Analiza mnijenja u podatkovnom skupu SentiGMapsCro

Gašparić, Brigita

SVEUČILIŠTE U ZAGREBU

FILOZOFSKI FAKULTET

ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI

SMJER NASTAVNIČKA INFORMATIKA

Ak. god. 2023./2024.

Brigita Gašparić

Analiza mnijenja u podatkovnom skupu SentiGMapsCro

Diplomski rad

Mentorica: prof.dr.sc. Nives Mikelić Preradović

Zagreb, rujan 2024.

Izjava o akademskoj čestitosti

Izjavljujem 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.

Za moju obitelj.

Sadržaj

1. Uvod

Jezik je jedna od temeljenih odrednica ljudskosti koja nas, ljude, razlikuje od ostatka živog svijeta. Često se kaže da je jezik živi organizam (Selimović, 1968), zbog čega je i samo proučavanje jezika neprestano sakupljanje informacija i promjena u korištenju istog.

Za razliku od zvučne komunikacije ili konkretnije, govora, pismena komunikacija nije vremenski ograničena (Malmberg, 1974), te su upravo zbog toga obavijesti prenesene pismenim putem savršena baza za promatranje suvremenog korištenja jezika. Naime, mi jezikom ne prenosimo samo obavijesti, već i svoja razmišljanja, osjećaje, stavove te emocionalno stanje; točnije, prenosimo naše sentimente.

Sentiment ili mnijenje jest, u suštini, subjektivni doživljaj naše stvarnosti – sve informacije koje dijelimo su na neki način označene našim osobnim i vrlo subjektivnim doživljajem, pa tako i informacije koje dijelimo putem Interneta. Otkad je internetska revolucija preobrazila naše živote i pružila nam cijeli svijet na dlanu uz pomoći naših elektroničkih uređaja, postepeno je rasla i potreba analizom ljudskog jezika na način koji bi računalo moglo procesuirati. Taj se postupak naziva računalna obrada prirodnog jezika (*Natural Language Processing – NLP*).

Obradom prirodnog jezika analiziramo ljudsku pismenu komunikaciju i određujemo subjektivni i kritički dio obavijesti, ili, u ovom slučaju, stava koji je iznesen. Mnijenja ljudi koji svoje stavove iznose javno, često nazivamo kritikom (Raos, 2012). Tako iznesenim mnijenjem, ljudi zauzimaju stav prema određenom entitetu koji može označavati događaj, poduzeće, proizvod, osobu i slično (Raos, 2012).

Mnoštvo takvih kritika možemo pronaći online u obliku recenzija kako bismo saznali što ljudi misle, ili kakvo je njihovo mnijenje, o određenom mjestu, robi i/ili usluzi. Jedna od takvih stranica je i https://www.top-rated.online na kojoj možemo odabrati određeni grad ili državu i filtrirati najpozitivnije recenzirana i najnegativnije recenzirana mjesta, npr.:

1. https://www.top-rated.online/countries/Croatia/cities/Zagreb/all/worst-rated
2. https://www.top-rated.online/countries/Croatia/cities/Zagreb/all/our-rank
3. https://www.top-rated.online/countries/Croatia/cities/Zagreb/all/top-rated

S tih su stranice preuzete recenzije za Grad Zagreb u Republici Hrvatskoj te je tako nastao podatkovni skup SentiGMapsCro. Nažalost, većina podataka ove vrste na Internetu je vrlo nestrukturirana, te je ponekad stvara izazov i ljudima, a kamoli računalima koja se najbolje

snalaze u jeziku koji je u obliku nula i jedinica. Kako bi računala mogla učiti, podatkovni skupovi poput skupa SentiGMapsCro moraju biti pregledani od strane čovjeka i proći kroz analizu sentimenta. Ukratko, ljudsko biće analizira recenzije i svaki dio koji se može označiti kao pozitivno ili negativno mnijenje označi brojem od -2 (vrlo negativno) do +2 (vrlo pozitivno) kako bi računalo ubuduće takve frazeme moglo prepoznati kao takve. Označavanje sentimenata nije nimalo lak zadatak, ne samo zbog kompleksnosti prirodnog jezika poput jezičnih figura (ironije i sarkazma), kolokvijalizama i žargona, već i zbog pravopisnih i gramatičkih pogrešaka koje ponekad onemogućuju pravilno označavanje i razumijevanje same recenzije.

Stoga, prvi se dio rada sastoji od teorijske pozadine glavnih pojmova te definicija metoda i pristupa za analizu sentimenata, kao i same analize kako se prikupljaju podatkovni skupovi sentimenata. U drugom dijelu fokus je na metodologiji istraživanja podatkovnog skupa SentiGMapsCro, kako je izgledao proces označavanja te analiza rezultata uz najzanimljivije primjere koji će ilustrirati poteškoće označavanja te prepreke s kojima se susrećemo tijekom analize sentimenata kad je u pitanju jezik i računalna obrada prirodnog jezika.

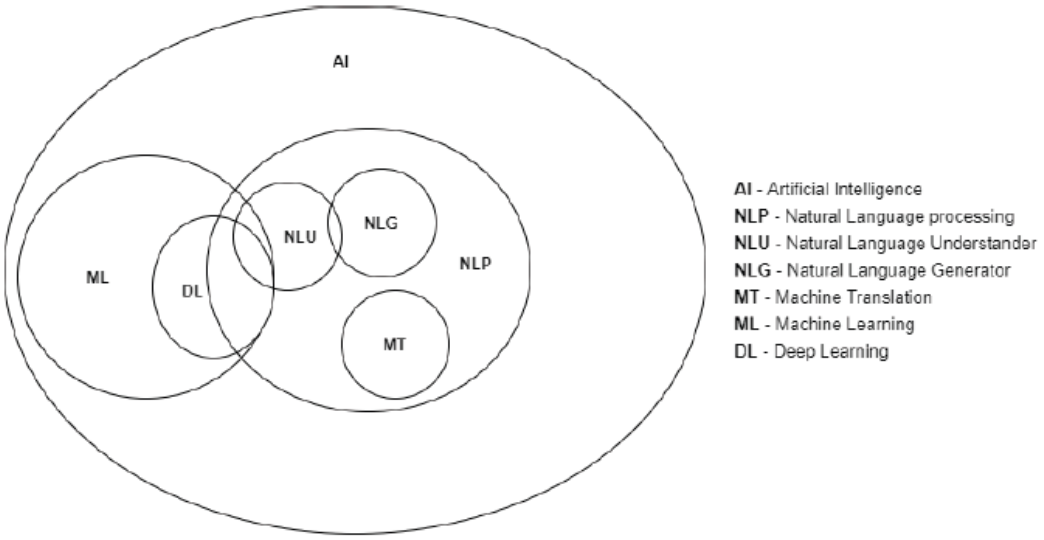
2. Obrada prirodnog jezika

Kako bismo mogli komunicirati i prenositi obavijesti, mora postojati sustav znakova koji je zajednički i govorniku i slušatelju (Škiljan, 1980). Unatoč eksponencijalnom rastu upotrebe tehnologije i umjetnik jezika, prirodni je jezik još uvijek najkorištenije sredstvo prijenosa informacija.

Jedna od ranije spomenutih karakteristika prirodnog jezika jest da kada se njime koristimo, prenosimo i vlastita mnijenja, bez obzira radi li se o razgovoru, poruci, objavi na društvenim mrežama, elektroničkoj pošti ili recenziji objavljenoj na Internetu. Upravo su iz tog razloga razumijevanje i analiza prirodnog jezika postali središnji fokus proučavanja istraživanja analize sentimenta.

U svrhu „komuniciranja“ s računalima, ili bolje rečeno, učenja računala, stvoren je niz formalnih ili umjetnih jezika (Kokan, 2021). To nažalost ne znači da računala mogu pojmiti i prirodni jezik. Naime, neke od osnovnih karakteristika prirodnog jezika, poput korištenja ironije, sarkazma, prenesenog značenja, žargona i slično, računalima uvelike otežavaju interpretaciju takvih emotivno obojenih izraza.

Ovdje ulazimo u područje umjetne inteligencije (engl. *Artificial intelligence*). Jedna od mogućih definicija umjetne inteligencije jest da je to „znanstveno područje koje istražuje načine kako postići da se računalo inteligentno ponaša“ (Šuman, 2021, str. 371). Točnije, umjetna inteligencija istražuje načine računalnog modeliranja (ljudske) inteligencije, uključujući sposobnost razumijevanja prirodnog ili ljudskog jezika (Silva i Fonseca, 2018). Obrada prirodnog jezika ili procesiranje prirodnog jezika (engl. *Natural Language Processing*) dio je umjetne inteligencije koji se bavi upravo time te predstavlja fokus ovog rada. U svrhu prikaza širine pojma „umjetna inteligencija“ i što sve podrazumijeva, ukratko analiziramo sliku 1.:



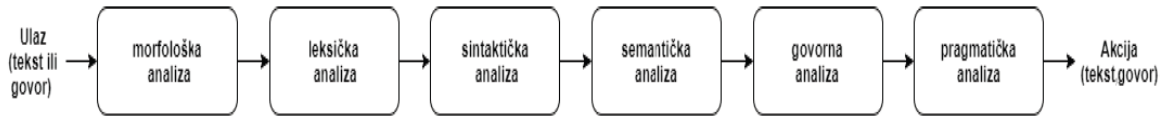
Slika 1. Prikaz odnosa pojmova povezanih s NLP-om (Šuman, 2021)

Umjetna inteligencija kao znanstveno područje obuhvaća obradu prirodnog jezika (NLP) i strojno učenje (ML). Obrada prirodnog jezika podrazumijeva strojno prevođenje (MT), razumijevanje prirodnog jezika (NLU), stvaranje prirodnog jezika (NLG), te duboko učenje (DL) koje zapravo spada i pod obradu prirodnog jezika i pod strojno učenje. Obrada prirodnog jezika usko je povezana s lingvistikom. Kao što je već spomenuto, računala „ne razumiju" ljudski jezik pa postoji potreba za njegovom obradom (Manning i Schutze, 1999). Prema Šuman (2021):

> „NLP (obrada prirodnog jezika) istražuje načine uporabe računala za obradu ili razumijevanje ljudskih – prirodnih jezika. Upotrebljava se za pretvaranje ili prevođenje podataka s prirodnog jezika na računalu razumljiv jezik – strojno razumljiv format. Nakon što procesiraju prirodni jezik, računala mogu komunicirati jezikom koji upotrebljavaju ljudi" (str. 373).

Najveći izazov obrade prirodnog jezika je dvosmislenost. Ta se dvosmislenost može iščitati u značenju riječi, morfologiji, sintaktičkim svojstvima te ulogama i povezanosti između tekstova. Taj je problem posebno uočljiv kod dužih i gramatički složenijih rečenica (Poibeau, 2017). Ljudska bića taj problem nemaju jer mogu uzeti u obzir kontekst cijelog teksta, svoje prethodno znanje i iskustvo. No, čak se i ljudi povremeno susreću s problemima u komunikaciji zbog dvosmislenosti. Računala uče problem dvosmislenosti rješavati na vrlo sličan način – uzimaju u obzir širi kontekst u neposrednoj blizini problematičnog frazema i svoj zaključak donose na temelju prošlih slučajeva (Šuman, 2019). Proces obrade prirodnog

jezika izvodi se kroz nekoliko faza koje određuju razine jezika (Liddy, 2003). Opći model razumijevanja prirodnih jezika prikazan je na slici 2.



Slika 2. Opći model razumijevanja prirodnih jezika (Šuman, 2021)

Na slici 2. prepoznajemo šest razina obrade prirodnog jezika. Na slici su prikazane odvojeno, no zapravo se često preklapaju. Proces počinje na morfološkoj razini koja se primarno bavi najmanjim jedinicama riječi koje nose neko značenje – morfemima. Na leksičkoj se razini riječi analiziraju i svrstavaju u leksičke kategorije poput vrsta riječi, te im se određuju obilježja poput broja, roda i padeža. Sintaktička razina bavi se sastavnim dijelovima rečenica poput predikata, objekta, priložnih oznaka te njihovih odnosa unutar rečenice. Semantička razina bavi se značenjem rečenica. Ova vrsta analize usko je povezana sa sintaktičkom razinom i te su dvije razine neodvojive. Sintaktička razina bavi se strukturom rečenice, dok semantička razina tu strukturu interpretira ovisno o kontekstu i pomaže nam donijeti zaključak o značenju rečenice. Na govornoj razini dana se rečenica stavlja u kontekst s ostalim rečenicama iz danog teksta ili konteksta (Šuman, 2021). Posljednja, pragmatična razina, bavi se analizom rečenice i mogućnostima njezinog korištenja ovisno o situaciji, vremenu i mjestu izgovaranja (Rajesh i dr., 2009). Nakon što u potpunosti savladamo razine obrade prirodnog jezika, možemo se fokusirati na analizu sentimenta u sljedećem poglavlju.

3. Analiza sentimenta

Kako bi uopće mogli govoriti o komunikaciji i prenošenju obavijesti, postoje određeni preduvjeti koji se moraju zadovoljiti. Jedan od tih preduvjeta jest već spomenuto postojanje sustava znakova koji je zajednički i govorniku i slušaocu (Škiljan, 1980). Od Škiljanovog rada iz 1980.-te ljudski ili prirodni jezik postao je globalno raširen putem Interneta i društvenih mreža te još uvijek predstavlja najmoćnije sredstvo prenošenja informacija. Gotovo svaki pojedinac danas svoje mnijenje prenosi putem poruka, komentara, objava na društvenim mrežama i recenzija. Kao što Roca (2022) ističe, činjenica da prirodni jezik sadrži mnijenja pojedinaca, razumijevanje prirodnog jezika postalo je glavni fokus proučavanja i istraživanja analize sentimenta.

Analiza sentimenta (engl. *sentiment analysis*), također poznata kao rudarenje mišljenja (engl. *opinion mining*), potpodručje je obrade prirodnog jezika (NLP) koje se usredotočuje na prepoznavanje i izdvajanje subjektivnih informacija iz tekstualnih podataka. Cilj je analize sentimenta odrediti osjećaj izražen u tekstu, koji se može iskazati u obliku jednostavne binarne razlike između pozitivnih i negativnih osjećaja do kompliciranijih klasifikacija, na primjer, prepoznavanje određenih emocija (radosti, ljutnje ili tuge). Analiza sentimenta naširoko se koristi za razumijevanje javnog mnijenja, povratnih informacija kupaca i interakcija na društvenim medijima, čime se pruža vrijedan uvid u ljudske emocije i stavove (Liu, 2012).

