TakeLab

Laboratorij za analizu teksta i inženjerstvo znanja
Text Analysis and Knowledge Engineering Lab
Sveučilište u Zagrebu, Fakultet elektrotehnike i računarstva
Unska 3, 10000 Zagreb, Hrvatska

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

DIPLOMSKI RAD br. 777

Analiza sentimenta u tvitovima na hrvatskom jeziku

Luka Krajcar

Zagreb, lipanj 2014.

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA
ODBOR ZA DIPLOMSKI RAD PROFILA

Zagreb, 10. ožujka 2014.

DIPLOMSKI ZADATAK br. 777

Pristupnik: **Luka Krajcar**
Studij: Računarstvo
Profil: Računarska znanost

Zadatak: **Analiza sentimenta u tvitovima na hrvatskome jeziku**

Opis zadatka:

Porastom količine korisnički generiranog sadržaja povećalo se zanimanje za strojnom analizom sentimenta. Osobito pogodan izvor podataka sačinjavaju tvitovi (engl. tweets), kratke poruke koje u stvarnome vremenu odašilju korisnici društvene mreže Twitter, a koje su dostupne javno i u velikim količinama. Poteškoću predstavalju kratkoća tekstova te nemogućnosti ručnog označavanja tako velikog broja poruka kroz različite teme. Problem se može ublažiti primjenom modela lagano nadziranog strojnog učenja, kod kojega se kao "oznake sa šumom" iskorištavaju elementi tvita koji upućuju na emocije korisnika.
U okviru diplomskoga rada potrebno je proučiti modele za analizu sentimenta u korisnički generiranome sadržaju s naglaskom na analizu sentimenta u porukama društvenih mreža. Razraditi model za analizu sentimenta u tvitovima na hrvatskome jeziku temeljen na modelu lagano nadziranog učenja. Izraditi odgovarajući ispitni skup s ručno označenim sentimentom. Razviti programsku implementaciju modela te ispitati rad nekoliko klasifikacijskih modela, uključivo stroja potpornih vektora, naivnog Bayesovog klasifikatora i modela maksimalne entropije. Ispitati utjecaj različitih načina predobrade teksta, odabira značajki i uporabe rječnika apriornog sentimenta. Razraditi i implementirati sustav koje će omogućiti grupiranje tvitova prema korisnički zadanome upitu te prikaz agregiranog sentimenta. Radu priložiti izvorni i izvršni kod razvijenog sustava, označene skupove podataka i potrebnu dokumentaciju te citirati korištenu literaturu.

Zadatak uručen pristupniku: 14. ožujka 2014.
Rok za predaju rada: 30. lipnja 2014.

Mentor:

Doc. dr.sc. Jan Šnajder

Djelovođa:

Doc. dr.sc. Tomislav Hrkać

Predsjednik odbora za diplomski rad profila:

Prof. dr.sc. Siniša Srbljić

SADRŽAJ

POPIS SLIKA

POPIS TABLICA

1. Uvod

U posljednjih nekoliko godina primjetan je značajan porast popularnosti društvene mreže Twitter[1] kako u svijetu, tako i u Hrvatskoj. Twitter kao najpopularniji mikro-blogging (engl. *micro-blogging*) servis s gotovo milijardu korisnika, te gotovo 60 milijuna tvitova (engl. *tweets*) dnevno predstavlja izvrstan izvor korisnički generiranog sadržaja. Tvitovi su poruke kojima korisnici javno ili unutar određene grupe objavljuju svoje interese, raspoloženja, stavove i sl. Ono što Twitter razlikuje od ostalih društvenih mreža je ograničenje duljine tvita koje ga i čini mikro-blogging servisom. Naime, svaki tvit može imati najviše 140 znakova. Ovakvo ograničenje iziskuje od korisnika da njihovi tvitovi budu sažeti te se tvitovi često sastoje od samo jedne ili dvije rečenice s prosjekom od 15 riječi po tvitu.

S obzirom na zastupljenost Twittera u gotovo svim sferama javnog ali i privatnog života, može se smatrati da tvitovi na jedan način predstavljaju refleksiju onoga što se događa u svijetu. U posljednje vrijeme pogotovo je zanimljiva promjena trendova kao jedan od aspekata koji se može pronaći u tvitovima. Upravo zbog toga raste interes za komercijalnom eksploatacijom tvitova i sve je veći broj tvrtki i marketinških agencija kojima tvitovi predstavljaju zlatni rudnik. Povratna informacija koja se dobiva od prosječnog korisnika može se pokazati kao presudan faktor u stvaranju prednosti u odnosu na konkurenciju, i čini se da analiza tvitova može donijeti tu prednost.

Najčešće analizirani trendovi tiču se proizvoda i usluga, tj. tvitova koji sadrže informacije o proizvodima i uslugama. Takvi tvitovi najčešće sadrže izraženo mišljenje ili stav korisnika koji se jednom riječju naziva sentimentom. Sentiment se može klasificirati u razne kategorije. Tipična je klasifikacija u tri klase, pozitivnu, negativnu i neutralnu. Ovakva podjela može se i dalje dijeliti na još preciznije klase, ali može se i pojednostaviti i tada se u obzir uzimaju samo pozitivna i negativna klasa. Analiza sentimenta je jedan od zadataka kojima se bavi obrada prirodnog jezika. U ovom radu naglasak je stavljen na analizu sentimenta u tvitovima na hrvatskom jeziku.

U području strojnog učenja i obrade prirodnog jezika razvijeni su brojni algoritmi

[1] www.twitter.com

s ciljem ekstrakcije ključnih značajki. Neki od najznačajnijih klasifikacijskih algoritama poput naivnog Bayesovog klasifikatora, stroja potpornih vektora (engl. *support vector machines - SVM*) i logističke regresije iskorišteni su za klasifikaciju tvitova na temelju izraženog sentimenta te je uspoređena njihova uspješnost. Glavni fokus ovog rada je upravo na usporedbi efikasnost raznih klasifikacijskih algoritama kod problema klasifikacije hrvatskih tvitova te traženja odgovora na pitanje: "Kako najuspješnije klasificirati tvit temeljem sentimenta izraženog u tom tvitu?" Kad je u pitanju označavanje tvitova, najčešće se radi o ogromnom broju tvitova te ručno označavanje postaje neprimjereno za ovakav problem. Zato je u radu ispitana mogućnost korištenja blagih oznaka (engl. *noisy labels*). Kod blagog označavanja koriste se razni elementi tvita kako bi se taj tvit uspješno klasificirao. Blage oznake temeljene su najčešće na emotikonima i hashtagovima. Postignuti rezultati lošiji su od onih postignutih u srodnim radovima zbog cijelog niza problema a najviše zbog .

Kako bi se rezultati istraživanja prikazali u nešto jasnijem stanju potrebno je razviti korisničku aplikaciju za prikaz agregiranog sentimenta. U ovakvoj aplikaciji korisniku se omogućuje da unosi upite pisane u prirodnom jeziku a kao rezultat prikazuju se tvitovi koji sadrže uneseni upit te razne statistike o relevantnim tvitovima. Ovakva aplikacija mora sadržavati veliku bazu označenih tvitova. Tvitovi su označeni na temelju najuspješnijeg klasifikacijskog algoritma.

Rad je organiziran na sljedeći način. U drugom poglavlju iznesen je pregled srodnih radova koji su korišteni i u izradi ovog rada. Kroz pregled srodnih radova izneseni su i ključni aspekti rada na kojima će biti najveći naglasak. U trećem poglavlju razrađena je teoretska osnova korištenih metoda predobrade tvitova i korištenih klasifikacijskih algoritama. U četvrtom poglavlju opisan je korišteni skup podataka, rezultati predobrade ulaznog skupa, rezultati klasifikacije korištenjem različitih klasifikacijskih algoritama te komentar i usporedba postignutih rezultata. U petom poglavlju opisana je korisnička aplikacija za prikaz agregiranog sentimenta, baza podataka koju koristi ta aplikacija te korištene tehnologije. U posljednjem poglavlju iznesen je zaključak na temelju postignutih rezultata u radu.

2. Srodni radovi

Kako bi se u ovom radu postigli što bolji rezultati te kako bi se unaprijed izbjegli mogući problemi bilo je potrebno proučiti srodne radove i analizirati postupke korištene u njima. Srećom, na temu analize sentimenta napisan je velik broj radova, onih koji se bave generalnom analizom sentimenta ali i onih koji se bave analizom sentimenta tvitova. U ovom poglavlju bit će izložen sažetak ključnih članaka korištenih u izradi te opisane sličnosti i razlike u odnosu na ovaj rad.

Prvi u nizu proučenih radova, (Pang i Lee, 2008) je zapravo i osnovni rad citiran u gotovo svim ostalim radovima kad je u pitanju analiza sentimenta. Za ovaj rad se može reći da je definirao pojam analize sentimenta kao i svu pripadajuću terminologiju, te na jednom mjestu prikupio i sistematizirao dotad objavljenu literaturu na ovu temu. U radu se analiza sentimenta definira kao problem klasifikacije teksta, najčešće u dvije ili tri klase (pozitivna-negativna ili neutralna-pozitivna-negativna). Govori se o ključnim izazovima i problemima pri klasifikaciji kao što su neformalni i često jezično i gramatički neispravni tekstovi, prisutnost sarkazma, ironije itd. Nadalje, istražuje se koje su značajke i koji klasifikacijski modeli optimalni za ovaj problem. Također, rad sadrži i pregled potencijalne namjene analize sentimenta kao što su npr. sažetak recenzija na raznim internetskim stranicama, povratne informacije za razne tvrtke i javne osobe, te kao podsustav nekog složenijeg sustava strojnog učenja. Ovaj rad predstavlja odličan uvod u tematiku, i iako ne utječe direktno na praktični dio rada poslužio je za upoznavanje sa svim negativnim i pozitivnim aspektima analize sentimenta.

Sljedeća dva rada bila su ključna u izradi rada zbog velikog tematskog preklapanja. Prvi od tih radova je (Kouloumpis et al., 2011). U radu se ispituju mogućnosti blagog označavanja tvitova temeljenog na emotikonima i hashtagovima te utjecaj različitih lingvističkih značajki na uspješnost klasifikacije. U radu su navedeni najčešći hashtagovi u njihovom korpusu tvitova a uz to i tipični predstavnici negativni, pozitivnih i neutralnih hashtagova. Ovo je bilo izuzetno korisno kod primjene blagog označavanja putem hashtagova. Korisne metode za redukciju broja značajki još su jedno značajno poboljšanje koje je otkriveno kroz ovaj rad te se tako ubrzala i poboljšala klasifika-

cija. Posljednja, ali ne i najmanje bitna mogućnost je mogućnost usporedbe rezultata u smislu količine tvitova korištenih u klasifikaciji. Razlike navedenog rada i diplomskog rada očituju se u vrsti klasifikatora i korištenju lingvističkih značajka (u navedenom radu se koriste značajke poput vrsta riječi, predefinirani rječnici sentimenta itd.).

Drugi značajni rad je (Go et al., 2009a). Ovaj rad ispituje mogućnosti označavanja tvitova samo na temelju emotikona. Rad je ponudio uvid u kvalitetniju predobradu i redukciju kako broja tvitova, tako i broja značajki te donio ideju za osnovni klasifikator temeljen na rječniku. Kod redukcije broja tvitova otkrivene su neke ključne poput redukcije duplih tvitova, retvitova i tvitova s obje vrste emotikona. Također, u radu se koriste naivni Bayesov klasifikator i stroj potpornih vektora koji su osnovni klasifikatori u diplomskom radu. Ograničenje na klasifikaciju u dvije klase, korištenja modela maksimalne entropije, podjela tvitova u tematske kategorije neki su od pristupa koji nisu iskorišteni u diplomskom radu te u tome leži ključna razlika ova dva rada.

Preostali radovi nisu imali ključan utjecaj na izradu diplomskog rada, međutim u njima su pronađene neke metode i pristupi koji su donijeli određeno poboljšanje rada. Prvi od takvih radova je (Go et al., 2009b) koji je dao ideju o uporabi Weka kolekcije implementacija algoritama strojnog učenja kao okruženja za razvoj i testiranje klasifikacijskih algoritama te korištenja n-grama za poboljšanje rada klasifikatora. (Sharma i Vyas) je dao detaljniji opis osnovnog klasifikatora temeljenog na rječniku te mogućih poboljšanja takvog klasifikatora. U (Pak i Paroubek, 2010) je uočena mogućnost klasifikacije tvitova u dva koraka. Prvi korak je klasifikacija tvita kao subjektivnog ili objektivnog, a drugi korak, koji se odvija ako je tvit klasificiran kao subjektivan je klasifikacija u pozitivnu ili negativnu klasu. Od ovakvog pristupa se ipak odustalo zbog potrebe za povećanjem ručno označenog skupa koji je ionako bio poprilično velik. U (Agarwal et al., 2011) pronađeno je proširenje liste pozitivnih i negativnih emotikona koje se pozitivno odrazilo na uspješnost klasifikacije. Također u radu su predstavljeni poprilično dobri rezultati klasifikacije putem stabla odluke, međutim u ovom radu stabla odluke nisu korištena zbog izuzetno loših rezultata na ispitnom skupu.

3. Klasifikacijski modeli

U ovom poglavlju iznesen je teoretski pregled blago nadziranog strojnog učenja, metoda korištenih u predobradi tvitova i korištenih klasifikacijskih modela. Metode korištene za predobradu tvitova dijele se na metode za redukciju broja tvitova i metode za redukciju broja značajki. Korišteni klasifikacijski modeli su osnovni klasifikator temeljen na rječniku, naivni Bayesov klasifikator, stroj potpornih vektora te logistička regresija.

