# Uloga računalne obrade prirodnog jezika i SQL-a u analizi sentimenta

**Brcković, Borna**

SVEUČILIŠTE U ZAGREBU
FILOZOFSKI FAKULTET
ODSJEK ZA INFORMACIJSKE I KOMUNIKACIJSKE ZNANOSTI
Ak. god. 2022./2023.

Borna Brcković

# Uloga računalne obrade prirodnog jezika i SQL-a u analizi sentimenta

Završni rad

Mentor: doc. dr. sc. Ivan Dunđer

Zagreb, rujan 2023.

# 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.

_____

(potpis)

# Sadržaj

# 1. Uvod

U današnjem svijetu naglasak je na informacijama te se sve više informacija prenosi putem blogova, komentara i različitih foruma. Zajednička karakteristika svih različitih tekstova koji se objave na internetu jest da tvorci tekstova, tj. ljudi koji su ih napisali prenose, odnosno izražavaju određene osjećaje, od potpunog nezadovoljstva do potpunog zadovoljstva. Posljedica ovih tekstova je upravo interes za stavovima izrečenim u tim tekstovima. Ovaj fenomen je sve naglašeniji pojavom interneta. Razlog tome je što je Internet otvorio put ka mnoštvu informacija koje smo do njegove pojave širili samo među prijateljima ili kolegama, a ukoliko nam je bila potrebna povratna informacija pitali smo iste ili proveli anketu. Ispitivanje, tj. analiziranje mišljenja je važan i zastupljen proces pri istraživanju stavova kupaca o različitim proizvodima, o izražavanju stavova, kod pisanja recenzija, traženja savjeta, analizi promjena u političkim orijentacijama ili političkim idejama, kod ispitivanja mišljenja o uslugama ili osobama te nekih općih i svakodnevnih komentara na društvenim mrežama. Na primjer; pružateljima proizvoda i usluga uvelike je olakšano pronaći ciljanu publiku, tj. koju vrstu sadržaja pojedinac preferira putem hashtagova. Jednako tako, korisnici se sve rjeđe obraćaju proizvođačima i objavljuju komentare online pa dolazi do određene vrste konflikta između dvije strane u kojoj niti kupac, niti pružatelj ne profitira osobito ako se radi o negativnim komentarima. U takvim slučajevima, kako bismo zaustavili štetu koja može nastati, moramo koristiti analizu sentimenta da bismo dobili bolji uvid u ono što, u ovom primjeru kupac, želi iskazati. Bitno je napomenuti da u analizi sentimenta ulogu može igrati i SQL (engl. *Structured Query Language*). Razlog tomu je upravo to što je analiza sentimenta dio računalne obrade prirodnog jezika te u svojim tehnikama i postupcima koristi različite jezične resurse (poput korpusa ili rječnika) te druge jezične alate koji su zapravo vrste baza podataka. Stoga će se ovaj rad fokusirati upravo na samu analizu sentimenta te ulogu računalne obrade prirodnog jezika te SQL-a u procesu analize sentimenta.

## 2. Računalna obrada prirodnog jezika

Da bismo mogli govoriti o računalnoj obradi prirodnog jezika (engl. *Natural Language Processing, NLP*), moramo prvo definirati dani pojam i sve ono što taj pojam obuhvaća unutar svoje domene. Tadić je zaključio kako sami naziv „računalna obrada prirodnog jezika“ je pojam jednako zanimljiv jezikoslovcima i informatičarima te da je na prvi pogled jasno kako se radi o izrazito interdisciplinarnom području (Tadić, 2003, str. 11). Dakle, sami shematski prikaz je suprotan od onoga u slučaju računalne lingvistike: *računalo* + *lingvistika* (Tadić, 2003, str. 11). Informatičarima je na prvome mjestu obrada podataka sa svim onim zahtjevima koji se postavljaju pred svaku obradu podataka. Ti zahtjevi se mogu sažeti na namjeru za što učinkovitijim djelovanjem i obradom – manji utrošak računalnih resursa. Jezične tehnologije, odnosno računalna obrada prirodnog jezika ima svoje sastavnice. Tako se načelno dijele na:

1. jezične resurse,
2. jezične alate,
3. komercijalne proizvode (Tadić, 2003, str. 27 ).

Za potrebe ovog rada najbitniji će biti jezični alati jer su upravo oni koji će biti glavni u analizi sentimenta iz aspekta računalne obrade prirodnog jezika. Međutim, ne smijemo isključiti jezične resurse jer su i oni sastavni u analizi sentimenta zato što su upravo oni ti koji služe za obradu podataka. Oni su računalno pribavljene, pohranjene i podržane zbirke jezičnih podataka, a sastoje se od korpusa, a potom i od rječnika (Seljan et al., 2015). Jezični resursi također služe za razvitak novih jezičnih resursa (potkorpusa na temelju većeg korpusa ili rječnika) ili za razvitak novih alata (npr. sustavi za segmentaciju na rečenice koji se temelje na evidenciji iz korpusa ili leksikonima, odnosno popisima riječi koji obvezatno poštuju sva lingvistička pravila) (Tadić, 2003, str. 28). Iz ovog je kuta jasno da su korpusi središte i ishodište jezičnih resursa, pa i alata i tehnologija. Korpuse je bitno metodološki razlikovati:

1. zbirka tekstova – svaki skup tekstova skupljen prema nekim kriterijima,
2. korpus – skup jezičnih odsječaka koji su odabrani i skupljeni prema eksplicitnim lingvističkim kriterijima s ciljem da čine jezični uzorak (Jaworski et al., 2023),
3. računalni korpus – korpus koji je kodiran na dosljedan i standardan način s ciljem da bude računalno pretraživ (Tadić, 2003, str. 28).

Također postoje i digitalni rječnici koji su pohranjeni, obrađeni i pretraživi računalom.

Jezični alati također imaju svoju razdiobu na jezične razine koja postoji i u lingvistici. Tako ih razlikujemo na:

1. glasovnoj, odnosno fonemskoj, odnosno grafemskoj razini,
2. razini riječi (morfologija i leksikografija),
3. sintaktičkoj razini,
4. semantičkoj razini,
5. pragmatičkoj razini,
6. diskursnoj razini,

kao i alate koji nerijetko objedinjuju većinu od ovih jezičnih razina, a to su:

7. sustavi za strojno prevođenje (engl. *machine translation, MT*) (Dunđer, 2015) i strojno potpomognuto prevođenje (engl. *machine-aided translation, MAT*) (Seljan et al., 2021),
8. strojno potpomognuto učenje jezika (engl. *Computer-aided language learning, CALL*) (Tadić, 2003, str. 29).

## 2.1. Alati prve razine (glasovna, fonemska i grafemska)

Alati na prvoj razini nisu toliko bitni u analizi sentimenta, ali neki njihovi dijelovi mogu se upotrijebiti u razvijanju pravopisnih provjernika koji itekako mogu pomoći računalnom razumijevanju analize sentimenta.

## 2.2. Alati za obradu na razini riječi

Alati za obradu na razini riječi pojavljuju se u dvama tradicionalnim jezikoslovnom područjima, a to su: leksikografija i morfologija. Ovdje valja istaknuti važnost alata za morfološku obradu koja se može odvijati na dvije razine: tvorbenoj i flektivnoj ovisno o tome koliko se želi modelirati morfološki sustav pojedinog jezika. Također im se može pristupati iz dva kuta: analize i sinteze, ovisno o tome želi li se od postojećih oblika dobiti segmentacija riječi na morfeme ili se žele generirati određene morfemske kombinacije (Tadić, 2003., str 30). Primjerice, engleski jezik ima izrazito jednostavnu morfologiju te se stoga može vrlo lako morfološki obraditi, posebice infleksijski dio morfologije. Međutim, ovo nije slučaj u hrvatskom, stoga treba s oprezom odrađivati morfološku segmentaciju. Svaki od ovih pristupa ima svoju primjenu u različitima sustavima, tj. alatima za obradu resursa:

1. označivači vrsta riječi (engl. *POS* (*part of speech) taggers*) koji svakoj pojavnici u tekstu pridjeljuju podatak o vrsti riječi; problemi nastaju kada se jedna ista riječ može iskoristiti u više konteksta, ali ovakvi slučajevi nisu karakteristični za hrvatski jezik, već su puno češći u engleskome jeziku,

2. morfosintaktički označivači (engl. *MSD taggers*) koji svakoj pojavnici uz vrstu riječi pridjeljuju i podatak o vrijednostima ostvarenih morfosintaktičkih kategorija (npr. rod, broj i padež za svaki imenični oblik u hrvatskim tekstovima),
3. lematizatori (engl. *Lemmatisers*) koji svakoj pojavnici u tekstu pridjeljuju njezinu lemu, tj. njezin polazni, kanonski, natuknički oblik,
4. stemeri (engl. *Stemmers*) koji procesom „stemanja, tj. korjenovanja riječi odbacuju infleksijski ili derivacijski nastavak riječ i daju korijen riječi (Tadić, 2003, str. 32).