Područje analize sentimenta znatno se proširilo posljednjih nekoliko desetljeća, potaknuto napretkom računalne lingvistike, strojnog učenja i sve većom dostupnošću digitalnih tekstualnih podataka. Rane faze analize sentimenta započinju 1950-ih i 1960-ih godina s pojavom računalne lingvistike. Istraživači su počeli istraživati mogućnost automatizacije analize teksta s pomoću računala. Najranija istraživanja uvelike su se oslanjala na ručno napisana pravila i leksikone kako bi se identificirale riječi i fraze koje u semantičkom smislu označavaju ili prenose mnijenje (Pang i Lee, 2008).

Tijekom 1990-ih i ranih 2000-tih godina došlo je do značajnog pomaka u području strojnog učenja. Uz dostupnost većih skupova podataka i snažnijih računalnih resursa, istraživači su počeli razvijati statističke modele za analizu sentimenta. Algoritmi kao što su Naive Bayes, Support Vector Machines (SVM) i logistička regresija korišteni su za klasifikaciju sentimenta na temelju određenih značajki izdvojenih iz teksta na prirodnom jeziku (Pang i Lee, 2008).

Pojava dubokog učenja nakon 2010.-e godine revolucionirala je analizu sentimenta. Modeli dubokog učenja, osobito rekurentne neuronske mreže (RNN) i konvolucijske neuronske mreže (CNN), pokazali su izvanredne rezultate u prepoznavanju složene i kontekstualne prirode sentimentalna u tekstu. Uvođenje umetanja riječi, kao što su Word2Vec i GloVe, dodatno je poboljšalo sposobnost modela da razumiju semantičke odnose između riječi (Mikolov i dr., 2013.).

Nedavno uvođenje modela temeljenih na transformatorima, kao što su BERT (engl. *Bidirectional Encoder Representations from Transformers*) i GPT (engl. *Generative Pre-trained Transformer*), postavilo je nova mjerila u analizi sentimenta. Ti modeli koriste mehanizme za samopozornost kako bi se učinkovitije obuhvatile ovisnosti velikog dometa u rečenicama i kontekstualne informacije, što dovodi do najsuvremenijih performansi u različitim zadaćama obrade prirodnog jezika, uključujući analizu sentimenta (Devlin i dr., 2018.; Radford i dr., 2018.).

3.1. Važnost analize sentimenta

Analiza sentimenta izrazito je važna u različitim područjima zbog svoje sposobnosti pružanja provedivih uvida iz nestrukturiranih tekstualnih podataka. U poslovnom svijetu razumijevanje osjećaja kupaca ključno je za poboljšanje njihovog zadovoljstva, poboljšanje proizvoda i usluga te osmišljavanje učinkovitih marketinških strategija. Analizom recenzija kupaca, povratnih informacija i interakcija na društvenim mrežama, tvrtke mogu procijeniti javno mnijenje, identificirati područja za poboljšanje i prilagoditi svoju ponudu potrebama kupaca (Liu, 2012).

S porastom broja društvenih mreža, korisnici tih mreža svakodnevno proizvode velike količine sadržaja. Analiza sentimenta omogućuje organizacijama, kreatorima politika tih društvenih mreža i istraživačima da prate javno mnijenje o različitim pitanjima, prate trendove i reagiraju na nova pitanja i probleme. Ta sposobnost analize sentimenta posebno dobiva na vrijednosti tijekom događaja kao što su izbori, lansiranje proizvoda i krize (Pang i Lee, 2008.).

U zdravstvu se analiza sentimenata može primijeniti za analizu povratnih informacija pacijenata, praćenje mentalnog zdravlja i procjenu javnih reakcija na vijesti i politike vezane uz zdravstvo. Razumijevanjem osjećaja pacijenata pružatelji zdravstvene skrbi mogu poboljšati skrb pacijenata, utvrditi potencijalne probleme i poboljšati cjelokupno iskustvo u području zdravstva (Liu, 2012).

Raspoloženje ulagača također ima ključnu ulogu na financijskim tržištima. Analiza sentimenta može se koristiti za analizu novinskih članaka, financijskih izvješća i rasprava na društvenim mrežama kako bi se procijenilo opće raspoloženje na tržištu. Ove informacije mogu pomoći investitorima u donošenju informiranih odluka, predviđanju tržišnih trendova i upravljanju rizicima (Pang i Lee, 2008).

U pravnom i političkom području, analiza sentimenta može pomoći u analizi javnog mnijenja o pravnim slučajevima, političkim kandidatima i političkim odlukama. Te su informacije vrijedne za političke kampanje, oblikovanje politika i razumijevanje društvenih reakcija na pravne presude (Liu, 2012.).

3.1.1. Klasifikacija analize sentimenta

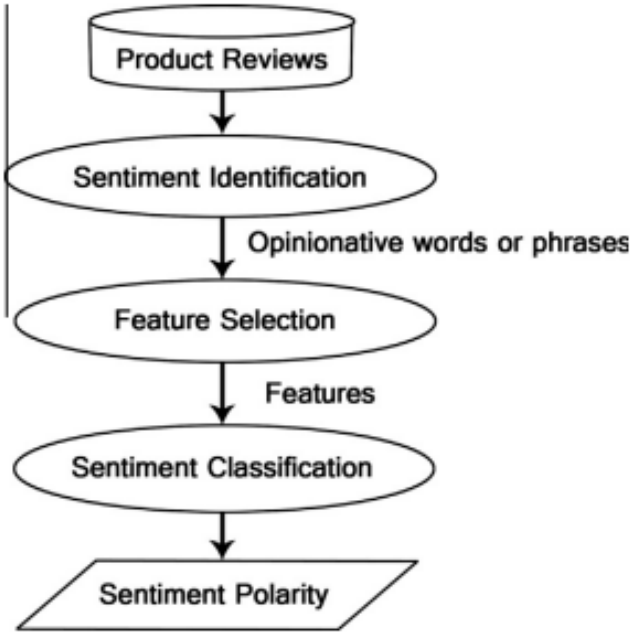
Pojam entiteta koji je spomenut u uvodu pojavljuje se i u sljedećoj definiciji: „Analiza sentimenta ili mnijenja podrazumijeva računalno proučavanje ljudskog mnijenja, stavova i emocija o određenom entitetu“ (Medhat i dr., 2014). Kako se ti entiteti mogu odnositi na bilo koje područje ljudskog djelovanja, stručnjaci koji koriste analizu sentimenta proučavaju društvo kao cjelinu i na svim razinama, od pojedinaca do proizvoda (Roca, 2022). Međutim, u literaturi se navodi da se u praktičnom korištenju analiza sentimenta i analiza mnijenja razilaze i prestaju biti sinonimi. Kako je Roca sažela, „analiza mnijenja se temelji na procesima ekstrakcije i analize ljudskog mnijenja, dakle analiza sentimenta prepoznaje i identificira emociju od koje je sačinjeno mnijenje“ (2022, str. 3).

U svrhu provedbe analize sentimenta prikupljamo podatkovne skupove koje anotiramo ili obilježavamo te ih zatim klasificiramo. Klasifikacija se može definirati kao razvrstavanje pojmova u razrede i njihove poddiobe kako bi se iskazali semantički odnosni među pojmovima (Tuđman, 1990). Također, te je skupove potrebno klasificirati prema širini sadržaja, a u svrhu ovog rada koristit će se troslojna klasifikacija prema autorima Medhat i dr. (2014):

- Analiza dokumenta
- Analiza rečenice
- Analiza aspekta

Prema samim nazivima svake razine, možemo razlučiti da je analiza dokumenta najširа i obuhvaća određeni dokument u cijelosti. Smatra se da je cijeli dokument osnovna informacijska jedinica jer govori o jednoj temi. Analiza rečenice se odnosi na obilježavanje unutar jedne gramatičke rečenice s obzirom na to je li mnijenje u njoj pozitivno ili negativno, a razlika između obilježavanja dokumenta i rečenice su minimalne. Klasifikacija teksta na

razini dokumenta ili na razini rečenice ne pruža potrebna detaljna mnijenja o svim aspekte entiteta koja su potrebna za njihovo korištenje u npr. marketingu, stoga se moramo fokusirati na analizu aspekta. Cilj je analize sentimenta na razini aspekta klasificirati sentiment ili mišljenje s obzirom na posebne aspekte entiteta o kojima je riječ (Medhat i dr., 2014). Na slici 3. možemo vidjeti kako proces analize sentimenta izgleda na primjeru recenzije proizvoda.

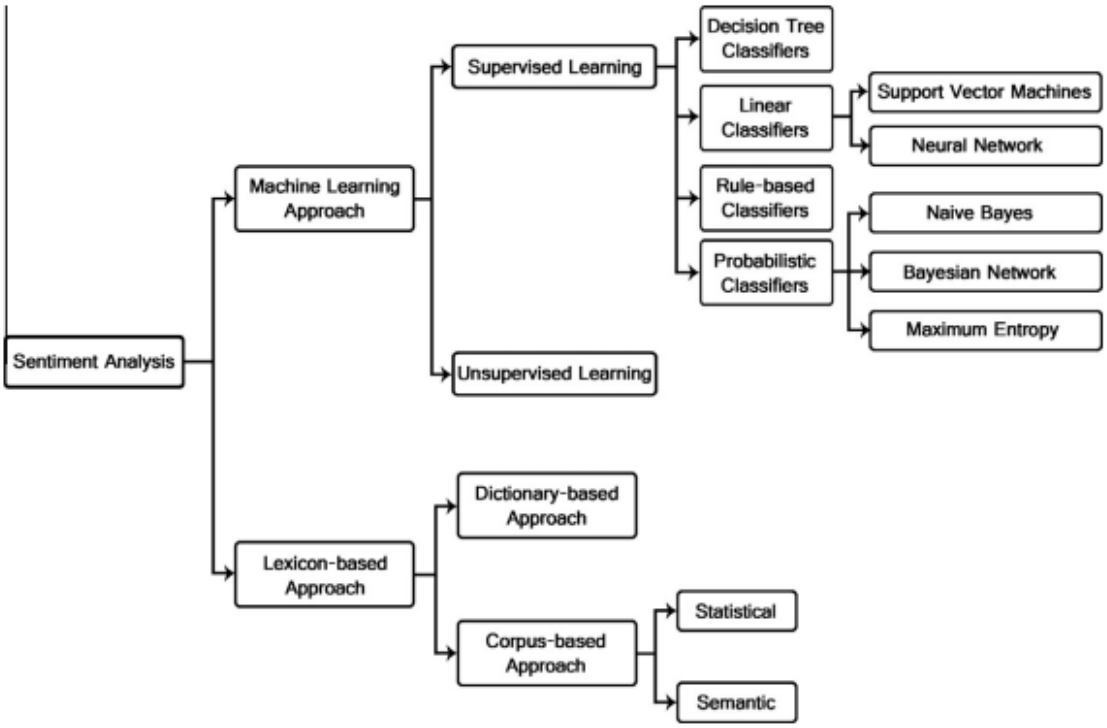


Slika 3. Proces analize sentimenta na primjeru recenzije proizvoda (Medhat i dr., 2014)

Rudarenje mišljenja izdvaja i analizira mnijenja ljudi o nekom entitetu, dok analiza sentimenta identificira mnijenje/osjećaj izražen u tekstu i zatim ga analizira, najčešće ovisno o tome je li mišljenje pozitivno ili negativno. Stoga je cilj analize sentimenta pronaći mišljenja, utvrditi osjećaje koje ta mišljenja izražavaju, te im klasificirati polaritet (pozitivan ili negativan). Ovaj se rad temelji na analizi aspekta, to jest, fokus obilježavanja su frazemi unutar rečenica koji su izražavali mišljenja i stavove recenzenata.

3.1.2. Klasifikacija sentimenta

Kako je klasifikacija teksta zapravo problem klasifikacije teksta, naš je fokus na klasifikaciji dokumenata prema određenoj temi tekstova. Najlakši način za odrediti temu određenog teksta jest da se izdvoje ključne riječi koje opisuju taj dokument te se često pojavljuju u samome tekstu (Roca, 2022). Na isti način možemo tretirati i klasifikaciju sentimenta. Na slici 4. vidimo mnoge tehnike koje se mogu koristiti za klasifikaciju sentimenta (Medhat i dr., 2014).



Slika 4. Tehnike klasifikacije sentimenta (Medhat i dr., 2014)

Ove klasifikacije tehnike koriste se kako bi se pripisalo određenje mnijenja ili sentimenta dokumentima i tako saznajemo vrijednost tih dokumenata; te se vrijednost temelji na dvije osnovne vrste klasifikacije sentimenta koja se koristi i za potrebe ovog rada, a to su pozitivna i negativna klasifikacija. Podatkovni skup nad kojim vršimo obilježavanje ili anotiranje, o kojem će biti više rečeno u poglavlju 4., sastoji se od podataka koji se obilježavaju sukladno toj osnovnoj pozitivno/negativnoj klasifikaciji (Medhat i dr., 2014).

3.1.2.1. Klasifikacija sentimenta na razini dokumenta

Klasifikacija sentimenta na razini dokumenta najčešća je metoda klasifikacije sentimenta kojom se pokušava odrediti sentiment cijelog dokumenta. Kao što je već spomenuto, klasifikacije sentimenta temelje se na metodama klasifikacije teksta, stoga se primjenjuju metode strojno nadziranog učenja. „Algoritmi strojnog učenja 'uče' informacije i odnose među njima izravno iz podataka ne oslanjajući se na teorijske jednadžbe i matematičke modele. Pri tome se modeli s vremenom unaprjeđuju porastom broja uzoraka dostupnih za učenje“ (Bolf, 2021, str. 591). Metode nadziranog učenja primjenjuju algoritme koji samostalno iz dokumenata prepoznaju vrstu riječi, koliko se često određena riječ pojavljuje, identificiraju sentimente, impliciraju ih, prepoznaju negaciju te razumiju sintaktičke ovisnosti (Bing, 2012).

Metoda nenadziranog učenja razlikuje se jer, kad govorimo o klasifikaciji sentimenta, ta metoda koristi klasifikaciju koja se temelji na prethodno prepoznatim sintaktičkim obrascima (Roca, 2022). Ti se obrasci baziraju na oznakama za vrste riječi, stoga se nenadzirano učenje sastoji od izdvajanja riječi koje se često pojavljuju zajedno te analize sintaktičke okoline u kojoj se te riječi zajedno pojavljuju, zatim određivanja njihove polarnosti (pozitivnog ili negativnog sentimenta) te se na temelju te klasifikacije s pomoću algoritma određuje prosječna polarnost tog sintaktičkog skupa riječi ili frazema u, primjerice, recenziji (Bing, 2012). Kako je metodu klasifikacije sentimenta na razini dokumenta vrlo teško primijeniti na platformama poput foruma i članaka te nije najpouzdanija metoda kada govorimo o zadacima koji obradu prirodnog jezika vrše s dubokim razumijevanjem teksta, u nastavku se fokusiramo na klasifikaciju sentimenta na razini rečenice i aspekta.

