijećeno je da su na nekim mjestima prilikom čišćenja nehotice obrisane ocjene te su na tim mjestima ručno dodane. Nadalje, greškom su na nekim mjestima tabulatori zamijenjeni s razmakom te se i to ručno ispravilo kako bi što točnije procesirali očišćeni korpus u Pythonu u kasnijim koracima. Sljedeće, nadodani su „]“ na kraju teksta recenzije na mjestima gdje su greškom izbrisana. Prilikom čišćenja su neki članovi tima dodali novi razmak nakon svake recenzije kako bi si olakšali čišćenje podataka. Takvi razmaci su u ovom koraku uklonjeni.

## Priprema neanotiranog korpusa

Nakon što su svi podaci očišćeni, započela je priprema neanotiranog korpusa. Odlučeno je da će se anotarati u proračunskim tablicama, točnije, u *Microsoft Excel* programu. Prvo je bilo potrebno segmentirati rečenice u korpusu, što se izvršilo programski u *PyCharmu*. Za segmentaciju rečenica korišten je biblioteka za obradu prirodnog jezika *spaCy* koji ima paket za hrvatski jezik. Segmentirane rečenice i linkovi recenzija pohranjeni su u rječnik te su se ti podaci pohranili u Excel tablicu uz pomoć Python biblioteke *openpyxl* koji se koristi za funkcije čitanja i pisanja xlsx/xlsm/xltx/xltm datoteka. Kod za segmentaciju rečenica te za izradu Excel tablice se nalazi u Prilogu 2. Ovakvu .xlsx datoteku je bilo potrebno očistiti jer *spaCy* nije točno segmentirao sve rečenice. Za anotaciju su pripremljene 2029 rečenica, odnosno 59 recenzija. Korpus je opisan brojanjem pojavnica i različnica u biblioteci *spaCy* u programskom jeziku *Python.* Kod se nalazi u Prilogu 3. Na kraju, dodan je padajući izbornik s anotacijama:

* 1 ‒ *negative*: negativan sentiment;
* 2 ‒ *neutral*: neutralan sentiment;
* 3 ‒ *positive*: pozitivan sentiment;
* 4 ‒ *mixed*: dio rečenice je pozitivan, a dio negativan;
* 5 ‒ *sarcasam*: sarkastični i ironični komentari.

# Rezultati projekta

Rezultati ovog projekta obuhvaćaju neoznačeni korpus dobiven *scrapeanjem* stranice *Recenzije filmova i serija,* kao i označene rečenice recenzija koje su rezultat analize sentimenta od strane 3 anotatora prema 5-stupanjskoj ljestvici. Iz označenih rečenica izračunat je koeficijent slaganja anotatora preko Krippendorffove alfe, a detaljna analiza provedena je usporedbom oznaka u svakoj pojedinoj kategoriji (*negative, neutral, positive, mixed, sarcasm*).

## Korpus

Dobiveni neoznačeni korpus se nalazi u .tsv datoteci i sastoji se od 2029 rečenica. U korpusu se nalazi ukupno 45330 pojavnica i 10688 različnica. Unutar neoznačenog korpusa, svaka recenzija je smještena u zaseban redak s tabulatorom odvojenim stupcima koji sadrže sljedeće informacije: *review, overall\_score, first\_impression, review\_date, reviewer i home\_url.* Podaci u završnoj verziji neanotiranog korpusa su pažljivo očišćeni. Međutim, očišćeni podaci sami po sebi nisu dovoljni za analizu, već je potrebno označiti rečenice. Nakon što se rečenice označe, dobivamo označeni korpus koji je spreman za analizu. Označeni korpus se nalazi u .xlsx datoteci. Svaka recenzija je segmentirana na rečenice, a svaka rečenica se nalazi u zasebnom retku. Redovi su poredani prema redoslijedu pojavljivanja kako bi označivači imali kontekstualni uvid. Svaki redak sadrži URL recenzije, samu rečenicu i padajući izbornik s ponuđenim pet sentimenta. Prazan redak se koristi kao graničnik između recenzija kako bi se održala jasnoća i preglednost korpusa.

## Analiza sentimenta

Kod procesa označavanja anotatori su naišli na određene izazove. Nadalje su navedeni primijećeni problemi kod anotacije.

Prvo, kod opisa radnje filmova ponekad je bilo teško razlučiti između činjenica i autorskih negativnih ili pozitivnih komentara o radnji. Niže su navedeni primjeri takvih komentara.

**Primjer 1:** Pronađite ih sve u ovom božićnom filmu prilagođenom djeci.

Dok se čini da je to objektivna činjenica o sadržaju filma, postoji mogućnost da je autor izrazio subjektivan stav o tome koliko je zanimljivo ili uspješno to što se u filmu trebaju pronaći određeni likovi.

