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. 1152

Model za analizu sentimenta u tvitovima na hrvatskome jeziku

Dolović Dino

Zagreb, srpanj 2015.

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

Zagreb, 6. ožujka 2015.

Predmet: **Analiza i pretraživanje teksta**

DIPLOMSKI ZADATAK br. 1152

Pristupnik: **Dino Dolović (0036457175)**
Studij: Računarstvo
Profil: Računarska znanost

Zadatak: **Model za analizu sentimenta u tvitovima na hrvatskome jeziku**

Opis zadatka:

Porastom količina korisnički generiranog sadržaja povećalo se zanimanje za strojnom analizom mišljenja izraženog u tekstu. Posebno je interestanta analiza mišljenja u tzv. mikroblogovima, primjerice porukama tvitera, zbog njihove dinamičnosti i izravnosti. Jedan od pristupa analizi mišljenja jest analiza sentimenta, kojom se utvrđuje je li tekst usmjeren pozitivno, negativno ili neutralno. Analiza sentimenta u mikroblogovima izazovan je problem zbog kratkoće i neformalnosti teksta.
U okviru diplomskoga rada potrebno je proučiti postupke za analizu sentimenta u mikroblogovima, s naglaskom na metode temeljene na nadziranom strojnom učenju. Razraditi model za analizu sentimenta u tvitovima na hrvatskome jeziku temeljen na nadziranome strojnom učenju. Model treba omogućiti predikciju sentimenta na razini cijele poruke ili na razini pojmova unutar poruke, po uzoru na modele razvijene za engleski jezik u okviru natjecanja SemEval-2013. Izraditi odgovarajući označeni skup podataka s ručno označenim sentimentom. Razviti programsku implementaciju modela te provesti iscrpno vrednovanje na odgovarajućem skupu podataka, uključivo analizu značajki i usporedbu sa referentnim modelima. Radu priložiti izvorni i izvršni kod razvijenog sustava, označene skupove podataka i potrebnu dokumentaciju te citirati korištenu literaturu.

Zadatak uručen pristupniku: 13. ožujka 2015.
Rok za predaju rada: 30. lipnja 2015.

Mentor:

Doc. dr. sc. Jan Šnajder

Djelovođa:

Doc. dr. sc. Tomislav Hrkać

Predsjednik odbora za diplomski rad profila:

Prof. dr. sc. Siniša Srbljić

Zahvaljujem Luki Bašeku, Marku Čulinoviću, Filipu Čulinoviću, Ivani Dolović, Mateju Jozići, Mihaelu Rašpergeru, Igoru Šoparu i Danijelu Živčecu na označavanju i izradi korpusa korištenog u ovom radu.

SADRŽAJ

POPIS SLIKA

POPIS TABLICA

1. Uvod

Knjige, blogovi, vijesti, recenzije proizvoda i ostali pisani tekstovi ogroman su izvor informacija i znanja. Velika većina takvih informacija je nestrukturirana te iz dana u dan više i više raste. Kako bismo učinkovito iskoristili takve informacije, odnosno pohranjeno znanje u njima, nužna je primjena računala, koje pružaju brzu obrade i pohranjivanje informacija te automatizaciju. No, računala sama po sebi ne mogu analizirati tekst niti pružati bilo kakve korisne informacije o nekom tekstu. U tu svrhu, razvijene su brojne metode i tehnike iz različitih područja. Jedno od područja jest obrada prirodnog jezika (engl. *natural language processing*), koja je interdisciplinarno područje na presjeku područja računarske znanosti, umjetne inteligencije i lingvistike. Neki od zadataka kojima se bavi ovo područje su sljedeći: automatsko zažimanje teksta (engl. *automatic summarization*), strojno prevođenje (engl. *machine translation*), prepoznavanje entitata u tekstu (engl. *named entity recognition*), analiza sentimenta (engl. *sentiment analysis*) i drugi.

Ovaj rad se bavi analizom mišljenja u mikroblogovima. Zbog velike popularnosti društvenih mreža te porasta količine korisnički generiranog sadržaja raste i zanimanje za ovo područje. Razlog tome jest to što korisnici prilikom generiranja sadržaja najčešće pišu o određenim temama, osobama, proizvodima i sl., izražavajući svoje mišljenje o njima. Primjerice, razumijevanje mišljenja korisnika o nekom proizvodu ili osobi može biti uvelike korisno za dionike o čijem se proizvodu piše, tj. iznosi mišljenje, odnosno osobama koje zastupaju određene osobe kao primjerice službe za odnose s javnošću (engl. *public relations, PR*).

Cilj ovog rada jest istražiti i proučiti postupke za analizu sentimenta u mikroblogovima, s naglaskom na metode temeljene na nadziranom strojnom učenju (engl. *supervised machine learning*). U fokusu rada je društvena mreža Twitter, poznata po generiranju vrlo kratkih poruka. Za potrebe izrade modela, potrebno je izraditi korpus sastavljen od tvitova (Twitter poruka) označenih određenim polaritetom koji određuje mišljenje korisnika u poruci. Tvitovi se odnose na dva pjevačka natjecanja, The Voice i XFactorAdria. The Voice – Najljepši glas Hrvatske, hrvatska je inačica svjetski

popularnog natjecateljskog pjevačkog showa koja je emitirana na hrvatskoj nacionalnoj radioteleviziji (HRT). Prva emisija emitirana je u siječnju, 2015. godine. Samo natjecanje sastoji se od različitih etapa natjecanja. Najprije mentori kroz audicije odabiru svaki po dvanaest natjecatelja koje će mentorirati kroz natjecanje. U drugoj fazi natjecanja, natjecatelji se kroz dvoboje bore za prolazak u daljnji krug natjecanja. Zadnju fazu natjecanja čine nastupi uživo gdje mentori uz gledatelje odlučuju o prolasku kandidata u sljedeću emisiju. XFactorAdria je međunarodno natjecanje po uzoru na britansku inačicu showa. Sama struktura natjecanja malo je drugačija od The Voica, no nije bitna za daljnje razumijevanje rada. Cilj oba natjecanja je pronaći najbolje pjevačke talente.
Kroz pauze natjecanja u emisijama uživo prikazuju se tvitovi korisnika koji tvitaju o natjecateljima, voditeljima, mentorima i o samoj emisiji. Temeljem broja tvitova o nekoj osobi te broja pregleda videa na Youtubeu prikazuju se neke grube statistike o popularnosti nekog od natjecatelja. Ovaj rad zamišljen je kako bi omogućio izradu takvih statistika, ali temeljem preciznijeg modela, odnosno razmatranjem i analizom sentimenta samog tvita.

Struktura rada je sljedeća: u drugom poglavlju opisani su srodni radovi značajni za ovo istraživanje koji se bave analizom sentimenta u standardnim tekstovima i mikroblogovima. Nakon toga slijedi poglavlje koje opisuje postupke izrade korpusa. Četvrto poglavlje opisuje tehničku stranu ovog rada, odnosno daje pregled metoda i tehnika korištenih pri implementaciji modela. Predzadnje poglavlje opisuje provedene eksperimente te prikazuje dobivene rezultata. Konačno, u zadnjem poglavlju dan je zaključan osvrt na cijelo istraživanje te su dane smjernice za buduće istraživanje.

2. Pregled područja

U uvodnom dijelu poglavlja dane su osnovne definicije koje se odnose na analizu sentimenta. Nakon toga napravljena je podjela područja i opisana je njegova primjena. Drugi dio donosi pregled samog područja analize sentimenta te opisuje najznačajnije radove za analizu sentimenta u standardnim tekstovima kao i u mikroblogovima.

2.1. Analiza sentimenta

Analiza mišljenja (engl. *opinion mining*) ili analiza sentimenta bavi se analizom ljudskih mišljenja, stavova te izraženih emocija prema nekom entitetu, primjerice proizvodu, osobi, događaju ili nekoj određenoj temi. Koristeći postojeće tehnike iz područja poput obrade prirodnog jezika, statistike i stojnog učenja, analizom sentimenta u tekstu nastoje se pronaći mišljenja te odrediti njihov sentiment. Mišljenje je definirano kao osobno uvjerenje ili presuda koja nije utemeljena na dokazima ili sigurnosti (WordNet, 2010). Mišljenje također možemo smatrati binomnim izrazom koji se sastoji od dviju komponenti: cilj, prema kome ili čemu se mišljenje odnosi te sentiment, kakvo je to mišljenje. Najčešća klasifikacija mišljenja jest na sljedeće tri klase sentimenta: pozitivna, negativna te neutralna. No, s obzirom na zahtjeve i željenu razinu granulacije, moguće je produbiti podjelu, odnosno odrediti brojčane vrijednosti za pojedini sentiment koje izražavaju jakost sentimenta, odnosno govore koliko je neki sentiment pozitivan odnosno negativan.

Sentiment je u tekstu moguće analizirati na više razina. Najniža razina jest analiza na razini fraza, gdje se nastoji odrediti ukupan sentiment nekog teksta koristeći isključivo fraze u tekstu. Razina iznad analize na razini fraza jest analiza na razini rečenice. Najviša razina jest analiza na razini cijelog dokumenta, koja je ujedno i najzahtjevnija jer uključuje kombinaciju i zbrojeve analiza prethodnih razina. Također, postoji i analiza sentimenta na razini aspekata nekog entiteta. Primjer takve analize uključuje određivanje sentimenta za pojedini aspekt nekog entiteta čime se dobiva detaljan uvid u

pojedine karakteristike entiteta. Kao primjer takve analize može se uzeti analiza goriva (tvar koja sagorijevanjem daje toplinu), gdje bi sustav za analizu sentimenta analizirao sentiment za pojedine aspekte koji se tiču goriva, kao npr. cijena, porijeklo, kvaliteta itd.