3.1. Blago nadzirano strojno učenje

Nadzirano strojno učenje je model strojnog učenja u kojemu se na klasifikator dovode primjeri iz ulaznog skupa koji su ručno označeni. U slučaju ručnog označavanja tvitova to bi značilo da je jedna ili više osoba dodijelila ručnu oznaku svakom korištenom tvitu tj. svrstala je taj tvit u pozitivnu, negativnu ili neutralnu klasu. Na temelju tako označenih tvitova moguće je izgraditi klasifikacijske modele.

Međutim kad se radi o velikom broju ulaznih podataka tj. u ovom slučaju tvitova, ručno označavanje postaje teško, tj. vremenski izrazito zahtjevno. Zato se u posljednje vrijeme sve više ispituje mogućnost korištenja blago nadziranog strojnog učenja. Kod takvog učenja za oznake se uzimaju određeni elementi ulaznih podataka. Kod tvitova se za "oznake sa šumom" uzimaju dvije vrste elemenata tvita, hashtagovi[1] i emotikoni[2] (engl. *hashtags and emoticons*).

Hashtag je zapravo riječ s prefixom "#" korištena najčešće u društvenim mrežama kao metapodatkovni tag sa svrhom grupiranja poruka. Ovakav pristup omogućuje korisnicima da pretražuju poruke koje sadrže određenih hashtag. Hashtagovi su često povezani s određenim tipom emocija, tako su npr. popularni pozitivni hashtagovi: "#love","#win", "#success", negativni hashtagovi: "#fail", "#ihate“, "#worst"“ i neutralni: "#job","#news", "#facts" itd. Na temelju hashtagova koje sadrži, pojedini tvit

[1] www.hashtags.org

[2] en.wikipedia.org/wiki/Emoticon

moguće je proglasiti pozitivnim, negativnim ili neutralnim. Često se koriste upravo hashtagovi na engleskom jeziku, neovisno o govornom području iz kojeg dolazi osoba koja objavljuje tvit. Ovo ipak ne vrijedi u potpunosti kad se radi o tvitovima na hrvatskom jeziku, kao što će se pokazati u budućem poglavlju. U ovom radu međutim nije korištena lista unaprijed određenih hashtagova, već su oni određeni na temelju ulaznih podataka.

Emotikoni su kratke slikovne reprezentacije izraza lica sastavljene najčešće od interpunkcijskih znakova. Emotikoni se često koriste u tekstualnoj komunikaciji na internetu kako bi se nadomjestio nedostatak govora tijela i prozodijskih svojstava govora te tako poboljšala interpretacija napisanog. Emotikoni tipično izražavaju ili pozitivnu ili negativnu emociju, dok neutralni emotikoni praktički ne postoje. Dakle, emotikoni se mogu iskoristiti za označavanje tvita koji ih sadrži kao pozitivnog ili negativnog. Slika 3.1 sadrži pozitivne i negativne emotikone koji su korišteni kao oznake u ovom radu.

Pozitivni emotikoni	Negativni emotikoni
:), :), :=) , :-),	:(, : (, :-(, :=(,
:D, =), :P, :p	:'(, -_-, :@, -.-

Slika 3.1: Primjeri emotikona

3.2. Redukcija broja tvitova

Ulazni skup podataka tj. korpus hrvatskih tvitova često sadrži tvitove pisane na velikom broju različitih jezika. Također, u tom korpusu ponekad se nađe više jednakih tvitova, tvitova koji sadrže više vrsta oznaka sa šumom itd. Sve ovakve tvitove potrebno je ukloniti iz ulaznog korpusa. Iz tog razloga provodi se nekoliko redukcija tvitova. Te redukcije su:

- Redukcija na temelju jezika
- Redukcija retvitova (engl. *retweet*)
- Redukcija duplih tvitova
- Redukcija tvitova s obje vrste emotikona
- Redukcija na temelju jezične vjerojatnosti

Redukcija na temelju jezika je prva redukcija koja se provodi nad korpusom. To je najjednostavnija od navedenih redukcija. Iz korpusa tvitova hrvatskih korisnika odabiru se samo tvitovi na hrvatskom jeziku.

Retvitovi nastaju kad korisnici Twittera ponovo objavljuju tj. prosljeđuju tuđi tvit. Ovakvi tvitovi započinju tekstom "RT" i u pravilu sadrže tvit sadržajem identičnim originalnom tvitu. To drugim riječima znači da su takvi tvitovi praktički duplikati osnovnog tvita i kao takvi ne nose nikakvu korisnu informaciju u odnosu na original. Ova redukcija iz korpusa izbacuje sve tvitove čiji tekst započinje s "R".

Sljedeća redukcija je redukcija duplih tvitova. Ponekad se u dohvaćanju tvitova s Twittera dohvati isti tvit više puta. Redukcijom duplih tvitova pokušava se korpus pročistiti od ovakvih duplikata. S obzirom na veličinu korpusa bilo bi memorijski zahtjevno i sporo učitati cijeli korpus u memoriju računala i zatim izbaciti sve duplikate. Iz ovog razloga duplikati se izbacuju pomoću privremenog spremnika veličine tisuću tvitova. Unutar spremnika čuva se posljednjih tisuću pročitanih tvitova. Ako se idući pročitani tvit nalazi unutar spremnika odbacuje se a u suprotnom pohranjuje se u spremnik umjesto najstarijeg pročitanog tvita.

Unutar korpusa se povremeno pojavljuju tvitovi koji sadrže obje vrste emotikona, tj. i pozitivne i negativne emotikone. Ovakve tvitove ne može se jednoznačno odrediti kao negativne ili pozitivne te ih se zato odbacuje.

Posljednja redukcija je redukcija na temelju jezične vjerojatnosti. Ova redukcija provodi se kako bi se povećala vjerojatnost da su tvitovi koji će se koristiti u klasifikaciji i označavanju uistinu na hrvatskom jeziku.

3.3. Redukcija broja značajki

U procesu izrade modela strojnog učenja nad tekstnim korpusom potrebno je ulazni tekst, tj. tvitove transformirati u oblik prikladan za pojedini model. Tako će vjerojatnosni modeli zahtijevati da se za svaku riječ koja se pojavljuje u korpusu izračunaju vjerojatnosti pojave te riječi u klasama. Kod drugih će se pak modela tekst tj. rečenica prezentirati kao vreća riječi[3] (engl. *bag of words*) u kojem će se tekst zamijeniti vektorom čiji su elementi najčešće binarne vrijednosti (0 ili 1) koje označuju pojavljuje li se neka riječ u tekstu ili ne. Međutim iz ulaznog teksta nije jednostavno izolirati samo riječi koje se pojavljuju u njemu. Tekstovi često sadrže posebne znakove, brojeve, datume i sl. Zato se u okviru izrade modela strojnog učenja za klasifikaciju teksta

[3] en.wikipedia.org/wiki/Bag-of-words_model

umjesto o riječima govori o značajkama (engl. *feature*). Značajke su zapravo dijelovi teksta koje dobijemo razdvajanjem teksta po prazninama, najčešće po razmacima. Tako dobiveni dijelovi teksta su najčešće stvarne riječi, ali često se događa da sadrže i neke dodatne znakove poput zareza, točaka, navodnika i sl. Tako će se dogoditi da se u postupku računanja frekvencije pojavljivanja pojedine značajke, riječi koje se pojavljuju s i bez npr. zareza gledaju kao različite, iako su u principu jednake. Kako bi se ovakvi i slični problemi izbjegli te kako bi se tako poboljšala uspješnost modela potrebno je provesti određene redukcije značajki. Te redukcije su:

- Redukcija URL-ova
- Redukcija korisničkih imena (engl. username)
- Redukcija višestrukog ponavljanja znakova
- Redukcija posebnih znakova
- Redukcija brojeva
- Redukcija na mala slova

U tvitovima se često nailazi na pojavu URL-ova[4] (engl. *URL - Uniform resource locator*). URL-ovi su često prisutni kako bi korisniku omogućili da pročita opširniju vijest od one objavljene u samom tvitu koja je ograničena na 140 znakova. Međutim ovakve značajke ne nose značajnu informaciju, tj. iako su tekstualno različite svođenje svih ovakvih značajki na jednu zajedničku neće se negativno odraziti na kvalitetu korpusa, štoviše samo će ju poboljšati. Zato su sve značajke koje formiraju valjani URL zamijenjene značajkom "URL". Sada će jedino ova značajka nositi određeni sentiment tj. moći će se proučiti njena pripadnost određenoj klasi (npr. URL značajka je najčešće prisutna u neutralnoj klasi).

Korisnička imena su još jedna stvar koja se redovito javlja u tvitovima. Kad se jedan korisnik obraća drugom kroz tvit, prije imena tog korisnika dodaje znak @. Kao i kod url-ova, svako korisničko ime samo za sebe ne nosi značajni sentiment i zato se pojava svakog korisničkog imena zamjenjuje značajkom "KORISNIK", te se tako reducira pojavnost korisnički imena.

Redukcijom višestrukog ponavljanja znakova cilj je određeni broj značajki koje su korisnici izmijenili kako bi posebno naglasili neki sentiment svesti na jednu značajku.

[4] en.wikipedia.org/wiki/Uniform_resource_locator

Tako se npr. javljaju značajke: "juuuuuhuuuuu!", "juuhuuuuuu!", "juuuuhuuu!". Potrebno je odrediti maksimalni dopušteni broj ponavljanja određenog znaka i u ovom slučaju je kao mjera uzet broj dva. Svođenjem na jedan znak riskiralo bi se da se izmjene neke validne riječi poput "Aaron", "Apple", "zoološki" i sl. Redukcijom višestrukog ponavljanja znakova na dva znaka postiže se da se sve gore navedene riječi svedu na jednu značajku: "juuhuu!" Iako tako dobivene riječi često nisu ispravne, ipak je sentiment koncentriran na jednu značajku za koju je povećana vjerojatnost pojave u tekstu.

Unutar svakog tvita postoji određen broj posebnih znakova, kao npr. interpunkcijskih znakova, emotikona, znakova valute itd. Takvi znakovi djeluju tako da prividno povećavaju broj značajki u korpusu. Zato je cilj redukcije takvih znakova pročistiti korpus i u njemu ostaviti u najboljem slučaju samo riječi. U postupku redukcije odbacuju se najprije emotikoni korišteni u označavanju. Kako se u svakom pozitivnom i negativnom tvitu označenom pomoću blagih oznaka javljaju emotikoni, takvi emotikoni imaju stopostotnu pripadnost određenoj klasi i kao takvi bi previše utjecali na klasifikaciju a to nije cilj. Ista stvar vrijedi za hashtagove korištene za označavanje neutralnih tvitova. Nakon toga slijedi uklanjanje svih posebnih znakova i svih interpunkcijskih znakova.

Sljedeća redukcija je redukcija brojeva. Brojevi često bez nekakvog dodatnog objašnjenja tj. konteksta u kojemu se nalaze ne nose nikakav sentiment. Iz tog razloga svi brojevi u tvitovima zamijenjeni su značajkom "BROJ".

Posljednja redukcija je svođenje svih značajki na mala slova. Ova redukcija ima jednu negativnu posljedicu a ta je da će poneke vlastite imenice koje bi inače bile zasebne značajke biti svrstane s nekim značajkama koje se isto pišu. Tako će npr. „Uma Thurman“ nakon ove redukcije postati „uma thurman“ te će značajka „uma“ biti ubrojena u značajku „uma“ (genitiv riječi um). Međutim ovakvih primjera nema mnogo i ne unose značajni šum u podatke dok istovremeno značajno smanjuju broj značajki.

3.4. Klasifikatori

Kod opisivanja procesa klasifikacije najprije je bitno odrediti osnovni klasifikator, tj. klasifikator s najlošijom uspješnosti, čije se rezultate klasifikacije korištenjem naprednijih tj. prikladnijih klasifikatora pokušava nadmašiti. Osnova gotovo svake klasifikacije su nasumični (engl. *random*) klasifikatori. Takvi klasifikatori nasumično pogađaju oznaku tj. klasu tvita. Potrebe za implementacijom ovakvih klasifikatora nema, jasno je da će takvi klasifikatori za problem klasifikacije u dvije klase rezultirati s oko 50 postotnom točnošću, dok će za problem klasifikacije u tri klase rezultirati

s oko 33 postotnom točnošću u slučaju uravnoteženih klasa. Međutim, ovakav nasumični klasifikator se teško može smatrati osnovom, zato se kao osnova uzima klasifikator temeljen na rječniku. Teoretski opisi klasifikatora, izuzevši klasifikator temeljen na rječniku, preuzeti su iz Murphy (2012), te pojednostavljeni za potrebe rada.

3.4.1. Klasifikator temeljen na rječniku

Klasifikator temeljen na rječniku radi na sljedeći način. Najprije se odaberu sve riječi koje sadrže tvitovi svih klasa (neovisno radi li se o dvije ili tri klase). Zatim se za svaku odabranu riječ računa frekvencija pojavljivanja u svakoj klasi, i rječniku klase za koju ta riječ ima najveću frekvenciju pojavljivanja pridodaje se ta riječ. Kada su obrađene sve riječi, rječnici klasa sortiraju se prema frekvencijama. Riječi s najvećom frekvencijom pojavljivanja nalaze se pri vrhu rječnika. Zatim se iz tih opširnih rječnika odabire određen broj riječi koje će predstavljati konačne rječnike za pojedine klase. Eksperimentalno je utvrđeno da se najbolji rezultati dobivaju kad se za svaku klasu uzme po 100 riječi. Nakon izrade rječnika započinje klasifikacija. Ulazni tvit rastavlja se na riječi, i za svaku riječ provjerava se njena prisutnost u rječnicima klasa. Klasa u čijem se rječniku pronađe najveći broj riječi iz rečenice postavlja se kao oznaka tom tvitu, tj. tvit se klasificira u tu klasu. Pretpostavka je da ovakav poprilično jednostavan klasifikator neće imati veoma dobre rezultate, međutim za vjerovati je da će biti uspješniji od nasumičnog klasifikatora. Ovaj klasifikator bi prema organizaciji i izvedbi spadao u modele nenadziranog strojnog učenja, jer se klasifikacija kod ovog klasifikatora provodi na temelju skupa riječi a ne na ručnoj ili blagoj oznaci tvita.