**Primjer 2:** Već smo nekoliko puta imali priliku da Zemlju unište kometi.

U ovom primjeru, anotatorima je bilo teško razlučiti da li je ova izjava samo objektivna činjenica o čestom motivu u filmskoj industriji ili je to negativan komentar autora koji implicira da je prevelika učestalost ovakvih scena loša ili dosadna.

**Primjer 3:** Iako, ako moram biti kritičan, akcije nisu bile baš spektakularne, ali svejedno su bile dobre.

Kod ovog primjera anotatori su teško razlučivali negativni i pozitivni sentimenta. Autor izražava kritičan stav prema akcijskim scenama, ali istovremeno izražava pozitivan stav o ukupnom dojmu filma. Treba pažljivo analizirati kontekst i ton izjave kako bi odredili dominantni sentiment i pravilno označili rečenicu.

Pored gore navedenih činjenica, nedostatak povezanosti u načinu pisanja stvara velike poteškoće u razumijevanju veze između rečenica. Na primjer, određena rečenica može se činiti pozitivnom, ali kad se uzme u obzir širi kontekst, zapravo nije. U donjem primjeru nalazi se rečenica koja stvara probleme zbog nedostatka povezanosti u pisanju, konkretno ta rečenica uopće nema glagola ili pridjeva koji bi ukazivali na njen sentiment.

**Primjer 4:** Alternativno, to je pretpostavka koja bi se mogla smatrati uspjehom.

U ovoj rečenici nedostaje izričit glagol ili pridjev koji bi nam dao jasno razumijevanje sentimenta. Izražava se neka vrsta pretpostavke, ali sam sentiment te pretpostavke ostaje neodređen. To stvara problem, jer je teško odrediti točan sentiment ili značenje ove rečenice bez jasnih indikacija.

Također, predstavljale su problem i rečenice s malo riječi. Ove rečenice sadrže vrlo malo informacija, a nedostatak dodatnih detalja, glagola ili pridjeva otežava anotatorima da odrede njihov sentiment. Iako se može naslutiti neki kontekst, sam tekst nije dovoljno opisan da bi se jasno utvrdilo je li sentiment pozitivan, negativan ili neutralan, kao što možemo vidjeti na sljedećim primjerima:

**Primjer 5:** OK, idemo dalje.

**Primjer 6:** Vaš izbor.

**Primjer 7:** I kako bi se priča završila?

**Primjer 8:** Otkrijmo, hoćemo li?

**Primjer 9:** Dobro?

Donji primjer dobro ilustrira situaciju gdje nedostatak konteksta često otežava razumijevanje sentimenta riječi koje mogu biti i pozitivne i negativne.

**Primjer 10:** U početku sam imao osjećaj da naslov filma zvuči poznato.

Izraz "u početku sam imao osjećaj" ne pruža dovoljno informacija o tome je li osjećaj pozitivan ili negativan. Riječ "poznato" također može biti interpretirana na više načina. Može ukazivati na pozitivno iskustvo prepoznavanja ili na negativno iskustvo koje sugerira da je nešto već viđeno ili predvidljivo. Bez daljnjeg konteksta, nije moguće jasno odrediti sentiment izražen u ovoj rečenici.

## Slaganje među anotatorima

Za evaluaciju usklađenosti među anotatorima koristila se statistička metoda poznata kao Krippendorffova alfa. Ova metoda omogućuje analizu podataka čak i kada postoje neki nedostatni podaci, što je važno u ovom istraživanju. Štoviše, Krippendorffova alfa se može primijeniti na različite kategorije i kada su u projekt uključena više od dva anotatora, kao što je slučaj u ovdje (Statistics How To, 2016). Prema Krippendorffu, vrijednost alfa pruža objektivan pokazatelj prihvatljivosti usklađenosti između anotatora. Ako je vrijednost alfa manja od 0.667, to ukazuje na nedovoljnu usklađenost. Kada je vrijednost alfa između 0.667 i 0.8, može se zaključiti da postoji djelomična prihvatljivost usklađenosti. Međutim, kada je vrijednost alfa veća od ili jednaka 0.8, s visokom sigurnošću je moguće reći da postoji pouzdana usklađenost među anotatorima (Batanović et al., 2020). Ova informacija pomaže u razumijevanju rezultata ovog istraživanja i u donošenju relevantnih zaključaka.

Kao rezultat, odabrana je Krippendorffova alfa jer nam omogućava usporedbu rezultata svih triju anotatora. Uz to, korišten je alat pod nazivom ReCal3: Reliability for 3+ Coders (dfreelon.org, n.d.) za izračunavanje Krippendorffove alfe. Ovaj alat, koji je razvio stručnjak dr. sc. Deen Freelon, ne samo da pruža Krippendorffovu alfu, već također nudi i Fleissov Kappa koeficijent i Cohenov Kappa koeficijent. Sve ove navedene funkcije alata omogućavaju da se rezultati ovog istraživanja pravilno i što točnije analiziraju i interpretiraju.