Kao i u svim zadacima koji uključuju obradu i rad s tekstovima, i u ovom području nailazimo na mnogo problema. Prvi problem jest način izražavanja mišljenja koji je složen, tj. mišljenja se često iznose vrlo detaljno, što je jasno čovjeku no ne i računalu. Uz to, česta je i pojava promjene tema, odnosno promjene cilja o kojem se iskazuje mišljenje. Nadalje, prisutan je i problem dvosmislenosti riječi. Naime, riječ se interpretira na različite načine ovisno o kontekstu u kojem se nalazi. Tako primjerice riječ *dobro* u većini slučajeva podrazumijeva pozitivan sentiment, no ovisno o kontekstu, riječ je moguće interpretirati kao negativnu kao npr. u sljedećem primjeru *Bilo je strašno, dobro da je nazvala*. Također, iskazivanje sarkazma i ironije vrlo je čest slučaj pri iznošenju mišljenja što je računalu gotovo nemoguće odrediti.

Primjene analize mišljenja su brojne, no najčešće se odnose na poslovnu domenu. Primjerice, istraživanje mišljenja o nekom proizvodu uvelike može biti od koristi nekoj kompaniji. Tako bi kompanija u slučaju negativnog trenda mišljenja mogla dublje analizirati kritike te ispraviti nedostatke proizvoda na koji se korisnici žale. S druge strane, pozitivne kritike dodatno bi utjecale na povećanje proizvodnje tog proizvoda. Druga primjena analize sentimenta jest u sustavima za preporučivanje (engl. *recommender systems*), gdje bi sustav temeljem analize recenzija automatski odredio mišljenje korisnika, odnosno sentiment prema onome što je recenzirao. Time bi ostalim korisnicima sustav preporučio samo one stvari koje su pozitivno ocijenjene. Primjeri preporučiteljskih sustava najviše se odnose na multimedijalni sadržaj poput glazbe, filmova, web-stranica i sl. Nadalje, naglim razvojem i masovnim korištenjem društvenih mreža poput Facebooka, Twittera te drugih, raste i zanimanje za analizom kratkih i dinamičnih poruka. Poruke koje pišu korisnici takvih mreža najčešće sadrže njihovo mišljenje o aktualnim temama i događajima u društvu. Njihovom bi se analizom i agregacijom jednostavno izračunali trendovi odnosno generalna mišljenja ljudi. Temeljem dobivenih trendova, bilo bi moguće poduzeti mjere kako bi se ti trendovi održavali u slučaju pozitivnog, odnosno promijenili u slučaju negativnog trenda.

2.2. Analiza sentimenta i mišljenja u standardnim tekstovima

U novije vrijeme, analiza sentimenta i mišljenja postaje sve popularnija u nestandardnim tekstovima, poput tvitova na Twitteru ili objava na Facebooku, no originalni radovi i istraživanja koja se bave ovim područjem započinju prije samog postojanja društvenih mreža te se bave analizom sentimenta u standardnim tekstovima. Pang et al. (2002) prvi analiziraju sentiment u filmskim recenzijama koristeći metode nadziranog strojnog učenja te ističu kako metode nadmašuju primitivne metode (engl. *baseline methods*). Također, isti autori u radu (Pang i Lee, 2005) istražuju drugačiji način klasifikacije samog teksta. Umjesto klasifikacije recenzije na pozitivnu i negativnu klasu, eksperimentiraju s dodjeljivanjem ocjena od 1 do 5.

Za razliku od Pang et al. (2002), Turney (2002) analiziraju metode nenadziranog strojnog učenja pri klasifikaciji sentimenta recenzije iz nekoliko različitih domena. Rezultat klasifikacije izračunava se temeljem prosječne semantičke orijentacije fraza u rečenicama koje sadrže pridjeve i priloge, a dobivaju se prethodnim označavanjem pomoću sustava za označavanje vrsta riječi (engl. *part of speech tagger, POS tagger*). Semantička orijentacija fraze izračunava se kao razlika uzajamne informacije (engl. *mutual information, MI*) između pozitivne fraze i riječi *odlično* (engl. *excellent*) te negativne fraze i riječi *los̆* (engl. *poor*). Recenzija se preporučuje ako je vrijednost izračunate semantičke orijentacije pozitivna.

Analizom sentimenta temeljenoj na aspektima entiteta bave se Snyder i Barzilay (2007). U njihovom radu proučava se sentiment restorana, koji se izračunava temeljem sentimenta određenih aspekata samog restorana, poput cijene jela, ambijenta, usluge itd. Za svaki od aspekata dodjeljuje se ocjena od 1 do 5 koja određuje sentiment tog aspekta restorana. Umjesto klasifikacije restorana temeljem pojedinačnih aspekata, pokušavaju se stvoriti zavisnosti između određenih aspekata pomoću relacije slaganja (engl. *agreement relation*). Relacija slaganja bilježi je li korisnik jednako zadovoljan svim aspektima restorana ili kroz recenziju izražava različite stupnjeve zadovoljstva restoranom.

2.3. Analiza sentimenta i mišljenja u mikroblogovima

Twitter je uz Facebook najpopularnija društvena mreža kojom se koristi gotovo milijardu korisnika. Omogućuje besplatnu registraciju korisnika te objavljivanje kratkih

poruka koji se nazivaju tvitovi (engl. *tweet*). Korisnici Twittera objavljuju tvitove koji su javni te za razliku od Facebooka omogućuje praćenje ostalih Twitter korisnika sa sličnim interesima bez ikakve dodatne povezanosti odnosno uspostavljanja virtualnog prijateljstva s njima. Tvitovi mogu biti objavljeni korištenjem neke od raznih aplikacija, od desktop klijenta, web-stranice do mobilnih aplikacija, što je u 80% i slučaj. Svakodnevno se objavljuje gotovo 500 milijuna novih tvitova (2015 Twitter, 2015). U nastavku su navedene neke specifičnosti karakteristične za tvitove:

- Maksimalna duljina svakog tvita je 140 znakova;
- Mogućnost korištenja *hashtagova* kako bi se najčešće označili metapodaci koji dodatno opisuju tvit. Metapodaci su najčešće nazivi tema ili događaja o kojima se piše, geolokacija itd. (npr., *#TheVoiceHRT, #xFactorAdria, #Zagreb*);
- Objavljivanje poveznica na multimedijski sadržaj, poput slika, videa ili pjesama;
- Korištenje opcije „re-tvita“ kojim se prikazuje tvit nekog drugog korisnika;
- Upotreba znaka @ ispred korisničkog imena omogućava direktno obraćanje nekom drugom korisniku.

Također, česta je i upotreba emotikona, čime korisnik najčešće želi iskazati svoju emocionalnu reakciju na temu ili neki događaj o kojem piše. Upravo zbog prethodno nabrojenih specifičnosti, tvitovi su vrlo ekspresivni i dinamični.

Kao što je rečeno, tvitovi se najčešće objavljuju korištenjem aplikacija na mobilnim uređajima. Korištenje mobilnih uređaja dodatno otežava zadatke obrade prirodnog jezika koji uključuju tvitove. Razlog tome jest taj što je korisnik zbog veličine mobilnog uređaja ograničen veličinom tipkovnice. Iz toga razloga, za razliku od standardnih tekstova, tvitovi su prepuni pravopisnih pogrešaka. Nadalje, tvitovi pisani na hrvatskom jeziku i srodnim jezicima koji koriste dijakritičke znakove, često su pisani bez njih. Primjerice, riječ *žao* u stvorenom korpusu spominje se jedanaest puta, dok se istoimena riječ bez dijakritika spominje deset puta. Ova činjenica dodatno otežava proces u predobradi tvitova.

Twitter je kao društvena mreža stvoren $2006.$ godine, a prvo zanimanje za analizu sentimenta započinje nekoliko godina nakon što sam servis doživljava ogroman rast. Go et al. (2009) u svojem radu prvi istražuju mogućnosti koje pružaju metode strojnog učenja u analizi sentimenta tvitova. Također, izrađuju i prvi automatski stvoren korpus označenih tvitova koji se temelji na emotikonima prisutnima u tvitu. Ako tvit sadrži emotikone poput :), :D i sl. tvit je označen kao pozitivan, odnosno kao negativan ako sadrži negativne emotikone kao npr. :(, :-(itd. Neutralni razred u tom istraživanju

je ignoriran. Korištenjem različitih klasifikatora (SVM, NB, MaxEnt) i jednostavnih značajki poput unigrama, bigrama te vrsta riječi, postignuti su rezultati od 80% preciznosti.

Slično kao i u prethodnom radu, Pak i Paroubek (2010) također stvaraju korpus koristeći pronađene emotikone, no u istraživanju uključuju i neutralnu klasu. Kao primjeri tvitova s neutralnim sentimentom uzimaju se tvitovi koje objavljuju razni novinski portali, npr. *New York Times*, *Wahsington Posts* itd. Korištenjem stvorenog korpusa evaluiraju se tri modela (NB, SVM, CRF) pomoću značajki poput unigrama, bigrama, trigrama. Također, provedena je i analiza distribucije vrsta riječi u označenom korpusu. Analizom distribucije dobiveni su neki zanimljivi i neki očekivani rezultati:

- korisnici češće koriste osobne zamjenice kada iznose svoje mišljenje, odnosno izražavaju subjektivnost, dok u objektivnim tekstovima više upotrebljavaju vlastita imena;
- glagoli u objektivnim porukama najčešće su u trećem licu jednine, dok se u subjektivnim više koristi prvo lice jednine;
- subjektivni glagoli najčešće sadrže pridjeve čime korisnici izražavaju snagu emocije;
- tvitovi s negativnim sentimentom često su pisani u prošlom vremenu;
- pozitivni tvitovi sadrže više superlativa od ostalih.