Ovi su alati izrazito bitni za dobivanje označenih korpusa, tj. korpusa koji su oplemenjeni dopunskim morfološkim podacima koji nam omogućuju oblike pretraživanja kakvi nad neoznačenim tekstom nisu mogući (Jaworski et al., 2017). Upravo zbog razvijanja korpusa je moguće da programski jezici poput SQL-a pronalaze pojavnice u korpusima koje su potrebne u analizi sentimenta (Tadić, 2003, str. 32).

## 2.3. Alati na sintaktičkoj razini

Naime, do sada je jasno da se alati jedne razine ne mogu razviti bez alata s njima prethodne razine. Stoga je smisleno da su alati na sintaktičkoj razini ovisili o onima na razini riječi jer se oslanjaju na dobivene rezultate prethodne razine. Upravo se zato većina današnjih alata za sintaktičku analizu nastavlja na alate za morfološku analizu, pa i opsežne leksikone. Obradi podataka na razini sintakse također je moguće pristupiti iz dva smjera, istima kao i na razini riječi, a to su: analiza i sinteza. Sustavi za analizu rečenica mogu biti parseri (engl. *parsers*) ili razdjelnici (engl. *chunkers*) (Tadić, 2003, str. 32, 33). Dok razdjelnici razdjeljuju rečenice na najčešće nerekurzivne sastavnice kojima je jednostavno odrediti granice (npr. predikatni skupovi, imenične fraze itd.), parseri ulaze dublje u analizu strukture rečenice, pa ih tako dijelimo na:

1. plitki (engl. *shallow*) parseri određuju odnose ovisnosti između dijelova u rečenici,
2. duboki (engl. *deep*) parseri obavljaju punu sintaktičku analizu do razine leksičkih unosaka,
3. robusni (engl. *robust*) parseri koji ne zastaju kada naiđu na neku nepostojeću kombinaciju nekih rečeničnih dijelova (npr. pri elipsi, pogrešnoj rekciji itd.) te su zbog toga izrazito pogodni za automatsku analizu stvarnog teksta u kojem su takve pojave izrazito češće nego u onim tekstovima koji su idealno formulirani za potrebe parsera i sličnih alata (Tadić, 2003, str. 33).

Nakon što parseri odrade svoje zadatke, podatci o rečeničnoj sintaktičkoj analizi vraćaju se u korpus i dobiva se sintaktički označen korpus koji nosi ime banka stabala (engl. *tree-bank*). Razlog tom imenu dolazi time što sintaktička analiza daje tzv. stabla parsiranja. Ova stabla zadovoljavaju sve okvire gramatičkih zahtjeva. Međutim, upravo zbog polisemije jezika i ponekad višeznačnih gramatičkih cjelina moguće je da nastanu stabla koja neće imati smisla čitatelju, ali čitatelji će ih često moći odbaciti jer mnoga od tih stabala neće imati smisla usprkos njihovoj točnosti u sintaktičkoj analizi. Upravo tako označeni korpusi dopuštaju korisnicima, pa i drugim automatskim programima ili sučeljima što raznovrsnije upite i pretrage nego što je slučaj u onim korpusima koji su označeni samo na morfološkoj razini. Naime, potpuna analiza nije uvijek potrebna jer nisu svi primjeri izrazito kompleksni pa zahtijevaju detaljno obrađene strukture, već su za neke dovoljni alati koji će u danom tekstu prepoznati pojedine karakteristične dijelove rečenica. Tako ćemo se npr. u zadacima crpljenja informacija, crpljenja termina i sličnim pothvatima koristiti sustavima koji su izgrađeni na temelju „lokalnih" gramatika, tj. gramatika koje opisuju zaokružene sintaktičke cjeline. Ove sintaktičke cjeline nerijetko čine također i zaokružene semantičke cjeline (Tadić, 2003, str. 33). Semantička cjelina je, gledajući po gore navedenoj listi, razinu iznad sintaktičke te stoga kako bismo ju dobili moramo imati adekvatno pripremljene alate i obrađene zadatke s razine sintaktičke analize i sinteze. Prema tome, sustavi za semantičke cjeline su sljedeći:

1. prepoznavanje i razvrstavanje naziva (engl. *named entity recogntion and classification, NERC*) – ovaj sustav spada u izrazito bitne sustave za semantičku analizu jer on dozvoljava da se nazivi (engl. *named entities*) unutar teksta lako povežu s onim iz vanjskog svijeta jer daju dodatne obavijesti, odnosno informacije, o tom entitetu, tj. nazivu. Najuobičajenija pitanja na koja možemo naići u tekstovima su pitanja poput *tko? kada? što? gdje? koliko? kamo? kuda?* i sl. te upravo ta pitanja daju važne informacije kojima se pojedini događaj ili osoba smješta u mjesto i vrijeme te povezuje s odgovarajućim činiteljima. Odgovori na ova pitanja najčešće su neki nazivi te stoga *NERC* igra nezamjenjivu ulogu u ovakvim strukturama i situacijama gdje je potrebna upravo pomoć tog sustava kako bi se razriješila referenca (Tadić, 2003, str. 63).
2. prepoznavanje vremenskih izraza (engl. *temporal expressions*) kao točke u vremenu ili vremenskog raspona
3. prepoznavanje prostornih izraza (engl. *spatial expressions*) kao točke u prostoru ili prostorne udaljenosti između dvije ili više točaka, uzduž pravca i sl.

4. prepoznavanje mjera (engl. *measure expressions*) kao izraz koji uključuje neku mjernu jedinicu, razmjer i sl. (Tadić, 2003, str. 34)

Postoje mnoge druge semantičke cjeline koje bi se mogle navesti kao primjeri. Međutim, ovo je dovoljno za potrebe razumijevanja ovoga rada. Sustavi za semantičku sintezu primjenjuju se u svim slučajevima kada je, na temelju neke semantičke strukture, potrebno generirati rečenicu prirodnog jezika. Ovo stvara tzv. kontrolirani jezik (engl. *controlled language*) u kojem se pod strogim nadzorom drže sve moguće sintaktičke varijacije i ograničava im se broj (Tadić, 2003, str. 34).

U ovu skupinu također spadaju i alati za razdvajanje rečenica u danom tekstu ukoliko se dogodi da su spojene. Najčešće su rečenice odvojene točkama, zarezima, uskličnicima, upitnicima itd. Međutim, točke se mogu koristiti i u druge svrhe; poput pisanja datuma, rednih brojeva i kratica, a pritom rečenica ne završava. Stoga je bitno razlikovati ova dva slučaja u analizi. Postoji i dosta sličan alat koji izvršava segmentaciju rečenice na riječi u određenom tekstu. Ovo se naizgled čini dosta banalno, ali mora biti jasno da nisu u svim jezicima, poput hrvatskog ili *lingue france*, tj. engleskog, riječi odvojene razmacima. Npr. u različitim azijskim jezicima nailazimo na granice koje ne naliče hrvatskom već koriste drukčije znakove (Tadić, 2003, str. 34).

## 2.4. Alati na semantičkoj razini

Alati za obradu jezika na semantičkoj razini mogu se, kao i prethodne razine, podijeliti u dva sloja: alati za semantiku rečenice te alati za semantiku riječi. Računalna semantika ima veze s računalnom leksikografijom s jedne strane, a s druge strane sa semantičkim mrežama u kojima se značenje riječi opisuje njihovim dovođenjem u međusobne semantičke odnose. Jedna od najpoznatijih semantičkih mreža je WordNet. Ona je projekt započet 1985. godine kako bi se razvila opća semantička mreža za engleski jezik pod vodstvom Georgea A. Millera nakon što je okupio psihologe i lingviste. WordNet je velika leksička baza engleskog jezika. Imenice, glagoli, pridjevi i prilozi grupiraju se u skupove kognitivnih sinonima (sinsetova, sinskupova) svaki izražavajući različit koncept. Sinsetovi su povezani putem pojmovno-semantičkih i leksičkih odnosa. Nastala mreža smisleno povezanih riječi i koncepata može se istraživati pomoću preglednika. WordNet je također besplatno i javno dostupan za preuzimanje. Struktura WordNeta čini ga korisnim alatom za računalnu lingvistiku i obradu prirodnog jezika. WordNet površno podsjeća na sinonimski rječnik, jer grupira riječi zajedno na temelju njihovih značenja. Međutim, postoje neke važne razlike. Prvo, WordNet ne povezuje samo oblike riječi i nizove slova, već i specifična značenja riječi. Kao rezultat toga, riječi koje su blizu jedna drugoj u

mreži, semantički su razdvojene. Drugo, WordNet označava semantičke odnose među riječima, dok grupiranje riječi u sinonimskom rječniku ne slijedi nikakav eksplicitan obrazac osim sličnosti značenja. Sustavi za analizu rečenične semantike zaduženi su za prepoznavanje semantičkih uloga ili npr. dubinskih padeža (u jeziku su to npr. agens, pacijens, instrument itd.) na temelju usko povezanih rezultata semantičke analize (Tadić, 2003, str. 68).