3.4.2. Naivni Bayesov klasifikator

Naivni Bayesov klasifikator temelji se na poznatom Bayesovom teoremu koji modelira uvjetnu vjerojatnost. Uvjetna vjerojatnost $P(C|D)$ u slučaju klasifikacije teksta tj. dokumenta označava vjerojatnost da dokument D pripada klasi C. Bayesov teorem u tom slučaju kaže sljedeće:

$$P(C|D) = \frac{P(C)P(D|C)}{P(D)} \tag{3.1}$$

Bayesovo pravilo zapravo daje uvid u to kako se uvjerenje u neku hipotezu mijenja u prisutnosti novog dokaza. Drugim riječima, pravilo omogućava izračunavanje nepoznate uvjetne vjerojatnosti iz poznatih uvjetnih i apriornih vjerojatnosti. Pridjev

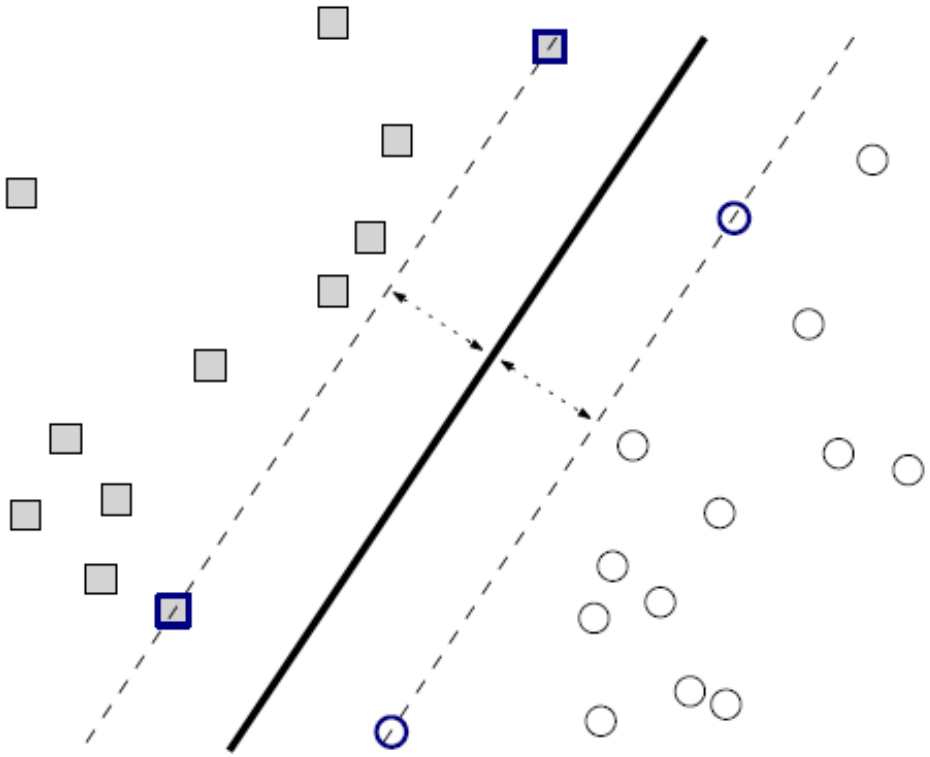
"naivan" pridjeljuje se Bayesovom klasifikatoru zbog naivne pretpostavke, a ta je da su sve značajke unutar jedne klase neovisne jedna o drugoj. U vidu tekstne klasifikacije to bi značilo da se izglednost klase može zamijeniti produktom uvjetnih vjerojatnosti pojedinih riječi, pod naivnom pretpostavkom da pojava jedne riječi ne utječe na pojavu iduće i obrnuto. Ako se riječ u rečenici označi sa R_i, gdje je i pozicija riječi u dokumentu, pretpostavka se zapisuje kao:

$$P(C|R_1 \dots R_n) \sim P(C) \prod_{i=1}^{N} P\left(R_i|C\right) \tag{3.2}$$

Iako se čini da je pretpostavka potpune neovisnosti vrlo vjerojatno pogrešna, pokazuje se da je u slučaju klasifikacije teksta i mnogih drugih klasifikacijskih problema zapravo vrlo uspješna. Uz to, naivna pretpostavka povećava jednostavnost algoritamske izvedbe te smanjuje memorijske i računske zahtjeve kod izgradnje klasifikacijskog modela. U radu je osim osnovnog modela naivnog Bayesovog klasifikatora koji u obzir uzima vjerojatnost pojave svake riječi zasebno korišten model u kojemu se promatra vjerojatnost pojave niza od dvije i niza od tri riječi. Kad se vjerojatnost računa za svaku riječ posebno govori se o unigramima tj. jednoj riječi, kad se vjerojatnost računa za dvije uzastopne riječi radi se o bigramima tj. nizu od dvije riječi i slično o trigramima za niz od tri riječi. Bigrami i trigrami posebno su dobri kod negacije u rečenici. Naivni Bayesov klasifikator neće biti sposoban prepoznati da niz "nije lijepo" ili "ne valja" sadrži negativan sentiment jer će svaku riječ gledati zasebno. Klasifikatori koji se temelje na bigramima i trigramima bit će sposobni modelirati i raditi s negacijom (pod pretpostavkom da se negacija nalazi dovoljno blizu riječi koja sadrži sentiment).

3.4.3. Stroj potpornih vektora

Stroj potpornih vektora je klasifikacijski model koji spada u skupinu modela nadziranog strojnog učenja. Postoje dvije vrste strojeva potpornih vektora, linearni i nelinearni. Za potrebe razumijevanja rada modela dovoljno se ograničiti na linearni model i problem dvije klase. Rad klasifikatora temelji se na razdvajanju uzoraka koji pripadaju različitim klasama. Svaki uzorak sadrži određen broj značajki koje ga smještaju na određeno mjesto u višedimenzionalnom prostoru. Stroj potpornih vektora pokušava odrediti granicu koja će najbolje razdvajati uzorke različitih klasa. U slučaju da se svaki uzorak može predstaviti s dvije značajke, radi se o dvodimenzionalnom prostoru, a granica među klasama bit će pravac. Ova situacija prikazana je na slici 3.2.



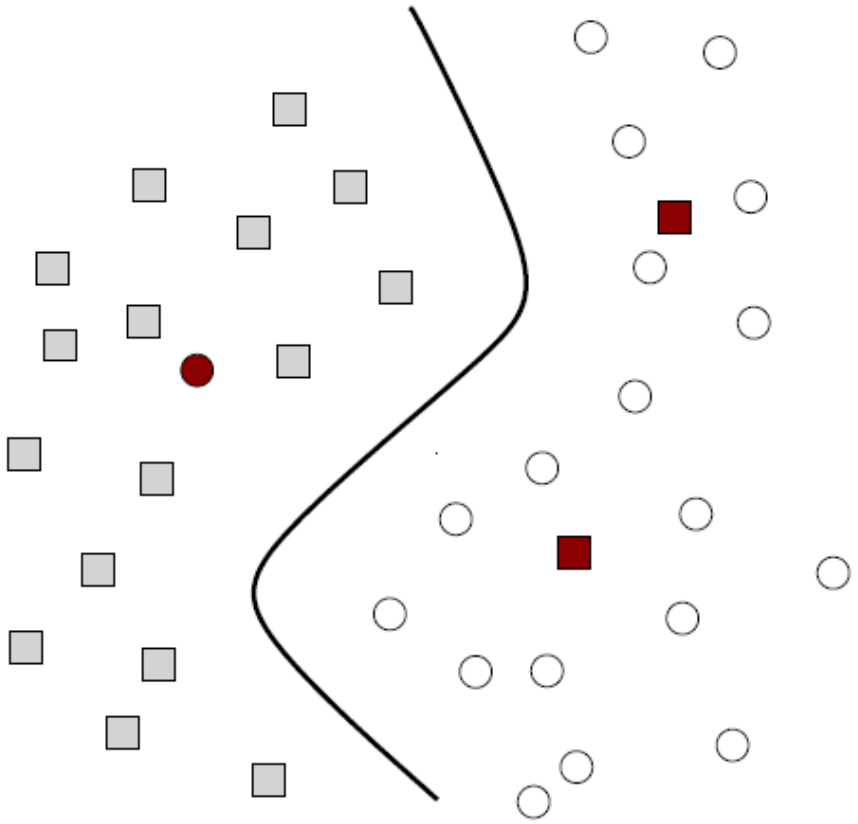
Slika 3.2: Stroj potpornih vektora za linearno odvojive klase

U ovom slučaju klase su linearno odvojive te su klase savršeno razdvojene. Međutim, među uzorcima tih klasa moguće je povući veliki broj pravaca koji bi ispravno klasificirali uzorke. Rad stroja potpornih vektora temelji se na traženju optimalne linije razdvajanja (u slučaju dvije značajke, općenito, radi se o hiper-ravnini). Optimalna linija pronalazi se maksimizacijom udaljenosti do najbližih uzoraka obje klase. Ti uzorci zovu se potporni vektori (vektori u multidimenzionalnom prostoru) te odatle i naziv stroj potpornih vektora.

Često se događa da klase ipak nisu linearno odvojive. U tom slučaju koristi se stroj potpornih vektora s jezgrenim funkcijama (engl. kernel). Kod ovakvog stroja potpornih vektora, ulazni uzorci se pomoću određenih matematičkih funkcija preslikavaju u linearni odvojive te u takvom prostoru izvršava standardni algoritam klasifikacije. Međutim kod problema klasifikacije teksta, kod kojeg se često radi o ogromnom broju ulaznih značajki, tj. svaki vektor je izuzetno velike dimenzionalnosti, primjena jezgrinih funkcija postaje prezahtjevna i najčešće nepotrebna kako su kod problema visoke dimenzionalnosti klase često linearno razdvojive. U ovom radu korišten je samo linearni stroj potpornih vektora, bez jezgrenih funkcija.

Ono što je posebno dobro kod stroja potpornih vektora je njegova prilagodljivost

na šum u podacima. Njegov rad dozvoljava greške kako bi se izbjegla pristranost modela što prikazuje slika 3.3. Bitno je naglasiti kako stroj potpornih vektora radi s numeričkim veličinama, svaka značajka ulaznog uzorka predstavljena je brojem u svojoj dimenziji. Ovakav pristup je naizgled neprikladan za rad s tekstom, međutim i za ovaj problem postoji rješenje. Radi se o vreći riječi (engl. bag of words), pristupu pretvaranja riječi u numeričke vrijednosti. Iz ulaznog skupa podataka odaberu se sve moguće različite riječi, te svaka riječ iz tog skupa dobije svoj redni broj tj. indeks. Svaki ulazni dokument tj. tvit razlaže se na riječi te se svaka riječ u dokumentu mijenja s njenim indeksom u skupu svih riječi. Na ovaj način ulazni dokument pretvara se u vektor numeričkih značajki. Ponekad se osim same pojavnosti riječi u dokumentu pohranjuje i broj pojavljivanja te riječi u dokumentu. Nedostatak ovakvog modela je izuzetno velika dimenzionalnost ulaznih vektora, međutim takvi vektori će ovisno o duljini teksta, najčešće na samo par mjesta imati zapisane brojeve, dok će sva ostala mjesta biti prazna (nule).



Slika 3.3: Stroj potpornih vektora za ulazne podatke sa šumom

3.4.4. Logistička regresija

Posljednji klasifikator korišten u ovom radu je logistička regresija. Logistička regresija je unatoč imenu, klasifikacijski model. Ime logistička regresija dolazi od toga što se taj algoritam nadograđuje na linearnu regresiju, tj. ideja je iskoristiti formulu linearne regresije za klasifikaciju. Kod linearne regresije se rezultirajuća funkcija zapisuje kao:

$$f(x) = w^T x \tag{3.3}$$

Gdje je $x = (1, x_1 ... x_n)$ vektor ulaznih značajki tj. ulazni uzorak, a $w = (w_0, w_1 ... w_n)$ je vektor koji određuje utjecaj pojedinih značajki na regresiju. Ovu formulu može se iskoristiti za procjenu vjerojatnosti ako se njena kodomena ograniči na vrijednosti između nula i jedan. Funkcija koja je prikladna kod ovog preslikavanja zove se logistička tj. sigmoidalna funkcija, a odatle i izraz logistička u imenu klasifikatora. Logistička funkcija predstavljena je formulom:

$$\sigma(\alpha) = \frac{1}{1 + e^{-\alpha}} \tag{3.4}$$

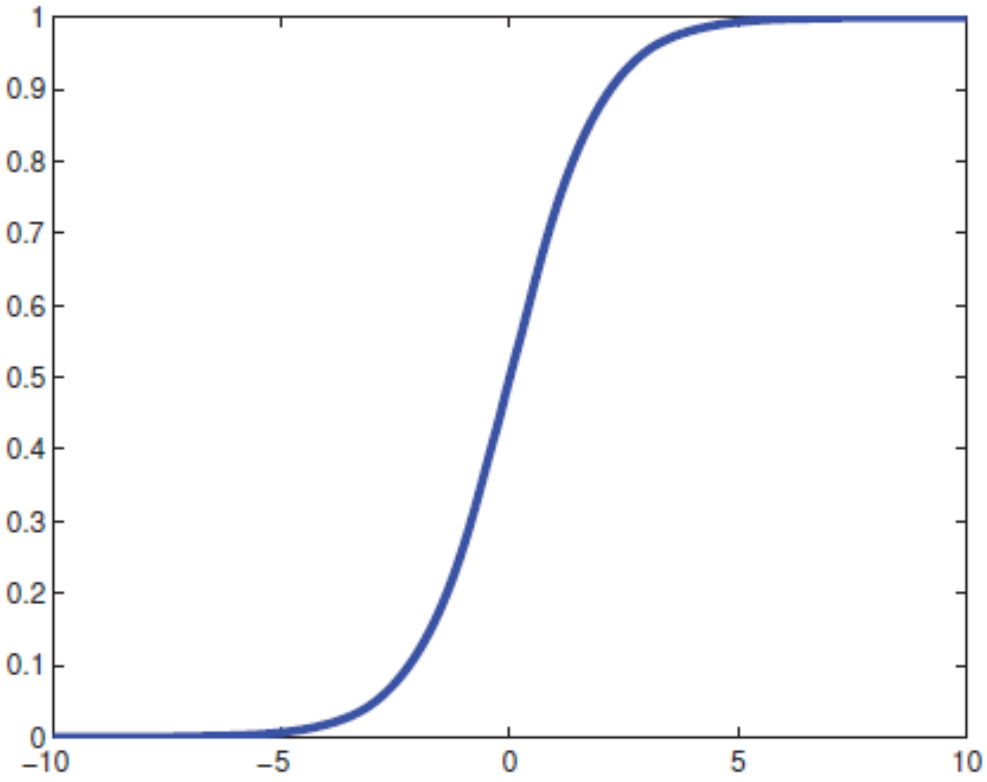
Logistička funkcija predstavljena je na slici 3.4.