Kouloumpis et al. (2011) istražuju korištenje različitih korpusa za učenje modela (SVM). Pri učenju modela koriste dva različita korpusa: prethodno opisan korpus stvoren na temeljem emotikona u (Go et al., 2009) te korpus stvoren temeljem *hashtagova*. Kao podatke za evaluaciju modela koriste poseban korpus tvitova koji su vezani uz određeni proizvod. Nadalje, koriste nove skupine značajki vezane posebno uz tvitove, kao npr. prisutnost pozitivnih i negativnih emotikona u tvitu, prisutnost skraćenica i sl. Dobivenim rezultatima pokazuju da nove značajke uz korištenje n-grama riječi kao i kombiniranje dvaju korpusa doprinose poboljšanju performansa njihovog sustava.

Primjene sentiment analize u tvitovima opisuju se u radovima poput (Bollen et al., 2011) te (Wang et al., 2012). U prvom radu, Bollen et al. (2011) istražuju može li se predvidjeti pad ili rast cijene dionica određene kompanije analizom javnog raspoloženja u tvitovima. U drugom radu, Wang et al. (2012) bave se analizom sentimenta tijekom kampanje za predsjedničke izbore 2012. godine u Sjedinjenim Američkim Državama. Nakon predobrade svakog tvita, tvit se veže na nekog od kandidata te se traži sentiment u tvitu. Sentimenti se agregiraju po osobama te se kasnije vizualiziraju kroz vrijeme.

U prethodnim godinama, istraživanje analize sentimenta u tvitovima postalo je i jednim od zadataka u okviru natjecanja SemEval (engl. *Semantic Evaluation*). SemEval jest skup na kojem se objavljuju zadaci vezani za istraživanje značenja jezika. Timovi znanstvenika rješavaju određene zadatke, te pomoću dobivenih podataka izrađuju sustav koji se potom evaluira. Neki od zadataka na nedavnim skupovima vezani su za analizu sentimenta u tvitovima. To su: (Nakov et al., 2013) koji će ovdje biti detaljnije opisan te noviji skupovi (Rosenthal et al., 2014) i (Rosenthal et al., 2015).
Glavni cilj skupa iz 2013. godine (Nakov et al., 2013) bio je napraviti kvalitetni korpus ručno označenih tvitova te ispitati rad sustava za analizu sentimenta nad dvama zadacima:

1. Za dani tvit i označene riječi ili fraze u njemu, odrediti sentiment označenih riječi ili fraza kao pozitivne, negativne ili neutralne;

2. Za dani tvit odrediti kakav je sentiment tvita: pozitivan, negativan ili neutralan.

Za izradu korpusa, najprije je korišten poseban sustav za ekstrakciju imenovanih značajki (engl. *named entity recognition*) specijalno namijenjen za tvitove, pomoću kojeg su dobiveni popularni entiteti koji su se smatrali temama. Nakon ekstrakcije tema, obavljen je postupak filtracije tvitova koristeći SentiWordNet[1] kako bi se na neki način odredio sentiment tvita. Tvitovi s vrijednošću sentimenta većeg od zadanog praga odabrani su kao tvitovi za označavanje. Za potrebe označavanja tvitova koristio se vanjski servis.[2]. Svaki tvit označavalo je pet različitih osoba, kako bi se u što većoj mjeri izbjegla subjektivnost osobe. Konačna oznaka sentimenta nekog tvita dodijeljena je tvitu temeljem većine oznaka, odnosno ako su tri osobe označile isti sentiment tvita, tvit je dobio konačnu klasu. U suprotnome, ako je broj različitih oznaka sentiment pri označavanju bio izjednačen, tvit je odbačen. Konačni korpus s označenim tvitovima sadrži oko deset tisuća tvitova te devet tisuća označenih fraza u tim tvitovima. Korpusi su korišteni i dalje dorađivani u daljnjim izdanjima natjecanja za neke novije zadatke, (Rosenthal et al., 2014) i (Rosenthal et al., 2015).
Kao mjere za evaluaciju izrađenih sustava korištene su standardne evaluacijske mjere: P (engl. *precision*), R (engl. *recall*) i F_1 (engl. F_1*-score*). Za prvi je zadatak samo jedan sustav koristio metode djelomično nadziranog strojnog učenja, dok su svi ostali timovi koristili metode nadziranog strojnog učenja: SVM, NB, MaxEnt. Najbolji rezultati za prvi i drugi zadatak dobiveni su korištenjem modela SVM s postignutim

[1] `http://sentiwordnet.isti.cnr.it/`

[2] *Mechanical Turk* - Amazonov servis koji omogućava jeftino i masovno obavljanje kao i pružanje poslova poput označavanja podataka

F_1 vrijednostima od 88.9% i 69% od strane istog tima (Mohammad et al., 2013). Iz dobivenih rezultata odmah je vidljivo kako je određivanje sentimenta na razini fraze ili riječi daleko jednostavniji zadatak od određivanja sentimenta cijele rečenice, odnosno tvita.

3. Predobrada i označavanje podataka

Ovo poglavlje bavi se podacima. Najprije je opisan postupak prikupljanja podataka za označavanje. Nakon toga opisane su smjernice prema kojima je bilo potrebno označavati podatke. Konačno, u zadnjem potpoglavlju napravljena je analiza i pregled označenog korpusa.

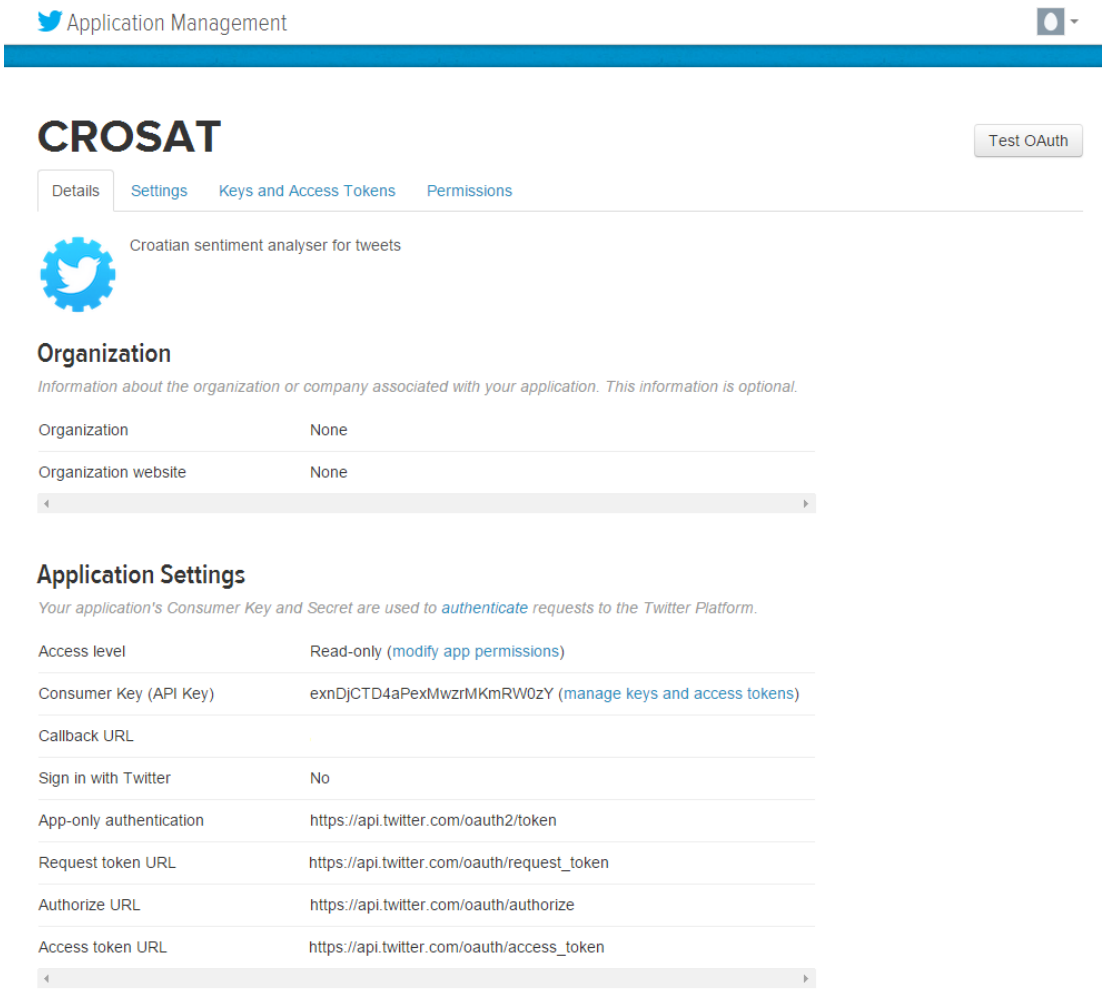
3.1. Prikupljanje podataka

Za potrebe izrade i evaluacije modela najprije je bilo potrebno prikupiti podatke, odnosno tvitove koji se odnose na prethodno spomenute teme u uvodnom dijelu rada, The Voice i XFactorAdria. Pri dohvaćanju tvitova korišten je programski jezik Python te Twitter API,[1] koji omogućava čitanje i pisanje tvitova te ostale napredne opcije pomoću računalnih programa. Twitter API je izrađen kao REST servis (*engl. representational state transfer*) koji dopušta Twitter-aplikacijama stvaranje novih tvitova, čitanje postojećih, dohvaćanje tvitova određenog korisnika i slične opcije. Također, Twitter API omogućuje pretraživanje i dohvaćanje tvitova u stvarnom vremenu (*engl. real time streaming*) prema određenim kriterijima pretrage, poput riječi u tvitovima, lokacije korisnika koji tvita i sl.

Za potrebe prikupljanja tvitova izrađena je Twitter-aplikacija nazvana *CROSAT* (engl. *croatian sentiment analyser for tweets*). Twitter-aplikacija se izrađuje tako da se ispune određeni podaci poput imena i opisa aplikacije. Nakon pristajanja na uvjete korištenja, aplikacija je izrađena te je potrebno izraditi potrebne pristupne značke (*engl. access token*) koje se koriste za potrebnu autentifikaciju i autorizaciju pri zahtjevu za resursima. Izgled sučelja s postavkama aplikacije prikazan je slici 3.1.