Konačno, dobiveni rezultati analize prikazani su na Slici 1 i Slici 2. Na njima je vizualno prikazana razina usklađenosti između anotatora, te pružaju dodatni uvid u istraživačke rezultate.

**Slika na kojoj se prikazuje tekst, Font, snimka zaslona, crta

Opis je automatski generiran**

Slika 1. Postotak slaganja anotatora.

**Slika na kojoj se prikazuje tekst, snimka zaslona, Font, crta

Opis je automatski generiran**

Slika 2. Krippendorffova alfa.

Na Slici 1 prikazani su rezultati koji pružaju uvid u postotak usklađenosti između anotatora. Ovi rezultati su bitni jer prikazuju suglasnost između anotatora u svojim anotacijama. Najveća razina podudarnosti, koja iznosi 77.329%, postoji između Anotatora 1 i Anotatora 2. Ova visoka razina suglasnosti između ta dva anotatora ukazuje na njihovu sličnu interpretaciju i razumijevanje anotiranog sentimenta rečenica. Najmanja razina suglasnosti, koja iznosi samo 70.133%, zabilježena je između Anotatora 1 i Anotatora 3. To ukazuje na manju podudarnost u njihovim interpretacijama sentimenta rečenica. Razlike u njihovim anotacijama mogu biti rezultat različitih stajališta, interpretacija ili razumijevanja. Suglasnost između Anotatora 2 i Anotatora 3 iznosi 74.421%. To ukazuje na umjerenu razinu suglasnosti između ta dva anotatora u pogledu sentimenta rečenica. Općenito, viša razina suglasnosti između anotatora ukazuje na veću konzistentnost i sličnost u njihovim anotacijama. Niža razina suglasnosti može ukazivati na veću varijabilnost ili različita tumačenja među anotatorima. Slika 2 prikazuje vrijednost Krippendorffove alfe, koja iznosi 0.605. Prema Krippendorffu, ova vrijednost označava neprihvatljivu razinu suglasnosti među anotatorima. Premda je razina suglasnosti nešto niža, takav rezultat nije iznenađujući s obzirom na izrazito nespretne napisane rečenice koje su ponekad stvarale velike poteškoće anotatorima, unatoč njihovom poznavanju hrvatskog jezika. Iako postoji prostor za poboljšanje, posebno u pogledu nespretnih rečenica, ova analiza nam pomaže da procijenimo pouzdanost i relevantnost dobivenih rezultata. S obzirom na ove rezultate, moguće je daljnje razvijanje strategije za poboljšanje usklađenosti među anotatorima, koje bi u budućim istraživanjima osigurali točnije i dosljednije rezultate analize.

Kod primjera istraživanja *Weighted Krippendorff’s alpha is a more reliable metrics for multi- coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation (2104)*, autori Jean-Yves Antoine, Jeanne Villaneau i Anaïs Lefeuvre istražuju upotrebu Krippendorffove alfa (α) kao mjere pouzdanosti za označavanje podataka kod više anotatora. Istraživanje je provedeno na korpusu filmskih recenzija, koji su relativno kratki tekstovi napisani na specijaliziranim francuskim mrežnim stranicama. Recenzije su se odnosile na isti francuski film, te korpus sadrži 183 rečenice. Anotacija je provedena od strane 25 sudionika, a  Krippendorffova alfa je dala zadovoljavajuće mjere pouzdanosti s vrijednostima od 0,69. Možemo uočiti da je u tom slučaju Krippendorffova alfa veća, te je usklađenost s anotatorima pouzdana. Tome pridodaje činjenica da Krippendorffova alfa ima tendenciju da bude veća kada je broj anotatora veći, kako mjeri suglasnost među anotatorima i uzima u obzir slučajno odstupanje i sistematske razlike među njihovim ocjenama. Kada ima više anotatora, postoji veća vjerojatnost da će slučajne varijacije između njihovih ocjena biti izbalansirane ili uklonjene, što može rezultirati većim vrijednostima Krippendorffove alfe. Također, slično kao i u ovom projektu, u istraživanju je korištena ljestvica od pet stupnjeva: vrlo negativno, umjereno negativno, bez emocija (neutralno),umjereno pozitivno i vrlo pozitivno. Ipak, potrebno je naglasiti da takva ljestvica može biti podložna pogreškama jer ljudi mogu različito percipirati i ocjenjivati razinu negativnosti i pozitivnosti. Razina subjektivnosti može rezultirati velikom varijacijom u interpretaciji i označavanju podataka.