Jednom kad je vrijednost izlazne funkcije ograničena na interval od nula do jedan, može se zapisati konačna formula za model logističke regresije. Ova formula vrijedi uz ograničenje na dvije klase, međutim lako se proširuje i na više klasa:

$$P(C_1|x) = \frac{1}{1 + e^{-w^T x}} \tag{3.5}$$

$$P(C_2|x) = 1 - P(C_1|x) \tag{3.6}$$

Proširenjem modela logističke regresije s dvije na više klasa dobiva se model sposoban klasificirati ulazne podatke u potrebni broj klasa. Kod logističke regresije kao i kod stroja potpornih vektora primjenjuje se model vreće riječi za stvaranje vektora iz ulaznih dokumenata tj. tvitova.



Slika 3.4: Logistička (sigmoidalna) funkcija

4. Evaluacija

Ključni dio rada je ustanoviti koliko se točno može odrediti sentiment tvitova na hrvatskom jeziku. U postupku određivanja sentimenta postoji nekoliko koraka obrade podataka kao što su redukcija broja tvitova i redukcija broja značajki. Obrađeni tvitovi koriste se kao tvitovi za učenje raznih klasifikacijskih modela poput naivnog Bayesa, stroja potpornih vektora i logističke regresije. U radu je cilj uz samu klasifikaciju tvitova procijeniti i uspješnost blago nadziranog strojnog učenja u problemu analize sentimenta. Kako bi se mogla dati procjena potrebno je rezultate klasifikatora učenih nad blago označenim podacima usporediti s rezultatima klasifikatora učenih nad ručno označenim podacima. Iz tog razloga ručno je označen primjeren skup tvitova. U ovom poglavlju će dakle biti opisani ulazni skup podataka, tj. tvitova, uspješnost metoda predobrade podataka te uspješnost pojedinih klasifikatora.

4.1. Predobrada

U ovom potpoglavlju bit će predstavljen ulazni skup tvitova tj. tekstni korpus tvitova korišten u ovom radu, izložen pregled skupova nadziranog i blago nadziranog strojnog učenja, te će se proučiti utjecaj metoda predobrade tvitova. Korpus tvitova opisan je u radu (Ljubešić et al., 2014). Korpus tvitova je sastavljen od tvitova hrvatskih, bosanskih, slovenskih i srpskih korisnika i sačinjen je od ukupno 17.5 milijuna tvitova. Tvitovi su pohranjeni u tekstualnoj datoteci u XML[1] formatu. Svaki tvit sadrži nekoliko bitnih informacija koje ga opisuju i omogućavaju da se njime lakše upravlja. Primjer jednog tvita dan je na slici 4.1.

Dakle, svaki tvit sadrži nekoliko opisnika. Ti opisnici su redom:

- id – jedinstveni identifikator tvita
- created_at – vrijeme objave tvita

[1] en.wikipedia.org/wiki/XML

```
<tweet id="394882498242838528" created_at="2013-10-28T17:44:58"
retrieved_at="2013-10-29T12:06:29.449085" favorite_count="0" retweet_count="0"
lang="hr" prob="0.980370280671" norm_length="45">
<screen_name>Bljesak</screen_name>
<text>Pobjednički povratak Marina Čilića u Parizu! http://t.co/Cykrxf7Tde</text>
</tweet>
```

Slika 4.1: Primjer tvita

- retrieved_at – vrijeme dohvaćanja (engl. download) tvita
- favourite_count - označuje koliko je korisnika tvit proglasilo omiljenim
- retweet_count – označuje koliko je korisnika ponovno objavilo tvit
- lang – upućuje na jezik kojim je pisan tvit
- prob – vjerojatnost da je tvit pisan navedenim jezikom
- norm_length - normalizirana duljina tvita (broja znakova)
- screen_name – ime korisnika koji je objavio tvit
- text – tekst tvita

Iako korpus sadrži tvitove hrvatskih, bosanskih, slovenskih i srpskih korisnika jasno je da ti korisnici mogu tvitove pisati i na drugim jezicima. Iz tog razloga prva analiza bit će analiza zastupljenosti pojedinih jezika u korpusu. Na slici 4.2 je prikazana podjela korpusa po jezicima.

Prema slici vidljivo je da je najviše tvitova pisano na hrvatskom, engleskom, slovenskom, bosanskom, srpskom i talijanskom jeziku. Ostali tvitovi pisani su na raznim, slabo zastupljenim jezicima te ih je ukupno oko 2.5 milijuna. Iako se čini da je hrvatski jezik najzastupljeniji jezik s gotovo 9 milijuna tvitova, što bi pogodovalo činjenici da se u radu analizira sentiment tvitova na hrvatskom jeziku, stvarna situacija nije takva. Uvidom u korpus vidi se da su zbog velike sličnosti hrvatskog, srpskog i bosanskog jezika, tvitovi pisani na bosanskom i srpskom jeziku redovito svrstani u tvitove na hrvatskom jeziku. Ovaj problem rješava se korištenjem datoteke s imenima hrvatskih korisnika od kojih su tvitovi prikupljeni. Datoteka sadrži imena 4500 hrvatskih korisnika i iskorištena je kako bi se filtriranjem postojećeg korpusa stvorio novi, prikladniji korpus. Nakon filtriranja na temelju korisnika novi korpus sadrži 1.7 milijuna tvitova. Primjetan je značajan pad broja tvitova u odnosu na početni korpus (veličina se smanjila na 10% ulaznog korpusa) međutim i dalje se radi o izuzetno velikom broju tvitova. Zastupljenost jezika u novonastalom korpusu prikazana je na slici 4.3.

Slika 4.2: Zastupljenost jezika u originalnom korpusu

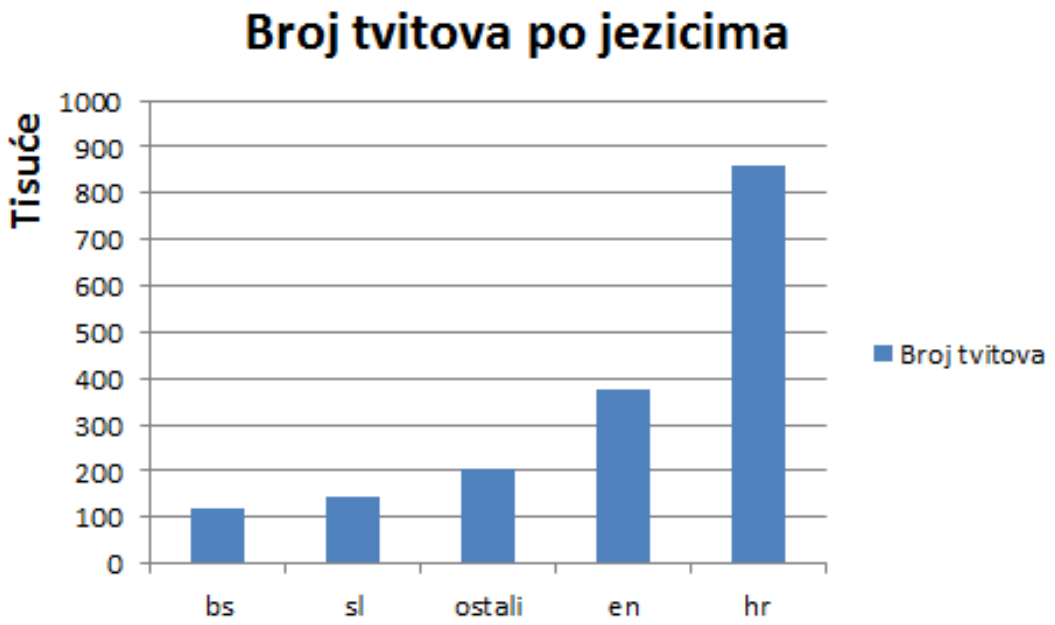
Iz slike je vidljivo da je hrvatski jezik ponovno najzastupljeniji s oko 850 tisuća tvitova, a nakon njega slijede engleski, slovenski i bosanski. Svi ostali jezici zastupljeni su s oko 200 tisuća tvitova. Početna lista emotikona pribavljena je s Wikipedije[2] a zatim je nadograđena i filtrirana na temelju empirijskih rezultata. Tablica 4.1 prikazuje zastupljenost emotikona i hashtagova u korpusima tvitova.

Tablica 4.1: Oznake sa šumom u korpusu tvitova

Korpus	Pozitivni emotikoni	Negativni emotikoni	Hashtagovi
Početni korpus	1468983	157767	2201674
Hrvatski korisnici	248968	29053	380088

Iz tablice je odmah uočljiv jedan problem, pozitivni emotikoni u puno su većoj mjeri zastupljeni od negativnih emotikona, naime ima ih gotovo 10 puta više. Jasno je da se smanjenjem broja tvitova smanjuje i broj emotikona i hashtagova međutim omjer se nije značajno promijenio.

[2] en.wikipedia.org



Slika 4.3: Zastupljenost jezika u korpusu tvitova hrvatskih korisnika

4.1.1. Redukcija broja tvitova

U ovom potpoglavlju prikazano je kako su redukcije broja tvitova utjecale na korpus tvitova hrvatskih korisnika, te kako se kroz redukcije mijenjao broj emotikona i hash-tagova. Tablica 4.2 prikazuje promjenu broja tvitova za svaku redukciju.

Tablica 4.2: Redukcija broja tvitova

Vrsta redukcije	Broj tvitova
Tvitovi hrvatskih korisnika	1710934
Redukcija na temelju jezika	861716
Redukcija retvitova	768902
Redukcija duplih tvitova	756484
Redukcija tvitova s obje vrste emotikona	755732
Redukcija na temelju jezične vjerojatnosti	370999

Iz tablice je vidljivo kako se najveće redukcije događaju kod redukcije na temelju jezika i kod redukcije na temelju jezične vjerojatnosti. Značajna redukcija je i redukcija retvitova, koja iz korpusa izbacuje gotovo sto tisuća tvitova. Redukcije duplih tvitova i tvitova s obje vrste emotikona ne utječu toliko značajno na broj tvitova koliko na kvalitetu korpusa. Vidljivo je da redukcije broja tvitova drastično smanjuju broj tvitova međutim i dalje se radi o izuzetno velikom broju tvitova nad kojim se provodi

analiza sentimenta. Uz sam broj tvitova bitno je provjeriti i što se događalo s oznakama sa šumom kroz provedene redukcije. Tablica 4.3 prikazuje kako se mijenjao broj emotikona i hashtagova u odnosu na provedene redukcije.

Tablica 4.3: Oznake sa šumom kroz redukcije

Pozitivni emotikoni	Negativni emotikoni	Hashtagovi
126432	15575	152228
118315	14384	124986
116780	14239	122962
116028	13487	122909
49973	5781	59023

Iz tablice je vidljivo da se kroz razne redukcije zadržava sličan omjer pozitivnih i negativnih emotikona, kao i omjer emotikona i hashtagova.

4.1.2. Skupovi podataka

Prije nego što se iz korpusa selektiraju tvitovi na temelju blagog označavanja potrebno je proučiti hashtagove koji se javljaju u korpusu. U tablici 4.4 je dan popis 10 najčešćih hashtagova u korpusu.

Tablica 4.4: Frekvencija pojavljivanja hashtagova

Hashtag	Frekvencija
#medvescak	387
#slavonija	396
#vladarh	401
#zagrebfacts	411
#zagreb	525
#osijek	555
#kckzg	568
#croatiaeu	628
#onokad	1302
#politikahr	1773

Iz tablice je možda teško zaključiti kakav sentiment nose određeni hashtagovi. Međutim pretraživanjem korpusa dolazi se do zaključka da su tvitovi koji sadrže navedene

hashtagove gotovo svi neutralni. Sljedećih najčešćih 20-ak hashtagova sadrži hashtagove koji upućuju na pozitivan ili negativan sentiment međutim njihova frekvencija je premala da bi se mogla uzeti u obzir tj. usporediti s brojem tvitova koje hashtagovi određuju kao neutralne. Međutim ova činjenica ne predstavlja veliki problem uzevši u obzir tvitove koje je moguće označiti pomoću emotikona. Broj tvitova po pojedinim klasama ograničen je brojem tvitova u kojima se javljaju odabrani neutralni hashtagovi. Zato se iz korpusa najprije selektiraju tvitovi koji sadrže neutralne hashtagove. Nakon toga odabiru se tvitovi koji sadrže negativne emotikone, i na kraju se među tvitovima koji sadrže pozitivne emotikone odabire isti broj tvitova jednak onome u neutralnoj i negativnoj klasi. Na ovaj način dobiva se uravnoteženi korpus tvitova, s jednakim brojem tvitova u svakoj klasi. Od preostalih tvitova odvaja se još 50 tisuća tvitova za potrebe testiranja i ručnog označavanja. Od tih 50 tisuća 10 tisuća sprema se za ručno označavanje. Konačna podjela korpusa dana je u tablici 4.5.