Kao što je već spomenuto, za dohvat tvitova koristi se REST servis Twitter API. Potrebni resursi dohvaćaju se slanjem HTTP-zahtjeva na sljedeći URL: `https://api.twitter.com/1.1/search/tweets.json`. HTTP-odgovor oblikovan je kao JSON objekt gdje je svaki tvit koji odgovara parametrima postavljenima u zahtjevu

[1] `https://dev.twitter.com/rest/public`



Slika 3.1: Prikaz postavka *CROSAT* aplikacija

predstavljen kao jedan JSON objekt sa svim potrebnim svojstvima. Isječak koda za dohvat tvitova nalazi se na slici 3.2. Program na slici 3.2 poziva se koristeći komandnu liniju ili pozivom funkcije *scrap* koju je potrebno uključiti u Python program iz kojeg se funkcija želi koristiti. Funkcija *scrap* kao jedini ulazni parametar prima listu pojmova prema kojima se obavlja pretraga za tvitovima. Funkcija uzastopno šalje HTTP-zahtjeve za traženim resursom, parsira vraćeni sadržaj te sprema dohvaćene tvitove. Nakon što je dohvaćeno više od deset tisuća tvitova, funkcija se prekida te vraća dohvaćene tvitove u pozivajući program. HTTP-zahtjevi koji se stvaraju i šalju za potrebnim resursima enkapsulirani su u metodi *oauth_req* čija implementacija nije prikazana. Primjer HTTP-zahtjeva nalazi se na slici 3.2 na liniji 20. Metoda *oauth_req* najprije oblikuje HTTP-zahtjev pomoću predanih parametara nakon čega se stvoreni zahtjev šalje servisu. Parametri koji se šalju u HTTP-zahtjevu su:

– q - popis pojmova temeljem kojih se obavlja pretraga. Tvit će biti vraćen ako u

```
# podaci potrebni za autentifikaciju
consumer_key = "..."
consumer_secret = "..."
access_token = "..."
access_token_secret = "..."

def scrap(terms):
    tweets = set()
    params = {
        'url_params': {
            'q': terms, # kriterij pretrage
            'count': 100 # broj dohvacenih tvitova po zahtjevu
        }
    }
    last_lowest = None

    while len(tweets) < 10000:
        # posalji zahtjev za dohvacanjem tvitova
        # npr., https://api.twitter.com/1.1/search/tweets.json?q=TheVoiceHRT&count=100
        content = oauth_req(params)

        # pohrani samo relevantene podatke (id, lang, status, user ...)
        new_tweets = parse_content(content)
        tweets.union(new_tweets)

        # pronadi id zadnje dohvacenog tvita koji sluzi za sljedeci zahtjev
        last_id = find_last_id(new_tweets)
        params['url_params']['max_id'] = last_id

        time.sleep(3)

    return tweets

if __name__ == '__main__':
    tweets = scrap(['TheVoiceHRT'])
```

Slika 3.2: Isječak koda za dohvaćanje tvitova

sebi sadrži barem jedan od predanih pojmova;

- count - broj koji označava koliko će tvitova biti dohvaćeno s jednim HTTP-zahtjevom;
- max_id - id tvita od kojeg će pretraga biti nastavljena. Naime, tvitovi se dohvaćaju određenim redoslijedom, odnosno od najnovijih prema starijim. Svaki zahtjev započinje pretragu tvitova od najnovijih, no kako su ti tvitovi već dohvaćeni prethodnim HTTP-zahtjevom, pretragu je potrebno obaviti od zadnje dohvaćenog tvita što specificiramo ovim parametrom.

Popis ostalih parametara koji se mogu koristiti u pretrazi dostupni su na ovoj poveznici `https://dev.twitter.com/rest/reference/get/search/tweets`. Zbog ograničenja postavljenih od strane Twittera moguće je da HTTP-odgovor ne sadrži traženi resurs. HTTP statusni kod u tom je slučaju 417, što označava da je poslano više HTTP-zahtjeva od dozvoljenog broja. Konkretno ograničenje[2] za ovaj resurs je postavljeno na slanje 450 HTTP-zahtjeva unutar 15 minuta, što je u ovom slučaju dovoljno za dohvaćanje svih 10000 tvitova odjednom. Implementacija funkcije *req_oauth* podržava i mogućnost izvršavanja dohvaćanja tvitova i nakon što je ograničenje prekoračeno, tako da pauzira program na 15 minuta. Popis ograničenja za različite resurse je različiti, a može se pregledati na sljedećoj poveznici `https://dev.twitter.com/rest/public/rate-limits`.
Tvitovi su dohvaćeni u tri navrata, odnosno svaki puta dan nakon emisije uživo. Konačan broj dohvaćenih tvitova za traženi pojam *TheVoiceHrt* iznosi oko 20 tisuća, dok za traženi pojam *XFactorAdria* iznosi oko 7 tisuća.

3.2. Odabir podskupa podataka za označavanje

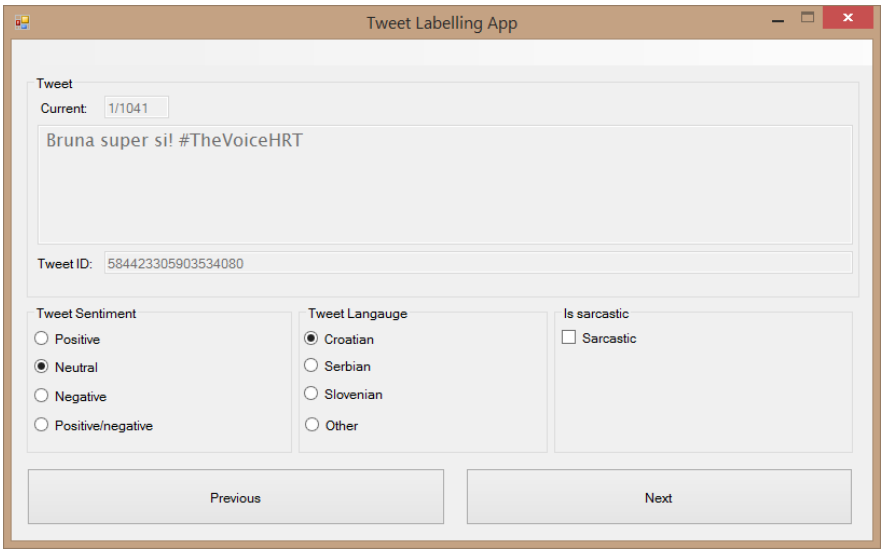
Nakon dohvata tvitova napravljena je filtracija i selekcija podskupa tvitova koji će biti označeni. Prvi korak filtracije bio je odbaciti sve tvitove koji se ponavljaju. Iako je svaki tvit jedinstven, odnosno za njega postoji jedinstveni identifikator, zbog opcije „re-tvitanja“ moguće je da se neki tvitovi u stvorenom korpusu ponavljaju. Takvi tvitovi u sebi sadrže znakove *RT* te su oni u ovom koraku odbačeni.
Sljedeći korak bio je izrada rječnika s riječima pozitivnog te negativnog sentimenta. Razlog tome jest taj što je korpus još uvijek preveliki te ga je potrebno smanjiti odnosno odabrati one tvitove za koje se smatra da će sadržavati neko mišljenje korisnika.

[2] `https://dev.twitter.com/rest/reference/get/search/tweets`

Najčešća vrsta riječi kojom se izražava sentiment su pridjevi, no korišteni su i glagoli kojima se često izražava mišljenje te u manjoj mjeri imenice. Primjeri riječi pozitivnog sentimenta: *karakteran, ljubazan, najbolji, odličan, ljubav, super, obožavam itd.* Primjeri riječi negativnog sentimenta: loš, glup, jadan, dosadan, mrzim, žao itd. Ukupan broj riječi u pozitivnom rječniku je 229, dok negativnih ima 206. Osim toga, svaka je riječ prije filtracije svedena na sve njezine oblike u svim licima jednine i množine. Također, ako riječ sadrži dijakritičke znakove, riječ je napisana i bez njih.
Nakon filtracije preostalo je oko 2700 tvitova iz prvog korpusa, The Voice HRT, dok je za drugi korpus preostalo njih 500. Ukupan broj tvitova koji je korišten u fazi označavanja jest 3123.

3.3. Podjela podskupa za označavanje

Kako bi rezultati označavanja bili što bolji, odnosno kako bi se maksimalno izbjegla subjektivnost pojedinog označivača, svaki tvit označavale su tri različite osobe. Pri označavanju je sudjelovalo devet osoba podijeljenih u tri skupine, pri čemu su dvije od skupina označavale 1041 tvita dok je jedna skupina označavala 1042 tvita. Za potrebe označavanja tvitova napisana je desktop aplikacija u programskome jeziku C#. Izgled aplikacije prikazan je na slici 3.3.



Slika 3.3: Aplikacija za označavanje

3.3.1. Upute za označavanje

Pri označavanju bilo je potrebno odrediti kakav je sentiment prisutan u tvitu. Temeljem pronađenog sentimenta tvit je bilo potrebno klasificirati u jedan od četiri moguća razreda: pozitivan, negativan, neutralan te pozitivan-negativan.
Sentiment je u tvitu najčešće izražen mišljenjem, kao npr. *Obozavam the voice*. Ako nema direktnog mišljenja u tvitu, potrebno je promatrati postoji li kakav emocionalni izražaj ili reakcija na neki događaj, kao primjerice žaljenje na nešto, ushićenost, iščekivanje i sl. Primjerice, sljedeći tvit izražava tugu korisnika zbog ispadanja njegova omiljenog kandidata: *Bogica evo nisam glasala,ali tako mi je žao sto neides dalje.. #TheVoiceHRT:(*. Ako u tvitu nije pronađen sentiment, tvit se smatra neutralnim. U tvitovima je također bilo potrebno označiti i jezik kojim je tvit pisan te postaviti oznaku ako je tvit sarkastičan. Detaljnije upute o označavanju uz primjere nalaze se u dodatku A.