Nadalje, postoji ukupno 85 rečenica u kojima se sva tri anotatora nisu složila. U Tablici 1 prikazani su primjeri nekih od tih rečenica zajedno s oznakama koje su im dodijeljene.

|  |  |  |  |
| --- | --- | --- | --- |
| **REČENICA** | **ANOTATOR 1** | **ANOTATOR 2** | **ANOTATOR 3** |
| Btw, film traje oko 2 sata s 1 scenom na sredini odjavne špice koja je bila zanimljiva i 1 scenom na kraju odjavne špice koja je završila osobno iznenađujućim posljednjim riječima prikazanim na ekranu. | *3 ‒ positive* | *2 ‒ neutral* | *4 ‒ mixed* |
| U početku sam pomalo sumnjao hoće li ovaj film biti ugodan kao i njegov prethodnik. | *1 ‒ negative* | *3 ‒ positive* | *4 ‒ mixed* |
| Ali onaj za koga sam mislio da će se poprilično istaknuti biti Max Malas koji je igrao Jakea Farmera. | *3 ‒ positive* | *4 ‒ mixed* | *2 ‒ neutral* |
| Kao moderna horor adaptacija, ali vjerojatno ne toliko nevjerojatna da bih imao dobre riječi na jeziku kako bi je promovirao. | *4 ‒ mixed* | *1 ‒ negative* | *3 ‒ positive* |
| Dapače, to je razlog koji se koristi za pokretanje pripovijesti i razlog koji stoji iza glavnih avantura. | *1 ‒ negative* | *2 ‒ neutral* | *3 ‒ positive* |

Tablica 1. Primjeri rečenica koje imaju različite oznake kod sva 3 anotatora.

U Tablici 2 prikazan je broj oznaka za svaku od pet anotacija (*negative, neutral, positive, mixed, sarcasm*) od sva tri anotatora.

|  |  |  |  |
| --- | --- | --- | --- |
| **OZNAKA** | **ANOTATOR 1** | **ANOTATOR 2** | **ANOTATOR 3** |
| *1 ‒ negative* | 189 | 155 | 85 |
| *2 ‒ neutral* | 700 | 766 | 656 |
| *3 ‒ positive* | 968 | 850 | 871 |
| *4 ‒ mixed* | 165 | 256 | 410 |
| *5 ‒ sarcasm* | 7 | 2 | 7 |

Tablica 2. Broj rečenica za svaku oznaku za pojedinog anotatora.

Vidljivo je da su sva tri anotatora najveći broj rečenica ocijenila kao pozitivne, dok je najmanji broj rečenica označen kao sarkazam. Sva tri anotatora imaju sličnu percepciju pozitivnog sentimenta, ali s manjim varijacijama u broju dodijeljenih anotacija. Također, primjećuje se da sva tri anotatora imaju sličnu percepciju negativnog sentimenta, ali s malim varijacijama u broju dodijeljenih anotacija. Anotator 1 je dodijelio najviše negativnih anotacija, dok je Anotator 3 dodijelio najmanje. Neutralni sentiment je najuravnoteženiji među sva tri anotatora.. Anotator 2 je dodijelio najviše neutralnih anotacija, dok je Anotator 1 dodijelio nešto manje, a Anotator 3 najmanje. Međutim, uočene su značajne razlike u broju rečenica označenih kao mješovite među anotatorima. Konkretno, Anotator 1 je dodijelio 165 rečenica s oznakom mješovito, dok je Anotator 2 dodijelio 256 rečenica, a Anotator 3 čak 410 rečenica s tom oznakom. Također, primijećene su značajne razlike u označavanju negativnih rečenica. Anotator 1 je označio 189 rečenica kao negativne, dok je Anotator 2 označio 155 rečenica, a Anotator 3 samo 85 rečenica s tom oznakom. Kod sarkazma vidljivi su mali brojevi anotacija, isti između Anotatora 1 i Anotatora 3, dok je Anotator 2 dodijelio najmanje anotacija za ovu kategoriju. Zanimljivo je napomenuti da nijedna rečenica označena kao sarkazam ne poklapa se između anotatora. Ove razlike u broju rečenica za određene oznake ukazuju na varijaciju u tumačenju i ocjenjivanju među anotatorima. Stoga je važno provesti daljnja istraživanja da bi se razumjeli uzroci ovih razlika i unaprijedili suglasnost među anotatorima, te tako postigli veću konzistentnost u procesu ocjenjivanja.