3.1.2.2. Klasifikacija sentimenta na razini rečenice

Prethodno je već rečeno da se u praksi rečenice u analizu sentimenta i obradi prirodnog jezika tretiraju kao kratki dokumenti, stoga ne postoji prevelika razlika između klasifikacije sentimenta na razini dokumenta i klasifikacije sentimenta na razini rečenice. Ono što stručnjaci predstavljaju kao jedinu značajnu razliku jest dubina analize: dokument u velikoj većini slučajeva sadrži više mnijenja, dok rečenice uglavnom sadrže samo jedno (Bing, 2012). Iako nije dovoljno ustanoviti sadrži li rečenica pozitivno ili negativno mnijenje, pokazalo se da stručnjacima uvelike pomaže razumijevanje na koga se sentimentalna orijentacija dane rečenice odnosi, pogotovo kada govorimo o recenzijama. Primjerice, velika je razlika u negativnoj rečenici nekog ugostiteljskog objekta odnosi li se negativna recenzija na cjelokupnu uslugu objekta ili samo na jednu osobu poput vlasnika/-ice ili konobara/-ice.

Na razini rečenice klasifikacija se vrši u dva koraka: prvo se određuje mnijenje, a zatim i polarnost tog mnijenja. Bitno je napomenuti da ne sadrže sve rečenice mnijenje, stoga je prije svega potrebno odrediti rečenice koje ga sadrže. Kako smo mnijenje definirali kao subjektivni stav korisnika o nekom proizvodu ili sadržaju, sljedeći je korak utvrditi povezanost rečenice s klasifikacijom subjektivnosti (Yu i Hatzivassiloglou, 2003). Takva se klasifikacija subjektivnosti temelji na određivanju subjektivnosti ili objektivnosti informacije u danoj rečenici: činjenice svrstavamo u objektivne rečenice, dok osobne stavove i razmišljanja klasificiramo kao subjektivne rečenice. Naravno, u analizi sentimenta od posebnog su nam značaja subjektivne rečenice koje iskazuju emocije, uvjerenja, stavove i slično, dok se objektivne rečenice najčešće ignoriraju.

Osim subjektivnosti i objektivnosti, također nam je bitno uočiti razlike u iskazivanu emocija na temelju vrste rečenica. U hrvatskom jeziku vrlo su česte rečenice koje se sastoje od zavisne i nezavisne surečenice, od kojih naš fokus pada na odnos zavisne rečenice čiji nam odnos u većini slučajeva otkriva i sentiment cijele rečenice. Sarkazam je također jezična figura koja je tijekom procesa anotiranja uz ironiju i druge vrste figurativnog jezika otežala autorici obilježavanje sentimenata određenih rečenica.

3.1.2.3. Klasifikacija sentimenta na razini aspekta

Kako bismo lakše odredili entitet na koji je mnijenje korisnika usmjereno, koristimo klasifikaciju sentimenta na razini aspekta. U posljednjem je desetljeću postalo opće prihvaćena činjenica da su klasifikacija sentimenta na razini dokumenta i rečenice preobuhvatne te zanemaruju nijansiranije emocije koje se zapravo mogu uočiti tek ako vršimo klasifikaciju sentimenta na razini aspekta (Wang i dr., 2021). Prije svega, bitno je razumjeti što aspekt predstavlja u bilo kojoj rečenici:

ASPEKT ENTITET

"Oštar zaslon *ovog računala* zaista je impresivan!"

Ova je rečenica klasičan primjer recenzije korisnika računala. Na ovom primjeru možemo uočiti da je oštar zaslon (podcrtano) aspekt entiteta koji je izrečen sintagmom „ovo računalo“ (u kurzivu).

Klasifikacija sentimenta na razini aspekta obično se sastoji od dva dijela: prepoznavanja aspekta te klasifikacije sentimenta (Wang i dr., 2021, str. 267). Bitno je napomenuti da sam aspekt nije uvijek nužno prisutan, već je možda implicitno izražen; recimo, u rečenici nakon rečenice iznad na kojoj smo analizirali aspekt i entitet bi se aspekt „oštar zaslon“ kao obilježje entiteta „ovo računalo“ moglo izraziti kroz zamjenicu „to“ („To je ono zbog čega sam se odlučio kupiti ga.“).

Drugi dio ovog procesa jest klasifikacija sentimenta tog aspekta. Već je spomenuto da analiza aspekta klasificira rečenice i dokumente na temelju polarnosti te se temelji na aspektu analiziranih rečenica i dokumenata (Sharma i dr., 2014). Dakle, procesom klasifikacije pokušavamo odrediti aspekte entiteta kojima su se autori odlučili koristiti kako bi izrazili svoje emocije: prvo se pronalaze glavni aspekti entiteta, zatim se određuje izražen sentiment za svaki aspekt te se ne kraju određuje sažetak njihovog polariteta (Kumar Laskari & Kumar Sanampudi, 2016). Za određivanje sentimenta aspekata najčešće se koristi ili binarni polaritet (pozitivno/negativno) ili trinomski polaritet (pozitivno, negativno ili neutralno), a polaritet se

definira pomoću Likertove ljestvice koja se sastoji od pet ili sedam kriterija za označavanje sentimenta gdje brojka 1 obično označava izrazito pozitivan sentiment, a brojke 5 ili 7 izrazito negativan sentiment (Tanujaya i dr., 2022). Primjer pitanja s Likertovom ljestvicom od pet kriterija dostupna je na slici 5.

Molimo Vas da iznesete svoje stavove o sljedećim tvrdnjama. Stav iskazujete zaokruživanjem broja SAMO JEDNOG od ponuđenih odgovora:

1 – uopće se ne slažem
2 – uglavnom se ne slažem
3 – nemam utvrđen stav (niti se ne slažem, niti sam suglasan)
4 – uglavnom sam suglasan
5 – sasvim sam suglasan

1.	NE osjećam se ugroženo dok se razgovara o računalima.	**1**	**2**	**3**	**4**	**5**
2.	Računala su nepouzdana.	**1**	**2**	**3**	**4**	**5**
3.	Poznavanje rada na računalu korisna je vještina.	**1**	**2**	**3**	**4**	**5**
4.	Volim upotrebljavati računalo.	**1**	**2**	**3**	**4**	**5**
5.	Način rada računala potpuno mi je nerazumljiv.	**1**	**2**	**3**	**4**	**5**
6.	Uporaba računala unapređuje posao.	**1**	**2**	**3**	**4**	**5**
7.	Povjerljivost podataka o bolesnicima ugrožena je uporabom računala.	**1**	**2**	**3**	**4**	**5**

Slika 5. Primjer pitanja s Likerovom ljestvicom od pet kriterija (Petrovečki i dr., 2008).

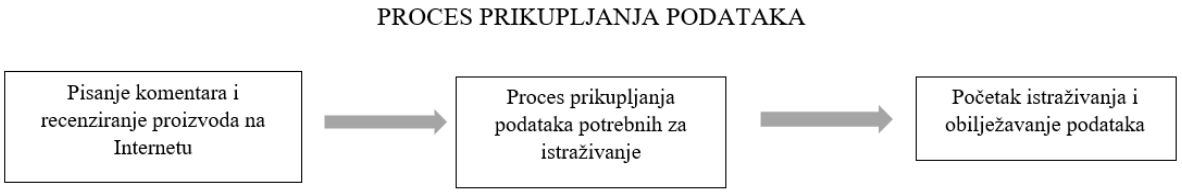
Budući da se cijeli navedeni proces klasificiranja odvija unutar podatkovnog skupa, u sljedećem se poglavlju orijentiramo upravo na podatkovni skup.

4. Podatkovni skup

Podatkovni skup je skup podataka preuzetih iz jednog izvora ili namijenjenih jednom projektu (Rosli i dr., 2016). U ovom ćemo poglavlju objasniti kako se podatkovni skupovi prikupljaju i klasificiraju.

4.1. Prikupljanje podatkovnih skupova

Prije prikupljanja podatkovnih skupova, potrebno je utvrditi o kakvom se istraživanju radi, kakvi su skupovi potrebni te gdje možemo pronaći podatke za takve skupove. To znači da sam proces prikupljanja podatkovnih skupova započinje određivanjem provjerenih izvora informacija te filtriranje podataka kako bi se skupio optimalan broj podataka za skup. Ako podatkovni skup ima previše podataka, ljudska interpretacija mogla bi potrajati puno duže, a strojno obilježavanje skupa moglo bi se pokazati nedovoljno detaljnim. Ako je podatkovni skup pak premalen, postoji šansa da stroj neće imati dovoljno primjera za svoj proces strojnog učenja (Canals, 2017). Proces prikupljanja podataka prikazan je na slici 6.



Slika 6. Proces prikupljanja podataka

4.2. Klasificiranje podatkovnih skupova

Klasificiranje podatkovnih skupova sentimenta odrađuje se po principu podjele entiteta prema unaprijed određenim kriterijima koji ovise o predmetu i svrsi istraživanja (Aggarwal, 2015, str. 286). Najčešći model klasifikacije podatkovnih skupova sastoji se od tri glavna koraka (Xu i dr., 2012):

1. određivanje subjektivnosti prikupljenih podataka
2. određivanje polarnosti prikupljenih podataka
3. određivanje intenziteta polarnosti prikupljenih podataka

Od svih prikupljenih podataka, istraživače zanima subjektivnost autora recenzija jer nam ona daje uvid u stavove autora prema entitetu koji recenzira. Iz tog razloga određivanje subjektivnosti podrazumijeva i ignoriranje objektivnih rečenica jer one nisu fokus istraživanja

(Aggarwal, 2015, str. 288). Razlikovanje subjektivnog mnijenja autora recenzije od objektivnih činjenica ponekad nije nimalo lako, stoga su autori poput Yu i Hatzivassiloglou (2003) te Niu i dr. (2005) pokušali razviti nove metode određivanja subjektivnosti na razini dokumenta i rečenice. Ti modeli još nisu savršeni, ali su obećavajući. Primjerice, u istraživanju Yua i Hatzivassilogloua (2003) određivanje subjektivnosti na razini dokumenta doseglo je točnost od 97% (str. 136). Analiza na razini rečenice mnogo je kompleksnija te još nije dosegla tako visok postotak. Bez obzira određuje li subjektivnost čovjek ili algoritam, određivanjem subjektivnosti dobivamo uvid u sažetak sentimentno orijentiranih sadržaja dokumenata što je od velike koristi i korisnicima i istraživačima (Pang i Lee, 2004).

Nakon utvrđivanja podatkovnog skupa koji sadrži subjektivne informacije slijedi određivanje polarnosti na razini rečenice što podrazumijeva označavanje ili anotiranje podatka kao pozitivnog, neutralnog ili negativnog, ovisno o mnijenju recenzenta izraženom u istom (Aggarwal, 2015). Taj podatak unutar podatkovnog skupa koji može biti dokument, rečenica, riječ ili izraz označavamo ovisno o tome izražava li recenzent svoj stav o entitetu u pozitivnom ili negativnom smislu.

Kako određivanje sentimenta podatka kao pozitivnog, neutralnog ili negativnog nije dovoljno za praktičke upotrebe analize sentimenta, tim je podatcima potrebno odrediti i intenzitet mnijenja. U mnogim slučajevima analize sentimenta ne možemo dobiti potpune podatke o zadovoljstvu ili nezadovoljstvu korisnika ako ne uvedemo intenzitet koji upućuje na to je li se retorika pojačala ili smirila (Tian i dr., 2018). Prirodni je jezik prepun snažnih i subjektivnih ekspresija koje autori recenzija često izražavaju figurativnim jezikom kojeg je ponekad vrlo teško analizirati i obilježiti. Upravo zbog toga istraživači poput Tian i dr. (2018) koriste polaritet i intenzitet polariteta koji sažimaju u četiri vrijednosti: neutralan intenzitet, niski, prosječan/središnji intenzitet te visoki intenzitet polariteta (str. 43). Kombinacija određivanja polariteta i intenziteta polariteta podataka pomaže u razumijevanju subjektivnog sadržaja podataka te uočavanju emocionalnog intenziteta figurativnog jezika autora recenzija.

5. Dosadašnja istraživanja

5.1. BERTić i BERT

Jedno od najznačajnijih istraživanja u području obrade prirodnog jezika jest BERTić, transformativni model temeljen na principu transformera Nikole Ljubešića i Davora Lauca (2021). BERTić je prethodno treniran na 8 milijardi tokena izvađenih s web stranica na hrvatskom, bosanskom, srpskom i crnogorskom jeziku (engl. BERTić — *The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian*). Ovaj se model usavršava kroz različite zadatke poput anotiranja dijelova govora, prepoznavanja entiteta, predviđanja geolokacije te razumijevanja uzročno-posljedičnih veza što trenutno čini najuspješnijim modelom za zadatke u području obrade prirodnog jezika. Za evaluaciju razumijevanja zdravorazumskih zaključivanja uveden je prijevod testa *Choice of Plausible* koji je nazvan COPA-HR (BERTić – model na hrvatskom jeziku, 2023).

Za treniranje BERTić-a koristio se pristup Electra (Clark i dr., 2020) koji je temeljen na treniranju manjeg generatora modela i većeg diskriminatornog modela. Diskriminatorni model razlikuje je li određena riječ izvorna riječ iz teksta ili je ta riječ generirana unutar modela (BERTić – model na hrvatskom jeziku, 2023). Autori poput Devlin i dr. (2019) tvrde da je pristup Electra ipak računalno učinkovitiji u usporedbi s BERT modelom koji se temelji na maskiranom modeliranju jezika. Ukratko, modeli koji se temelje na maskiranom modeliranju jezika prvo „maskiraju“ ili prikrivaju dio ulaznih slika/teksta te uče generirati maskirane dijelove u sličnom kontekstu i rekonstruirati slike/tekstove. Diskriminacija modela, u kontekstu računalne znanosti, odnosi se na proces određivanja točnosti različitih matematičkih modela u predviđanju ponašanja sustava. Diskriminatorno se modeliranje temelji na zadacima rekonstruiranja tekstova i slika s ciljem prepoznavanja koji su dijelovi tekstova izvorni, a koji su zamijenjeni (Zhang i dr., 2024).

Od ranijih istraživanja na modelu BERT od posebne je važnosti rad autora Xu, Liu, Shu i Yu (2019) u kojem je model BERT korišten kako bi se lakše razumjele recenzije korisnika. Autori su predložili novi zadatak koji su nazvali „čitanje recenzija s razumijevanjem“ (engl. *review reading comprehension* - RRC) i istražili mogućnost korištenja recenzija kao vrijedan resurs za odgovaranje na pitanja korisnika. Prihvatili su BERT kao osnovni model i predložili zajedničku naknadnu obuku modela kao pristup unapređenju područja obrade prirodnog jezika i zadatka čitanja recenzija s razumijevanjem. Dodatno su istražili primjenu tog pristupa u drugim dvjema zadaćama koje se temelje na recenzijama: izvlačenje aspekta iz recenzija i

klasifikacija sentimenta na razini aspekta. Eksperimentalni rezultati pokazuju da je pristup nakon dodatnog uvježbavanja modela prije fine prilagodbe učinkovit (Xu i dr., 2019, str. 2332).