## 2.5. Alati za obradu jezika na pragmatičkoj razini

Dolazimo do skupine alata za obradu jezika, a to su alati za obradu jezika na pragmatičkoj razini. Oni se uključuju u istraživanja inteligentnih sustava, tj. inteligentnih sučelja za komunikaciju čovjeka i računala. Upravo iz tog razloga nećemo ulaziti u detalje tih alata jer za potrebe ovog rada potrebni su oni alati koji se mogu upotrijebiti na relaciji računalo-računalo.

## 2.6. Alati za obradu jezika na diskursnoj razini

Iduća skupina alata su alati na diskursnoj i posljednjoj razini ove strukture. Diskurs označava različite tipove teksta u lingvistici (diskurs. Hrvatska enciklopedija http://www.enciklopedija.hr/Natuknica.aspx?ID=15415). U ovo spadaju alati koji izvršavaju zadatke na razini danog teksta pa tako imamo sljedeće:

1. automatsko generiranje sažetka (engl. *automatic summarization*) – automatsko stvaranje sažetka iz danog teksta putem dva različita postupka (ekstrakcije ili apstrakcije),
2. razrješavanje unakrsnih referenci (engl. *coreference resolution*) – zadatak koji se odnosi na razrješavanje problema razumijevanja koji se imenski skupovi ili riječi odnose na isti entitet unutar teksta. Npr. ukoliko napišemo: „Kada je Marina došla u školu, shvatila je da je zaboravila kišobran. Sjetila se da ga je ostavila doma“. U tom slučaju ovdje će zadatak biti da program razumije kako se zamjenica „ga“ u drugoj rečenici odnosi na kišobran iz prve rečenice, tj. da obje riječi označavaju jedan te isti entitet u ekstralingvističkom svijetu (Ng i Cardie, 2002),
3. analiza diskursa (engl. *discourse analysis*) – posljednji i najveći zadatak koji inkorporira sve prethodne razine, pa čak i druge zadatke s diskursne razine kako bi se analizirale lingvističke strukture više tekstova na različitim razinama. Ovo uključuje identificiranje teme u tekstu, povezanost strukture, prethodno navedeno razrješavanje unakrsnih referenci i analize teksta za razgovorni diskurs. Samim time i svojom kompleksnosti, analiza diskursa je interdisciplinarno područje jer se bavi lingvističkim i nelingvističkim elementima. Cilj joj je dublje razumijevanje kako se jezik koristi za

stvaranje novih značenja i kako se kulturni ili društveni kontekst odražava u jezičnom izražavanju. Objedinjujući sve ove zadatke zajedno mogu se dati informacije jezičnom prevođenju, automatskom generiranju sažetaka, automatskom odgovaranju i, ovom radu najbitnijoj stvari, analizi sentimenta. Samim time nalazi primjene u mnogim poljima, koji ne uključuju samo lingvistiku ili računalnu obradu prirodnog jezika, poput sociologije, političkih znanosti, komunikacije itd. (Joty et al., 2019).

## 2.7. Alati različitih namjena

Postoji još vrsta alata, ali iduća dva više ne slijede striktno raslojavanje na jezične razine, već ovisno o njihovoj namjeni, na različite načine u zaokružene sustave koji su rezultat kombiniranja i koordiniranja postojećih alata za obradu jezika na različitim jezičnim razinama. Ta dva alata su:

1. strojno prevođenje (engl. *machine translation, MT*) (Dunđer, 2021a) ,
2. strojno potpomognuto učenje jezika (engl. *computer-aided language learning, CALL*).

Strojno je prevođenje bilo je jedan od najranije zamišljenih primjena računala uopće, a teorijski nacrti potiču još iz 1949. iz znamenitog Weaverovog memoranduma (Dunđer, 2021b). Ukoliko definiramo prevođenje kao postupak prijenosa značenja iz jednog jezika u drugi ili kao postupak transkodiranja jednog teksta u drugi, onda možemo razlikovati strojno prevođenje (*MT*) i strojno potpomognuto prevođenje (engl. *machine-aided translation, MAT*) (Dunđer, 2020). Strojno je prevođenje ono prevođenje koje obavlja isključivo stroj, odnosno računalo, bez intervencije čovjeka, a strojno potpomognuto prevođenje koje obavlja čovjek uz pomoć računala (Tadić, 2003).

## 2.8. Komercijalni proizvodi

Naposljetku dolazimo do alata koji se mogu smatrati i komercijalnim proizvodima jer su konačan produkt svake jezične tehnologije koji se mogu prodavati ili su čak inkorporirani u kupljenim programima. Tako imamo sljedeće primjere:

1. provjernici (engl. *checkers*) pravopisa (engl. *spelling-checkers*),
2. provjernici gramatike (engl. *grammar-checkers*),
3. provjernici stila (engl. *style-checkers*).

Za ove provjernike bilo je potrebno razviti određene jezične alate. Tako je npr. za pravopisne provjernike bilo dovoljno razviti jezične alate za obradu na razini višeslova ili riječi, dok je za gramatičke provjernike i provjernike stila bilo potrebno razviti jezične alate za obradu na razini rečenice, pa čak i nadrečenične sintakse i semantike.

Nadalje, prethodno spomenuti rječnici (opći ili specijalni, objasnidbeni, tezaurusi, leksičke baze itd.), koji su ujedno i vrste jezičnih resursa također spadaju i u komercijalne proizvode jer su se pojavili na internetu. Oni spadaju pod online rječnike. Postoje i online korpusi koji se prodaju kao gotovi proizvodi određenim organizacijama i sl. Međutim, najbitnije je napomenuti da ovi komercijalni alati/resursi imaju jako važno svrhu zato što oni služe kao velike baze podataka iz kojih se crpe sve informacije, nazivlje i slično kako bi se mogle obavljati duboke i detaljne analize, npr. analiza sentimenta.

Još neki od sustava koje vrijedi spomenuti su sustavi za crpljene informacija (engl. *information extraction, IE*) te crpljenje nazivlja (engl. *terminology extraction, TE*) (Seljan et al., 2017; Seljan et al., 2013), strojevi za diktiranje (engl. *speech-to-text* sustavi), sustavi za automatsko spikiranje (engl. *text-to-speech* sustavi) (Dunđer, 2013) i sustavi za automatsko odgovaranje (Tadić, 2003, str. 41-44).

Prije analize procesa koji su ukomponirani u svim ovim zadacima kod analize sentimenta, potrebno je objasniti još i sami pojam analize sentimenta te pojam SQL-a.

# 3. Analiza sentimenta

Analiza sentimenta ili istraživanje mišljenja (engl. *sentiment analysis, SE* ili *opinion mining, OP*) je informatičko proučavanje i tehnika obrade prirodnog jezika koja se koristi u svrhu analize ljudskih osjećaja, stavova i misli o nekom entitetu o kojemu je riječ u danom tekstu (Dunđer et al., 2017). Analiza sentimenta određuje jesu li podaci, odnosno ton, mišljenje, stavovi i osjećaji korisnika u komentarima ili sličnim tekstovima pozitivni, negativni ili neutralni. Ova je tehnika izuzetno važna u marketingu pri donošenju pojedinih poslovnih odluka koje su bazirane na mišljenjima i komentarima korisnika i/ili kupaca (Kovačević i Kovačević, 2021). Međutim, bitno je napomenuti da neki stručnjaci govore kako postoji razlika između analize sentimenta i istraživanja mišljenja. Naime, tvrdnja govori kako istraživanje mišljenja pronalazi, izvlači i analizira mišljenja ljudi o određenom entitetu, dok analiza sentimenta identificira sentiment u tekstu te ga zatim analizira (Tsytsarau i Palpanas, 2012).