# Aktivnosti projekta

U sklopu ovog projekta, Anotator 1, Anotator 2 i Anotator 3 imali su glavni zadatak anotirati sentimente u tekstualnim recenzijama. Osim anotacije sentimenta, svaki anotator je obavljao i dodatne zadatke. Anotator 1 je bio odgovoran za planiranje projekta, organizaciju projektnih aktivnosti, čišćenje podataka, kontrolu očišćenog korpusa i izradu završnog izvještaja. Anotator 2 je također sudjelovao u planiranju projekta i organizaciji aktivnosti, te je obavio čišćenje podataka i izradu završne prezentacije, uključujući i izlaganje. Anotator 3 je sudjelovao u planiranju projekta, organizaciji projektnih aktivnosti i čišćenju podataka.

Prvi korak u provedbi projekta bio je detaljno planiranje. Definirani su ciljevi projekta, postavljen je vremenski okvir i dodijeljene su odgovornosti članovima tima. Također su osigurani potrebni resursi, kao što su pristup relevantnima podacima i alatima za anotaciju. Stvorena je *Google Docs* datoteka za završni izvještaj u koju je svaki član tima opisao svoj rad. Također je datoteka za evidenciju sati pojedinog člana podijeljena sa svim članovima tima pomoću servisa *Microsoft 365.* Nakon planiranja, uslijedila je organizacija projektnih aktivnosti prema postavljenom rasporedu. Koordinirana je suradnja između članova tima i redovito su održavani sastanci s ciljem praćenja napretka projekta i rješavanjem eventualnih izazove. Sljedeći korak bio je čišćenje podataka. Proveden je postupak uklanjanja nepotrebnih znakova, ispravljanja grešaka i standardizacije formata podataka. Kroz ovaj proces, osigurana je kvaliteta i konzistentnost podataka koji su korišteni za anotaciju sentimenata. Nakon čišćenja podataka, započet je proces anotacije sentimenata. Proces označavanja sentimenata sastojao se od dva potprocesa: razumijevanja sadržaja rečenice i procjene sentimenta te pridavanja odgovarajuće oznake sentimenta rečenici. Sentiment je anotiran na 5-stupanjskoj ljestvici: negativan, neutralan, pozitivan, miješan (dio rečenice je pozitivan, a dio negativan) te sarkastični i ironični komentari. Anotacija dokumenta je izvršena u *Microsoft Excel* datoteci, gdje su rečenice bile segmentirane u zasebne redove. Također, provela se kontrola kvalitete kojom su identificirane eventualne pogreške ili nedosljednosti i na vrijeme ispravljene. Nakon završetka provjere kvalitete, tim je pripremio izvještaj koji detaljno opisuje sve provedene korake, metodologiju anotacije, korištene alate i postignute rezultate. Također je pripremljena i izrađena prezentacija koja je obuhvatila ciljeve projekta, metodologiju, postignute rezultate i zaključke. Vizualni elementi prezentacije su pažljivo odabrani kako bi se projektni rad i rezultati prezentirali na jasan i pregledan način.

# Aktivnosti svakog člana projektnog tima

U sljedećoj *Tablici 3* možemo vidjeti zadatke koje je obavljao svaki član tima.

|  |  |
| --- | --- |
| **ČLAN TIMA** | **ZADACI** |
| **Mila Bikić** | - Planiranje projekta  - Organizacija projektnih aktivnosti  - Prikupljanje podataka  - Čišćenje podataka  - Razvoj neanotiranog korpusa  - Kontrola očišćenog korpusa  - Izrada završnog izvještaja |
| **Valerija Bočkaj (Anotator 1)** | - Planiranje projekta  - Organizacija projektnih aktivnosti  - Čišćenje podataka  - Kontrola očišćenog korpusa  - Anotacija  - Izrada završnog izvještaja |
| **Tea Krolo (Anotator 2)** | - Planiranje projekta  - Organizacija projektnih aktivnosti  - Čišćenje podataka  - Anotacija  - Izrada završne prezentacije  - Izlaganje prezentacije |
| **Lorena Šubić (Anotator 3)** | - Planiranje projekta  - Organizacija projektnih aktivnosti  - Čišćenje podataka  - Anotacija |

Tablica 3. Obavljeni zadaci svakog člana tima.

# Evidencija radnih sati svakog člana projektnog tima

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TJEDAN** | **I** | **II** | **III** | **IV** | **V** |  |
| Opis projektnih aktivnosti | 08.05-14.05. | 15.05.-21.05. | 22.05.-28.05. | 29.05.-04.06. | 05.06.-12.06. |  |
| 1. FAZA PROJEKTA |  |  |  |  |  | 3 |
| Planiranje projekta | 2 |  |  |  |  | 2 |
| Organizacija projektnih aktivnosti | 1 |  |  |  |  | 1 |
| 2. FAZA PROJEKTA |  |  |  |  |  | 14 |
| Prikupljanje podataka | 3 |  |  |  |  | 3 |
| Čišćenje podataka |  | 3 |  |  |  | 3 |
| Razvoj neanotiranog korpusa |  | 3 | 3 |  |  | 6 |
| Anotacija korpusa |  |  |  |  |  | 0 |
| Kontrola kvalitete |  |  | 2 |  |  | 2 |
| 3. FAZA PROJEKTA |  |  |  |  |  | 3 |
| Analiza projekta |  |  |  |  |  | 0 |
| Završni izvještaj |  |  |  | 3 |  | 3 |
| Izrada završne prezentacije |  |  |  |  |  | 0 |
| **UKUPNO** | **6** | **6** | **5** | **3** | **0** | **20** |