5.2. Primjeri istraživanja recenzija u području turizma

U posljednjih četrdeset godina mnoga su se istraživanja bavila analizom kvalitete turističkih destinacija u svrhu poboljšanja kvalitete usluge za domaće i strane turiste (Borrajo-Millán i dr., 2021). S porastom novih tehnologija u području obrade prirodnog jezika i umjetne inteligencije te globalnim korištenjem društvenih mreža, istraživači su razvili nove modele s pomoću kojih mogu analizirati iskustva korisnika mnogo detaljnije nego što je to prije nekoliko godina bilo moguće. Jedan od primjera takvog istraživanja jest „Sentiment Analysis to Measure Quality and Build Sustainability in Tourism Destinations“ autora Fernanda Borrajo-Millána, María-del-Mar Alonso-Almeide, María Escat-Cortes i Liu Yi iz 2021. godine koje analizira privlačnost turističkih destinacija u Španjolskoj ljudima iz Kine. Autori su na kraju istraživanja zaključili da je analiza sentimenta izuzetno koristan alat u području turizma jer negativni i neutralni komentari recenzenata turističkih destinacija daju prostora za poboljšanje i planiranje budućih investicija (Borrajo-Millán i dr., 2021).

Slična istraživanja koja koriste obradu prirodnog jezika i analizu sentimenta za poboljšanje turizma su „Sentiment Analysis to Determine Accommodation, Shopping and Culinary Location on Foursquare in Kupang City“ (2015) autorice Pauline Aliandu, „Tourist Place Reviews Sentiment Classification Using Machine Learning Techniques“ (2020) autora Apeksha Arun Wadhe i Shraddha S. Suratkara , „A Recommendation Mechanism for Under-Emphasized Tourist Spots Using Topic Modeling and Sentiment Analysis“ (2020) autora Wafa Shafqata i Yung-Cheol Byuna te „Sentiment analysis of tourist reviews from online travel forum for improving Indonesia tourism sector“ (2022) autora Muhammada Abdula Jabbaar, Arinde Dwi Okfantia, Anggraini Widjanarti i Alvina Andhika Zulen te Bank Indonesiae.

6. Podatkovni skup SentiGMapsCro i INCEpTION

Prije nego li objasnimo metodologiju istraživanja analize mnijenja u podatkovnom skupu SentiGMapsCro, potrebno je objasniti kakav je to skup, kako je nastao i s pomoću kakvog su alata recenzije u tom skupu bile označene.

6.1. SentiGMapsCro

Već smo ustanovili da podatkovni skup nastaje sakupljanjem podataka iz jednog izvora ili namijenjenih jednom projektu (Rosli i dr., 2016). Tako je i ovaj skup nastao traženjem i filtriranjem internetskih stranica koje se bave isključivo recenzijama korisnika za objekte, lokacije i slično.

Istraživanje je započelo odabirom stranice „Top rated“ na adresi https://www.top-rated.online. Na sučelju same web stranice može se odabrati opcija „blizu mene“ koja koristi trenutnu lokaciju korisnika za filtriranje grada ili se mjesto ili grad od interesa može upisati u tražilicu na vrhu sučelja. Sukladno tim opcijama, odabran je Grad Zagreb te su za podatkovni skup SentiGMapsCro odabrane sljedeće stranice s najpozitivnije recenziranim i najnegativnije recenziranim mjestima i objektima na području Grada Zagreba:

- https://www.top-rated.online/countries/Croatia/cities/Zagreb/all/worst-rated
- https://www.top-rated.online/countries/Croatia/cities/Zagreb/all/our-rank
- https://www.top-rated.online/countries/Croatia/cities/Zagreb/all/top-rated

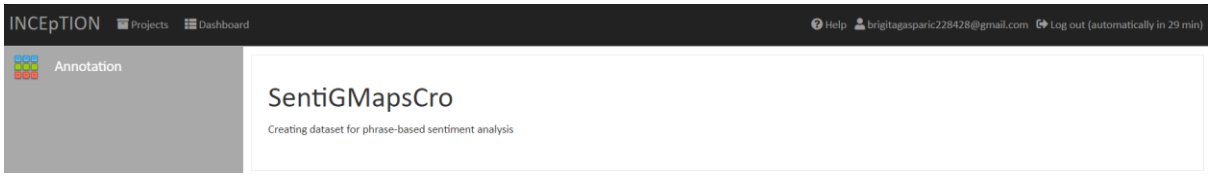
Više o označavanju samog podatkovnog skupa SentiGMapsCro bit će rečeno u sljedećem poglavlju.

6.2. INCEpTION

INCEpTION je relativno novi alat koji pomaže u anotiranju/označavanju zadataka koji podrazumijevaju interaktivno i semantičko obilježavanje, na primjer, povezivanje koncepata, povezivanje činjenica, populacija baze znanja, označavanje semantičkog okvira (Klie i dr., 2018, str. 5). Takvi zadaci mogu biti vrlo dugotrajni i zahtjevni za ljude, stoga je razvijena platforma za anotiranje ili označavanje podataka koja uključuje i mogućnost strojnog učenja koje aktivno pomaže tijekom procesa označavanja u smislu da alat nudi preporuke za označavanje te uči što je korisnik odbio označiti (Klie i dr., 2018, str. 5). Sama je platforma i generička i modularna, što znači da može biti i vrlo obuhvatna, ali i vrlo specifična za određene kategorije podataka. Ova je platforma usmjerena na niz područja istraživanja kojima je potrebna analiza sentimenta i njihovo označavanje kao što su digitalna humanistika,

bioinformatika i lingvistika (Klie i dr., 2018). Ovaj alat održava Tehničko Sveučilište u Darmstadtu, točnije njihov Odsjek za informatiku. INCEpTION je javno dostupan putem Interneta kao softver otvorenog koda, a izgled same platforme dostupan je na slikama 7 i 8.

Slika 7. Izgled platforme INCEpTION



Slika 8. Izgled platforme INCEpTION unutar skupa SentiGMapsCro

Ova se platforma pokazala izuzetno korisnom istraživačima diljem svijeta jer ispunjava tri glavna kriterija oko kojih se istraživači u području obrade prirodnog jezika slažu, a to su:

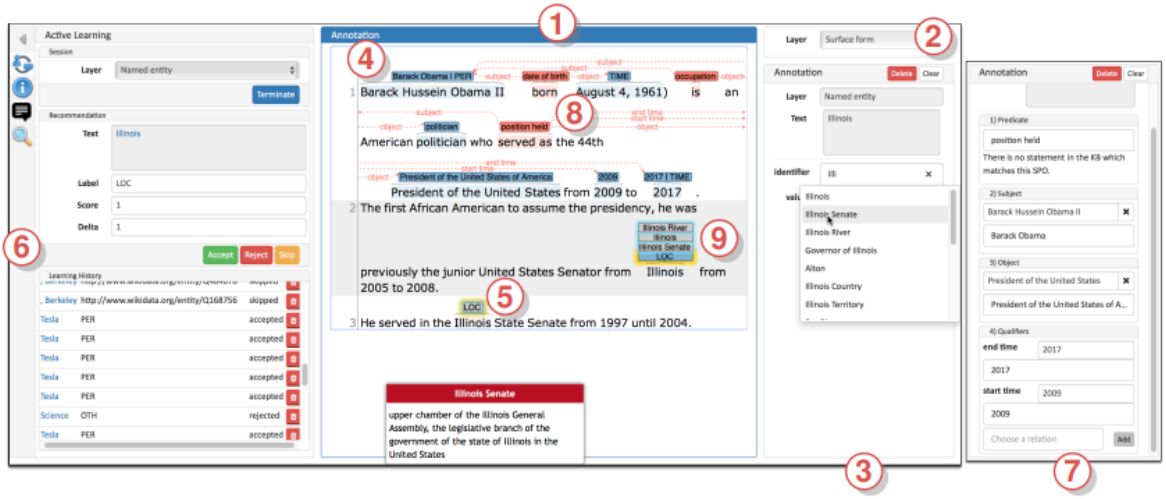
1. Pomoć u anotaciji/obilježavanju/označavanju – INCEpTION pruža pomoć istraživačima te uči iz njihovog procesa označavanja;
2. Upravljanje znanjem – procesom označavanja stvara se novo znanje koje je potrebno adekvatno pohraniti za buduće zadatke;
3. Prilagodljivost i proširivost – zbog različite prirode svakog projekta, a ponekad i zadataka unutar projekta, izuzetno je važno da si istraživači mogu prilagoditi alat sukladno novim zahtjevima (Klie i dr., 2018).

INCEpTION ispunjava ova tri kriterija na nekoliko načina. Kako bi se poboljšala učinkovitost i brzina semantičkih zadataka označavanja, ovaj alat nudi preporuke koje korisniku nude razne mogućnost obilježavanja podataka. Ako se ova mogućnost želi maksimalno iskoristiti, korisnik može uključiti opciju „aktivnog modula učenja“ koja korisnika vodi kroz preporuke za anotiranje na brz i učinkovit način. Upravljanje znanjem u cijelosti je integrirano; baze znanja mogu se stvarati i uređivati, a podržano je i povezivanje entiteta i činjenica. Modularna arhitektura INCEpTION-a omogućuje korisnicima da povećaju svoju učinkovitost i brzinu označavanja prilagođenim algoritmima strojnog učenja, formatiranjem podataka, bazama znanja, vrstama označavanja te vizualizacijom.

INCEpTION nudi niz funkcionalnosti koje se očekuju od generičke platforme za označavanje: svestrano i intuitivno korisničko sučelje, fleksibilnu konfiguraciju sustava označavanja, mogućnost istodobnog pokretanja više projekata za više korisnika te podršku tijekom samog procesa označavanja (Klie i dr., 2018).

6.2.1. Korisničko sučelje za označavanje

Sustav označavanja INCEpTION-a organiziran je u slojevima koji definiraju skup atributa koje oznaka može nositi. Korisnici mogu definirati proizvoljan broj slojeva koji su svaki ili rasponi ili odnosi između raspona. Svaki sloj može imati proizvoljan broj značajki koje mogu biti nizovi, brojevi, konceptualne reference ili reference na druge oznake. Korisničko sučelje za označavanje na slici 9 prikazuje tekst dokumenta u središnjem dijelu (označeno brojkom 1). Označavanje raspona teksta ovdje stvara novu anotaciju na sloju koji je odabran u desnoj bočnoj traci (označeno brojkom 2), npr. imenovani entitet. Primjedbe o rasponima prikazuju se kao mjehurići iznad teksta. Kada se izradi nova oznaka ili se odabere postojeća oznaka, njezine se značajke prikazuju na desnoj bočnoj traci i tamo se mogu uređivati (označeno brojkama 3 i 4). Ovisno o vrsti značajke, prikazuje se specijalizirani uređivač. Na primjer, uređivač za dodjelu koncepata iz baze znanja polje je za unos s automatskim dovršetkom u kojem se prikazuju subjekti iz baze znanja koji odgovaraju unosu korisnika. Lijeva bočna traka omogućuje pristup dodatnim funkcijama, jedna od kojih je i aktivni način učenja. (Klie i dr., 2018).



Slika 9. Sustav anotiranja u INCEpTION-u

Legenda: 1 – područje za označavanje; 2 – odabir sloja označavanja; 3 – uređivač značajki za povezivanje subjekata; 4 – imenovani subjekti povezani s Wikidata; 5 – prijedlozi za označavanje; 6 - bočna traka za aktivno učenje; 7 – uređivač za povezivanje činjenica; 8 – označeni dio rečenice; 9 – preporuke za oznake.

6.2.2. Preporuke oznaka

Kako bi se poboljšala učinkovitost označavanja, INCEpTION nudi i preporuke oznaka. To su algoritmi koji upotrebljavaju resurse strojnog učenja i/ili znanja kako bi mogli ponuditi prijedloge za anotiranje tijekom samog procesa. Korisniku se prikazuju s već izrađenim napomenama u drugoj boji (na slici 9 označeno brojkom 5). Korisnik može prihvatiti prijedlog klikom na njega. Time se prijedlog pretvara u odgovarajuću oznaku koja se po želji može dodatno uređivati. Korisnik također može odbiti prijedlog dvostrukim klikom na njega. Podsustav za preporučivanje osmišljen je za kontinuirano praćenje aktivnosti korisnika, ažuriranje, to jest preinaku modela preporuka i pružanje uvijek ažuriranih prijedloga. Istodobno se može upotrebljavati više preporuka kao, na primjer, preporuke visoke preciznosti/niskog odaziva (primjerice upotrebom dinamičkog rječnika) koje su korisne tijekom ranih faza označavanja te klasifikatori usmjereni na odaziv koji su osjetljivi na kontekst (primjerice klasifikatori sekvenci) za kasnije faze. Kako bi se izbjeglo da klasifikatori daju previše pogrešnih prijedloga, prag kvalitete može se konfigurirati prema željama korisnika. INCEpTION podržava dvije vrste preporuka: unutarnje i vanjske. Algoritmi za interne preporuke izravno se integriraju u platformu implementacijom Java sučelja, dok se algoritmi za vanjske preporuke koriste jednostavnim protokolom temeljenom na HTTP-u za razmjenu UIMA CAS XMI (XML prikaz UIMA oznaka). Algoritmi za vanjske preporuke omogućuju korisnicima da iskoriste postojeće i prethodno osposobljene modele strojnog učenja ili biblioteke iz drugih programskih jezika (Klie i dr., 2018).

6.2.3. Aktivno učenje

Cilj je aktivnog učenja brzo postići dobru kvalitetu prijedloga za označavanje traženjem povratnih informacija od korisnika za koje se očekuje da će biti najinformativnije osnovnom algoritmu strojnog učenja. Trenutačno se koristi strategija uzorkovanja nesigurnosti (Lewis i Gale, 1994) za pokretanje aktivnog učenja jer zahtijeva samo da algoritmi za preporuke daju ocjenu pouzdanosti za svaki prijedlog. Način rada aktivnog učenja (na slici 9 označen brojkom 6) radi samo za aktivni sloj u danom trenutku kako bi se izbjegle zabune. Nakon odabira sloja, sustav ističe prijedlog za koji se traži unos u području za označavanje te

prikazuje detalje o prijedlogu u bočnoj traci za aktivno učenje. Korisnik tada može prihvatiti, odbiti ili preskočiti prijedlog. Preskočeni prijedlozi ponovno se predstavljaju korisniku kada više nema prijedloga za prihvaćanje ili odbijanje. Izbori su pohranjeni u povijesti učenja gdje ih korisnik može pregledati i poništiti ako je potrebno. Kada je omogućen način rada aktivnog učenja, korisnik i dalje može odstupiti od svojih smjernica te proizvoljno stvarati i mijenjati oznake. Sve učinjene promjene preko bočne trake za aktivno učenje ili u glavnom uređivaču odmah se preuzimaju u prijedloge i smjernice aktivnog učenja koje treba ažurirati (Klie i dr., 2018).