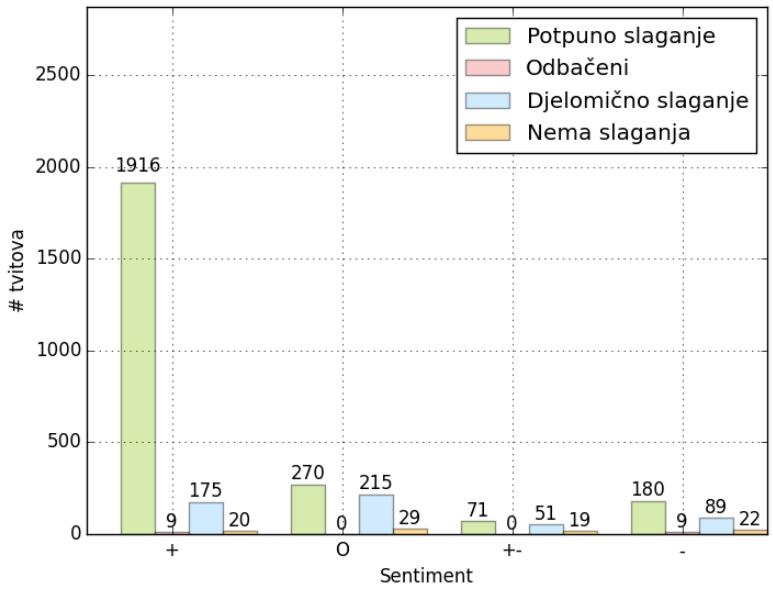
3.4. Rezultati i analiza označavanja

Prosječno trajanje označavanja skupa podataka po osobi iznosilo je oko 2 sata i 30 minuta. Ukupan broj označavanje iznosi 9363 od čega je svaki tvit označen od tri različite osobe što ukupno daje 3121 označenih tvitova.
Prije detaljnije analize najprije su odbačeni tvitovi koji su pisani na jeziku različitom od hrvatskog, srpskog i slovenskog. Takvih tvitova bilo je ukupno 115 te preostali broj tvitova iznosi 3006. U nastavku analize promatraju se samo preostali tvitovi.
Kako bi se odabrali samo najbolje označeni podaci, potrebno je izbaciti neke tvitove. Takvi tvitovi nastaju zbog neslaganja označivača: ako su sve tri osobe različito označile tvit, tvit se odbacuje jer nema slaganja. Takvih tvitova bilo je ukupno 33. Nadalje, ako je dvoje označivača označilo tvit kao pozitivan, a jedan kao negativan, tvit se također odbacio. Vrijedi i obrnuti slučaj. Takvih primjera bilo je ukupno 9. Analizom odbačenih tvitova ustanovljeno je da je do pogrešaka u označavanju došlo uglavnom zbog graničnih slučajeva ili nepažnje označivača. Detaljni prikaz odbačenih te preostalih tvitova prikazan je u tablici 3.1.

Tablica 3.1: Statistika označenog korpusa

	# tvitova	% u korpusu
Ukupno preostalih	3006	96.31
Poptuno slaganje (3/3)	2437	81.07
Djelomično slaganje (2/3)	530	17.63
Nema slaganja (0/3)	30	1.00
Odbačeni (+ + -, - - +)	9	0.30
Ukupno	2967	95.06

Rezultati označavanja grupirani po sentiment klasi tvita i načinu slaganja označivača prikazani su grafom na slici 3.4. Stupci koji prikazuju odbačene tvitove i tvitove za koje nije postojalo slaganje prikazani su za više klasa sentimenta istodobno. Npr., ako su označivači označili sentiment tvita redoslijedom kao pozitivan (+), neutralan (O) i pozitivan-negativan (+-), svakoj klasi sentimenta pridodana je vrijednost da je bilo neslaganja. Isto vrijedi i za odbačene tvitove, no oni su samo promatrani za klasu pozitivnih i negativnih. Iz grafa na slici 3.4 vidljivo je da je najviše neslaganja bilo za neutralnu klasu (O).



Slika 3.4: Rezultati grupirani po klasi sentimenta i slaganju označivača

Slaganje označivača izračunato je pomoću mjere Fleiss' kappa, statističke mjere koja daje procjenu pouzdanosti slaganja između dva ili više označivača pri dodjeljivanju kategoričkih oznaka koje su u ovom slučaju sentiment oznake tvitova. Fleiss' kappa mjera izračunata je pojedinačno za svaku skupinu te je uprosječena za konačan rezultat. Rezultati su prikazani u tablici 3.2. Iako ne postoje točno određene upute za procjenu rezultata, ipak su dane neke smjernice (Landis i Koch, 1977). Dobiveni rezultat $\kappa = 0.725$ može se svrstati u drugu po redu najbolju kategoriju označavanja $[0.61 - 0.81]$ koja označava znatno slaganje između označivača (*engl. substantial agreement*).

Tablica 3.2: Rezultati slaganja koristeći Fleiss' kappa

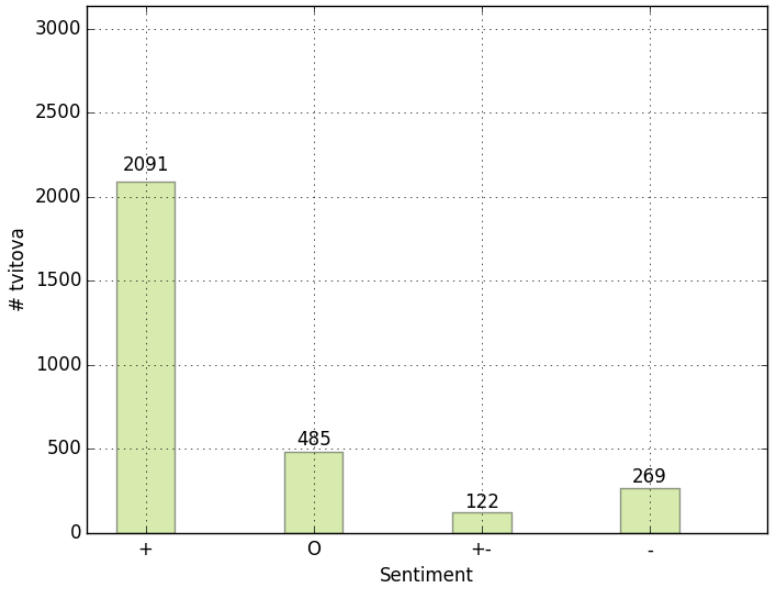
	P_{avg}	P_e	κ
$g1$	0.860	0.549	0.742
$g2$	0.869	0.534	0.720
$g3$	0.852	0.515	0.696
Prosj.	0.869	0.532	0.721

Rezultati označavanja za pojedini sentiment prikazani su matricom zabune u tablici 3.3. Reci matrice prikazuju oznaku sentimenta tvita koji je dodijelila osoba, dok se u stupcima nalaze konačne oznake sentimenta tvita. U tablici nisu prikazana označavanja tvitovi koji su odbačeni, odnosno gdje je slaganje označavanja tvitova bilo djelomično ili ga nije bilo jer takvim tvitovima prema uspostavljenim pravilima nije moguće odrediti konačnu oznaku sentimenta i ovdje su redundantni. Iz matrice zabune vidljivo je da je najviše djelomičnih slaganja bilo između pozitivne i neutralne klase sentimenta.

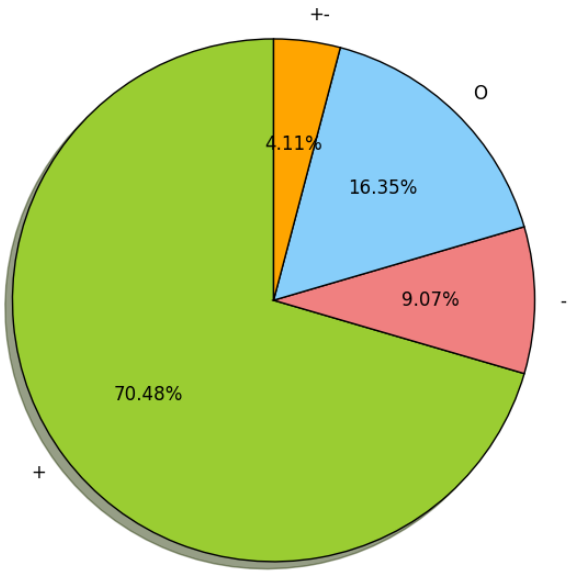
Tablica 3.3: Matrica zabune označenog korpusa

		Oznaka klasifikatora			
		$+$	O	$-$	$+-$
Oznaka označivača	$+$	6098	164	0	29
	O	146	1240	67	11
	$-$	0	49	718	11
	$+-$	29	2	22	315

Konačni rezultati za pojedinu sentiment klasu prikazani su brojčano na slici 3.5 te u postocima na slici 3.6. Iz grafova na slikama 3.5 i 3.6 vidljivo je da je korpus pretežito sastavljen od pozitivnih tvitova.



Slika 3.5: Rezultati označavanja po klasi sentimenta



Slika 3.6: Rezultati označavanja po klasi sentiment izraženi u postocima

4. Modeli za analizu sentimenta i mišljenja

U ovom poglavlju opisan je programski paket korišten u implementaciji modela za analizu sentimenta u tvitovima. Nakon toga dane su osnovne teorijske postavke korištenih metoda nadziranog strojnog učenja te su opisane značajke korištene u pri implementaciji.