Tablica 4.5: Podjela korpusa na skupove tvitova

Skupovi tvitova	Broj tvitova
Neutralni tvitovi	5500
Pozitivni tvitovi	5500
Negativni tvitovi	5500
Tvitovi za ručno označavanje	10000
Tvitovi za potrebe testiranja	40000
Tvitovi za klasifikaciju	300000

Iz tablice je vidljivo da je u korpusu pronađeno 5500 neutralnih tvitova, i prema toj brojci je ograničen broj pozitivnih i negativnih tvitova (kojih inače ima tek 200 više) kako bi se dobio uravnoteženi broj tvitova po klasama. Za označavanje je odvojeno 10000 tvitova. Dio ovih tvitova bit će korišten za testiranje modela učenih nad blago označenim tvitovima, a nad ostatkom će se također izgraditi modeli strojnog učenja radi usporedbe efikasnosti nadziranog i blago nadziranog strojnog učenja. Određeni broj tvitova odvojen je za potrebe raznih testiranja, analiza itd. U zadnjem skupu tvitova nalazi se 300 tisuća tvitova. Ovi tvitovi bit će klasificirani pomoću najuspješnijeg klasifikatora i pohranjeni u bazu podataka za naknadni prikaz kroz korisničku aplikaciju.

Osam tisuća tvitova bilo je potrebno ručno označiti kako bi se stvorio primjeren ispitni skup, te primjeren skup za treniranje modela. Ovdje će ukratko biti objašnjeno

kako se uopće označavaju tvitovi. Poprilično je očito kakvi će se tvitovi svrstavati u pozitivnu i negativnu klasu, bit će to tvitovi s jasno izraženim pozitivnim mišljenjem, stavom ili emocijom za pozitivnu klasu tj. negativnim za negativnu klasu.

Primjeri negativnih tvitova:

- "Dan je tek počeo a već jedva čekam da završi -_- #diemondaydie"
- "Kiša i ružno vrijeme. Osjetno zahlađenje i još puše. Ništa od kupanja danas."
- "neponovilo se vise nikada..nažalost takva je povjest do sada bila :("

Primjeri pozitivnih tvitova:

- "Uvijek mi je drago kada pišem novo priopćenje - novi launch je u zraku :) #startup"
- "Želimo vam ugodan još jedan radni ponedjeljak. #prvasmjena"
- "Do ovog tvita - nisam znala! Zakon! Dobre stvari u životu treba ponavljati!"

Dok je poprilično jasno i lagano prepoznati pozitivne i negativne tvitove, najviše problema postoji s klasifikacijom neutralnih tvitova. Iako je i dobar dio neutralnih tvitova jednostavno prepoznati, ponekad nije toliko jasno treba li tvit stvarno svrstati u neutralnu klasu. Često se i u potpuno činjeničnom tekstu može prepoznati negativni ili pozitivni kontekst, iako je autor takvog teksta imao namjeru jednostavno prenijeti objektivno stanje stvari. Zato je pravilo u označavanju tvitova sljedeće : Ako bi se tvit ikada mogao naći na naslovnici novina, ili kao rečenica na Wikipediji tada je taj tvit neutralan. Na ovaj način će svi tvitovi koji sadrže gotovo isključivo činjenice govorili oni o npr. velikoj prirodnoj katastrofi ili pak o velikom napretku neke tvrtke biti svrstani u neutralne, osim ako u njima nije prepoznato negativno ili pozitivno mišljenje, stav ili emocija autora. Primjeri neutralnih tvitova:

- "neredi u Venezueli traju već danima, sve veći broj žrtava #venezuela "
- "liječnicima manje plaće, sud poništio Milinovićev kolektivni ugovor #politikahr #hzzo"
- "Kako građani vide članstvo u #EU? Saznajte danas na konf. Europsko građanstvo."

Dakle vidljivo je da neki tvitovi mogu imati negativnu ili pozitivnu interpretaciju u generalnom smislu ali da namjera autora teksta nije bila takva, i takvi tvitovi svrstavaju se u neutralnu kategoriju. Tijekom označavanja pojavilo se nekoliko potencijalnih problema. Prisutnost sarkazma i ironije, implicitnost onog što se govori, referenciranje na nekog korisnika ili prethodni tvit, a ponekad i loše koncipirana rečenica otežavaju klasifikaciju takvog tvita . Svi navedeni problemi unose dodatan šum u označene podatke te tako uzrokuju pogrešku kako testiranja tako i učenja modela. U tablici 4.6 prikazana je raspodjela osam tisuća označenih tvitova po klasama:

Tablica 4.6: Raspodjela ručno označenih tvitova po klasama

Neutralna klasa	Pozitivna klasa	Negativna klasa
3884	2258	1858

Dakle iz tablice je vidljivo da neutralnih tvitova ima gotovo kao negativnih i pozitivnih zajedno, međutim ne radi se o toliko velikoj razlici i ova neujednačenost ne predstavlja prevelik problem u učenju modela, a gotovo nikakav problem za testiranje kod kojeg će biti odvojeno po n tvitova svake klase za testni skup. Razdvajanje korpusa na navedene skupove bio je posljednji korak u manipulaciji brojem tvitova, međutim nije bio zadnji korak u manipulaciji nad tvitovima. U sljedećem poglavlju bit će opisan proces kojim će se pokušati smanjiti broj značajki tj. riječi u korpusu za treniranje modela.

4.1.3. Redukcija broja značajki

Tablica 4.7 prikazuje utjecaj svake pojedine redukcije i zajednički utjecaj svih redukcija na blago označenom skupu:

Iz tablice je vidljivo da su sve redukcije zajedno smanjile početni broj značajki za gotovo 40 posto. Najveći utjecaj u redukciji prema tablici ima redukcija posebnih znakova, a nakon toga slijede redukcija na mala slova i redukcija URL-ova. Brojke i postotci prikazani u tablici odnose se na direktno primjenjivanje redukcije na ulazni skup, međutim ove redukcije primjenjuju se na ulazni skup jedna za drugom kod svih redukcija, te pročišćenjem ulaznog teksta u jednoj redukciji dodatno se poboljšava utjecaj iduće redukcije i zato je ukupan postotak smanjenja značajki 39 posto umjesto 35 posto (drugi postotak dobije se ako se pomnoži postotak redukcije svih pojedinih redukcija) dakle vrijedi da je cjelina uspješnija od sume pojedinih redukcija. Tablica

Tablica 4.7: Rezultati redukcije broja značajki - blage oznake

Vrsta redukcije	Broj značajki	Dio ulaznog skupa
Početni skup	74889	100 %
Redukcija URL-ova	69334	92 %
Redukcija korisničkih imena	72306	96 %
Redukcija ponavljanja znakova	74652	99 %
Redukcija posebnih znakova	59974	80 %
Redukcija brojeva	74378	99 %
Redukcija na mala slova	69579	92 %
Sve redukcije	45412	61 %

4.8 prikazuje utjecaj redukcija na ručno označeni skup:

Tablica 4.8: Rezultati redukcije broja značajki - ručne oznake

Vrsta redukcije	Broj značajki	Dio ulaznog skupa
Početni skup	52678	100 %
Redukcija URL-ova	48698	92 %
Redukcija korisničkih imena	50855	97 %
Redukcija ponavljanja znakova	52340	99 %
Redukcija posebnih znakova	43155	82 %
Redukcija brojeva	52126	99 %
Redukcija na mala slova	49151	93 %
Sve redukcije	33573	64 %

Iz tablice je primjetno da su postotci redukcija podjednaki kao i kod blago označenog skupa, razlika je u svega 3 posto. Dominantna redukcija je i ovdje redukcija posebnih znakova, a slijede ju redukcija URL-ova i redukcija na mala slova. Jedina redukcija koja ovdje nije spomenuta je redukcija hrvatskih stop-riječi (engl. stopwords). Stop-riječi su česte riječi u jeziku koje su bitne za strukturu rečenica međutim same po sebi ne nose nikakvu bitnu informaciju. U hrvatskom su to riječi kao: "i, pa, ili, nego, biti, ćemo", itd. Ove riječi uklanjanju se pri postupku klasifikacije. Datoteka s hrvatskim stop riječima sadrži ukupno 2024 riječi. Upravo je ta brojka dobra procjena za smanjenje broja značajki s obzirom na to da tekstni korpusi korišteni u ovom radu sadrže gotovo sve takve riječi.

4.2. Implementacija i vrednovanje klasifikatora

U ovom poglavlju izložen je kratki osvrt na korištene implementacije klasifikatora strojnog učenja i odgovarajuće vrednovanje tih klasifikatora. Kao okruženje za izgradnju i testiranje klasifikatora korišten je sustav Weka. [3] Weka (engl. *Waikato Environment for Knowledge Analysis*) je zapravo skup implementacija velikog broja algoritama strojnog učenja koja omogućava brzu i jednostavnu obradu podataka, razvijena na sveučilištu Waikato na Novom Zelandu. Također, Weka je slobodan softver, dostupan unutar GNU General Public[4] licence. Sustav se može pokretati na dva načina, kao grafičko sučelje i kao zasebna konzola. Kod pokretanja grafičkog sučelja, učitavanje modela i pokretanje klasifikacije je sporije i memorijski zahtjevnije te je u ovom radu Weka korištena kroz pozive u konzoli. Svaka klasifikacija podataka u Weki može se razložiti u nekoliko ključnih koraka i svaki korak sadrži pripadajući Weka model podataka za procesiranje. Ključni koraci bitni za proces klasifikacije su:

- Stvaranje modela ARFF za trening i za testiranje
- Pretvaranje modela iz tekstnog oblika u oblik vreće riječi
- Treniranje klasifikatora nad ulaznim podacima
- Testiranje modela kros-validacijom
- Testiranje modela nad testnim podacima

Prvi korak u radu s Wekom je stvaranje modela AREF za trening i testiranje. U prethodnom poglavlju su opisani skupovi podataka i jasno je da se ručno označene tvitove i blago označene tvitove dijele svaki na skup za trening i skupa za testiranje. Kako bi se započeo rad s Wekom, potrebno je najprije dovesti ulazne skupove podataka u prikladan oblik. Svi tvitovi svake pojedine klase moraju se zapisati u jednu tekstnu datoteku. Tako se za svaki skup tvitova stvara direktorij s tri datoteke, od kojih svaka sadrži sve tvitove jedne od tri klase. Weka ovakav direktorij pretvara u model ARFF, interni model podataka prilagođen strukturom za bržu obradu i manju memorijsku zahtjevnost kod izvođenja klasifikacijskih zadataka. Zapravo, model ARFF je obična tekstna datoteka, koja sadrži strukturirane i prikladno označene ulazne podatke.

Nakon što je stvoren model ARFF za sve skupove tvitova, potrebno je taj model pretvoriti u model vreće riječi. U ovom koraku može se odabrati veliki broj opcija ko-

[3] www.cs.waikato.ac.nz/ml/weka

[4] www.gnu.org/

ristan za različite vrste klasifikacija. Može se odabrati da li u vreći riječi uz svaku riječ koristiti samo njenu pojavnost ili frekvenciju, na koji način razdvajati ulazne rečenice na riječi, izbacivati stop riječi, koristiti unigrame, bigrame ili trigrame i još mnoge druge. Cilj je pronaći najbolju kombinaciju za trenutni klasifikator tj. kombinaciju za koju će klasifikator najtočnije klasificirati tvitove.

Nakon što su sve metode obrade ulaznog modela provedene i stvoren je željeni model vreće riječi može se započeti s klasifikacijom. Weka nudi veliki broj različitih klasifikatora, a u ovom radu testirane su tri implementacije, naivni Bayesov klasifikator, SVM i logistička regresija. Prije klasifikacije potrebno je (kod nekih klasifikatora) podesiti određene parametre bitne za svaki model. Nakon pokretanja procesa treninga klasifikatora Weka stvara interni model u koji se pohranjuje dobiveni klasifikator. Tako dobiveni model može se koristiti više puta za testiranje, bez potrebe za ponovnim treniranjem.

Nakon što je klasifikacijski model stvoren u prošlom koraku vrijeme je za testiranje. Prva vrsta testiranja je testiranje k-strukom unakrsnom validacijom. Kod ovakvog testiranja ulazni skup podataka dijeli se u dva skupa. Jedan od tih skupova predstavljat će skup za trening a drugi skup za testiranje. Nad ovakvim podacima provodi se klasifikacija te se ocjenjuje uspješnost modela. Ovaj postupak podijele ulaznog skupa i testiranja ponavlja se k puta i na kraju se procjenjuje pogreška klasifikatora kao prosječna pogreška k izvođenja. Ovakav pristup daje vrlo realnu procjenu mogućnosti i uspješnosti testiranog klasifikatora.

Posljednji korak je testiranje modela nad testnim tj. ručno označenim podacima. Na model stvoren u trećem koraku dovode se ručno označeni tvitovi, te se za svaki tvit provjerava ispravnost klasifikacije. Bitno je naglasiti da su svi podaci za testiranje pretvoreni u prikladan oblik vreće riječi s obzirom na podatke za trening. Kako je moguće da testni tvitovi sadrže riječi koje se ne pojavljuju u tvitovima za trening, ovakvi slučajevi uzrokovali bi probleme kod klasifikacije. Međutim, Weka nudi opciju izgradnje modela vreće riječi za testni skup u odnosu na skup za treniranje, te se može eksplicitno odrediti što napraviti s takvim riječima, da li ih odbaciti ili im dodijeliti određenu vrijednost.