Tablica 4. Evidencija radnih sati Mile Bikić.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TJEDAN** | **I** | **II** | **III** | **IV** | **V** |  |
| Opis projektnih aktivnosti | 08.05-14.05. | 15.05.-21.05. | 22.05.-28.05. | 29.05.-04.06. | 05.06.-12.06. |  |
| 1. FAZA PROJEKTA |  |  |  |  |  | 3 |
| Planiranje projekta | 2 |  |  |  |  | 2 |
| Organizacija projektnih aktivnosti | 1 |  |  |  |  | 1 |
| 2. FAZA PROJEKTA |  |  |  |  |  | 11 |
| Prikupljanje podataka |  |  |  |  |  | 0 |
| Čišćenje podataka |  | 3 |  |  |  | 3 |
| Razvoj neanotiranog korpusa |  |  |  |  |  | 0 |
| Anotacija korpusa |  | 2 | 3 |  |  | 5 |
| Kontrola kvalitete |  |  | 2 | 1 |  | 3 |
| 3. FAZA PROJEKTA |  |  |  |  |  | 4 |
| Analiza projekta |  |  |  | 2 |  | 2 |
| Završni izvještaj |  |  |  | 2 |  | 2 |
| Izrada završne prezentacije |  |  |  |  |  | 0 |
| **UKUPNO** | **3** | **5** | **5** | **5** | **0** | **18** |

Tablica 5. Evidencija radnih sati Valerije Bočkaj.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TJEDAN** | **I** | **II** | **III** | **IV** | **V** |  |
| Opis projektnih aktivnosti | 08.05-14.05. | 15.05.-21.05. | 22.05.-28.05. | 29.05.-04.06. | 05.06.-12.06. |  |
| 1. FAZA PROJEKTA |  |  |  |  |  | 3 |
| Planiranje projekta | 2 |  |  |  |  | 2 |
| Organizacija projektnih aktivnosti | 1 |  |  |  |  | 1 |
| 2. FAZA PROJEKTA |  |  |  |  |  | 8 |
| Prikupljanje podataka |  |  |  |  |  | 0 |
| Čišćenje podataka |  | 3 |  |  |  | 3 |
| Razvoj neanotiranog korpusa |  |  |  |  |  | 0 |
| Anotacija korpusa |  | 2 | 3 |  |  | 5 |
| Kontrola kvalitete |  |  |  |  |  | 0 |
| 3. FAZA PROJEKTA |  |  |  |  |  | 2 |
| Analiza projekta |  |  |  |  |  | 0 |
| Završni izvještaj |  |  |  |  |  | 0 |
| Izrada završne prezentacije |  |  |  | 2 |  | 2 |
| **UKUPNO** | **3** | **5** | **3** | **2** | **0** | **13** |

Tablica 6. Evidencija radnih sati Tee Krolo.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **TJEDAN** | **I** | **II** | **III** | **IV** | **V** |  |
| Opis projektnih aktivnosti | 08.05-14.05. | 15.05.-21.05. | 22.05.-28.05. | 29.05.-04.06. | 05.06.-12.06. |  |
| 1. FAZA PROJEKTA |  |  |  |  |  | 3 |
| Planiranje projekta | 2 |  |  |  |  | 2 |
| Organizacija projektnih aktivnosti | 1 |  |  |  |  | 1 |
| 2. FAZA PROJEKTA |  |  |  |  |  | 8 |
| Prikupljanje podataka |  |  |  |  |  | 0 |
| Čišćenje podataka |  | 3 |  |  |  | 3 |
| Razvoj neanotiranog korpusa |  |  |  |  |  | 0 |
| Anotacija korpusa |  | 2 | 3 |  |  | 5 |
| Kontrola kvalitete |  |  |  |  |  | 0 |
| 3. FAZA PROJEKTA |  |  |  |  |  | 1 |
| Analiza projekta |  |  |  |  |  | 0 |
| Završni izvještaj |  |  |  | 1 |  | 1 |
| Izrada završne prezentacije |  |  |  |  |  | 0 |
| **UKUPNO** | **3** | **5** | **3** | **1** | **0** | **12** |

Tablica 7. Evidencija radnih sati Lorene Šubić.