6.2.4. Upravljanje znanjem

Za upravljanje znanjem INCEpTION podržava baze znanja temeljene na RDF-u (engl. *Resource Description Framework*). Dok se manje interne baze znanja mogu koristiti za znanje specifično za određeno područje, velikim se vanjskim bazama znanja može pristupiti putem SPARQL-a (engl. *SPARQL Protocol and RDF Query Language*). Fleksibilni konfiguracijski mehanizam koristi se za podršku različitim prikazima znanja, kao što su Wikidata, DBPedia, OWL, CIDOC-CRM, SKOS, itd. Međutim, njime se ne nastoji ponuditi potpuna potpora naprednim značajkama programa kao što je OWL. Baze znanja omogućuju korisniku da stvara oznake temeljene na znanju, na primjer, oznake o subjektima baze znanja u dokumentima (na slici 9 povezivanje subjekata označeno brojkom 3) ili stvaranje novih baza znanja označavanjem subjekata, predikata i objekata u tekstu (na slici 9 povezivanje činjenica označeno brojkama 7 i 8). Korisnici također mogu istražiti i urediti sadržaj baze znanja unutar INCEpTION-a. Kako bi se olakšao postupak povezivanja subjekta, INCEpTION može neobavezno uzeti u obzir kontekst navođenja subjekta kako bi korisniku dostavio rangirani popis potencijalnih kandidata. Isti se pristup koristi za pokretanje algoritma za preporuke koji prikazuje visokorangirane kandidate kao prijedloge za označavanje (na slici 9 označeno brojkom 9) u području za označavanje gdje ih korisnik može prihvatiti jednim klikom (Klie i dr., 2018).

6.2.5. Prilagodba i proširenje

Postoje dva pristupa za prilagodbu i proširenje INCEpTION-a:

a) Unutarnja proširenja

Mehanizmi za događaje i injektiranje ovisnosti Spring Boot3 koriste se kako bi interno modularizirali INCEpTION. Točke proširenja omogućuju registriranje novih vrsta oznaka i svojstva oznaka, novih urednika ili novih internih algoritama za preporuke. Moduli mogu

koordinirati svoje zadatke jedni s drugima kroz događaje. Na primjer, glavno područje za označavanje izdaje događaj kada je oznaka stvorena ili promijenjena. Algoritmi za preporuke i način rada aktivnog učenja reagiraju na ovaj događaj kako bi se ažurirali. Funkcionalnost se stoga može ne samo dodati, već i ukloniti kako bi se stvorile prilagođene verzije INCEpTION-a. Modularni pristup koji se temelji na događajima također omogućuje sustavu sveobuhvatno bilježenje korisničkih i sistemskih radnji. Te podatke, na primjer, mogu upotrebljavati voditelji anotacijskih projekata kako bi ocijenili uspješnost procesa označavanja (Klie i dr., 2018).

b) Vanjska proširenja

Trenutačno su podržani vanjski algoritmi za preporuke i baze znanja. Prednosti korištenja vanjskih usluga uključuju povećanu stabilnost (neuspješne usluge ne ruše cijelu platformu), skalabilnost (uvođenje usluga resursa kojima trebaju podaci na različitim strojevima) i slobodan izbor programskog jezika (primjerice, većina okvira za duboko učenje ne provodi se u Javi). Osim toga, INCEpTION koristi (de facto) standarde kao što su UIMA i RDF za anotiranje te OWL i SPARQL za baze znanja kako bi se postigla visoka razina interoperabilnosti s postojećim alatima i resursima (Klie i dr., 2018).

6.2.6. Primjeri upotrebe

Kako bi se osiguralo da INCEpTION ostane generički, korišten je u suradnji s istraživačima iz raznih područja. Slijedi nekoliko primjera suradnji:

a) FAMULUS

Schulz i dr. (2017) koristili su INCEpTION za označavanje izvješća medicinskih studija slučaja s argumentacijskim dijelovima. Oni se koriste za obuku modela strojnog učenja koji procjenjuje dijagnostičku kompetenciju budućih liječnika. Unaprijed obučeni model dubokog učenja integriran je kao vanjski algoritam za preporuke i koristi se tijekom označavanja. Korisnici INCEpTION-a s algoritmima za preporuke izvješćuju o korisnosti i poboljšanju brzine i kvalitete oznaka.

b) EDoHa

Stahlhut i dr.. (2018.) izradili su alat za validaciju hipoteza s pomoću alata INCEpTION. EDoHa sadrži uređivač hipoteza ili dokaza koji korisnicima omogućuje stvaranje hipoteza i povezivanje dokaza u obliku tekstualnih odlomaka.

c) Rangiranje subjekata temeljenih na znanju

Kako bi se korisnicima pružila potpora tijekom povezivanja sa subjektom, pristup ponovnog rangiranja opisan u radu Sorokin i Gurevych (2018) prilagođen je i integriran u INCEpTION. Koristi se kao preporuka u okviru za automatsko predlaganje za imenovani entitetski sloj.

U okviru suradnje s navedenim alatima INCEpTION bilježi aktivnosti korisnika, na primjer, koje pomoćne značajke najbolje odgovaraju određenim zadacima, uvode li pristranost u rezultate označavanja te kako poboljšati korisničko sučelje za poboljšano korisničko iskustvo. Korisnik, naravno, ne mora koristiti sve značajke ovog alata. Više o tome kako se ovaj alat za označavanje koristio u svrhu izrade ovog diplomskog rada i anotiranje podatkovnog skupa SentiGMapsCro u sljedećem poglavlju.

7. Metodologija istraživanja

U području obrade prirodnog jezika postoji mnogo istraživanja, jedno od kojih je i završni rad autorice ovog diplomskog rada pod nazivom „Leksikon emocija hrvatskog jezika“ (Gašparić, 2020). Za potrebe svog završnog rada, autorica je ispravljala automatski generirane prijevode Google Prevoditelja s engleskog jezika na hrvatski jezik kako bi ispravan leksikon za hrvatski jezik postao dio *EmoLexa* (engl. *NRC Word-Emotion Association Lexicon;* Mohammad, 2013), rječnika od 14 182 riječi od kojih je svaka obilježena sentimentom i emocijom koju iskazuje. Najveći problemi s kojima se autorica susrela tijekom ispravljanja prijevoda i emocija za završni rad ponavljaju se i diplomskom radu, a tu su problem višeznačnosti i subjektivnosti, te nekoliko novih poteškoća koje se javljaju tijekom označavanja recenzija metodom analize sentimenta – pravopisna i gramatička točnost recenzija u podatkovnom skupu SentiGMapsCro te prepoznavanje sarkazma ironije i ostalih figurativnih načina izražavanja kojih je prirodni jezik prepun.

Izrada ovog diplomskog rada započela je istraživanjem doktoranda Gaurisha Panduranga Thakkara pod mentorstvom zajedničke mentorice prof.dr.sc. Nives Mikelić Preradović s Filozofskog fakulteta Sveučilišta u Zagrebu. U suradnji s kolegom Thakkarom stvoren je podatkovni skup SentiGMapsCro koji je uvezen u alat za označavanje tekstualnih dokumenata zvan INCEpTION. Sam je podatkovni skup sastavljen od pozitivnih i negativnih recenzija izvučenih s internetskih stranica za recenziranje mjesta, objekata i slično pod nazivom „Top rated“ (https://www.top-rated.online).

Nakon što je mnoštvo recenzija uvezeno u alat INCEpTION, filtrirane su isključivo recenzije na hrvatskom jeziku te je 4. ožujka 2022. godine nastao podatkovni skup SentiGMapsCro koji sadržava tisuće recenzija. Tada je počela druga faza istraživanja u kojoj je autorica do početka srpnja 2022. godine označavala riječi (posebice pridjeve), izraze, a ponekad i cijele surečenice metodom analize sentimenta temeljene na aspektu. U ovoj fazi istraživanja definirana je analiza sentimenta podijeljena u pet kategorija, a grafički prikaz s primjerima dostupan je na slici 10:

- -2 (vrlo negativan sentiment)
- -1 (negativan sentiment)
- 0 (neutralan sentiment)
- 1 (pozitivan sentiment)
- 2 (vrlo pozitivan sentiment)

BROJČANA VRIJEDNOST	-2	-1	0	+1	+2
VRIJEDNOST IZRAŽENA RIJEČIMA	VRLO NEGATIVNO	NEGATIVNO	NEUTRALNO	POZITIVNO	VRLO POZITIVNO
PRIMJER	„užasno neprikladno“	„neprikladno“	“Došli smo i otišli.“	„prikladno“	„zaista prikladno“

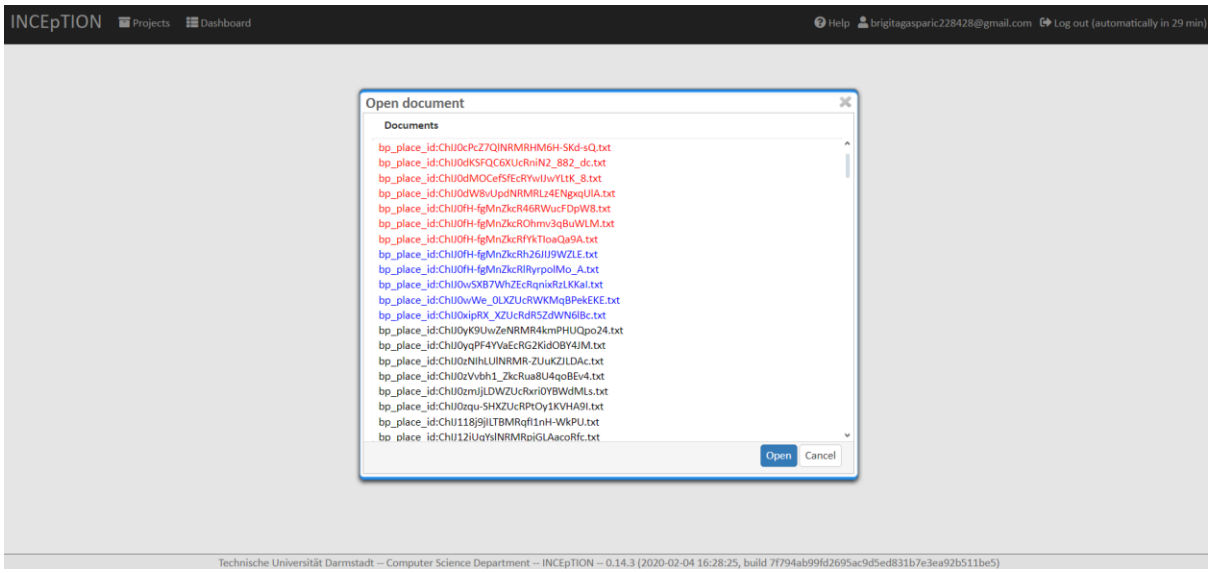
Slika 10. Primjer označavanja sentimenta

U alat INCEpTION uvezena su 2482 dokumenta te je obilježeno ukupno 82 dokumenata. Mnogi su dokumenti bili na drugim jezicima ili su sadržavali samo brojčanu oznaku od jedan do pet. Ti su dokumenti i recenzije preskočeni kako bi se obilježilo što više opisnih recenzija na hrvatskom jeziku. Proces je rezultirao označavanjem sentimenta više od 500 recenzija po uzoru na „Tweet Sentiment Extraction Data“ (Kaggle, 2020). Navedeni kriteriji bit će detaljno objašnjeni i potkrijepljeni primjerima u idućem poglavlju.

U posljednjoj je fazi istraživanja fokus bio na analizi označenih riječi, izraza i rečenica, bilježenje problema na koje je autorica naišla tijekom označavanja te izdvajanje reprezentativnih primjera za svaki od njih. Neke od najizraženiji poteškoća uključuju korištenje nestandardnog pisma, gramatičku i pravopisnu netočnost, spajanje interpunkcije s kraja rečenice s prvom riječi sljedeće rečenice, tumačenje višeznačnosti na razini leksika i razini sintakse, segmentaciju teksta te prepoznavanje korištenih jezičnih figura.

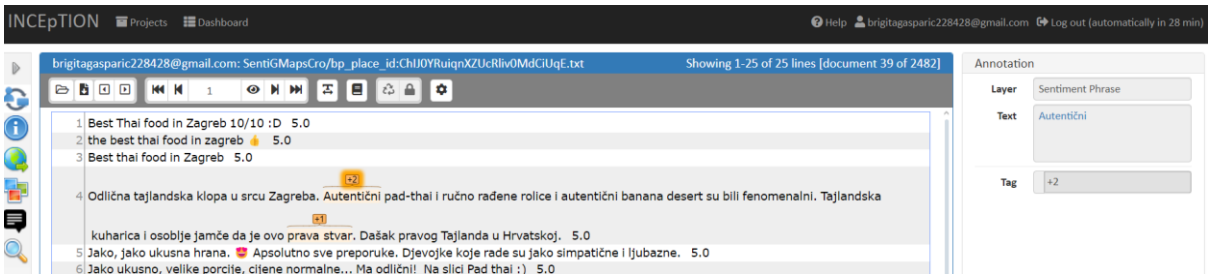
8. Analiza mnijenja u podatkovnom skupu SentiGMapsCro

Nakon upoznavanja s INCEpTION-om i uvođenja odabranih najboljih i najgorih recenzija sa stranice „Top rated“ za Grad Zagreb, započelo je označavanje/obilježavanje sentimenta. Nakon ulaska u alat INCEpTION, potrebno je odabrati opciju „Annotation“ u gornjem desnom kutu koja je vidljiva na slici 8. Nakon odabira opcije „Annotation“, otvara se popis dokumenata (podatkovni skup SentiGMapsCro) koje treba označiti. Izgled tog prozora vidljiv je na slici 11.