Svaki postupak testiranja klasifikatora prikazuje razne vrijednosti uspješnosti klasifikacije. Osnovni podaci o uspješnosti klasifikatora su broj točno klasificiranih tvitova, dakle broj tvitova koje je klasifikator svrstao u klasu kojoj stvarno pripada i broj netočno klasificiranih tvitova koji su svrstani u klasu kojoj ne pripadaju. Svako testiranje donosi i matricu konfuzije. Matrica zabune u redcima sadrži stvarne klase podataka a u stupcima klase koje je odredio klasifikator. Tako se za svaku klasu može vidjeti koliki

je broj točno klasificiranih primjera za tu klasu, te u koje je druge klase klasifikator krivo svrstao određene primjere. Za svaki klasifikator u ovom radu bit će prikazana matrica zabune kako bi se moglo detaljnije komentirati rezultate. Mjera uspješnosti klasifikatora koja se koristi u ovom radu je točnost. Točnost se određuje kao omjer točno klasificiranih primjera i svih testiranih primjera. Ovo je jednostavna mjera međutim daje poprilično dobar uvid u rad klasifikatora. Iako se u drugim radovima koriste i druge mjere poput preciznosti, odziva i f-mjere, bitno je napomenuti da te mjere za problem tri i više klasa imaju istu vrijednost i ne samo to nego imaju isto vrijednost kao i točnost. Iz tog razloga u radu će biti uz svaki klasifikator izložena matrica konfuzije, broj točno i netočno klasificiranih primjera te točnost klasifikatora.

4.3. Uspješnost klasifikatora

U ovom poglavlju bit će izneseni rezultati klasifikacije za sve vrste klasifikatora korištenih u radu. Rezultati koji će biti ispitani su rezultat klasifikacije za k-struku unakrsnu validaciju i rezultat klasifikacije testnih tvitova na klasifikatoru treniran na blago označenim podacima te rezultat k-struke unakrsne validacije za klasifikator treniran na ručno označenim podacima. Jasno je da u slučaju klasifikatora treniranog na ručnim podacima nema smisla izdvajati skup za testiranje jer je taj korak zapravo dio k-struke unakrsne validacije. Cilj k-struke kros validacije klasifikatora treniranih na blago označenim tvitovima bit će ustanoviti sposobnost klasifikatora da se prilagodi ulaznim podacima. Kod testiranja tog klasifikatora na ručno označenim primjerima cilj će biti ustanoviti sposobnost klasifikatora da se prilagodi neviđenim podacima. K-struka validacija klasifikatora treniranog na ručno označenim podacima objedinjuje navedene dvije stavke ali uz to pruža mogućnost usporedbe modela nadziranog i blago nadziranog strojnog učenja.

4.3.1. Klasifikator temeljen na rječniku

Prvi rezultati klasifikacije prikazani su za klasifikator temeljen na rječniku. Oznaka "S" (stvarno) u tablici stoji za stvarnu klasu tvita, a oznaka "K" (klasificirano) za klasu tvita koju je klasificirao algoritam. Tablica 4.9 prikazuje rezultate klasifikacije za klasifikator treniran na blago označenom skupu tvitova. Klasifikator je testiran na tri tisuće ručno označenih tvitova.

Tablica 4.9: Klasifikator temeljen na rječniku - blage oznake

S \ K	Negativno	Neutralno	Pozitivno
Negativno	529	247	224
Pozitivno	575	310	115
Neutralno	713	192	94
Broj točnih	Broj netočnih	Točnost	Netočnost
1247	1753	41.58 %	58.42 %

Tablica 4.10 prikazuje rezultate klasifikacije za klasifikator treniran na ručno označenim tvitovima. Klasifikator je testiran na tisuću petsto ručno označenih tvitova.

Tablica 4.10: Klasifikator temeljen na rječniku - ručne oznake

S \ K	Negativno	Neutralno	Pozitivno
Negativno	452	22	26
Neutralno	476	14	10
Pozitivno	452	22	26
Broj točnih	Broj netočnih	Točnost	Netočnost
539	961	35.93 %	64.07 %

Rezultati klasifikacije pomalo su iznenađujući. Klasifikator treniran na blago označenim primjerima uspješniji je u klasifikaciji od onog treniranog na ručno označenim podacima. Međutim postoji objašnjenje za takve rezultate a ono se krije u činjenici da ručno označeni skup nije uravnotežen i u njemu je najviše neutralnih tvitova. Zato se velik broj riječi koje su dobar predstavnik pozitivne i negativne klase nalazi u neutralnom rječniku. Ovo se vidi prema stanju u tablici u kojoj je većina tvitova klasificirana u neutralnu klasu. Ovo činjenica upućuje na glavni nedostatak ovog jednostavno klasifikatora a to je nemogućnost skaliranja ulaznih podataka, tj. prilagodbe na neuravnoteženi ulazni skup tvitova. Međutim ono što zadovoljava kod ovog klasifikatora je činjenica da su rezultati klasifikacije, iako loši, bolji od rezultata nasumične klasifikacije i to za 10 posto u slučaju blago označenih tvitova. Međutim očito je da kvaliteta rezultata nije zadovoljavajuća i potrebno je ustanoviti mogu li složeniji klasifikatori postići bolje rezultate.

4.3.2. Naivni Bayesov klasifikator

U ovom potpoglavlju prikazani su rezultati klasifikacije za naivni Bayesov klasifikator. Rezultati su prikazani redom za klasifikator temeljen na unigramima, bigramima i trigramima. Prvo će biti prikazni rezultati za klasifikator temeljen na unigramima. U tablici 4.11 prikazani su rezultati k-struke unakrsne validacije za klasifikator temeljen na unigramima i treniran na blago označenim tvitovima.

Tablica 4.11: K-struka unakrsna validacija za NB unigram klasifikator - blage oznake

S \K	Negativno	Neutralno	Pozitivno
Negativno	3444	379	1677
Neutralno	216	4962	322
Pozitivno	1680	578	3242
Broj točnih	Broj netočnih	Točnost	Netočnost
11648	4852	70.59 %	29.41 %

U tablici 4.12 prikazani su rezultati testiranja na testnom skupu za klasifikator temeljen na unigramima i treniran na blago označenim tvitovima.

Tablica 4.12: Rezultati na testnom skupu za NB unigram klasifikator - blage oznake

S \K	Negativno	Neutralno	Pozitivno
Negativno	501	212	287
Neutralno	234	405	360
Pozitivno	261	230	509
Broj točnih	Broj netočnih	Točnost	Netočnost
1415	1585	47.18 %	52.82 %

U tablici 4.13 prikazani su rezultati k-struke unakrsne validacije za klasifikator temeljen na unigramima i treniran na ručno označenim tvitovima.

Iz rezultata prikazanih je vidljivo da se naivni Bayesov klasifikator uspijeva prilagoditi blago označenim tvitovima i postiže poprilično dobre rezultate kod k-struke unakrsne validacije samo na tom skupu. Veliki problem nastaje kada se na klasifikator dovedu ručno označeni tvitovi. U ovom slučaju točnost klasifikacije pada na ispod pedeset posto, što nije rezultat koji zadovoljava. Problem je u najvećoj mjeri uzrokovan klasifikacijom tvitova koji pripadaju u neutralnu klasu. Ti tvitovi se klasificiraju u sve tri klase s gotovo podjednakom zastupljenošću. Međutim uvidom u rezultate k-struke

Tablica 4.13: K-struka unakrsna validacija za NB unigram klasifikator - ručne oznake

S \K	Negativno	Neutralno	Pozitivno
Negativno	702	678	478
Neutralno	550	2437	897
Pozitivno	255	869	1134
Broj točnih	Broj netočnih	Točnost	Netočnost
4273	3727	53.41 %	46.59 &

unakrsne validacije klasifikatora treniranog na ručno označenim podacima postaje vjerojatnije da uzrok ovim slabijim rezultatima klasifikacije leži u samim tvitovima. Takav klasifikator ne uspijeva se prilagoditi podacima i griješi najviše tako da klasificira pozitivne i negativne tvitove u neutralnu klasu. Ovaj problem bit će nešto detaljnije razrađen u usporedbi rezultata klasifikatora.

Nakon unigrama, prikazuju se rezultati za klasifikator temeljen na bigramima. U tablici 4.14 prikazani su rezultati k-struke unakrsne validacije za klasifikator temeljen na bigramima i treniran na blago označenim tvitovima.

Tablica 4.14: K-struka unakrsna validacija za NB bigram klasifikator - blage oznake

S \K	Negativno	Neutralno	Pozitivno
Negativno	3648	337	1515
Neutralno	277	4874	349
Pozitivno	1798	540	3162
Broj točnih	Broj netočnih	Točnost	Netočnost
11684	4816	70.81 %	29.19 %

U tablici 4.15 prikazani su rezultati testiranja na testnom skupu za klasifikator temeljen na bigramima i treniran na blago označenim tvitovima.

Tablica 4.15: Rezultati na testnom skupu za NB bigram klasifikator - blage oznake

S \K	Negativno	Neutralno	Pozitivno
Negativno	524	212	264
Neutralno	275	415	309
Pozitivno	285	227	488
Broj točnih	Broj netočnih	Točnost	Netočnost
1427	1573	47.58 %	52.42 %

U tablici 4.16 prikazani su rezultati k-struke unakrsne validacije za klasifikator temeljen na bigramima i treniran na ručno označenim tvitovima.

Tablica 4.16: K-struka unakrsna validacija za NB bigram klasifikator - ručne oznake

S \K	Negativno	Neutralno	Pozitivno
Negativno	907	378	573
Neutralno	870	1800	1214
Pozitivno	384	543	1331
Broj točnih	Broj netočnih	Točnost	Netočnost
4038	3962	50.47 %	49.53 %

Rezultati klasifikatora temeljenog na bigramima pokazuju maleno poboljšanje u odnosu na klasifikator temeljen na unigramima (oko 1 posto veća točnost). Kao i kod modela unigrama problem kod k-struke unakrsne validacije klasifikatora treniranog na blagom skupu problem se javlja kod klasifikacije pozitivnih i negativnih tvitova koji se često klasificiraju u suprotnu klasu. Kod testiranja klasifikatora na testnom skupu najveću pogrešku ponovno unosi raspršenost tvitova neutralne klase. Rezultati klasifikatora treniranog na ručnom skupu lošiji su za oko tri posto nego što je to slučaj kod unigrama i čini se da povećanje složenosti modela u slučaju ručno označenih podataka smanjuje točnost klasifikacije.

Posljednji klasifikator koji je ispitan je klasifikator temeljen na trigramima. U tablici 4.17 prikazani su rezultati k-struke unakrsne validacije za klasifikator temeljen na trigramima i treniran na blago označenim tvitovima.

Tablica 4.17: K-struka unakrsna validacija za NB trigram klasifikator - blage oznake

S \K	Negativno	Neutralno	Pozitivno
Negativno	3571	342	1587
Neutralno	278	4864	358
Pozitivno	1739	562	3199
Broj točnih	Broj netočnih	Točnost	Netočnost
11634	4866	70.5 %	29.5 %

U tablici 4.18 prikazani su rezultati testiranja na testnom skupu za klasifikator temeljen na trigramima i treniran na blago označenim tvitovima.

Tablica 4.18: Rezultati na testnom skupu za NB trigram klasifikator - blage oznake

S \K	Negativno	Neutralno	Pozitivno
Negativno	520	215	265
Neutralno	264	426	309
Pozitivno	274	239	487
Broj točnih	Broj netočnih	Točnost	Netočnost
1433	1567	47.78 %	52.22 %

U tablici 4.19 prikazani su rezultati k-struke unakrsne validacije za klasifikator temeljen na trigramima i treniran na ručno označenim tvitovima.

Tablica 4.19: K-struka unakrsna validacija za NB trigram klasifikator - ručne oznake

S \K	Negativno	Neutralno	Pozitivno
Negativno	1039	184	635
Neutralno	1122	1202	1560
Pozitivno	475	277	1506
Broj točnih	Broj netočnih	Točnost	Netočnost
3747	4253	46.83 % 53.17 %	

Iz rezultata klasifikatora temeljenog na trigramima vidljivo je da su u svakom od tri prikazana slučaja rezultati lošiji ne samo od rezultata klasifikatora temeljenog na bigramima nego i od onoga temeljenog na unigramima. Čini se da ovoliko povećanje složenosti modela ipak ne donosi željeno poboljšanje u smislu točnosti klasifikatora. Međutim ovakvi rezultati dobiveni su i u srodnim radovima gdje se također pokazalo da su unigrami ili bigrami najpogodniji za naivni Bayesov klasifikator u smislu složenosti modela.

4.3.3. Stroj potpornih vektora

U ovom potpoglavlju bit će prikazani rezultati klasifikacije za stroj potpornih vektora. Kao što je navedeno u poglavlju o klasifikacijskim modelima koristi se linearni stroj potpornih vektora. Zbog memorijske i procesorske zahtjevnosti izrađeni su jedino klasifikatori temeljeni na unigramima. Složenije modele nije bilo moguće pokrenuti na jednom računalu. U tablici 4.20 prikazani su rezultati k-struke unakrsne validacije za stroj potpornih vektora treniran na blago označenim tvitovima.

Tablica 4.20: K-struka unakrsna validacija za SVM - blage oznake

S \K	Negativno	Neutralno	Pozitivno
Negativno	3587	154	1759
Neutralno	376	4611	513
Pozitivno	1766	258	3476
Broj točnih	Broj netočnih	Točnost	Netočnost
11674	4826	70.75 %	29.25 %

U tablici 4.21 prikazani su rezultati testiranja na testnom skupu za stroj potpornih vektora treniran na blago označenim tvitovima.

Tablica 4.21: Rezultati na testnom skupu za SVM - blage oznake

S \K	Negativno	Neutralno	Pozitivno
Negativno	577	107	316
Neutralno	343	192	464
Pozitivno	336	104	560
Broj točnih	Broj netočnih	Točnost	Netočnost
1329	1670	44.31 %	55.69 %

U tablici 4.22 prikazani su rezultati k-struke unakrsne validacije za stroj potpornih vektora treniran na ručno označenim tvitovima.