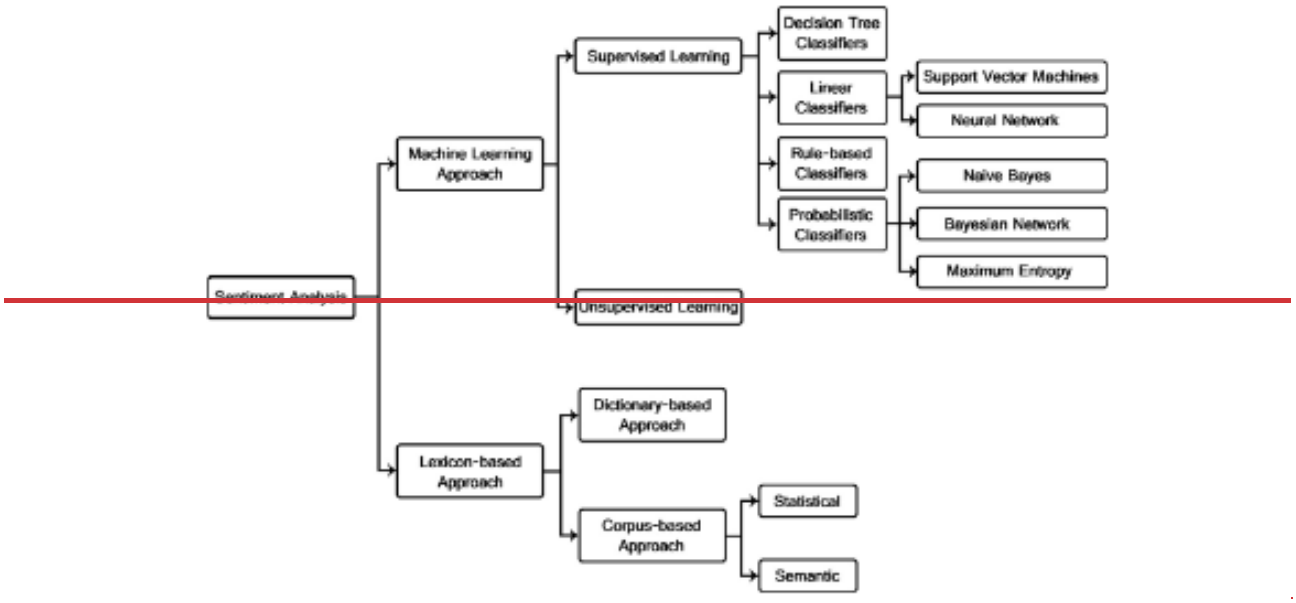
Analiza sentimenta može se smatrati i procesom klasifikacije, kao što je ilustrirano na slici 1. Postoje tri glavne razine klasifikacije u analizi sentimenta: analiza na razini dokumenta, analiza na razini rečenice i analiza na razini aspekta. Analiza na razini dokumenta ima za cilj klasificirati mišljenje izraženo u dokumentu kao pozitivno ili negativno mišljenje ili sentiment. Razmatra cijeli dokument kao osnovnu informacijsku jedinicu (govoreći o jednoj temi). Analiza na razini rečenice ima za cilj klasificirati sentiment izražen u svakoj pojedinoj rečenici. Prvi korak je utvrditi je li rečenica subjektivna ili objektivna. Ako je rečenica subjektivna, analiza na razini rečenice će odrediti je li rečenica izražava pozitivno ili negativno mišljenje. Wilson et al. (2005) su istaknuli da izrazi sentimenta nisu nužno subjektivne prirode. Međutim, nema temeljne razlike između klasifikacija na razini dokumenta i klasifikacija na razini rečenice jer su rečenice samo kratki dokumenti (Liu, 2012). Klasificiranje teksta na razini dokumenta ili na razini rečenice ne pruža potrebne pojedinosti potrebne za mišljenja o svim aspektima entiteta, što je potrebno u mnogim aplikacijama, za dobivanje tih pojedinosti, trebamo se uputiti na razinu aspekta. Analiza na razini aspekta ima za cilj klasificirati sentiment u vezi s određenim aspektima entiteta. Prvi korak je prepoznati entitete i njihove aspekte. Osobe koje izražavaju mišljenje mogu dati različite ocjene za različite aspekte istog entiteta, kao što je slučaj u ovoj rečenici: „Kvaliteta glasa ovog telefona nije dobra, ali vijek trajanja baterije je dug“.

Alati za obradu prirodnog jezika mogu se koristiti kako bi olakšali proces analize sentimenta. Pružaju bolje razumijevanje prirodnog jezika i time mogu pomoći u dobivanju preciznijih rezultata analize sentimenta. Ti su alati korišteni kako bi pomogli u zadacima obrade teksta,

ekstrakciji informacija te također u analizi sentimenta (Dunđer i Pavlovski, 2019). To otvara novi trend istraživanja koji se odnosi na korištenje obrade prirodnog jezika kao pretprocesiranja prije analize sentimenta.

U mnogim aplikacijama važno je uzeti u obzir kontekst teksta i korisničke preferencije. Zato je potrebno provesti više istraživanja o analizi sentimenta temeljenoj na kontekstu. Korištenjem tehnika prijenosnog učenja, možemo koristiti povezane podatke iz domene u pitanju kao podatke za treniranje. Korištenje alata za obradu prirodnog jezika (NLP) kako bi se ojačao proces analize sentimenta privuklo je relativno nedavno pažnju istraživača i još uvijek zahtijeva neka poboljšanja.

Tehnike klasifikacije analize sentimenta prikazane su na slici 1.




Slika 1. Tehnike klasifikacije analize sentimenta (Medhat et al., 2014)

S obzirom na to da postoje mnoge tehnike klasifikacije analize sentimenta, ovaj rad bazirat će se isključivo na onima koje u svojim operacijama koriste NLP. U ovom slučaju to će biti pristup koji se bazira na leksikonima, sukladno tome, dijeli se na pristup baziran na rječnicima te pristup baziran na korpusima. Pristup baziran na korpusima se dijeli na statistički i semantički. Metode temeljene na leksikonima zahtijevaju ljudsku anotaciju, dok su statističke metode automatske metode koje se češće koriste. Metode temeljene na leksikonima obično započinju s malim skupom „početnih" riječi. Zatim proširuju taj skup kroz detekciju sinonima ili korištenjem online resursa kako bi dobili veći leksikon. To se pokazalo kao težak zadatak, kako su izvijestili Whitelaw et al. (2005). Statistički pristupi, s druge strane, su potpuno automatski.

Tehnike odabira značajki tretiraju dokumente ili kao skup riječi (tzv. vreća riječi (engl. *Bag of Words, BOWs*)), ili kao niz koji zadržava redoslijed riječi u dokumentu (Medhat et al., 2014). Vreća riječi se češće koristi zbog jednostavnosti u procesu klasifikacije. Najčešći korak odabira značajki je uklanjanje zaustavnih riječi i vraćanje riječi na korijen ili osnovni oblik riječi.

Zadatak analize sentimenta smatra se problemom klasifikacije sentimenta. Prvi korak u problemu klasifikacije sentimenta je izdvajanje i odabir značajki teksta. Neki od trenutnih značajki su: prisutnost i frekvencija pojmova, dijelovi govora (POS), riječi i fraze mišljenja te negacije (Aggarwal i Zhai, 2012).

Prisutnost i frekvencija pojmova su značajke pojedinačne riječi ili n-grami riječi i njihovi brojači frekvencija. Ovo ili daje binarnu težinu riječima (nula ako se riječ pojavljuje, ili jedan ako nije), ili koristi težine učestalosti pojma kako bi označio relativnu važnost značajki (Mejova i Srinivasan, 2011). Proces identifikacije dijelova govora (POS) uključuje pronalaženje pridjeva, budući da su oni važni pokazatelji mišljenja. Riječi i fraze mišljenja su riječi i fraze koje se često koriste za izražavanje mišljenja, uključujući pozitivna i negativna mišljenja i stavove. S druge strane, neke fraze izražavaju mišljenje bez korištenja riječi mišljenja. Na primjer: koštalo me ruku i nogu. Naposljetku, nalazimo negacije što podrazumijeva da pojava negativnih riječi može promijeniti orijentaciju mišljenja, npr. „nije dobro“ znači isto što i „loše“ (Medhat et al., 2014).

Riječi mišljenja koriste se u mnogim zadacima klasifikacije sentimenta. Pozitivne riječi mišljenja koriste se kako bi izrazile neka željena stanja, dok se negativne riječi mišljenja koriste kako bi izrazile neželjena stanja. Postoje također i izrazi mišljenja i idiomi koji zajedno čine leksikon (liste) mišljenja. Postoje tri glavna pristupa za sastavljanje ili prikupljanje liste riječi mišljenja. Ručni pristup je vremenski vrlo zahtjevan i obično se ne koristi samostalno. Obično se kombinira s druga dva automatizirana pristupa kao konačna provjera kako bi se izbjegle pogreške koje su rezultat automatiziranih metoda (Medhat et al., 2014).