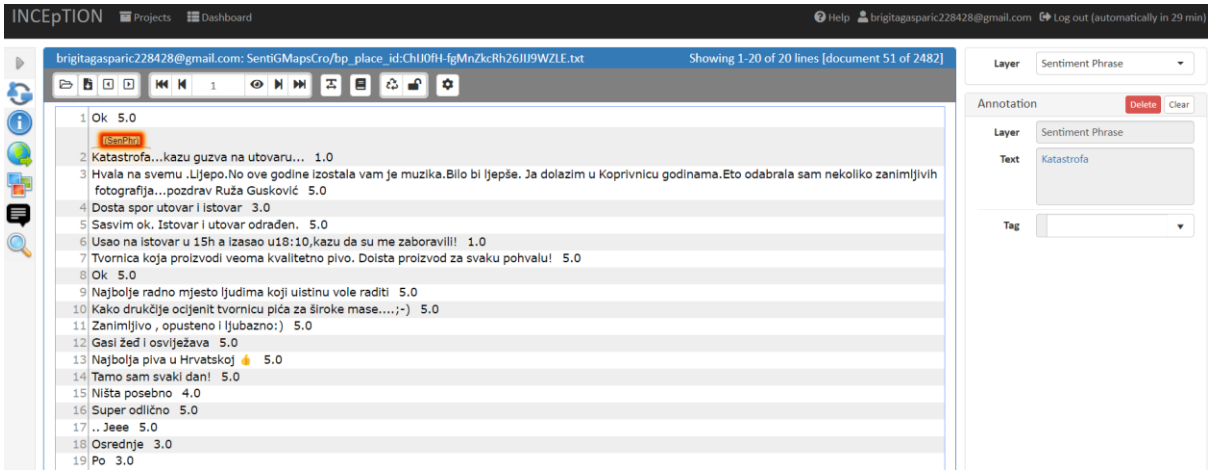
Slika 11. Popis dokumenata podatkovnog skupa SentiGMapsCro u INCEpTION-u

Dokumenti unutar podatkovnog skupa SentiGMapsCro na slici 11 obojeni su u tri boje: crvenu, plavu i crnu. Crvena boja označava tekstualnu datoteku unutar koje su sve recenzije označene te je datoteka zaključana i više se ne može uređivati. Izgled jedne zaključane datoteke obojene crvenom bojom dostupan je na slici 12:



Slika 12. Izgled zaključane datoteke u INCEpTION-u

Ako je datoteka obojena plavom bojom, to znači da je ta datoteka otvorena i označavanje je započeto, ali nije zaključana te se unutar same datoteke još uvijek mogu vršiti izmjene u označavanju recenzija. Primjer otključane datoteke obojene plavom bojom vidljiv je na slici 13:

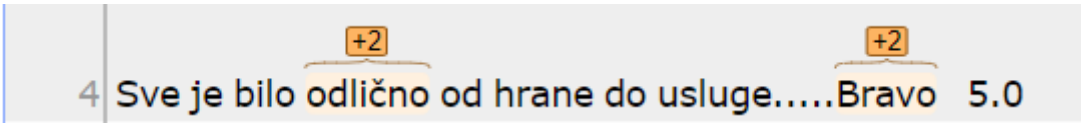


Slika 13. Izgled otključane datoteke u INCEpTION-u

Dok na je slici 12 desna bočna traka za označavanje siva te se vrijednosti ne mogu mijenjati, datoteka na slici 13 ima otključanu bočnu traku za označavanje s obojenim slovima odabrane riječi što znači da se unutar dokumenta u recenzijama još uvijek mogu mijenjati oznake. Datoteke obojene crnom bojom daju korisniku do znanja da datoteka nikad nije otvorena od strane bilo kojeg korisnika na trenutnom projektu.

8.1. Obilježavanje recenzija

Za potrebe ovog rada definirana je analiza sentimenta podijeljena u pet kategorija (od -2 do +2). Osoba koja je obilježavala sentimente ovih recenzija bila je autorica ovog rada, stoga su sve recenzije obilježene samo i isključivo prema njezinoj interpretaciji oznaka od vrlo negativnog do vrlo pozitivnog sentimenta. Uzmimo za primjer sljedeću recenziju na slici 14:



Slika 14. Primjer 1 – obilježavanje

Prema dokumentu „Tweet Sentiment Extraction Data“ (Kaggle, 2020), nakon što imamo definirane kategorije za obilježavanje sentimenta, možemo početi obilježavati. Bitno je naglasiti da je autorica morala prilagoditi način obilježavanja potrebama svog rada i hrvatskom jeziku. Dok upute za „Tweet Sentiment Extraction Data“ traže od istraživača za pronađu primjere za određene sentimente, ovaj se rad fokusirao na obrnuto: odabranim se recenzijama obilježavao sentiment prema unaprijed određenim kategorijama. Na Primjeru 1 vidljiva je recenzija unutar koje su riječi „odlično“ i „bravo“ označene vrlo pozitivnim sentimentom u obliku oznake +2. Cilj ovakvog označavanja sentimenta jest da se obilježenim riječima pridruži sentiment koji u najvećem broju slučajeva odražava njihovo značenje neovisno o kontekstu. Dakle, riječi poput „odlično“ i „bravo“ će u gotovo svakom kontekstu (izuzev određenih jezičnih figura poput ironije i sarkazma) izražavati vrlo pozitivan sentiment.

Također je bitno napomenuti da se izbjegavalo višestruko označavanje istih riječi i izraza te obilježavanje segmenata unutar recenzija neutralnima (oznakom 0). Iz tog je razloga samo osamnaest riječi i izraza označeno sentimentom oznake 0, a neke od njih su: *prolazno, OK, tako, solidno, tipično, u redu, usput, tako-tako.* Primjer recenzije obilježene neutralnim sentimentom dostupan je na slici 15:

Slika 15. Primjer 2 – neutralna recenzija

8.2. Kategorije obilježenih segmenata recenzija

Kako bi se olakšala analiza obilježenih segmenata recenzija, određene su tri kategorije od kojih će svaka biti objašnjena i potkrijepljena primjerima. Kategorije uključuju obilježenu jednu riječ te obilježenu rečenicu ili izraz. Kao treću kategoriju također treba istaknuti jezične figure metaforu, ironiju i sarkazam.

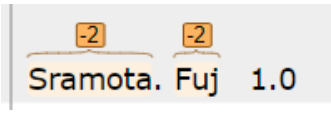
8.2.1. Jedna riječ

Ovo je najjednostavnija kategorija koja je ujedno i najčešća. Mnoge su recenzije sadržavale pozitivne ili negativne pridjeve i imenice koje odaju sentiment čak i izvan konteksta same recenzije. Neki od primjera su:

- *OK* – neutralan sentiment (0),

- *ljubaznost, profesionalnost, odlično, bravo, ukusno, dobro, super, fantazija, gostoljubivo, najbolji, vrhunski, uredno, vrh*, *top* – pozitivan ili vrlo pozitivan sentiment (+1 ili +2),
- *neodržavano, musavo, zadimljeno, neukusno, bezobrazno, prljavo, fuj, strašno, odron, sramota* – negativan ili vrlo negativan sentiment (-1 ili -2).

Reprezentativan primjer ovakve recenzije vidljiv je na slici 16:



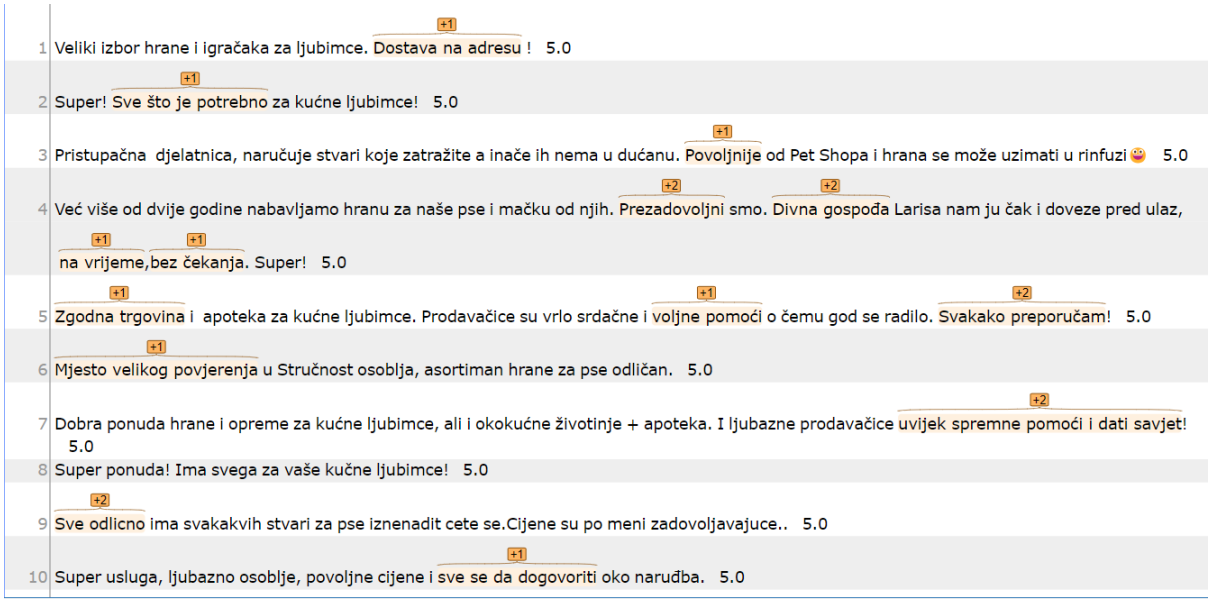
Slika 16. Primjer 3 – jedna riječ

8.2.2. Rečenica ili izraz

One recenzije koje su bile više opisnog karaktera bilo je vrlo izazovno obilježiti označavanjem sentimenta za samo jednu riječ, stoga je bilo potrebno obilježiti više riječi koje zajedno odaju određeno značenje. Neki od primjera su:

- *svidjelo nam se, za svaku preporuku, sve pohvale, jako dobro, dobar asortiman, jako čisto, pristupačne cijene, usluga na nivou, nikad nisam doživio neugodnosti* - pozitivan ili vrlo pozitivan sentiment (+1 ili +2),
- *treba bolje održavanje, prevelika gužva, puno smeća, prekratko radno vrijeme, vrlo iritantno, spora usluga, nisam oduševljen* - negativan ili vrlo negativan sentiment (-1 ili -2).

Primjer obilježenih izraza možemo vidjeti na slici 17:



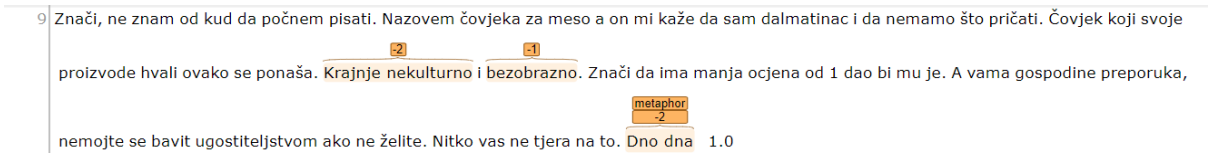
Slika 17. Primjer 4 – izrazi

8.2.3. Jezične figure – metafora i ironija

Od jezičnih figura, recenzenti su se najviše koristili metaforom i ironijom. Metafora se koristila kako bi se slikovito izrazio i vrlo pozitivan i vrlo negativan sentiment:

- *sunce na obzoru, pun pogodak, vrh, čista petica, prva liga, kolači mame na grijeh, zakon, ludilo, tolko dobar da ti mozak stane* - vrlo pozitivan sentiment (+2),
- *pretpotopna oprema, trgnite se malo, Bože sačuvaj, dno dna, rak rana, odron; konobar jednom nogom u grobu, poljubili smo vrata, raspad sistema, banana, ponudom i izgledom zastao negdje u 20. stoljeću, duša me boli, klinički mrtav, žali Bože* - vrlo negativan sentiment (-2).

Primjer jedne negativne recenzije s metaforom možemo vidjeti na slici 18:

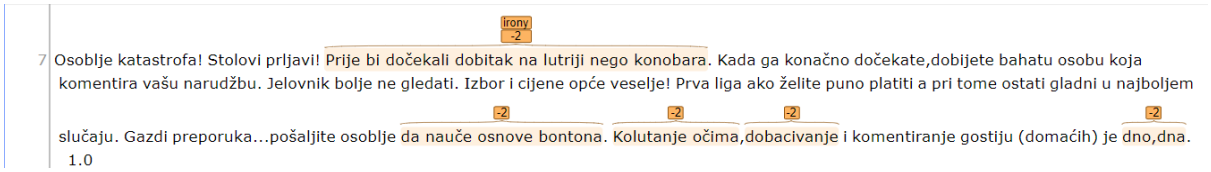


Slika 18. Primjer 5 – metafora

Ironija se s druge strane koristila kako bi se na vrlo oštar način izrazilo nezadovoljstvo (vrlo negativan sentiment s oznakom -2):

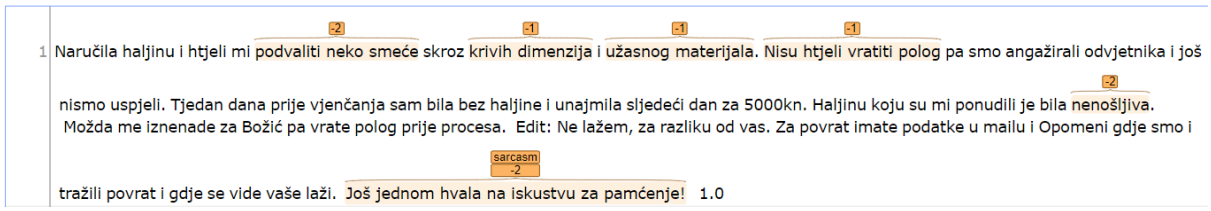
- super vam je tu, za ubiti se;
- nemam komentara;
- doživljaj je super ako želite ostati gladni;
- još da i radi.