Tablica 4.22: K-struka unakrsna validacija za stroj potpornih vektora - ručne oznake

S \K	Negativno	Neutralno	Pozitivno
Negativno	680	812	366
Neutralno	520	2561	803
Pozitivno	291	931	1036
Broj točnih	Broj netočnih	Točnost	Netočnost
4277	3723	53.46 %	46.54 %

Iz rezultata je vidljivo da se stroj potpornih vektora odlično prilagođava blagim podacima, čak i malo bolje nego naivni Bayesov klasifikator. Međutim testiranje na ručno označenim primjerima ponovno donosi izrazito loše rezultate. Za razliku od naivnog Bayesovog klasifikatora koji je griješio najviše na neutralnoj klasi, kod stroja potpornih vektora uz taj problem neutralne klase postoji i problem klasifikacije pozitivnih i negativnih tvitova u nasuprotne klase. Što se tiče klasifikatora treniranog nad

ručno označenim primjerima, rezultati se ne razlikuju previše od onih dobivenih kod naivnog Bayesa.

4.3.4. Logistička regresija

Posljednji klasifikator čiji se rezultati prikazuju u ovom potpoglavlju je logistička regresija. Logistička regresija je kao i stroj potpornih vektora u korištenoj implementaciji imala veće memorijske i procesorske zahtjeve od naivnog Bayesovog klasifikatora, te je zato treniranje provedeno samo na unigramima. U tablici 4.23 prikazani su rezultati k-struke unakrsne validacije za logističku regresiju treniranu na blago označenim tvitovima.

Tablica 4.23: K-struka unakrsna validacija za logističku regresiju - blage oznake

S \K	Negativno	Neutralno	Pozitivno
Negativno	3507	204	1789
Neutralno	406	4541	543
Pozitivno	1836	288	3376
Broj točnih	Broj netočnih	Točnost	Netočnost
11424	5076	69,23 %	30,77 %

U tablici 4.24 prikazani su rezultati testiranja na testnom skupu za logističku regresiju treniranu na blago označenim tvitovima.

Tablica 4.24: Rezultati na testnom skupu za logističku regresiju - blage oznake

S \K	Negativno	Neutralno	Pozitivno
Negativno	547	117	336
Neutralno	353	172	474
Pozitivno	346	134	520
Broj točnih	Broj netočnih	Točnost	Netočnost
1239	1761	41.3 %	58.7 %

U tablici 4.25 prikazani su rezultati k-struke unakrsne validacije za logističku regresiju treniranu na ručno označenim tvitovima.

Tablica 4.25: K-struka unakrsna validacija za logističku regresiju - ručne oznake

S \K	Negativno	Neutralno	Pozitivno
Negativno	610	852	396
Neutralno	530	2521	833
Pozitivno	321	951	986
Broj točnih	Broj netočnih	Točnost	Netočnost
4117	3883	51.4 %	48.6 %

Iz tablica je vidljivo da su u slučaju unakrsne validacije modela logističke regresije treniranog na blago označenim podacima rezultati nešto lošiji nego kod naivnog Bayesovog klasifikatora i stroja potpornih vektora. Međutim drastično lošiji rezultati su izraženi kod testiranja modela na ručno označenim podacima. U tom slučaju točnost jedva prelazi onu klasifikatora temeljenog na rječniku. Glavni uzrok tome je izuzetno loša klasifikacija neutralnih tvitova. Konačno, točnost modela treniranog nad ručno označenim tvitovima je nešto lošija od točnosti ostalih modela, također zbog svrstavanja pozitivnih i negativnih tvitova u neutralnu klasu.

4.4. Usporedba rezultata klasifikatora

U ovom poglavlju bit će ukratko iznesen osvrt na rezultate klasifikacije različitih klasifikacijskih modela. Klasifikator temeljen na rječniku poslužio je za definiranje donjeg praga točnosti klasifikacije i cilj je bio korištenjem složenijih klasifikatora poboljšati uspješnost klasifikacije. Ključni problem klasifikatora temeljenog na rječniku bila je nemogućnost prilagodbe na neuravnoteženi skup tvitova. Ovaj problem su svi složeniji klasifikatori korišteni u radu uspješno savladali kao što su pokazali rezultati k-struke unakrsne validacije modela treniranih na blago označenim tvitovima. Iz tih se rezultata moglo zaključiti da su klasifikatori sposobni poprilično dobro prilagoditi se ulaznom skupu tvitova. Točnost se kod svih klasifikatora u ovom slučaju kretala oko sedamdeset posto. Problem je predstavljala klasifikacija pozitivnih i negativnih tvitova koji su završavali u suprotnoj klasi od one kojoj stvarno pripadaju. Najveću točnost za ovakvu klasifikaciju ostvario je model naivnog Bayesovog klasifikatora temeljenog na bigramima. Ključan problem pojavio se kad su se prethodno istrenirani klasifikatori susreli s ručno označenim tvitovima. Točnost klasifikacije se u ovom slučaju kretala oko četrdeset sedam posto, što je tek pet posto bolji rezultat od onog postignutog u jednostavnom klasifikatoru temeljenom na rječniku. Uvid u matrice zabune otkrio

je dva ključna problema kod klasifikacije ručno označenih tvitova. Prvi problem je klasifikacija neutralnih tvitova kod koje su tvitovi često klasificirani kao pozitivni ili negativni. Do ovog problema dolazi iz više razloga. Prvi razlog je taj da su blago označeni tvitovi koji pripadaju neutralnoj klasi označeni na temelju desetak "neutralnih" hashtagova. Ti tvitovi su zapravo najčešće tvitovi novinskih kuća, web-portala, sportske vijesti i sl. Vrlo rijetko su tvitovi koji sadrže neutralne hashtagove oni koje su objavili prosječni tj. nekomercijalni korisnici. Iz tog razloga skup neutralnih tvitova je zapravo ograničen tematski i vokabularom. S druge strane ručno označeni tvitovi su najčešće oni osobnih korisnika, koji vrlo često sadrže neformalni način izražavanja. Drugi razlog zbog kojeg je otežana klasifikacija neutralne klase je subjektivnost neutralne klase. U neutralnu klasu se svrstavaju tvitovi koji sadrže obje vrste sentimenta, tvitovi koji sadrže sarkazam, tvitovi koji sadrže riječi koje nose pozitivan ili negativan sentiment ali su zapravo objektivni i ne sadržavaju u sebi stav ili mišljenje. Sve navedene činjenice uzrokuju "raspršenost" tvitova neutralne klase po svim ostalim klasama, tj. navode klasifikator da na temelju riječi koje sadrže jak sentiment svrstaju neutralni tvit u pozitivnu ili negativnu klasu. Drugi problem s kojim se susreće kod testiranja klasifikatora je klasifikacija negativnih i pozitivnih tvitova u suprotne klase od onih kojima stvarno pripadaju. Do ovoga dolazi zbog korištenja sarkazma i ironije u tvitovima gdje najčešće negativni tvitovi sadrže riječi s pozitivnim sentimentom i obrnuto te zbog korištenja negacije u rečenicama koje se ne uspijeva modelirati s jednostavnijim modelima (iako je model bigrama donio određeno poboljšanje). Do sada se moglo zaključiti da se klasifikatori uspješno prilagođavaju ulaznim skupovima tvitova ali da imaju problema s klasifikacijom prije neviđenih, ručno označenih tvitova. Međutim do sad dobiveni rezultati nisu dovoljni da se procjeni da li problem s klasifikacijom uzrokuje korištenje neprikladnih blagih oznaka ili pak složenost i neformalnost ulaznog skupa tvitova. Kako bi se utvrdila kvaliteta korištenja blago nadziranog strojnog učenja bilo je potrebno ručno označiti dovoljan broj tvitova kako bi se klasifikatori mogli trenirati nad takvim tvitovima te bi se tako dobila mogućnost usporedbe dva modela strojnog učenja. Točnost postignuta kod klasifikatora treniranih nad ručno označenim tvitovima nije previše odudarala od točnosti postignute u testiranju klasifikatora treniranih na blago označenim tvitovima. I kod takve klasifikacije javili su se prije navedeni problemi. Drugim riječima ostvareni rezultati upućuju na generalni problem klasifikacije korisničkih tvitova tj. na složenost ulaznih podataka. S obzirom na sličnu točnost postignutu kod nadziranog i blago nadziranog strojnog učenja može se zaključiti da su blage oznake dobar način za označavanje velikog broja ulaznih podataka. Ovo se pogotovo odnosi na emotikone dok s druge strane primjena hashtagova kao oznake sa

šumom još uvijek nema maksimalnu iskoristivost najviše zbog nezastupljenosti hashtagova u tvitovima. Hashtagovi su nešto što još uvijek dobiva na popularnosti i tek polako ulazi u opću primjenu kod tvitova, barem kad se radi o tvitovima hrvatskih korisnika.

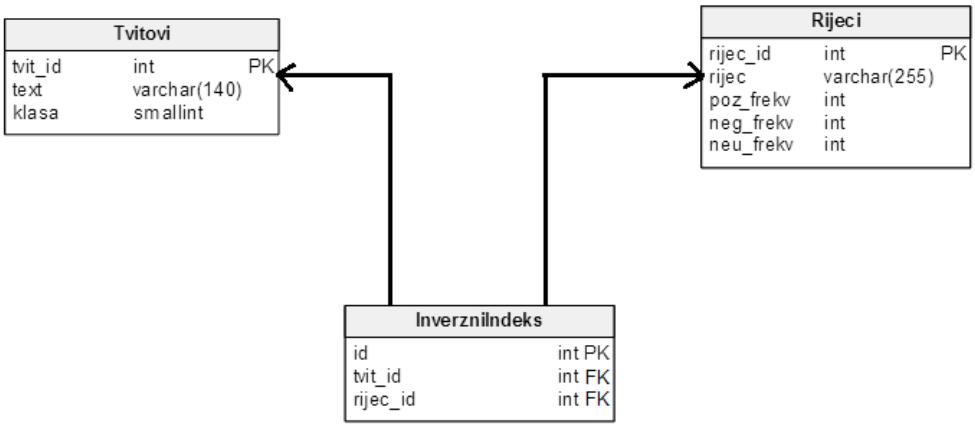
5. Agregacija sentimenta

Kako bi se izbjegla samo suhoparna predstavljanja rezultata analize sentimenta u tvitovima potrebno je te rezultate predstaviti na razumljiviji i intuitivniji način. S time u vidu razvijena je jednostavna aplikacija koja omogućuje unos upita u prirodnom jeziku o raznim temama i područjima interesa, a kao rezultat u grafičkom obliku predstavlja statistike unesenom upitu. Drugim riječima agregacija sentimenta predstavlja ukupni sentiment vezan uz neki entitet, pojam i sl. Tako će npr. agregacija za korisnički upit "sport" sadržavati statistiku o pojavnosti riječi "sport" u pozitivnoj, negativnoj i neutralnoj klasi. Za složenije upite ta će se statistika izračunavati na nešto složeniji način ali će u konačnici opet predstavljati istu stvar - zastupljenost upita po klasama. U sljedećim potpoglavljima opisani su ključni dijelovi i prikazan je izgled i način upotrebe aplikacije.

5.1. Model baze podataka

Osnova korisničke aplikacije je baza podataka koja sadrži sve podatke bitne za prikazivanje korisniku. Baza je izgrađena nad tristo tisuća tvitova koji su označeni putem najuspješnijeg klasifikatora (naivni Bayesov klasifikator sa bigramima). Baza podataka za ovu aplikaciju je relativno jednostavna i sastoji se od samo tri tablice. Prva tablica je tablica "Tvitovi", u toj tablici je pohranjeno tristo tisuća obrađenih tvitova. Svaki redak u toj tablici sadrži jedinstveni identifikator, tekst tvita i oznaku klase kojoj tvit pripada. Iduća tablica je tablica "Riječi" u kojoj su pohranjene sve različite riječi koje su se pojavile u navedenih tristo tisuća tvitova. Svaki redak u ovoj tablici sadržava jedinstveni identifikator riječi, samu riječ, te frekvenciju pojavljivanja riječi u pozitivnoj, negativnoj i neutralnoj klasi. Posljednja je tablica "InverzniIndeks". Svaki redak u ovoj tablici sadrži jedinstveni identifikator, jedinstveni identifikator riječi, te polje jedinstvenih identifikatora tvitova u kojima se ova riječ javlja. Naziv inverzni indeks dolazi upravo od imena metode koja se primjenjuje u ovom slučaju. Kako se korisnički upiti unose u prirodnom jeziku, za očekivati je da će oni tipično sadržavati

nekoliko riječi. Ideja će biti za te riječi pronaći sve tvitove u kojima se javljaju i provjeriti postoji li presjek za skup tvitova svake od tih riječi. Zato se umjesto pohranjivanja svih indeksa riječi koje sadrži određeni tvit, pohranjuju svi indeksi tvitova u kojima se javlja određena riječ, te odatle i naziv inverzni indeks. Slika 5.1 prikazuje model baze podataka.