Hu i Liu (2004) te Kim i Hovy (2004) su predstavili glavnu strategiju pristupa temeljenog na rječniku. Ručno se prikuplja mali skup riječi mišljenja s poznatim orijentacijama. Zatim se taj skup proširuje pretraživanjem poznatih korpusa kao što su WordNet (Miller et al. 1990) ili tezaurus (Mohammad et al., 2009) za sinonime i antonime tih riječi. Nove pronađene riječi se dodaju na početni popis, nakon čega započinje sljedeća iteracija. Iterativni proces završava kada se više ne pronađu nove riječi. Nakon završetka procesa, može se provesti ručna provjera radi uklanjanja ili ispravljanja pogrešaka.

Pristup temeljen na rječniku ima glavnu manu koja se sastoji u nemogućnosti pronalaženja riječi mišljenja s orijentacijama specifičnim za domenu i kontekst. Qiu et al. (2010) koristili su pristup temeljen na rječniku kako bi identificirali rečenice s mišljenjem u kontekstualnom oglašavanju (2010). Predložili su strategiju oglašavanja kako bi poboljšali relevantnost oglasa i korisničko iskustvo. Koristili su sintaktičko parsiranje i rječnik sentimenta te predložili pristup temeljen na pravilima kako bi se nosili s izdvajanjem riječi vezanih uz temu i identifikacijom stajališta potrošača u izdvajanju ključnih riječi oglasa na jednom internetskom forumu. Njihovi rezultati su pokazali učinkovitost predloženog pristupa u izdvajanju ključnih riječi oglasa i odabiru oglasa.

Pristup temeljen na korpusu pomaže rješavanju problema pronalaženja riječi mišljenja s orijentacijama specifičnim za kontekst. Njegove metode ovise o sintaktičkim obrascima ili obrascima koji se pojavljuju zajedno, zajedno s početnim popisom riječi mišljenja, kako bi pronašli druge riječi mišljenja u velikom korpusu. Jedna od ovih metoda bila je predstavljena od strane Hatzivassilogloua i McKeowna (1997). Oni su započeli s popisom početnih pridjeva mišljenja i koristili ih zajedno s nizom jezičnih ograničenja kako bi identificirali dodatne pridjeve mišljenja i njihove orijentacije. Ograničenja se odnose na veznike poput I, ILI, ALI, ILI-ILI... Primjerice, veznik I implicira da spojeni pridjevi obično imaju istu orijentaciju. Ova ideja naziva se dosljednost sentimenta, koja nije uvijek dosljedna u praksi. Također postoje izrazi protivljenja kao što su ALI, međutim, koji označavaju promjene mišljenja. Kako bi se odredilo jesu li dva spojena pridjeva iste ili različite orijentacije, primjenjuje se učenje na velikom korpusu. Zatim se veze između pridjeva oblikuju u graf i izvodi se klasteriranje na grafu kako bi se dobila dva skupa riječi: pozitivne i negativne. Metoda uvjetnih slučajnih polja (engl. *conditional random fields, CRF*) korištena je kao tehnika učenja slijeda za izdvajanje izraza mišljenja (Lafferty et al., 2001).

Također je korištena i od strane Jiaoa i Zhoua (2011) kako bi se razlikovali polariteti sentimenta pomoću višeslojnog algoritma usklađivanja uzoraka. Njihov algoritam primijenjen je na kineske internetske recenzije. Ustanovili su mnogo rječnika emocija. Radili su na recenzijama automobila, hotela i računala putem interneta. Njihovi su rezultati pokazali da je njihova metoda postigla visoku učinkovitost.

Xu et al. (2011) koristili su dvoslojni CRF model s nesigurnim međuovisnostima kako bi izvukli usporedne odnose. To je postignuto upotrebom složenih međuovisnosti između odnosa, entiteta i riječi te nesigurnih međuovisnosti među odnosima. Njihov je cilj bio stvoriti grafički

model za izdvajanje i vizualizaciju usporednih odnosa između proizvoda iz recenzija kupaca. Rezultate su prikazali kao mape usporednih odnosa za podršku donošenju odluka u upravljanju rizikom tvrtke. Radili su na recenzijama mobilnih uređaja s internetskih portala blogova, društvenih mreža i e-pošte. Njihovi su rezultati pokazali da njihova metoda može točnije izdvojiti usporedne odnose od drugih metoda i da je njihova mapa usporednih odnosa potencijalno vrlo učinkovit alat za podršku upravljanju rizikom tvrtke i donošenju odluka.

Pristup temeljen na korpusu sam po sebi nije toliko učinkovit kao pristup temeljen na rječniku jer je teško pripremiti ogroman korpus koji bi pokrio sve engleske riječi, ali ovaj pristup ima veliku prednost koja može pomoći u pronalaženju riječi mišljenja specifičnih za domenu i kontekst te njihovih orijentacija koristeći korpus domene.

Računalna obrada prirodnog jezika se koristi uz pristup temeljen na leksikonima kako bi se pronašla sintaktička struktura i pomoglo u pronalaženju semantičkih odnosa (Bolshakov i Gelbukh, 2004). Moreo et al. (2012) koristili su tehnike obrade prirodnog jezika kao pretprocesiranje prije nego što su koristili svoj predloženi algoritam temeljen na leksikonima za analizu sentimenta. Njihov predloženi sustav sastoji se od modula za automatsko otkrivanje fokusa i modula za analizu sentimenta sposobnog za procjenu korisničkih mišljenja o temama u vijestima, koristeći leksikon taksonomije (leksikon koji ima klasificirane riječi ovisno o njihovim međusobnim svojstvima) specifično dizajniranu za analizu vijesti. Njihovi rezultati bili su obećavajući u scenarijima gdje prevladava kolokvijalni jezik.

Pristup za analizu sentimenta koji su predstavili Caro i Grella (2012) temeljio se na dubokoj analizi rečenica pomoću NLP analize, koristeći analizu ovisnosti kao korak pretprocesiranja. Njihov algoritam za analizu sentimenta oslanjao se na koncept propagacije sentimenta, koji pretpostavlja da svaki jezični element poput imenice, glagola itd. može imati intrinzičnu vrijednost sentimenta koja se širi kroz sintaktičku strukturu analizirane rečenice. Predstavili su skup sintaktičkih pravila koja su imala za cilj obuhvatiti značajan dio sentimentne važnosti izražene tekstom. Predložili su sustav za vizualizaciju podataka u kojem je trebalo filtrirati neke podatkovne objekte ili kontekstualizirati podatke kako bi se korisniku prikazale samo informacije relevantne za korisnički upit. Kako bi to postigli, predstavili su kontekstualnu metodu za vizualizaciju mišljenja mjerenjem udaljenosti, u tekstualnim procjenama, između upita i polariteta riječi sadržanih u samim tekstovima. Proširili su svoj algoritam tako da izračunava bodove polariteta temeljene na kontekstu. Njihov pristup je pokazao visoku učinkovitost nakon primjene na ručnoj korpusu od 100 recenzija restorana.

Min i Park (2012) koristili su NLP iz drugačije perspektive. Koristili su NLP tehnike kako bi identificirali vremenske i vremenske izraze zajedno s tehnikama rudarenja i algoritmom rangiranja. Njihova predložena metrika ima dva parametra koji „hvataju“ vremenske izraze vezane uz upotrebu proizvoda i entiteta proizvoda tijekom različitih razdoblja kupnje. Identificirali su važne jezične znakove za parametre kroz eksperiment s prikupljenim podacima recenzija, uz pomoć NLP tehnika. Radili su na recenzijama proizvoda s Amazona. Njihovi rezultati su pokazali da je njihova metrika bila korisna i slobodna od nepoželjnih pristranosti.