4.1. Programski paket Sklearn

Za izradu sustava za analizu sentimenta u tvitovima odabran je programski paket scikit-learn. Scikit-learn ili skraćeno Sklearn je programski modul otvorenog koda koji sadrži brojne implementacije metoda strojnog učenja kao i ostale često korištene metode za predobradu teksta, ekstrakciju značajki iz teksta i slika, evaluacije modela i sl. Modul je pisan u programskom jeziku Python, no sadrži brojne omotače (engl. *wrappers*) prema efikasno implementiranim metodama u programskom jeziku C++. Primjer takvog omotača je implementacija stroja s potpornim vektorima te logističke regresije koji su zapravo omotači oko LIBSVM, odnosno LIBLINEAR modula pisanih u programskom jeziku C++. Također, kako bi što efikasnije implementirali metode, koriste se moduli poput NumPy i SciPy. NumPy je nadogradnja na programski jezik Python te služi kao podrška za višedimenzionalna polja i matrice i omogućuje brojne i efikasne numeričke operacije nad njima. SciPy je kolekcija programa otvorenog koda koji se koriste u znanstvene svrhe, različite analize te istraživanja u inženjerstvu. Sadrži programe za optimizaciju, linearnu algebru, integraciju i interpolaciju, procesiranje signala i slika te brojne druge.

4.2. Korišteni modeli

4.2.1. Naivan Bayesov klasifikator

Bayesov klasifikator jest klasifikacijski model. Klasifikacija primjera ostvaruje se koristeći Bayesovo pravilo koje određuje vjerojatnost pripadanja primjera za svaku klasu. Model pripada skupini generativnih modela koji pretpostavljaju da je vjerojatnost pripadanja nekog primjera određenoj klasi proporcionalna zajedničkoj vjerojatnosti te klase i primjera. To zapravo znači da takvi modeli modeliraju zajedničku vjerojatnost primjera i klase. Također, model je parametarski, tj. pretpostavlja da se primjeri u ulaznom prostoru pokoravaju nekoj razdiobi te se učenje modela svodi na pretraživanje parametara razdiobe (Šnajder i Dalbelo Bašić, 2012).

Klasifikacija primjera zasniva se na izračunu aposteriorne vjerojatnosti $p(C_j|\mathbf{x})$, odnosno vjerojatnosti pripadanja primjera $\mathbf{x}$ klasi C_j. Vjerojatnost se izračunava posredno, temeljem zajedničke gustoće $p(\mathbf{x}|C_j)$ koristeći Bayesovo pravilo:

$$P(C_j|\mathbf{x}) = \frac{p(\mathbf{x}, C_j)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|C_j)P(C_j)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|C_j)P(C_j)}{\sum_{k=1}^{K} p(x|C_k)p(C_k)}, \tag{4.1}$$

gdje je $P(C_j)$ apriorna vjerojatnost klase, a $p(\mathbf{x}|C_j)$ klasom uvjetovana gustoća, odnosno izglednost klase.

Optimalna klasifikacijska odluka jest ona koja maksimizira aposteriornu vjerojatnost $P(C_j|\mathbf{x})$, odnosno klasificira primjer $\mathbf{x}$ u najizgledniju klasu C_j. Ova hipoteza još se i naziva maksimalna aposteriorna hipoteza (engl. *maximum a posteriori probability, MAP*):

$$h(\mathbf{x}) = \underset{C_k}{\operatorname{argmax}}\, p(\mathbf{x}|C_k)P(C_k). \tag{4.2}$$

U slučaju diskretnih ulaznih varijabli, što je i slučaj u ovome radu, potrebno je procijeniti parametre diskretnih razdioba $P(\mathbf{x}|C_j)$ gdje je $\mathbf{x} = (x_1, \ldots, x_n)$ te $P(C_j)$ kako bi model bio ostvaren. Broj procjenitelja koje je potrebno pronaći za prvu razdiobu jest $2^n - 1$, dok je ukupan broj parametara koje je potrebno procijeniti za drugu razdiobu $K - 1$, ako je ukupan broj različitih klasa K. Dakle, ukupno je potrebno procijeniti $(2^n - 1)K + K - 1$ parametara, što eksponencijalno raste s brojem dimenzija n. Takav model biti će sklon prenaučenosti, odnosno imat će visoku varijancu te će loše generalizirati. Stoga je problem potrebno pojednostavniti uvođenjem pretpostavki.

Naivan Bayesov klasifikator (engl. *naive bayes classifier*) pojednostavljuje problem tako što pretpostavlja uvjetnu nezavisnost varijabli za zadanu klasu, odnosno:

$$P(x_i|x_j, C) = P(x_i|C). \tag{4.3}$$

Također, potrebno je primijetiti da je vjerojatnost $P(x_1, \ldots, x_n|C_j)$ moguće faktorizirati primjenom pravila lanca:

$$\begin{aligned} P(x_1, \ldots, x_n|C_j) &= P(x_1)P(x_2|x_1)P(x_3|x_1, x_2) \ldots P(x_n|x_1, \ldots, x_n - 1) \\ &= \prod_{k=1}^{n} P(x_k|x_1, \ldots, x_k - 1). \end{aligned} \tag{4.4}$$

Primjenom izraza (4.3) te (4.4) dobiva se izraz:

$$P(x_1, \ldots, x_n|C_j) = \prod_{k=1}^{n} P(x_k|x_1, \ldots, x_k - 1) = \prod_{k=1}^{n} P(x_k|C_j). \tag{4.5}$$

Uvevši pretpostavku o uvjetnoj nezavisnosti varijabli za danu klasu, broj parametara koje je potrebno procijeniti smanjen je na samo $K - 1 + \sum_{k=1}^{n}(K_k - 1)K$ gdje je K_k broj mogućih vrijednosti značajke x_k te sada linearno ovisi o n. Iako pretpostavka o uvjetnoj nezavisnosti uglavnom ne vrijedi, naivan Bayesov klasifikator vrlo dobro funkcionira. Glavna prednost ovog modela jest jednostavnost učenja modela pošto se nepoznati parametri jednostavno procjenjuju metodom najveće izglednosti čija rješenja leže u zatvorenoj formi, što znači da nema potrebe za korištenjem iterativnih optimizacijskih metoda za pronalazak rješenja. Konačan model dan je sljedećim izrazom:

$$h(x_1, \ldots, x_n) = \operatorname*{argmax}_j P(C_j) \prod_{k=1}^{n} P(x_k|C_j). \tag{4.6}$$

4.2.2. Logistička regresija

Nasuprot Bayesovog klasifikatora, logistička regresija (engl. *logistic regression*) je diskriminativan model. Takvi modeli izravno modeliraju granicu između klasa, što ih čini puno jednostavnijima u odnosu na generativne modele u smislu broja parametara koje je potrebno procijeniti. Model logističke regresije dan je sljedećim izrazom:

$$h(\mathbf{x}|\mathbf{w}) = \sigma(\mathbf{w}\phi(\mathbf{x})) = \frac{1}{1 + \exp(-\mathbf{w}\phi(\mathbf{x}))}. \tag{4.7}$$

Logistička regresija također je probabilistički model čiji se izlazi mogu interpretirati kao aposteriorne vjerojatnosti klase $P(C_j|\mathbf{x})$, odnosno kao vjerojatnosti pripadanja primjera određenoj klasi. No, ta vjerojatnost temelji se na pretpostavci da su klase normalno distribuirane te da imaju dijeljenu kovarijacijsku matricu. Ako je ta pretpostavka pogrešna, vjerojatnosti procjene neće biti točne. Izraz probabilističkog izlaza logističke regresije dan je sljedećim izrazom:

$$h(\mathbf{x}|\mathbf{w}) = \sigma(\mathbf{w}\phi(\mathbf{x})) = P(C_1|\mathbf{x}). \tag{4.8}$$

Osim toga, za razliku od linearnog poopćenog modela temeljenog na metodi najmanjih kvadrata, logistička regresija ne kažnjava ispravno klasificirane primjere koji su daleko od granice već će za takve primjere na izlazu modela vjerojatnost biti blizu ili jednaka 1.0.

Kako bismo naučili model logističke regresije, potrebno je pronaći $\mathbf{w}$, vektor težina koje definiraju hiperravninu koja razdvaja granice klasa. Kao i kod drugih algoritama nadziranog strojnog učenja, određivanje odnosno optimizacija parametara svodi se na minimizaciju funkcije pogreške na skupu za učenje. U ovom slučaju, rješenje nije u zatvorenoj formi, stoga ga je potrebno pronaći koristeći neke od iterativnih metoda optimizacije kao npr. gradijentni spust.

Diskriminativni modeli skloni su prenaučenosti – previše se prilagođavaju podacima za učenje te stoga loše generaliziraju, tj. loše klasificiraju ostale još neviđene podatke. Kako bi se spriječila prenaučenost modela, u model se dodatno ugrađuje mjera složenosti kojom se kontrolira složenost modela. Takav postupak naziva se regularizacija. Ovim postupkom kontrolira se ujedno i broj značajki. Ako je njihov broj veliki, regularizacijskim parametrom potisnut će se nebitne značajke, odnosno njihove vrijednosti biti će svedene ka nuli. Nadalje, kod linearno odvojivih problema, regularizacijom se sprečava prestrmi nagib sigmoide, te se nastoji zadržati blagi prijelaz između klasa (Šnajder i Dalbelo Bašić, 2012).

4.2.3. Stroj potpornih vektora

Kao i logistička regresija, stroj potpornih vektora (engl. *support vector machine, SVM*) također je diskriminativan model, ali bez probabilističkog izlaza. Za razliku od perceptrona, metode koji za linearno odvojivi problem uvijek pronalazi proizvoljnu hipotezu koja je konzistentna s primjerima za učenje, SVM uvodi kriterij maksimalne margine kako bi riješio problem proizvoljnosti hipoteze. Granica između dviju klasa koju SVM pronalazi postavljena je tako da je prostor između pozitivnih i negativnih primjera što veći. Nalaženje maksimalne margine koja je dana izrazom (4.9), svodi se na problem konveksne optimizacije uz ograničenja. Točnije, riječ je o problemu kvadratnog programiranja.