Slika 5.1: Model baze podataka

5.2. Rad aplikacije

U prošlom potpoglavlju spomenut je pojam inverznog indeksa. Na toj ideji temelji se rad aplikacije. Rad aplikacije započinje u trenutku kad korisnik unese upit za koji želi prikaz agregiranog sentimenta. Ovaj upit prolazi kroz sve metode redukcije značajki navedene u prethodnim poglavljima rada. Nakon svih redukcija početni upit je transformiran u listu značajki tj. listu riječi. Za svaku riječ traži se njen jedinstveni identifikator u tablici "Riječi". Za svaku pronađenu riječ, na temelju tablice "InverzniIndeks" koja sadrži indekse svih tvitova u kojima se riječ javlja, dohvaćaju se tvitovi iz tablice "Tvitovi". U ovom trenutku za svaku riječ postoji lista tvitova u kojima se ta riječ javlja, te se pokušava pronaći presjek svih lista tvitova. Ako takav presjek ne postoji pokušava se tražiti presjek svih podskupova riječi dok se ne pronađe presjek koji postoji ili se dostigne razina samih riječi. Rezultat ove radnje je najveći mogući presjek lista tvitova ulaznog skupa. Ako takav presjek ne postoji, rezultat će biti najveći mogući presjek jednog od podskupova, te konačno ako niti takav presjek ne postoji, rezultat radnje bit će ulazne riječi. Ovisno o prethodnom rezultatu korisniku

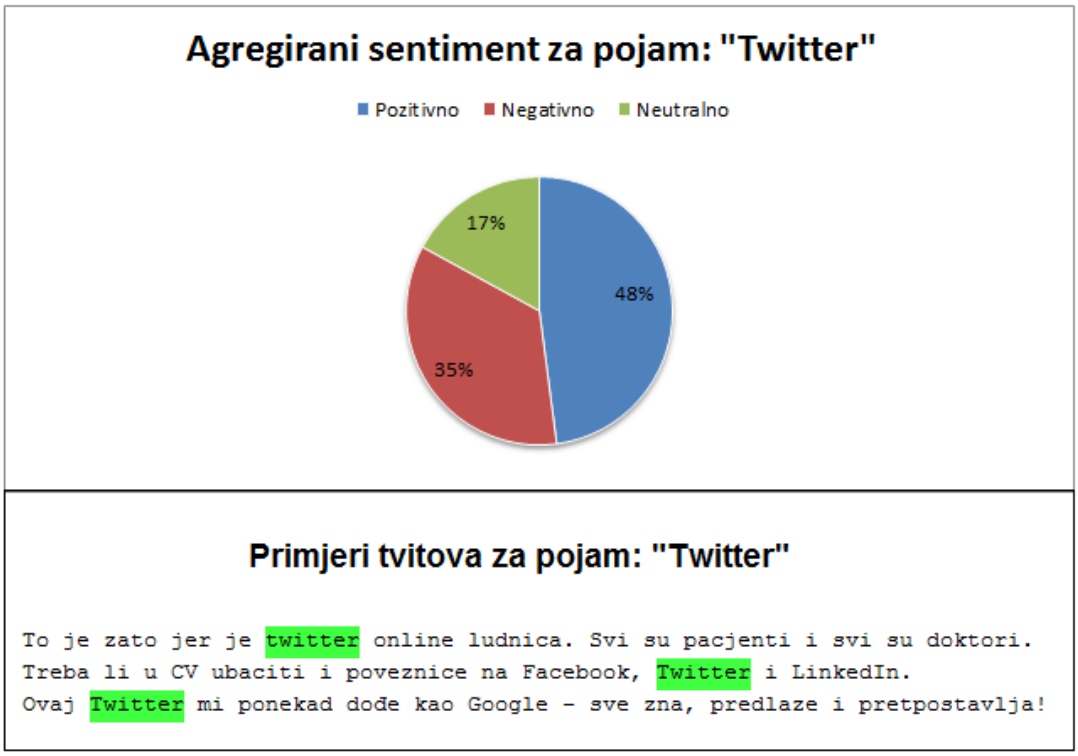
će se prikazati različite statistike. U slučaju da navedeni presjek postoji, korisniku će biti prikazana grafička statistika o broju pozitivnih, negativnih i neutralnih tvitova u kojima se upit javlja a u slučaju da presjek ne postoji korisniku će biti prikazan agregirani sentiment za svaku pojedinu riječ, tj. bit će prikazani podaci iz tablice "Riječi" za svaku riječ upita.

5.3. Upotreba aplikacije

U ovom poglavlju je ukratko objašnjena upotreba i prikazan izgled korisničke aplikacije. Nakon otvaranja aplikacije korisniku se prikazuje početna stranica s jednostavnim tekstualnim poljem u koje korisnik upisuje upit. Nakon upisivanja upita korisnik klikom miša na tipku "Analiziraj!" pokreće pretraživanje tvitova i prikupljanje statistika. Početni ekran prikazan je na slici 5.2.

Slika 5.2: Početna stranica korisničke aplikacije

Kada se prikupe svi potrebni podaci korisniku se prikazuju rezultati pretrage. U gornjem dijelu stranice pokazuju se statistike u grafičkom obliku. Ovisno o rezultatu pretrage postojat će ili jedna grafika u slučaju pronađenog presjeka ili onoliko grafika koliko postoji riječi u upitu u slučaju da presjek ne postoji. U doljnjem dijelu stranice bit će ispisani tvitovi u kojima se javlja upit s podcrtanim riječima iz upita u prikazanom tvitu. Također, koji tvitovi će se korisniku prikazati ovisi o rezultatima pretrage sentimenta. Primjer rezultata jednog upita prikazan je na slici 5.3:



Slika 5.3: Rezultat provođenja upita

6. Zaključak

U posljednjih nekoliko godina primjetan je nagli porast korisnički generiranog sadržaja unutar društvenih mreža. Jedna od takvih mreža osobito pogodna za strojnu analizu sentimenta je Twitter i poruke koje odašilju njegovi korisnici - tvitovi. Cilj strojne analize sentimenta je automatski odrediti mišljenje, stav ili emociju izraženu u tekstu klasifikacijom u jednu od tri klase: pozitivnu, negativnu ili neutralnu. U ovom radu analiziran je sentiment tvitova na hrvatskom jeziku.

U prvom poglavlu iznesen je pregled srodnih radova i utvrđene su smjernice za izradu rada. U drugom poglavlju opisani opisane su metode predobrade tvitova i korišteni klasifikacijski modeli. U trećem pogavlju opisan je korišteni skup podataka, utjecaj metoda predobrade na taj skup, implementacija klasifikatora i ono ključno, rezultati klasifikacije. U posljednjem pogavlju opisana je korisnička aplikacija za prikaz agregiranog sentimenta. Ključni dio rada je bilo postići što veću točnost klasifikacije i s obzirom na srodne radove u ovom radu ipak nije postignuta očekivana točnost. Uzroka nešto manje točnosti u odnosu na srodne radove ima nekoliko i potrebno ih je ovdje iznijeti.

U početku izrade ovog rada činilo se da klasifikacija tvitova koji su ograničeni duljinom na 140 znakova, te se vrlo često događa da su sastavljeni od samo jedne rečenice, ne bi trebala biti problematična. Međutim ubrzo se pokazalo da to neće biti slučaj. Prvi problem koji se pojavio bio je jezični problem, tj. kako iz ulaznog korpusa tvitova izolirati samo tvitove pisane na hrvatskom jeziku. Problem se djelomično riješio korištenjem dodatnih podataka o tvitu no i u konačnom korpusu ostao je određeni postotak tvitova koji nisu pisani hrvatskim jezikom. Sljedeći problem odnosio se na redukciju broja tvitova. U korpusu je postojao određen broj duplih tvitova, retvitova itd. Ovaj problem je izuzetno dobro riješen upotrebom metoda korištenih u drugim radovima. Sljedeći korak bila je redukcija broja značajki. U drugom i trećem poglavlju je u detalje iznesen postupak redukcije broja značajki koji se pokazao veoma dobar te je njime smanjen broj značajki gotovo u pola. Iako su dosad navedeni problemi predstavljali određene zapreke u izradi rada, nisu bili toliko krucijalni kao

problem koji se pojavio u samoj klasifikaciji, a to je problem koji bi se mogao opisati kao problem neutralne klase. Kako je rečeno, kako bi se dobio potreban broj neutralnih tvitova korišteni su neutralni hashtagovi kao neutralna oznaka. Međutim pokazalo se da su ovako označeni tvitovi, iako neutralni, tematski ograničeni (najčešće se radi o tvitovima koji sadrže nekakve vijesti, vremenske prognoze i sl.) te gotovo u potpunosti nedostaju neutralni tvitovi privatnih korisnika. Također tematska ograničenost automatski implicira i ograničenost rječnika, tj. tipičnog vokabulara korištenog u pisanju vijesti. Samo učenje klasifikatora na temelju kros validacije postizalo je solidne rezultate, usporedive s rezultatima u srodnim radovima. Međutim problem se pojavio kod testiranja klasifikatora na ručno označenim primjerima. Zamijećen je drastičan pad uspješnosti klasifikacije, a uvidom u matrice konfuzije pokazalo se da je problem uzrokovala neutralna klasa, kod koje je velik broj tvitova bio klasificiran kao pozitivan tj. negativan. Kako bi se usporedila efikasnost blago označenih tvitova označen je i primjeren ručni skup. Međutim testiranje je pokazalo podjednaku uspješnost. Nameće se pitanje koji je uzrok ovog problema a odgovor je ponovno neutralna klasa, međutim kod ručno označenih primjera problemi su najčešće prisutnost sarkazma i ironije, implicitno značenje (korisnik piše tvit kao odgovor drugom korisniku te se često ključni dio značenja gubi), te polarnost neutralnih tvitova (sadrže i pozitivni i negativni dio). Kad su se u obzir uzele samo dvije klase, rezultati su se značajno poboljšali, međutim za praktične svrhe nije primjereno izostaviti neutralnu klasu te je korisnička aplikacija napravljena na temelju klasifikacije u tri klase.

Rezultati su nažalost lošiji od rezultata dobivenih u srodnim radovima, a uzroci tome leže u gore navedenim problemima. Međutim, ovo je (po trenutnim spoznajama) prvi pokušaj analize sentimenta hrvatskih tvitova te kao takav služi za probijanje leda i otkrivanja svih problema koje to nosi sa sobom. Kako je ovaj rad temeljen na nekoliko koraka poput redukcije tvitova, redukcije broja značajki, te konačno same klasifikacije, napredak je moguće ostvariti u svakom od tih koraka npr. uočavanjem novih metoda za redukciju broja značajki, testiranjem većeg broja klasifikacijskih algoritama, ali i uzimanjem u obzir nekih novih vrsta značajki poput npr. vrsta riječi. Također bolji rezultati možda bi se mogli postići kvalitetnijim filtriranjem blagih oznaka. Za očekivati je da će se rezultati poboljšati i samim protokom vremena, prvenstveno zbog činjenice da su hashtagovi relativno nov pojam na Twitteru te hrvatski korisnici tek počinju usvajati korištenje ovog elementa mikro-blogginga što se vidi na temelju statistike prisutnosti hashtagova u tvitovima. Konačno može se reći da iako mjesta za napredak svakako ima, rezultati prvog klasifikatora hrvatskih tvitova generalno gledano nisu loši te mogu poslužiti kao vodilja svih budućih radova koji će se nadovezivati na ovu temu.

LITERATURA

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, i Rebecca Passonneau. Sentiment analysis of twitter data. U *Proceedings of the Workshop on Languages in Social Media*, stranice 30–38. Association for Computational Linguistics, 2011.

Alec Go, Richa Bhayani, i Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, stranice 1–12, 2009a.

Alec Go, Lei Huang, i Richa Bhayani. Twitter sentiment analysis. *Entropy*, 17, 2009b.

Efthymios Kouloumpis, Theresa Wilson, i Johanna Moore. Twitter sentiment analysis: The good the bad and the omg! *ICWSM*, 11:538–541, 2011.

Nikola Ljubešić, Darja Fišer, i Tomaž Erjavec. Tweetcat: a tool for building twitter corpora of smaller languages. U *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may 2014. European Language Resources Association (ELRA).

Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2012.

Alexander Pak i Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. U *LREC*, 2010.

Bo Pang i Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

Jayant Sharma i Aniruddh Vyas. Twitter sentiment analysis. *Indian Institute of Technology unpublished report (2010 http://home. iitk. ac. in/˜ jaysha/cs365/projects/report. pdf)*.

Analiza sentimenta u tvitovima na hrvatskom jeziku

Sažetak

U posljednjih nekoliko godina primjetan je nagli porast korisnički generiranog sadržaja unutar društvenih mreža. Jedna od takvih mreža osobito pogodna za strojnu analizu sentimenta je Twitter, mikro-blogging servis unutar kojega korisnici odašilju kratke poruke - tvitove. Cilj strojne analize sentimenta je automatski odrediti mišljenje, stav ili emociju izraženu u tekstu klasifikacijom u jednu od tri klase: pozitivnu, negativnu ili neutralnu. Zbog nemogućnosti ručnog označavanja velikog broja poruka koristi se model blago nadziranog strojnog učenja. U radu je ispitan utjecaj raznih metoda predobrade podataka te rad nekoliko klasifikacijskih modela poput naivnog Bayesovog klasifikatora, stroja potpornih vektora i logističke regresije. Prikazana je i obrazložena usporedba modela nadziranog i blago nadziranog strojnog učenja. Prikaz agregiranog sentimenta i statistike omogućen je kroz korisničku aplikaciju.

Ključne riječi: Twitter, obrada prirodnog jezika, analiza sentimenta, strojno učenje, hrvatski jezik

Sentiment analysis of tweets in Croatian language

Abstract

In recent years there has been a sharp rise in user-generated content within social networks. One of these networks particularly suitable for machine analysis of sentiment is Twitter, a micro-blogging service within which users broadcast short messages - tweets. The goal of machine sentiment analysis is to automatically determine the opinion, attitude or emotion expressed in text by classification into one of three categories: positive, negative or neutral. Due to the inability of manual labeling of a large number of messages a model of distant supervised machine learning. This paper investigates the impact of different methods of preprocessing data and effectiveness of several classification models such as naive Bayes classifier, support vector machines and logistic regression. Comparison of supervised and distant supervised machine learning is presented and explained in detail. Representation of aggregate sentiment and statistics is provided via a user application.

Keywords: Twitter, natural language processing, sentiment analysis, machine learning, croatian language