# 4. SQL

SQL je jezik za organiziranje, upravljanje i dohvaćanje podataka pohranjenih u računalnoj bazi podataka. Naziv „SQL“ je akronim za *Structured Query Language*. Iz povijesnih razloga, SQL se obično izgovara kao „sikvel“, ali također se koristi i alternativno izgovaranje „S-Q-L“. Kao što ime implicira, SQL je programski jezik koji koristite za interakciju s bazom podataka. Zapravo, SQL radi s jednim određenim tipom baze podataka nazvanom relacijska baza podataka. Računalni sustavi imaju bazu podataka u kojoj se pohranjuju važne informacije. Ako je računalni sustav u poslovnom okruženju, baza podataka može pohranjivati inventar, proizvodne podatke ili podatke o prodaji ili plaćanju. Na osobnom računalu baza podataka može sadržavati podatke o čekovima, popise ljudi i njihove telefonske brojeve ili podatke izvučene iz većeg računalnog sustava. Računalni program koji upravlja bazom podataka naziva se sustav za upravljanje bazom podataka, ili DBMS (engl. *Database Management System*). Kada treba dohvatiti podatke iz baze podataka, koristi se SQL jezik za postavljanje zahtjeva. DBMS obrađuje SQL zahtjev, dohvaća tražene podatke i vraća ih. Ovaj proces postavljanja zahtjeva za podacima iz baze podataka i primanje rezultata naziva se upit baze podataka stoga i naziv strukturirani upitni jezik. Naziv strukturirani upitni jezik zapravo je donekle pogrešno ime. SQL je puno više od alata za upite, iako je to bila njegova prvotna svrha, a dohvaćanje podataka i dalje je jedna od njegovih najvažnijih funkcija (Groff et al., 2002, str. 8-11).

Postavlja se pitanje upotrebe i korisnosti SQL-a u analizi sentimenta. Al-Khafaji i Habeeb (2017) proveli su istraživanje u kojemu su htjeli istražiti korisnost MS-SQL servera i SQL-a kao jezika za analizu sentimenta na velikom broju tweetova. Koristili su pojam pretprocesiranje te ga objasnili kao upotrebu tehnika za pripremu podataka za analitički proces koje uključuje nekoliko koraka, pri čemu svaki korak proizvodi podatke spremne za sljedeći korak sve dok se završi transformacijski proces i podaci budu u najboljoj formi za analizu. Sukladno tome, istražili su tri najbitnija koraka u prikupljanju podataka na temelju pretprocesiranja, a to su: tokenizacija, stemanje te „uklanjanje zaustavnih riječi“ (Al-Khafaji i Habeeb, 2017). Ovi koraci štede vrijeme analize i prostor za pohranu, posebno kod velikih skupova podataka, uz smanjenje podataka; povećava se točnost sustava analize rezultata. Prva dva termina objašnjena su u poglavlju 2.1. Uklanjanje zaustavnih riječi (engl. *stop words removal*) uklanja riječi koje su beskorisne u procesu pretraživanja informacija, a poznate su kao zaustavne riječi. Ove riječi nemaju vrijednost (pozitivnu ili negativnu) u sustavu za analizu sentimenta. Stoga se trebaju ukloniti iz skupa podataka. Primjeri zaustavnih riječi uključuju „the“, „as“, „of“, „and“, „or“,

„to“ itd. Uklanjanje zaustavnih riječi je bitno u pretprocesiranju, ima neke prednosti poput smanjenja veličine spremljenog skupa podataka te poboljšanja ukupne učinkovitosti i djelotvornosti analitičkog sustava (Al-Khafaji i Habeeb, 2017).

Kako bi se tweet pretvorio u tokene (zasebne riječi), tweet prolazi kroz dvije faze: segmentaciju riječi i čišćenje. Kao što je prethodno spomenuto, segmentacija riječi je proces razdvajanja rečenice (odnosno rečenica) pisanog jezika na njegove riječi koje čine strukturu te rečenice (odnosno rečenica). Predloženi sustav analizira tweetove napisane na engleskom jeziku (Slika 2). Tweetovi su kratke rečenice, pa predloženi algoritam razdvaja rečenicu na riječi i simbole (koji su razdvojeni razmacima) i sprema svaku riječ ili simbol u zaseban redak u tablici. Ovaj algoritam dohvaća tweet iz tablice tweetova. U koraku 101, funkcija sljedećeg razmaka određuje sljedeći razmak u tweetu. U koraku 104, funkcija odrezivanja podniza reže podniz koji zauzima položaj od P do I kao token, koji će se pohraniti u bazu podataka. Ovaj proces će se nastaviti sve do kraja tweeta T. Na primjer, ako je tweet bio „It”, „will”, „be”, „a”, „very”, „hot”, „summer,”, „I”, „think”, „@The-knight” tokeni su „bit“, „će“, „vrlo“, „vruće“, „ljeto“, „mislim“, „@The-knight“. Primjećujemo da se riječ „summer“ i zarez „„ smatraju jednim tokenom jer nisu odvojeni razmakom. Slično tome, izraz „@The-knight“ (Al-Khafaji i Habeeb, 2017).

```
Word segmentation algorithm
Input   Tweets_table
                                Output Tokens_table
100   while (more tweets exist in tweets table) do
101   {   T= next tweet; P=1;
102       while not end of T
103       {     I= next_space (T); // determines the position of next space in T
104             token= trim_substring (T, P, I, rest_string);
105             insert into tokens_table (tweet number, token);
106             T= rest_string ;}
107   }
```

Slika 2. Algoritam koraka tokenizacije (Al-Khafaji i Habeeb, 2017)

Slijedi tzv. *cleaning-phase*, tj. faza čišćenja s obzirom na to da prethodna faza proizvodi zbirku riječi, simbola, interpunkcijskih znakova i brojeva (Slika 3). Sustav analize zahtijeva značajne riječi koje se odnose na vrijednost (pozitivnu ili negativnu), stoga je potrebno ukloniti nevrijedne riječi i simbole iz pohranjenih riječi dobivenih iz prethodne faze. Neke od tih riječi su izrazi koji započinju znakovima @ i #. Oni nemaju vrijednost jer se koriste u posebne svrhe

na Twitteru. Na primjer, znak @ se koristi za označavanje korisničkih imena u tweetovima, poput „Dobro jutro @twitter!“. Korisnik na Twitteru može koristiti nominirano korisničko ime @korisničko_ime kako bi spomenuo drugog korisnika u tweetovima, poslao mu poruku ili stvorio vezu do njegovog profila. Također, u ovoj fazi bit će uklonjene jednoslovne riječi, brojevi i interpunkcijski znakovi. Ova faza će smanjiti pohranu korištenu za skup podataka i zadržati samo one podatke koji će se koristiti za sustav analize sentimenta. Uklanjanje jednoslovnih riječi i riječi s dva slova, osim riječi „no“, jer ona mijenja značenje rečenica, što utječe na analitički proces. Slika 3 prikazuje algoritam ove faze. Ovaj algoritam dohvaća token iz tablice tokena. Ako token sadrži jedan ili više simbola kao prefiks, koraci 205-206 će ukloniti te simbole. Funkcija trimleft() uklanja prvo slovo iz tokena. Slovo se uklanja ako je član skupa simbola prikazanog u koraku 200. Ako token sadrži jedan ili više simbola kao sufiks, koraci označeni brojevima 209-210 će ukloniti te simbole. Funkcija trimright() uklanja zadnje slovo iz tokena. Slovo se uklanja ako je član skupa simbola. Tada će token ostati kao riječ koja sadrži samo slova. Dva simbola (#) i (@) se ne uklanjaju jer će unos tokena koji započinju s njima biti uklonjen iz tablice tokena, uz dodatne tokene koji se sastoje od jednog ili dva slova, ili ako token sadrži samo simbole bez slova. U tim slučajevima unos tokena također će biti uklonjen iz tablice tokena, što se radi u koraku broj 212. U prethodnom primjeru, tokeni će biti očišćeni kako bi dali ovaj rezultat: „It“, „will“, „be“, „very“, „hot“, „summer“, „think“. Riječ „summer,“ se smatra tokenom nakon uklanjanja zareza. Fraza „@The-knight“ se uklanja iz tablice tokena jer se odnosi na korisničko ime „The-knight“ i neće imati vrijednost kao pozitivna ili negativna u analitičkom procesu. Slova „a“ i „I“ se uklanjaju jer su jednoslovna (Al-Khafaji i Habeeb, 2017).

```
Cleaning algorithm
Input      Tokens _table
Output   updated   Tokens_table
200 symbols = {,,.,;,:,?,~,`,+,=,%,!,$,^,&,*,",-,_,(,),{,},[,],|,\,/}
201  while (more tokens exist in tokens table) do
202 {T= next token;
203      F= the first letter of T;
204      while F in symbols do // remove symbols from the beginning of token
205      {trimleft (T, F, T);
206        F= the first letter in T ;}
207      E= the last letter of T;
208      while E in symbols do // remove symbols from the end of token
209       {trimright (T, F, T);
210         E=the last letter in T ;}
211      concatenation (T, E, T);
212      if T starts with # or @ or the length of T<=2   then delete T from tokens table}
```

Slika 3. Algoritam koraka uklanjanja zaustavnih riječi (Al-Khafaji i Habeeb, 2017)

Do sada su podaci pohranjeni u tablici velikog skupa podataka i sadrže mnogo riječi koje nisu korisne za analitički sustav, a koje su poznate kao zaustavne riječi. Zaustavne riječi engleskog jezika pohranjene su u tablici, a u ovoj fazi uspoređuju se stavke u tablici tokena sa svakom riječju u tablici zaustavnih riječi kako bi se iz tablice tokena izbrisale zaustavne riječi za svaki tweet. Slika 4 prikazuje algoritam za uklanjanje zaustavnih riječi. Isti prethodni primjer dat će ovakav rezultat: preostali tokeni su „hot“, „summer“, „think“. Riječi „It“, „will“, „be“, „very“ uklanjaju se kao zaustavne riječi (Al-Khafaji i Habeeb, 2017).

```
Stop words removing algorithm

Input    Tokens _table

Output updated tokens_table

300   while (more tokens exist in tokens table) do

301  {T= next token;

302       if T in stop_words_table

303        then delete the entry of T from tokens table}
```

Slika 4. Algoritam koraka uklanjanja zaustavnih riječi (Al-Khafaji i Habeeb, 2017)

Tablica 1 prikazuje tekstove nekih tweetova kao primjere u procesima.