Primjer jedne recenzije s ironijom dostupan je na slici 19:



Slika 19. Primjer 6 – ironija

Sarkazma nije bilo pretjerano mnogo, ali se jedna recenzija posebno ističe sarkazmom koji je u zasebnoj rečenici i ne bi se na isti način shvatila recenzija da se tu rečenicu izvuče iz konteksta:



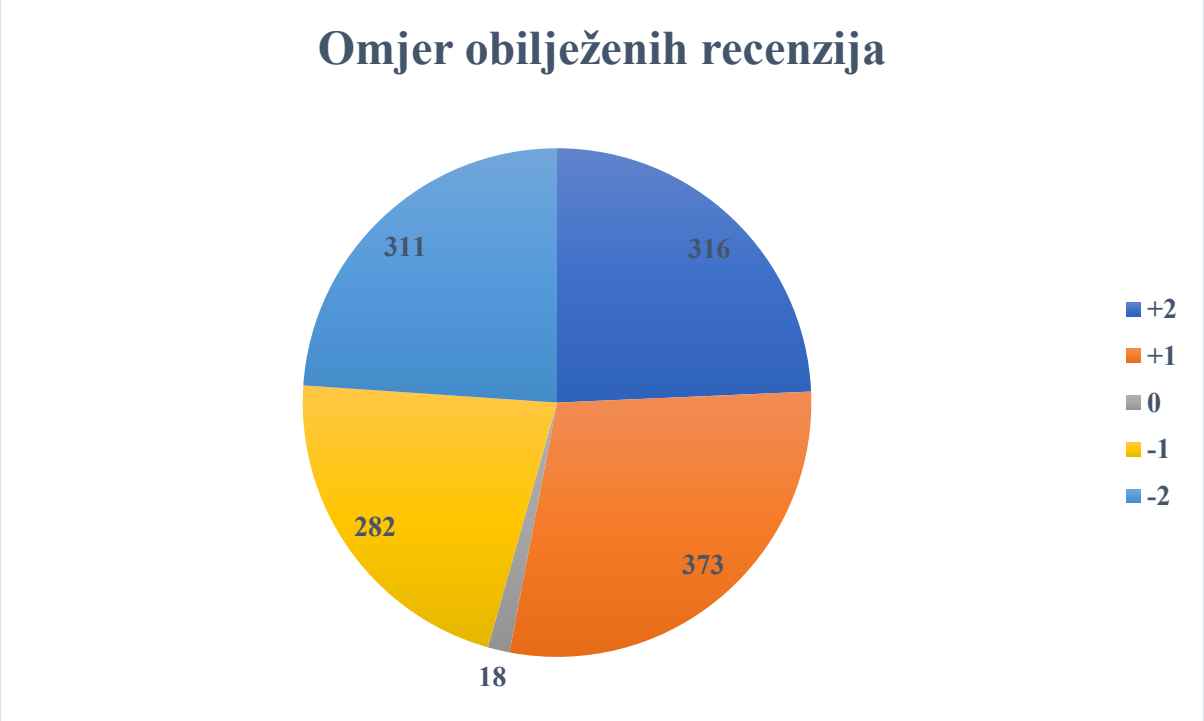
Slika 20. Primjer 7 – sarkazam

8.3. Rezultati

Nakon završetka obilježavanja sentimenata recenzija te kategoriziranja istih, zanimljivo je i ustanoviti koliko je točno puta svaka oznaka bila iskorištena. Iako se u korisničkom priručniku za INCEpTION dostupnom na internetskoj stranici https://inception-project.github.io/releases/33.4/docs/user-guide.html#_getting_started (INCEpTION Project, n.d.) navodi da bi INCEpTION trebao imati opciju statističkih podataka, oni u ovom projektu nisu bili dostupni. Iz tog je razloga autorica još jednom prošla sve obilježene recenzije i u Excelu zapisivala koliko je puta svaka oznaka korištena u svakom dokumentu. Taj je proces dostupan na slici 21:

Slika 21. Proces brojanja iskorištenih oznaka

Ovaj je proces rezultirao sljedećim podacima vidljivima na slici 22:



Slika 22. Omjer obilježenih recenzija

Kao što je vidljivo na slici iznad, najviše puta se koristila oznaka +1 za pozitivan sentiment koja je iskorištena 373 puta. Zatim slijedi oznaka +2 za vrlo pozitivan sentiment koja je iskorištena 316 puta te odmah potom oznaka -2 koja označava vrlo negativan sentiment, a

iskorištena je 311 puta. Pretposljednja je oznaka -1 za negativan sentiment koja je iskorištena 282 puta te je, očekivano, najmanje korištena oznaka 0 za neutralan sentiment.

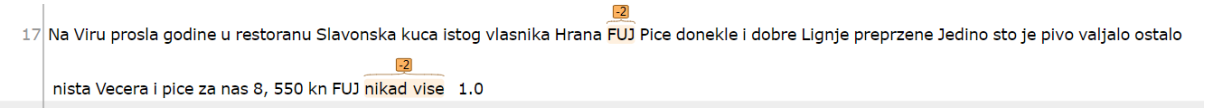
8.4. Problemi tijekom obilježavanja

Prirodni je jezik poput živog organizma, mijenja se i adaptira potrebama ljudi; točnije, ljudi koriste jezik vrlo individualno i subjektivno za svoje potrebe kako uživo tako i pismeno putem Interneta (Toppelberg i Collins, 2010). Iako mnogi istraživači već desetljećima pokušavaju prevesti prirodni jezik u oblik koji računala mogu razumjeti korištenjem raznih metoda, taj proces još uvijek nije u potpunosti moguć zbog mnogostrukih problema s kojima se susrećemo, pogotovo u procesu analize sentimenta podataka koji su izvorno pisani od strane ljudskih korisnika. Neke od tih metoda su dosegle zaista zavidnu razinu, pa se tako mnogi algoritmi i metode spomenuti u prethodnim poglavljima koriste za obradu prirodnog jezika u slučaju svjetskih jezika s mnogo govornika s minimalnom intervencijom istraživača.

To nažalost nije slučaj s hrvatskim jezikom koji nema ni približan broj govornika poput, recimo, mandarinskog, engleskog ili španjolskog jezika. Stoga nas ne čudi činjenica da proces analize prirodnog jezika u slučaju hrvatskog ne može odraditi bez istraživača kojima je hrvatski materinji jezik. Uzimajući u obzir činjenicu da su internetske recenzije dostupne gotovo svima te ih korisnici često pišu u žurbi i pod pozitivnim ili negativnim dojmom, analizirat ćemo nekoliko najčešćih problema na koje je autorica naišla prilikom obilježavanja podatkovnog skupa SentiGMapsCro.

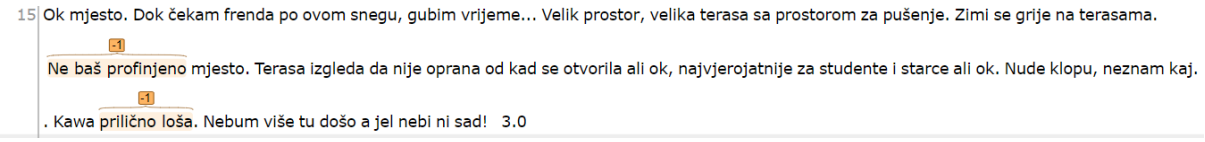
8.4.1. Gramatička i pravopisna netočnost

Jedan od prvih problema na koji se naišlo u procesu obilježavanja jest gramatička i pravopisna netočnost u tekstu recenzija. Svrha ovog procesa analize sentimenta na temelju aspekta i samog obilježavanja jest svojevrsno „prevođenje“ sentimenata prirodnog jezika u oblik koji se kasnije metodom strojnog učenja može obraditi pomoću računala te se na taj način koristiti u razne druge svrhe unutar područja obrade prirodnog jezika (primjerice, stvaranja algoritama koji će samostalno moći određivati sentimente recenzija na odabranim internetskim stranicama). Upravo su iz tog razloga gramatička i pravopisna netočnost u tekstovima recenzija nepoželjne: cilj je obilježavati standardne strukture i forme hrvatskog jezika na koje bi računala u kasnijim procesima analize sentimenta mogla naići, što sljedeći primjeri nažalost nisu.



Slika 23. Primjer 8 - dijakritički znakovi

Na primjeru 8 možemo vidjeti reprezentativan primjer recenzije u kojoj se izbjegavaju hrvatski dijakritički znakovi (Babić, Finka i Moguš, 1995, str. 23-29). Ovo izbjegavanje dijakritika zaista nije neuobičajeno na Internetu, zbog čega su neke riječi poput *vise* [više] i *moze* [može] obilježene, ali se u pravilu izbjegavalo obilježavati netočno napisane riječi.



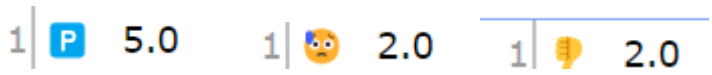
Slika 24. Primjer 9 - gramatika i pravopis

Također se izbjegavalo obilježavati netočno napisane riječi poput *ljepo* umjesto *lijepo* te *odlićan* umjesto *odličan*. Na primjeru 9 na slici iznad možemo vidjeti jednu recenziju koja je napisana s nekoliko gramatičkih i pravopisnih pogrešaka (*neznam* umjesto *ne znam*), ali se još uvijek mogla iskoristiti za obilježavanje točno napisanih dijelova.

8.4.2. Nestandardno pismo

8.4.2.1. EMOTIKONI

Mnoge recenzije uopće ne sadrže riječi, ali sadrže emotikon (engl. *emoji*). Ovisno o vrsti emotikona koji je korisnik odabrao i brojčanoj recenziji koja je dana, možemo odrediti sentiment te recenzije. Slijedi nekoliko primjera:



Slika 25. Primjeri emotikona

Iz primjera na slici 25 možemo ustanoviti da su druga i treća recenzija negativne i bez brojčane oznake ocjene pored njih. No, prva recenzija koja sadržava samo emotikon koji označava parkirno mjesto ili parkiralište ne bismo mogli zaključiti je li to pozitivna ili negativna recenzija bez ocjene 5.0 pored same recenzije. Dakle, možemo zaključiti da je korištenje nestandardnog pisma u obliku emotikona vrlo često i općeprihvaćeno na

Internetu jer može odati sentiment korisnika bez korištenja riječi ili amplificirati ton recenzije ako je emotikon korišten u sklopu same recenzije. Kako bilo, emotikoni nisu predmet proučavanja ovog rada, stoga nisu obilježeni oznakama za sentiment.

8.4.2.2. Žargonizmi

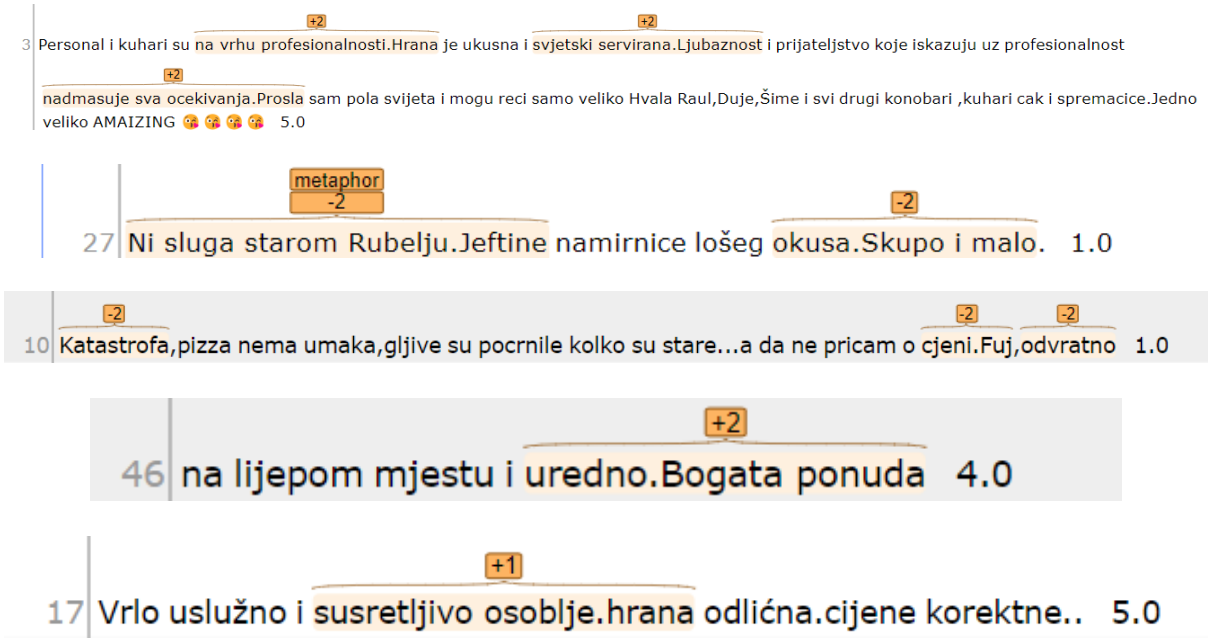
Uporaba žargonizama ili slenga spada u nestandardno pismo jer su to riječi jezičnog idioma koji nije dio hrvatskog standardnog jezika te se tretiraju kao leksičke figure (Škarić, 2008). U ovom su radu neki žargonizmi obilježeni jer odaju sentiment recenzije te su prema mišljenju autorice dovoljno česti da postoji velika mogućnost da se ponovno pojave u drugim recenzijama. Korištenje žargonizama je vrlo često, pogotovo na Internetu, stoga sve veću uporabu slenga treba uzeti u obzir prilikom određivanja sentimenta, pogotovo ako je cilj analize dobivanje korisničkih povratnih informacija o uslugama i proizvodima (Manuel, Indukuri i Krishna, 2010). Neki od najčešćih primjera su:

- Baš je *lit;*
- top topova;
- bedara.

8.4.2.3. Problemi nadrečeničnog označavanja sentimenta

Još jedan problem tijekom obilježavanja sentimenta odnosi se na segmentaciju teksta. To je zadatak izdvajanja relevantnih jedinica u tekstu poput riječi, fraza, rečenica, a ponekad i odlomaka iz tekstova. Segmentacijom odabranih dijelova iz teksta istraživači se mogu detaljnije fokusirati na zadatke poput segmentacije riječi, detekcije granica rečenice, sažimanja teksta i analize osjećaja (Karahan, 2021).

U ovom se radu ovaj problem manifestira u vidu pogrešnog korištenja interpunkcija. Konkretnije, neki korisnici u svojim recenzijama iza interpunkcije na kraju rečenice nisu stavili razmak, pa je INCEpTION automatski odabirao riječi i prije i poslije interpunkcije, što se nažalost u samom alatu ne može korigirati. Ti su dijelovi teksta obilježeni bez obzira na pogrešne interpunkcije u nadi da će se to u nekoj novoj verziji INCEpTION-a moći ispraviti. Primjeri takvih recenzija dostupni su na slikama ispod:



Slika 26. Primjeri problema sa segmentacijom teksta

9. Zaključak

Područje obrade prirodnog jezika vrlo je široko te nudi mnoge mogućnosti istraživanja, pogotovo za jezike s malim brojem govornika poput hrvatskog jezika. Internet nam danas nudi mnoštvo podataka u obliku strukturiranih i nestrukturiranih jezičnih jedinica iz kojih se mogu iščitavati sentimenti. Unatoč svim metodama i pristupima kojima istraživači pokušavaju dekodirati i brojčano obilježiti prirodni jezik kroz subjektivni stav pojedinca, područje analize sentimenta još uvijek pokušava pronaći rješenja za lingvističke prepreke unutar računalne obrade prirodnog jezika.

Podatkovni skup SentiGMapsCro sadržava najbolje i najgore recenzije objekata i lokacija na području Grada Zagreba te je obilježen uz pomoć alata INCEpTION. Unatoč alatu za obilježavanje sentimenata odabranih recenzija, proces je trajao duže od očekivanog te je sam proces obilježavanja bio otežan problemima od kojih su neki bili očekivani, ali ne svi. Primjerice, lingvistički problemi računalne obrade prirodnog jezika vrlo su česti prilikom analize sentimenta, stoga je bilo za očekivati da će recenzije od strane ljudskih korisnika biti gramatički i pravopisno netočne te da će sadržavati žargon ili sleng. S druge strane, sam je alat INCEpTION razočarao jer nije imao mogućnost naknadne segmentacije teksta te statističke analize obilježenih dokumenata. Obje bi mogućnosti uvelike olakšale i ubrzale proces obilježavanja jer bi označeni dijelovi rečenica bili točno segmentirani (sa razmakom nakon interpunkcije pri kraju rečenice) te bi se izbjeglo ponovno pregledavanje i prebrojavanje oznaka svih obilježenih recenzija kako bi se dobili statistički podaci o ovom podatkovnom skupu.