# Završna evaluacija

Projektni tim sastoji se od četiri člana, od kojih je jedan zadužen za tehnički aspekt projekta i koordinaciju, dok su preostala tri člana anotatori. Tim je pratio rad putem Excel tablica i ispisanog izvještaja.

Prikupljeno je 231 recenzija pomoću biblioteke *BeautifulSoup* u .tsv formatu. Za čišćenje podataka, svaki od tri anotatora dobio je 25 recenzija, ukupno 75 recenzija, koje su očišćene u aplikaciji *Notepad++.* Nakon čišćenja, korpus je segmentiran u rečenice pomoću *spaCy* biblioteke za obradu prirodnog jezika. Segmentirane rečenice su pohranjene u rječnik i spremljene u Excel tablicu korištenjem biblioteke *openpyxl*.

Nakon pripreme neoznačenog korpusa, započela je anotacija u Microsoft Excelu. Pripremljeno je 2029 rečenica iz 59 recenzija za anotaciju. Dodan je padajući izbornik s pet sentimenta: negativan, neutralan, pozitivan, mješovit i sarkastičan. Analiza sentimenta je izazvala određene poteškoće za anotatore, uključujući razlučivanje činjenica od autorskih komentara o radnji filma, razumijevanje konteksta i tona izjava te nedostatak povezanosti u načinu pisanja. Također, rečenice s manjim brojem riječi bile su problematične zbog nedostatka konteksta za određivanje sentimenta.

Za evaluaciju usklađenosti među anotatorima korištena je statistička metoda Krippendorffova alfa, koja je rezultirala vrijednošću od 0.605, ukazujući na djelomičnu prihvatljivost usklađenosti. Najveća razina suglasnosti (77.329%) postignuta je između Anotatora 1 i Anotatora 2. Od ukupno 2029 rečenica, 85 rečenica nije dobilo suglasne oznake od sva tri anotatora. Najveći broj rečenica označen je kao pozitivan sentiment, dok je najmanji broj označen kao sarkazam. Primijećene su značajne razlike u označavanju mješovitih i negativnih rečenica među anotatorima.

Ovi rezultati pružaju uvid u varijaciju u tumačenju i ocjenjivanju među anotatorima te naglašavaju potrebu za daljnjim istraživanjem kako bi se postigla veća konzistentnost u procesu ocjenjivanja.

# Literatura

Batanović, V., Cvetanović, M., & Nikolić, B. (2020). A versatile framework for resource-limited sentiment articulation, annotation, and analysis of short texts. *PLOS ONE*, *15*(11), e0242050.

Jean-Yves Antoine, Jeanne Villaneau, Anaïs Lefeuvre. Weighted Krippendorff’s alpha is a more reliable metrics for multi- coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation.. EACL 2014, Apr 2014, Gotenborg, Sweden. 10 p. ffhal-00998303f.

*ReCal3: Reliability for 3+ Coders – Deen Freelon, Ph.D.* (n.d.). Dfreelon.org. Pristupljeno 30. svibnja, 2023, sa <http://dfreelon.org/utils/recalfront/recal3/>.

Statistics How To (2016). Krippendorff’s Alpha Reliability Estimate: Simple Definition. Statistics How To. Pristupljeno 31. svibnja 2023, sa <https://www.statisticshowto.com/krippendorffs-alpha/>.

# Popis slika i tablica

[Slika 1. Postotak slaganja anotatora. 6](#_Toc136774514)

[Slika 2. Krippendorffova alfa. 7](#_Toc136774515)

[Tablica 1. Primjeri rečenica koje imaju različite oznake kod sva 3 anotatora. 9](#_Toc136774521)

[Tablica 2. Broj rečenica za svaku oznaku za pojedinog anotatora. 9](#_Toc136774522)

[Tablica 3. Obavljeni zadaci svakog člana tima. 13](#_Toc136774523)

[Tablica 4. Evidencija radnih sati Mile Bikić. 14](#_Toc136774524)

[Tablica 5. Evidencija radnih sati Valerije Bočkaj. 14](#_Toc136774525)

[Tablica 6. Evidencija radnih sati Tee Krolo. 15](#_Toc136774526)

[Tablica 7. Evidencija radnih sati Lorene Šubić. 16](#_Toc136774527)