Ako je dan skup primjera za učenje i njihovih oznaka $(x_i, y_i)_{i=1,\ldots,k}$ pri čemu je $x_i \in \mathbb{R}$ te $y_i \in \{1, -1\}^k$, tada je potrebno pronaći ono rješenje problema koje zadovoljava

sljedeća ograničenja:

$$\begin{aligned} \underset{\mathbf{w},w_0,\eta_i}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{k}\eta_i \\ y_i(\mathbf{w}^T x_i + b) \geq 1 - \eta_i \\ \eta_i \geq 0 \end{aligned} \tag{4.9}$$

Kao što je rečeno, SVM pronalazi hiperravninu s maksimalnom marginom koja linearno odvaja pozitivne od negativnih primjera. Sve što je potrebno zapamtiti nakon pronađene hiperravnine jesu primjeri koji se nalaze na margini, odnosno koji se nazivaju potporni vektori. Nadalje, parametrom C kontrolira se kazna za krivo klasificirane primjere. Veća vrijednost parametra C označava strožu kaznu za neispravno klasificirane primjere, što znači da će model biti kompleksniji, tj. sklon prenaučenosti. Parametar C je zapravo hiperparametar modela kojeg je potrebno optimirati, najčešće n-terostrukom validacijom.

Također, moguće je i koristiti jezgrene funkcije. Jezgrene funkcije omogućavaju stvaranje nelinearnog modela, odnosno preslikavaju značajke u više dimenzije, gdje je izglednije da će primjeri biti linearno razdvojivi. Osnovne jezgrene funkcije su: linearna, polinomijalna te radijalna bazna funkcija. Linearna jezgra definirana izrazom (4.10) daje linearan model. Polinomijalna jezgra uključuje sve kombinacije ulaznih varijabli do uključivo stupnja p. Radijalne bazne funkcije ovise samo o udaljenosti između primjera. Poseban slučaj radijalne bazne funkcije jest Gaussova jezgra dana izrazom (4.12). Gaussova jezgra mjeri sličnost dvaju primjera temeljem njihove udaljenosti u ulaznom prostoru. Za slične primjere vrijedi $\kappa(\mathbf{x}, \mathbf{x}') \to 1$ te su oni blizu jedan drugome u prostoru značajki. Za potpuno različite primjere vrijedi $\kappa(\mathbf{x}, \mathbf{x}') \to 0$. Kod radijalnih baznih funkcija potrebno je dodatno optimirati i parametar γ, koji kontrolira kojom brzinom $\kappa(\mathbf{x}, \mathbf{x}')$ teži k nuli u ovisnosti o njihovoj udaljenosti. Stoga je potrebno provesti iscrpno pretraživanje parametara γ i C kako bi se dobio optimalan model.

$$\kappa(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\mathbf{T}\mathbf{x}'. \tag{4.10}$$

$$\kappa(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\mathbf{T}\mathbf{x}' + 1)^p. \tag{4.11}$$

$$\kappa(\mathbf{x}, \mathbf{x}'') = \exp\{-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\} = \exp\{-\gamma\|\mathbf{x} - \mathbf{x}'\|^2\}. \tag{4.12}$$

4.3. Predobrada podataka

4.3.1. Tokenizacija

S obzirom na kratkoću tvitova, postupak rastavljanja tvita na rečenice nije potreban zato što se tvit u najvećem broju slučajeva sastoji samo od jedne rečenice. Stoga je prvi korak predobrade podataka tokenizacija, odnosno rastavljanje rečenice na niz znakova. U ovom slučaju korištena je jednostavna metoda tokenizacije, tj. razdvajanje tokena po razmaku između pojedinih znakova. Zbog nekorištenja sadržaja poveznica (npr. URL poveznice na slike i video) te „re-tvitanja“, odnosno direktne komunikacije s nekim drugim korisnikom (npr. @TheVoice), svi takvi tokeni zamijenjeni su riječima LINK i USER. Također, potrebno je bilo i identificirati posebno pozitivne i negativne emotikone. Lista emotikona preuzeta je s poveznice[1]. Svaki emotikon zamijenjen je konstantom EMOT_POS te EMOT_NEG, no uz pamćenje samog emotikona, odnosno njegovog sadržaja.
Izlaz iz tokenizatora je lista tokena koji se koriste kod ekstrakcije značajki ili se dodatno normaliziraju u sljedećoj fazi predobrade.

4.3.2. Normalizacija

Sljedeći korak predobrade podataka jest normalizacija tvitova. Kao ulazni parametar funkcija za normalizaciju prima listu tokena, dobivenih pomoću funkcije za tokenizaciju. Za svaki token u listi, ako token nije emotikon, on se normalizira. Najprije se sva velika slova u tokenu zamjenjuju malim slovima. Primjerice token *BRAVOOOO* se pohranjuje kao *bravoooo*. Nakon toga, ako token sadrži uzastopno ponavljanje nekih od znakova, npr. *bravoooo*, tokenu se odbacuju sva uzastopna ponavljanja istog znaka, npr. *bravo*.
S obzirom na kasnije korištenje tokena s i bez zaustavnih riječi,[2] napravljene su dvije inačice funkcije za normalizaciju, odnosno ona koja ostavlja takve riječi za kasniju upotrebu te ona koja ih odbacuje.
Proces predobrade podataka završava normalizacijom tvitova nakon čega se obavlja ekstrakcija značajki.

[1] `https://en.wikipedia.org/wiki/List_of_emoticons`

[2] Zaustavne riječi su riječi koje imaju gotovo isključivo gramatičku funkciju, ili se javljaju vrlo često u nekom dokumentu. Zaustavne se riječi ne prikazuju u popisu lema, niti u popisu različnica. Primjeri zaustavnih riječi su engleske riječi *the, a, I* i hrvatske riječi *je, pa, to*. (2006 FER, 2006)

4.3.3. Ekstrakcija značajki

Svaka korištena značajka implementirana je kao zasebna funkcija. S obzirom na to da se sentiment traži na razini cijele poruke odnosno tvita, svaka funkcija kao ulazni parametar prima listu tokena tvita te vraća par vrijednosti – ime značajke te vrijednost značajke. Vrijednosti značajke izražene su cijelim pozitivnim brojevi (engl. *integer*) ili kao Booleove vrijednosti istine ili laži (engl. *boolean*). Primjer jedne značajke jest *broj pozitivnih emotikona* koji vraća broj tokena koji se nalaze u listi pozitivnih emotikona. Lista ulaznih tokena može biti sastavljena od normaliziranih tokena ili od običnih tokena dobivenih tokenizacijom. To naravno ovisi o samoj značajki. Primjerice, značajka *broj tokena pisanih svim velikim slovima* (engl. *all caps*), prima nenormalizirane tokene kako bi vratila ispravnu vrijednost. Ako bi značajka primila normalizirane tokene, vratila bi pogrešnu vrijednost, odnosno 0 zbog toga što su u postupku normalizacije sva slova tokena zamijenjena malim slovima. U nastavku se nalazi popis korištenih značajki grupiranih u logične cjeline:

1. Značajke specifične za tvitove – značajke koje su karakteristične za tvitove, a koje ne nalazimo u standardnim tekstovima:
   - broj hashtagova (cijeli broj) – npr. *#TheVoiceHRT, #timIvan*;
   - broj riječi sa uzastopnim ponavljanje istog slova (cijeli broj) – npr. *bravooo*;
   - broj veznika (cijeli broj) – broj veznika pronađenih u tvitu. Značajka je uvedena prvenstveno zbog pozitivno/negativne sentiment klase koja najčešće sadrži veznike poput: a, ali, nego, iako, osim i sl.;
   - broj tokena (cijeli broj);
   - broj rečenica (cijeli broj) – broj rečenica u tvitu. Značajka je implementirana na primitivan način, koristeći točku kao znak odvajanja između rečenica, stoga nije vrlo precizna;
   - emotikoni – značajke vezane uz emotikone dobri su indikatori sentimenta u tvitovima jer ih korisnici najčešće koriste kako bi izrazili svoju emocionalnu reakciju prema nekome ili nečemu. Korištene značajke vezane uz emotikone su:
     (a) broj pozitivnih emotikona (cijeli broj) – npr. :), :D, =), 8) itd.;
     (b) zadnji token pozitivan emotikon (Booleova vrijednost);
     (c) broj negativnih emotikona (cijeli broj) – npr. :(, ;(, :‘(, :S itd.;
     (d) zadnji token negativan emotikon (Booleova vrijednost) ;

(e) sadrži li tvit i pozitivan i negativan emotikon (Booleova vrijednost);

(f) sadrži li tvit emotikon (Booleova vrijednost).

- broj ponavljajućih znakova ! (cijeli broj);
- broj ponavljajućih znakova ? (cijeli broj);
- broj ponavljajućih znakova !? (cijeli broj);
- zadnji znak !? (Booleova vrijednost);
- broj riječi pisanih velikim slovima (cijeli broj) – *BRAVOOOO, SUPEEERR*.

2. Značajke vezane uz leksikon
   - broj pozitivnih riječi (cijeli broj) – dobar, najbolja, divan, drag, lijepa itd.;
   - broj negativnih riječi (cijeli broj) – bezvezan, čudan, jadno, preloše itd.;
   - sadrži li tvit i pozitivne i negativne riječi (Booleova vrijednost);
   - sadrži li tvit i pozitivne i negativne riječi i veznik (Booleova vrijednost);
   - sentiment zadnjeg tokena (cijeli broj) – ako je zadnji token pozitivan vrati 0, ako je negativan vrati 1, inače vrati 2.