Također, sam alat INCEpTION bi trebalo modernizirati i učiniti prilagodljivijim korisniku. Dizajn alata izgleda zastarjelo, ali alat ispunjava svoju svrhu. Podatkovni je skup SentiGMapsCro djelomice obilježen te se ti podaci mogu koristiti u mnoge svrhe unutar područja računalne obrade prirodnog jezika, posebice strojno učenje. To se područje još uvijek susreće s mnogim lingvističkim problemima, od kojih su neki specifični za hrvatski jezik, stoga je potrebno dalje razvijati metode i alate koji će se usredotočiti na rješavanje te problematike.

Literatura

1. Aggarwal, C. (2015). Data classification. U *Data Mining* (str. 285-344). Springer, Cham. https://doi.org/10.1007/978-3-319-14142-8_10.
2. Aliandu, P. (2015). Sentiment analysis to determine accommodation, shopping and culinary location on Foursquare in Kupang City. *Procedia Computer Science, 72*, str. 300-305. https://doi.org/10.1016/j.procs.2015.12.144.
3. Babić, S., Finka, B. i Moguš, M. (1995). *Hrvatski pravopis*. Školska knjiga.
4. *BERTIĆ – model na hrvatskom jeziku*. (2023). EkonInfoChecker. Pristupljeno 16. srpnja 2024. Dostupno na https://ekoninfochecker.efri.uniri.hr/?p=735.
5. Bing, L. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool.
6. Bolf, N. i Bolf (ur.), N. (2021). Osvježimo znanje: Strojno učenje. *Kemija u industriji*, 70 (9-10), 591-593. Preuzeto s https://hrcak.srce.hr/263495.
7. Borrajo-Millán, F., Alonso-Almeida, M.-d.-M., Escat-Cortes, M. i Yi, L. (2021). Sentiment analysis to measure quality and build sustainability in tourism destinations. *Sustainability, 13*(11), 6015. https://doi.org/10.3390/su13116015.
8. Canals, L. (2017). *Instruments for gathering data*. Research-publishing.net. https://eric.ed.gov/?id=ED573582
9. Clark, K., Luong, M.-T., Le, Q. V. i Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. *arXiv*. https://arxiv.org/abs/2003.10555.
10. Devlin, J., Chang, M. W., Lee, K. i Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805. https://arxiv.org/abs/1810.04805.
11. Devlin, J., Chang, M.-W., Lee, K. i Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. https://arxiv.org/abs/1810.04805.
12. Gašparić, B. (2020). *Leksikon emocija hrvatskog jezika* (završni rad). ODRAZ - open repository of the University of Zagreb, Faculty of Humanities and Social Sciences. https://urn.nsk.hr/urn:nbn:hr:131:577940.
13. INCEpTION Project. (n.d.). *User guide: Getting started (Version 33.4)*. Dostupno na https://inception-project.github.io/releases/33.4/docs/user-guide.html#_getting_started.

14. Jabbar, M. A., Okfantia, A. D., Widjanarti, A. i Zulen, A. A. (2022). Sentiment analysis of tourist reviews from online travel forum for improving Indonesia tourism sector. U *IFC-Bank of Italy Workshop on "Data Science in Central Banking: Applications and tools"*. Bank Indonesia. https://www.bis.org/ifc/publ/ifcb59_18.pdf.
15. Kaggle. (2020). *Tweet Sentiment Extraction Data.* Pristupljeno 16. srpnja 2024. https://www.kaggle.com/c/tweet-sentiment-extraction/data.
16. Karahan, S. (2021). Text segmentation and its applications to aspect-based sentiment analysis. *Medium*. Dostupno na https://medium.com/artiwise-nlp/text-segmentation-and-its-applications-to-aspect-based-sentiment-analysis-fb115f9ab4e9.
17. Klie, J.-C., Bugert, M., Boullosa, B., de Castilho, R. E. i Gurevych, I. (2018). The INCEpTION platform: Machine-assisted and knowledge-oriented interactive annotation. U *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations* (str. 5–9). Santa Fe, NM: Association for Computational Linguistics.
18. Kokan, R. (2021). *Obrada prirodnog jezika - chatbot programi (Diplomski rad)*. Zagreb: Sveučilište u Zagrebu, Prirodoslovno-matematički fakultet. https://urn.nsk.hr/urn:nbn:hr:217:017085.
19. Kumar Laskari, N. i Kumar Sanampudi, S. (2016). Aspect Based Sentiment Analysis Survey. *IOSR Journal of Computer Engineering (IOSR-JCE)*, str. 24 - 28.
20. Lewis, D. D. i Gale, W. A. (1994). A sequential algorithm for training text classifiers. U *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (str. 3–12).
21. Liddy, E. D. (2003). Natural language processing. U *Encyclopedia of Library and Information Science* (2. izdanje, str. 2126-2136). Marcel Decker, Inc.
22. Liu, B. (2012). *Sentiment analysis and opinion mining*. Morgan & Claypool.
23. Ljubešić, N. i Lauc, D. (2021). BERTić - The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. U *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing* (str. 37–42). Kyiv, Ukraine: Association for Computational Linguistics.
24. Malmberg, B. (1974). *Fonetika*. Sarajevo: Svjetlost.
25. Manning, C. D. i Schutze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.
26. Manuel, K., Indukuri, K. V. i Krishna, P. R. (2010). Analyzing internet slang for sentiment mining. U *Proceedings of the 2010 Second Vaagdevi International*

Conference on Information Technology for Real World Problems (str. 9–11). https://doi.org/10.1109/ICITRWP.2010.16478997.

27. Medhat, W., Hassan, A. i Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal, 5*(4), str. 1093-1113. https://doi.org/10.1016/j.asej.2014.04.011.
28. Mikolov, T., Chen, K., Corrado, G. i Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint* arXiv:1301.3781. https://arxiv.org/abs/1301.3781.
29. Mohammad, S. (2013*). NRC Emotion Lexicon*. [online] saifmohammad.com/WebPages/NRC-Emotion-Lexicon. Dostupno na: saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.
30. Niu, Y., Zhu, X., Li, J. i Hirst, G. (2005). Analysis of polarity information in medical text. *AMIA Annual Symposium Proceedings, 2005*, str. 570-574. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1560818/pdf/amia2005_0570.pdf.
31. Pang, B. i Lee, L. (2004). A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. *Proceedings of the 42nd ACL,* str. 271-278.
32. Pang, B., i Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 2*(1-2), str. 1-135.
33. Petrovečki, M., Bilić-Zulle, L. i Pupovac, V. (2008). *Prikupljanje podataka i mjerenje*. Katedra za medicinsku informatiku, Medicinski fakultet Sveučilišta u Rijeci. Retrieved from http://mi.medri.hr/assets/opuz/P2.pdfv.
34. Poibeau, T. (2017). *Machine translation*. Cambridge, MA: MIT Press Essential Knowledge Series.
35. Radford, A., Narasimhan, K., Salimans, T. i Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.
36. Raos, N. (2012). *Mišljenja i komentari: Bofl*. Institute for Medical Research and Occupational Health.
37. Rosli, M. M., Tempero, E. i Luxton-Reilly, A. (2016). What is in our datasets? Describing a structure of datasets. U *Proceedings of the Australasian Computer Science Week Multiconference* (Članak broj 28). Association for Computing Machinery. https://doi.org/10.1145/2843043.2843059.

38. Schulz, C., Sailer, M., Kiesewetter, J., Meyer, C. M., Gurevych, I., Fischer, F. i Fischer, M. R. (2017). Fallsimulationen und automatisches adaptives Feedback mittels Künstlicher Intelligenz in digitalen Lernumgebungen. *e-teaching.org Themenspecial "Was macht Lernen mit digitalen Medien erfolgreich?"*, str. 1–14.
39. Shafqat, W. i Byun, Y.-C. (2020). A recommendation mechanism for under-emphasized tourist spots using topic modeling and sentiment analysis. *Sustainability, 12(*1), 320. https://doi.org/10.3390/su12010320.
40. Sharma, R., Nigam, S. i Jain, R. (2014). Opinion Mining of Movie Reviews at Document Level. *International Journal on Information Theory (IJIT)*, str. 13 - 21.
41. Silva, C. S. R., i Fonseca, J. M. (2018). Artificial intelligence and algorithms in intelligent systems: Advanced analytics: Moving forward artificial intelligence (AI), algorithm intelligent systems (AIS) and general impressions from the field. U R. Silhavy (Ed.), *Artificial Intelligence and Algorithms in Intelligent Systems - Proceedings of 7th Computer Science On-line Conference, 2018* (str. 308–317). Springer Verlag. https://doi.org/10.1007/978-3-319-91189-2_30.
42. Sorokin, D. i Gurevych, I. (2018). Mixing context granularities for improved entity linking on question answering data across entity categories. U *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (*SEM)* (str. 65–75).
43. Stahlhut, C., Stab, C. i Gurevych, I. (2018). Pilot experiments of hypothesis validation through evidence detection for historians. U O. Alonso i G. Silvello (ur.), *Proceedings of the First Biennial Conference on Design of Experimental Search & Information Retrieval Systems (DESIRES)*, Bertinoro, Italy, 28.-31. kolovoza 2018. (vol. 2167, str. 83-89). CEUR-WS.org. https://ceur-ws.org/Vol-2167/paper7.pdf.
44. Škarić, I. (2008.). *Temeljci suvremenoga govorništva*. Školska knjiga.
45. Škiljan, D. (1980). *Pogled u lingvistiku*. Zagreb: Školska knjiga.
46. Šuman, S. (2019). *Sustav za prevođenje poslovnih opisa u model podataka entiteta i veza.* Doktorska disertacija, Odjel Informatike, Rijeka.
47. Šuman, S. (2021). Pregled metoda obrade prirodnih jezika i strojnog prevođenja. *Zbornik Veleučilišta u Rijeci, 9*(1), str. 371-384. https://doi.org/10.31784/zvr.9.1.23.
48. Tanujaya, B., Prahmana, R. i Mumu, J. (2023). Likert scale in social sciences research: Problems and difficulties. *FWU Journal of Social Sciences, 16* (Winter 2022), str 89-101. https://doi.org/10.51709/19951272/Winter2022/7.

49. Tian, L., Lai, C. i Moore, J. (2018). Polarity and Intensity: the Two Aspects of Sentiment Analysis. U A. Zadeh, P. P. Liang, L.-P. Morency, S. Poria, E. Cambria i S. Scherer (ur.), *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)* (str. 40-47). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-3306.
50. *Top Rated Online*. (n.d.). Pristupljeno 16 .srpnja, 2024. https://www.top-rated.online.
51. Toppelberg, C. O. i Collins, B. A. (2010). Language, culture, and adaptation in immigrant children. *Child and Adolescent Psychiatric Clinics*, *19*(4), str. 697–717. https://doi.org/10.1016/j.chc.2010.07.008.
52. Tuđman, M. (1990). *Obavijest i znanje*. Zagreb: Zaklada za Informacijske znanosti filozofskog fakulteta Sveučilišta u Zagrebu.
53. Wadhe, A. A. i Suratkar, S. S. (2020). Tourist place reviews sentiment classification using machine learning techniques. U *2020 International Conference on Industry 4.0 Technology (I4Tech)* (str. 1-6). Pune, India. https://doi.org/10.1109/I4Tech48345.2020.9102673.
54. Wang, J., Xu, B., & Zu, Y. (2021). Deep learning for aspect-based sentiment analysis. In *2021 International Conference on Machine Learning and Intelligent Systems Engineering (MLISE)* (str. 267-271). Chongqing, China. https://doi.org/10.1109/MLISE54096.2021.00056.
55. Xu, H., Liu, B., Shu, L. i Yu, P. S. (2019). BERT post-training for review reading comprehension and aspect-based sentiment analysis. U *Proceedings of NAACL-HLT 2019* (str. 2324–2335). Minneapolis, MN: Association for Computational Linguistics.
56. Xu, T., Peng, Q. i Cheng, Y. (2012). Identifying the semantic orientation of terms using S-HAL for sentiment analysis. *Knowledge-Based Systems, 35*, str. 279-289. https://doi.org/10.1016/j.knosys.2012.04.011.
57. Yu, H. i Hatzivassiloglou, V. (2003). Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. U *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing* (str. 129-136). https://aclanthology.org/W03-1017.pdf.
58. Zhang, S., Wang, Q., Fu, J., Bian, J., & Xiong, H. (2024). Vision ELECTRA: Adversarial masked image modeling with hierarchical discriminator. *OpenReview*. https://openreview.net/forum?id=yKC6Jd0CsP.

Popis slika

Analiza mnijenja u podatkovnom skupu SentiGMapsCro

Sažetak

Ovaj diplomski rad istražuje analizu sentimenta na temelju aspekta unutar podatkovnog skupa SentiGMapsCro koji je nastao sakupljanjem najbolje i najgore recenziranih mjesta i objekata na području Grada Zagreba u Hrvatskoj s internetske stranice https://www.top-rated.online. Jezik, temeljni element ljudske komunikacije, prenosi informacije, emocije i mnijenje. Ovaj rad naglašava važnost analize sentimenta u razumijevanju subjektivnih iskustava podijeljenih na Internetu. Opisuje se postupak ručnog označavanja podatkovnog skupa SentiGMapsCro unutar alata za označavanje pod nazivom INCEpTION, s oznakama sentimenata u rasponu od -2 (vrlo negativan sentiment) do +2 (vrlo pozitivan sentiment) za obuku modela strojnog učenja. Rad naglašava izazove u analizi sentimenta, kao što su složenost prirodnog jezika, prisutnost ironije i metafora, problemi segmentacije te gramatičke pogreške. U radu se ocrtavaju teoretski okviri i istraživačke metodologije korištene u procesu označavanja, ilustrirajući poteškoće s kojima se istraživači susreću u računalnoj obradi prirodnog jezika.

Ključne riječi: analiza sentimenta, podatkovni skup, SentiGMapsCro, INCEpTION

Sentiment Analysis in the SentiGMapsCro Data Set

Summary

The thesis explores sentiment analysis based on aspect within the SentiGMapsCro dataset, focusing on the best and worst reviews of places in Zagreb, Croatia extracted from https://www.top-rated.online. Language, a fundamental element of human communication, conveys information, emotions, and opinions. The study emphasizes the importance of sentiment analysis in understanding subjective experiences shared online. It describes the manual labeling process of the dataset within the annotation tool called INCEpTION, with sentiments ranging from -2 (very negative sentiment) to +2 (very positive sentiment), to train machine learning models. The thesis highlights challenges in sentiment analysis, such as natural language complexities and the presence of irony, metaphors, segmentation issues, and grammatical errors. It outlines theoretical frameworks and research methodologies employed in the labeling process, illustrating the difficulties encountered in computer-based natural language processing.

Key words: sentiment analysis, dataset, SentiGMapsCro, INCEpTION