Izvršavanje koraka tokenizacije nad skupom podataka doveo je do rezultata u kojem je pohranjeno 5.372.158 tokena u tablici tokena. Tablica 2 prikazuje uzorak podataka nakon segmentacije riječi i čišćenja riječi. Prema tablici 2, tokeni (vrlo, lijep, film i ljubav) pripadaju tweetu br. 1 i isto vrijedi za sljedeće tweetove (Al-Khafaji i Habeeb, 2017).

U ovom koraku uklanjaju se zaustavne riječi koje su već pohranjene u tablici tokena. Time se smanjuje broj tokena na 3.162.653 tokena. Tablica zaustavnih riječi sadrži 403 riječi, dobivene sa web stranice Onix Text Retrieval Toolkit (Al-Khafaji i Habeeb, 2017).

Tablica 1. Tekst tweetova (Al-Khafaji i Habeeb, 2017)

| Broj tweeta | Tekst tweetova |
|---|---|
| 1 | „very nice film I love it“ |
| 2 | „Nice time I love this film“ |
| 3 | „I loved it very much nice film“ |
| 4 | „not nice film, so long, it should be shorter“ |
| 5 | „long film but I love the hero, it is nice“ |

Tablica 2. Token iz određenih tweetova (Al-Khafaji i Habeeb, 2017)

| Broj tweeta | Token |
|---|---|

| | |
|---|---|
| 1 | Very |
| 1 | Nice |
| 1 | Film |
| 1 | Love |
| 2 | Nice |
| 2 | Time |
| 2 | Love |
| 2 | This |
| 2 | film |

U predloženom sustavu koristi se Porter stemmer. Ovaj korak će provesti smanjenje broja riječi. Na primjer, tokeni „loved“ u tweetu br. 3 i „shorter“ u tweetu br. 4 bit će reducirani na riječi „love“ i „short“ (Al-Khafaji i Habeeb, 2017). Tablica 3 prikazuje tokene nakon uklanjanja zaustavnih riječi.

Tablica 2~~3~~. Tokeni nakon uklanjanja zaustavnih riječi (Al-Khafaji i Habeeb, 2017)

| Broj tweeta | Token |
|---|---|
| 1 | Nice |
| 1 | Film |
| 1 | Love |
| 2 | Nice |
| 2 | Time |
| 2 | Love |
| 2 | film |

Vrijeme izvođenja predloženog sustava bilo je sat i 32 minute za skup podataka koji sadrži milijun tweetova. Izvršeno je na računalu s 8,00 GB RAM-a, procesorom Intel(R) Core(TM)

i7 s 1,87 GHz i hard diskom od 450 GB. Algoritmi pretprocesiranja implementirani su korištenjem SQL server pohranjenih procedura (Al-Khafaji i Habeeb, 2017).

Priprema podataka za sustav analize sentimenta zahtijeva nekoliko koraka. Ovi koraci igraju ulogu smanjenja podataka u rudarenju podataka kako bi se sačuvalo vrijeme rudarenja i prostor za pohranu bez gubitka podataka, posebno kod ogromnog skupa podataka. Svaki korak smanjuje količinu podataka uklanjanjem nekorisnih stavki. Te stavke mogu biti simboli, riječi ili fraze. Zatim se koristi algoritam za korjenovanje kako bi se podaci učinili djelotvornima za analitički sustav, što smanjuje varijabilne riječi koje imaju slično značenje. Korjenovanje je vrlo važan korak za sustav analize sentimenta. Sustav pretprocesiranja implementiran je korištenjem MS-SQL servera, što može biti razlog za relativnu sporost (Al-Khafaji i Habeeb, 2017).

Zaključak provedenog istraživanja jest, SQL nije pogodan za takve aplikacije, ali autori su inzistirali na njegovoj upotrebi kako bi dizajnirali potpuno integriran sustav rudarenja podataka. Međutim, autori su osmislili plan za implementaciju sustava korištenjem segmentacije podataka i višenitnog izvršavanja kako bi se povećala vremenska učinkovitost sustava (Al-Khafaji i Habeeb, 2017).

# 5. Zaključak

Područje računalne obrade prirodnog jezika, jezika SQL-a u kontekstu analize sentimenta izrazito je složeno područje s jako produbljenom interakcijom među područjima. Ovim radom produbili su se temeljni koncepti računalne obrade prirodnog jezika. Obuhvaćeni su njeni dijelovi koji uključuju jezične alate, jezične resurse te komercijalne proizvode koji također mogu biti svrstani pod jezične alate. Jednako tako objašnjen je i SQL u kontekstu korisnosti u analizi sentimenta. Zahvaljujući tehnikama računalne obrade prirodnog jezika u mogućnosti smo „dešifrirati“ široke nijanse ljudskog jezika i izražavanja te izvlačiti osjećaje, mišljenja i stavove pojedinaca iz mnoštva tekstualnih izvora. Nadalje, sjedinjenje računalne obrade prirodnog jezika i SQL-a te sinergija između razumijevanja jezika, manipulacije i dohvaćanja podataka dovela je do razvoja korisnih alata za analizu sentimenta s praktičnom primjenom u različitim područjima, od istraživanja tržišta pa do praćenja društvenih mreža (npr. Twittera koji je naveden kao primjer u radu). Međutim, nužno je priznati da izazovi i dalje postoje. Upravo složenost ljudskog jezika, različitih konteksta i kulturnih varijacija mogu predstavljati velike prepreke za postizanje optimalne analize sentimenta. Nadalje, kako volumen podataka nastavlja rasti, jednako tako raste i potreba za učinkovitim rješenjima i što bržim i efektivnijim djelovanjem SQL-a pri velikim količinama podataka. Također, koliko se za sada čini, pristupi putem korpusa i rječnika su manje zastupljeni od onih pristupa baziranim na strojnom učenju. Izgledi za daljnji napredak NLP-a i SQL-a ostaju obećavajući. Dok istraživanje i razvoj u području strojnog učenja i umjetne inteligencije nastavljaju evoluirati, može se očekivati poboljšana točnost i prilagodljivost alata za analizu sentimenta. Slično, napredak u sustavima upravljanja bazama podataka nastavit će poboljšavati sposobnosti SQL-a, omogućujući usklađeniju obradu podataka. Ovaj rad istaknuo je ključne uloge koje NLP i SQL imaju u području analize sentimenta. Njihov zajednički utjecaj nije samo preoblikovao način na koji analiziramo tekstualne podatke, već je i pružio dragocjene uvide u ljudske emocije i stavove.