# Prilozi

## Prilog 1: Kod za *scrapeanje* mrežne stranice

from bs4 import BeautifulSoup as bs  
import requests  
import csv  
  
MAINURL = "https://www.recenzijefilmova.com/komedije-filmovi/"  
PAGENUMBER = 25  
  
  
def get\_links():  
 *"""Funkcija get\_links() dohvaća sve linkove sa prvih BROJ\_STRANICA sa MAIN\_URL-a te ih vraća u obliku liste."""* links = []  
 for i in range(1, PAGENUMBER + 1):  
 url = MAINURL + "page/" + str(i)  
 print("Fetching links from page " + str(i))  
 response = requests.get(url)  
 html = response.content  
 soup = bs(html, "lxml")  
 items = soup.find\_all("div", {"id": "tdi\_39"})  
 for item in items:  
 for h3 in item.find\_all("h3", {"class": "entry-title td-module-title"}):  
 link = h3.find("a", href=True)["href"]  
 if "recenzija" in link:  
 links.append(link)  
  
 return links  
  
  
def get\_review\_data(url):  
 *"""Funkcija vraća sljedeće sa određenog URL-a u obliku dictionary: author, date, score, impression, review."""* review\_dict = {}  
 response = requests.get(url)  
 html = response.content  
 soup = bs(html, "lxml")  
 print("Fetching review from " + url)  
  
 author = ""  
 author\_elem = soup.find("a", {"class", "tdb-author-name"})  
 if author\_elem is not None:  
 author = author\_elem.getText()  
 review\_dict["author"] = author  
  
 time = soup.find("time", {"class": "entry-date updated td-module-date"})  
 date = time.getText()  
 review\_dict["date"] = date  
  
 score\_elem = soup.find("div", {"class": "td-review-final-score"})  
 score = ""  
 if score\_elem is not None:  
 score = score\_elem.getText()  
 review\_dict["score"] = score  
  
 impression = ""  
 impression\_elem = soup.find("div", {"class": "td-review-summary-content"})  
 if impression\_elem is not None:  
 impression = impression\_elem.getText()  
 review\_dict["impression"] = impression  
  
 div = soup.find("div", {"data-td-block-uid": "tdi\_68"}).find("div", {"class": "tdb-block-inner td-fix-index"})  
 review = ""  
 for child in div.findChildren(recursive=False):  
 if child.name == "p":  
 review += child.getText(strip=True).replace("<br>", "").replace("</br>", "")  
  
 if child.name == "iframe":  
 break  
 review = "[" + review + "]"  
 review\_dict["review"] = review  
 review\_dict["url"] = url  
  
 return review\_dict  
  
  
with open("osvrti\_opj.tsv", "w", newline = "", encoding = "utf-8") as review\_file:  
 review\_writer = csv.writer(review\_file, delimiter='\t', quotechar='"', quoting=csv.QUOTE\_MINIMAL)  
 review\_writer.writerow(["review", "overall\_score", "first\_impression", "review\_date", "reviewer", "home\_url"])  
  
 links = get\_links()  
 print("Found " + str(len(links)) + " links")  
 for link in links:  
 review = get\_review\_data(link)  
 review\_writer.writerow([review.get("review"), review.get("score"), review.get("impression"), review.get("date"), review.get("author"), review.get("url")])

## Prilog 2: Kod za segmentiranje rečenica i izradu Excel tablice

import spacy  
import csv  
import openpyxl  
  
nlp = spacy.load('hr\_core\_news\_sm')  
datoteka = "C:\\Users\\milab\\OneDrive - FFUNIZG\\FFZG\\IV\\Obrada prirodnog jezika\\Projekt\\osvrti\_opj.tsv"  
  
def extract\_first\_column(tsv\_file):  
 first\_column = []  
 with open(tsv\_file, "r", encoding="utf-8") as file:  
 reader = csv.reader(file, delimiter="\t")  
 for row in reader:  
 if row:  
 row = row[0].strip("[]")  
 first\_column.append(row)  
 return first\_column  
  
recenzije = extract\_first\_column(datoteka)  
  
recenice = []  
  
for row in recenzije:  
 doc = nlp(row)  
 rec\_recenice = [sent.text for sent in doc.sents]  
 for recenica in rec\_recenice:  
 recenice.append(recenica)  
 recenice.append("\n")  
  
workbook = openpyxl.load\_workbook("korpus\_anotacija\_final.xlsx")  
worksheet = workbook.active  
  
column = "A"  
for i, value in enumerate(recenice, start=1):  
 cell=f"{column}{i}"  
 worksheet[cell] = value  
  
workbook.save("korpus\_anotacija\_final.xlsx")

## Prilog 3: Brojanje pojavnica i različnica

import spacy  
import pandas as pd  
nlp = spacy.load("hr\_core\_news\_sm")  
  
data\_frame = pd.read\_excel("korpus\_anotacija\_final.xlsx")  
second\_column = data\_frame.iloc[:, 1].astype(str).values.tolist()  
second\_column\_string = " ".join(second\_column)  
  
print(len(second\_column\_string))  
  
doc = nlp(second\_column\_string)  
  
token\_count = len(doc)  
  
print("Token count: ", token\_count)  
  
out = []  
  
seen = set()  
  
for word in doc:  
 if word.text not in seen:  
 out.append(word)  
 seen.add(word.text)  
  
print(f"Unique tokens: {len(out)}")