# 6. Literatura


1. Aggarwal, C. C., & Zhai, C. X. (2012). Mining Text Data. Springer New York Dordrecht Heidelberg London: Springer Science+Business Media, LLC.
2. Al-Khafaji, D. H. K., & Habeeb, A. T. (2017). Efficient algorithms for preprocessing and stemming of tweets in a sentiment analysis system. *IOSR J. Comput. Eng*, *19*(3), 44-50.
3. Bolshakov, I. A., & Gelbukh, A. (2004). Computational Linguistics (Models, Resources, Applications)
4. Caro, L. D., & Grella, M. (2012). Sentiment analysis via dependency parsing. Computer Standards & Interfaces.
5. diskurs. *Hrvatska enciklopedija, mrežno izdanje.* Leksikografski zavod Miroslav Krleža, 2021. Pristupljeno 17. 8. 2023. <http://www.enciklopedija.hr/Natuknica.aspx?ID=15415>.
6. Dunđer, I. (2013). CroSS: Croatian Speech Synthesizer-design and implementation. In *Proc. 16th International Multiconference INFORMATION SOCIETY-IS* (pp. 257-260).
7. Dunđer, I. (2015). *Sustav za statističko strojno prevođenje i računalna adaptacija domene (Statistical Machine Translation System and Computational Domain Adaptation)* (Doctoral dissertation, Doctoral dissertation, University of Zagreb, Zagreb).
8. Dunđer, I. (2020). Machine translation system for the industry domain and Croatian language. *Journal of information and organizational sciences*, *44*(1), 33-50.
9. Dunđer, I. (2021a). Analiza modela sustava za automatsko statističko strojno prevođenje. *Politehnika: Časopis za tehnički odgoj i obrazovanje*, *5*(2), 39-47.
10. Dunđer, I. (2021b). PREGLED RAZVOJA TEHNOLOGIJE AUTOMATSKOG STROJNOG PREVOĐENJA. *POLYTECHNIC&DESIGN*, *9*(02), 90-100. DOI: 10.19279/TVZ.PD.2021-9-2-03
11. Dunđer, I., Horvat, M., & Lugović, S. (2017). Exploratory Study of Words and Emotions in Tweets of UK Start-up Founders. In *Proceedings of the Second International Scientific Conference "Communication Management Forum" (CMF2017). The Edward Bernays College of Communication Management. Zagreb* (pp. 201-224).

12. Dunđer, I., & Pavlovski, M. (2019). Behind the dystopian sentiment: a sentiment analysis of George Orwell's 1984. In *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 685-690). IEEE.
13. Groff, J. R., Weinberg, P. N., & Oppel, A. J. (2002). *SQL: the complete reference* (Vol. 2). McGraw-Hill/Osborne.
14. Hatzivassiloglou, V., & McKeown, K. (1997). Predicting the semantic orientation of adjectives. Proceedings of the annual meeting of the Association for Computational Linguistics (ACL '97).
15. Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of ACM SIGKDD international conference on Knowledge Discovery and Data Mining (KDD '04).
16. Jaworski, R., Seljan, S., & Dunđer, I. (2017). Towards educating and motivating the crowd–a crowdsourcing platform for harvesting the fruits of NLP students' labour. In *Proc. 8th Language & Technology Conference–Human Language Technologies as a Challenge for Computer Science and Linguistics* (pp. 332-336).
17. Jaworski, R., Seljan, S., & Dunđer, I. (2023). Four Million Segments and Counting: Building an English-Croatian Parallel Corpus through Crowdsourcing Using a Novel Gamification-Based Platform. *Information*, *14*(4), 226.
18. Jiao, J., & Zhou, Y. (2011). Sentiment Polarity Analysis based multi-dictionary. Presented at the 2011 International Conference on Physics Science and Technology (ICPST '11).
19. Joty, S., Carenini, G., Ng, R., & Murray, G. (2019). Discourse analysis and its applications. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts (pp. 12-17).
20. Kim, S., & Hovy, E. (2004). Determining the sentiment of opinions. Proceedings of the international conference on Computational Linguistics (COLING '04).
21. Kovačević, A. i Kovačević, Ž. (2021). ALATI ZA ANALIZU SENTIMENTA. *Polytechnic and design, 9* (3), 167-174. https://doi.org/10.19279/TVZ.PD.2021-9-3-02
22. Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Proceedings of the International Conference on Machine Learning (ICML '01).

23. Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies.
24. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams engineering journal*, *5*(4), 1093-1113.
25. Mejova, Y., & Srinivasan, P. (2011). Exploring feature definition and selection for sentiment classifiers. Proceedings of the fifth international AAAI conference on weblogs and social media.
26. Miller, G., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). WordNet: An on-line lexical database. Oxford University Press.
27. Min, H. J., & Park, J. C. (2012). Identifying helpful reviews based on customer's mentions about experiences. Expert Systems with Applications, 39, 11830–11838.
28. Mohammad, S., Dunne, C., & Dorr, B. (2009). Generating high-coverage semantic orientation lexicons from overly marked words and a thesaurus. Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP '09).
29. Moreo, A., Romero, M., Castro, J. L., & Zurita, J. M. (2012). Lexicon-based comments-oriented news sentiment analyzer system. Expert Systems with Applications, 39, 9166–9180.
30. Ng, V., & Cardie, C. (2002). Improving machine learning approaches to coreference resolution. Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 104-111).
31. Qiu, G., He, X., Zhang, F., Shi, Y., Bu, J., & Chen, C. (2010). DASA: Dissatisfaction-oriented advertising based on sentiment analysis. Expert Systems with Applications, 37, 6182–6191.
32. Seljan, S., Dunđer, I., & Gašpar, A. (2013). From digitisation process to terminological digital resources. In *2013 36th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (pp. 1053-1058). IEEE.
33. Seljan, S., Dunđer, I., & Stančić, H. (2017). Extracting terminology by language independent methods. In *the Proceedings of the 2nd International Conference on Translation and Interpreting Studies, Innsbruck, Austria, Peter Lang* (pp. 141-147).
34. Seljan, S., Erdelja, N. Š., Kučiš, V., Dunđer, I., & Bach, M. P. (2021). Quality Assurance in Computer-Assisted Translation in Business Environments. In *Natural Language Processing for Global and Local Business.* IGI Global, 247-270. DOI: 10.4018/978-1-7998-4240-8.ch011

35. Tadić, M. (2003). *Jezične tehnologije i hrvatski jezik.* Zagreb: Ex Libris.
36. Tsytsarau, M., & Palpanas, T. (2012). Survey on mining subjective data on the web. Data Mining and Knowledge Discovery, 24(3), 478–514.
37. Whitelaw, C., Garg, N., & Argamon, S. (2005). Using appraisal groups for sentiment analysis. Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM), 625-631.
38. Wilson, T., Wiebe, J., & Hoffman, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. Proceedings of HLT/EMNLP.
39. Xu, K., Liao, S. S., Li, J., & Song, Y. (2011). Mining comparative opinions from customer reviews for competitive intelligence. Decision Support Systems, 50, 743–754.

# 7. Popis slika i tablica

# 8. Sažetak


## Uloga računalne obrade prirodnog jezika i SQL-a u analizi sentimenta

U ovom završnom radu fokus je na primjeni i važnosti računalne obrade prirodnog jezika i SQL-a u analizi sentimenta, što uključuje određivanje osjećaja, stavova i mišljenja, različitih tekstova iz različitih područja ljudskog djelovanja i situacija u stvarnom životu. Ponajprije su objašnjeni svi krovni pojmovi koji su potrebni kako bi se lakše prodrlo u srž teme. Zatim su predstavljeni različiti aspekti računalne obrade prirodnog jezika, tj. neke osnovne zadaće računalne obrade prirodnog jezika, poput tokenizacije, lematizacije, analize riječi, semantičke analize i drugih te pojedini algoritmi i tehnike unutar same računalne obrade prirodnog jezika. Također je objašnjeno na koji je način SQL koristan i bitan za analizu sentimenta, posebice pri pronalaženju velike količine podataka iz velikih baza podataka. Objašnjeni su svi procesi koji se koriste i koji su prisutni pri analizi sentimenta te kako SQL i NLP izvršavaju svoje prethodno navedene zadatke pri analizi sentimenta. Naposljetku, dan je zaključak o korisnosti NLP-a i SQL-a u analizi sentimenta te hipotetskim problemima koji im stoje na putu za još uspješnije izvršavanje analize.



Ključne riječi: računalna obrada prirodnog jezika, NLP, SQL, analiza sentimenta, jezični alati

## 9. Summary

# The Role of Computational Natural Language Processing and SQL in Sentiment Analysis

This bachelor thesis focuses on the application and importance of Natural Language Processing (NLP) and SQL in sentiment analysis, which involves determining the feelings, opinions, and attitudes expressed in various texts from different areas of human activity and real-life situations. Firstly, all the key concepts necessary to delve into the core of the topic are explained. Subsequently, various aspects of Natural Language Processing are introduced, including fundamental tasks such as tokenization, lemmatization, word analysis, semantic analysis, and other related algorithms and techniques within the realm of NLP. The manner in which SQL is useful and essential for sentiment analysis is also explained, especially in dealing with substantial amounts of data from large databases. The processes used and involved in sentiment analysis will be elucidated, detailing how SQL and NLP execute their aforementioned tasks within sentiment analysis. Finally, a conclusion is drawn regarding the utility of NLP and SQL in the sentiment analysis, along with hypothetical challenges they may encounter in their pursuit of more successful analysis.

Keywords: natural language processing, NLP, SQL, sentiment analysis, language tools