3. N-grami – značajke se odnose na prisustvo određenog broja uzastopnih tokena (n-grami tokena) ili znakova tokena (n-grami znakova) u tvitu. N-grami se implementiraju kao vektori te se najprije stvaraju na skupu za učenje. Nakon što su stvoreni nad skupom za učenje, moguće je generirati n-gram značajke za pojedini tvit. Npr. ako se niz riječ *dobar nastup* pojavljuje u skupu za učenje kao značajka bigrama, tada se za trenutni tvit na poziciji odgovornoj za taj niz riječi postavlja vrijednost istine, odnosno da tvit sadrži taj bigram. N-grami su efikasno implementirani koristeći slabo popunjene matrice (engl. *sparse matrix*) iz paketa Scipy. Korištene su sljedeće n-gram značajke:
   - 1-grami tokena (vektor Booleovih vrijednosti) – prisutnost pojedinog tokena, npr. sadrži li tvit riječ *bezvezan*;
   - 2-grami tokena (vektor Booleovih vrijednosti) – prisutnost niza od dva tokena, npr. *bezvezan nastup*;
   - 3-grami tokena (vektor Booleovih vrijednosti) – prisutnost niza od tri tokena, npr. *stvarno bezvezan nastup*;
   - 4-grami tokena (vektor Booleovih vrijednosti) – prisutnost niza od 4 tokena, npr. *stvarno bezvezan nastup !!!*;

– 3-grami znakova (vektor Booleovih vrijednosti) – prisutnost niza od 3 znakova, npr. *bez, ezv, zve, vez, eza, zan*;
– 4-grami znakova (vektor Booleovih vrijednosti) – prisutnost niza od 4 znakova, npr. *bezv, ezve, zvez, veza, ezan*;
– 5-grami znakova (vektor Booleovih vrijednosti) – prisutnost niza od 5 znakova, npr. *bezve, ezvez, zveza, vezan*.

5. Eksperimentalno vrednovanje

Ovo poglavlje započinje opisom evaluacijskih mjera korištenih pri vrednovanju modela. Sljedeće potpoglavlje opisuje način na koji je provedeno iscrpno učenje modela, dok su rezultati i vrednovanje modela prikazani u idućem potpoglavlju. Na kraju je dana analiza pogrešaka klasificiranih primjera od strane najboljeg modela.

5.1. Evaluacijske mjere

Za potrebe evaluacije modela u ovom radu korištene su standardne evaluacijske mjere koje se koriste pri evaluaciji klasifikacijskih modela. U nastavku je dana definicija i formula za izračun svake od mjera.
Preciznost (engl. *precision,* P) je mjera koja opisuje koliki je udio ispravno klasificiranih primjera (engl. *true positives,* TP) u skupu pozitivno klasificiranih primjera koji čine ispravno klasificirani primjeri (TP) i pozitivno klasificirani primjeri koji su zapravo pogrešno klasificirani (engl. *false positives,* FP). Mjera preciznosti definirana je sljedećim izrazom:

$$P = \frac{TP}{TP + FP}. \tag{5.1}$$

Suprotno mjeri preciznosti, mjera odziva (engl. *recall*, R) definirana je kao udio broja pozitivnih primjera u skupu svih pozitivnih primjera skupa nad kojim se model evaluira. Takav skup čine ispravno klasificirani primjeri, TP te primjeri koji su pozitivni, ali su modelom klasificirani kao negativni (engl. *false negatives,* FN). Mjera odziva dana je formulom:

$$R = \frac{TP}{TP + FP}. \tag{5.2}$$

Kombinacijom prethodno opisanih mjera, preciznosti i odziva, definira se F_1-mjera

(engl. F_1 *score*). F_1-mjera definirana je kao harmonijska sredina između preciznosti i odziva. Važnost pojedine komponente, odnosno preciznosti i odziva, u izrazu (5.3) kontrolira se parametrom β, koji se najčešće postavlja na vrijednost $\beta = 1$ što označava podjednaku važnost obje komponente (Van Rijsbergen, 1979).

$$F_\beta = \frac{(1+\beta^2)PR}{\beta^2 P + R} \underset{\beta=1}{=} \frac{2PR}{P+R} \tag{5.3}$$

Prethodno spomenute mjere najčešće se izračunavaju koristeći matrice zabune tako da se uvrste potrebne vrijednosti iz matrice u formule pojedinih mjera. Primjer takve matrice prikazan je u tablici 5.1. Primjerice, kako bi se izračunala mjera preciznosti potrebno je iz matrice zabune iščitati potrebne vrijednosti: broj ispravno klasificiranih primjera TP, koji se nalazi u prvom retku i prvom stupcu matrice te broj lažno pozitivno klasificiranih primjera FP koji su zapravo trebali biti klasificirani kao negativni. Nakon što su potrebne vrijednosti odčitane iz matrice, koristeći formulu za izračun preciznosti (5.1) te odgovarajuće vrijednosti izračunava se preciznost.

Tablica 5.1: Primjer matrice zabune

		Stvarna klasa	
		p	**n**
Predviđena klasa	**p′**	True Positive	False Postive
	n′	False Negative	True Negative

U slučaju višeklasne klasifikacije, kao što je to slučaj u ovome radu, moguće je definirati mikro i makro inačice svih spomenutih mjera. Makro-mjere izračunavaju se najprije posebno za svaku klasu te se konačna vrijednost makro-mjera izračunava kao prosjek svih vrijednosti pojedine klase. Slučaj je obrnut kod mikro-mjera, gdje se konačna vrijednost mjere izračunava odmah koristeći sve klase. U tom slučaju uz parametar $\beta = 1$, P, R i F_1 imaju jednake vrijednosti. Razlog različitih inačica mjera jest taj što u slučaju neujednačenog skupa klase s manjim brojem primjera uopće nemaju

utjecaja na mikro rezultat. To ujedno i ukazuje da je makro-mjera stroža, odnosno u pravilu poprima manje vrijednosti.

5.2. Učenje modela

Cilj ovog rada jest izrada klasifikacijskog modela koji klasificira tvit temeljem sentimenta na razini cijele poruke, odnosno mišljenja korisnika u tvitu u jednu od predefiniranih klasa sentimenta. Tvit može biti klasificiran u jednu od četiri moguće klase sentimenta: pozitivnu, negativnu, neutralnu te pozitivno/negativnu. Za izradu modela korištene su tri različite klasifikacijske metode strojnog učenja: naivan Bayesov klasifikator, logistička regresija te stroj s potpornim vektorima. Sve metode opisane su u prethodnom poglavlju 4.2. Pri izradi modela iskorištene su postojeće implementacije svih metoda u programskom paketu sklearn, opisanom u poglavlju 4.1. Primarni fokus stavljen je na pronalazak najboljeg modela te najboljih značajki za modele te na njihovom iscrpnom vrednovanju.

Za potrebe učenja i vrednovanja modela, korišten je označeni skup opisan u poglavlju 3.4. Skup za učenje sastoji se od 80% označenog skupa podataka, odnosno 2373 primjera, dok je ostatak korpusa od 594 primjera iskorišten za konačno vrednovanje modela. Označeni skup podataka je nasumično ispremiješan prije same podjele na skup za učenje te vrednovanje kako bi se izbjegle moguće pristranosti. Također, zastupljenost klasa sentimenta u oba skupa gotovo je podjednaka te je ujedno i jednaka izvornom skupu. Primjerice, udio tvitova označenih s neutralnom klasom sentimenta u označenom skupu je 16.35%, dok isti udio u skupovima za učenje i vrednovanje iznosi 16.64%, odnosno 15.15%.

Nakon podjele primjera na skup za učenje te validaciju svaki je primjer najprije predobrađen. Svaki tvit iz skupa je tokeniziran i normaliziran zbog potrebe određivanja značajki iz tvita. Također, kako bi se odredio utjecaj zaustavnih riječi na klasifikaciju tvita, napravljena su dva zasebna učenja modela. U prvom slučaju u postupku normalizacije zaustavne riječi se odbacuju, dok se u drugom ostavljaju.

U sljedećem koraku napravljena je ekstrakcija značajki. Značajke su grupirane prema logičnom smislu. Nadalje, kako bi se razmotrio samostalni utjecaj nekih od značajki, one su ostavljene same u grupi, kao primjerice unigrami. Popis od ukupno 13 grupa značajki nalazi se u tablici 5.2. U recima tablice popisane su sve značajke korištene u ovom radu, te su detaljno opisane i potkrijepljene primjerima u poglavlju 4.3.3. U stupcima tablice nalaze grupe značajki. Pripadnost pojedine značajke nekoj grupi pri-

kazana je simbolom $+$. Pojedine značajke razmatrane su u izvornom obliku, što znači da nisu normalizirane, poput značajki *# riječi pisanih velikim slovima* i *# riječi s ponavljajućim slovima*. Nenormalizirane značajke korištene su kako bi se dobila ispravna vrijednost značajke koju ta značajka predstavlja. Primjerice, ako se u tvitu nalazi riječ *BRAVOOO*, nakon normalizacije dobila bi se riječ *bravo*, čime bi prethodno spomenute značajke bile postavljene na neispravnu vrijednost. Značajke su grupirane logički. Primjerice, grupa značajki $f1$ sastoji se od značajki koje su specifične za tvitove, $f2$ sadržava značajke vezane uz emotikone, grupa $f3$ sastavljena od značajki vezanih uz stvoreni leksikon riječi pozitivnog i negativnog sentimenta dok grupa $f4$ sadržava sve značajke vezane uz n-grame, tj. n-grame tokena te n-grame slova. Nadalje, u grupi $f5$ nalaze se sve značajke, dok se u preostalim grupama razmatraju neki zanimljivi slučajevi i kombinacije grupa značajki. Primjerice, grupa $f8$ sastoji se samo od unigrama riječi što se često koristi i kod primitivnih metoda klasifikacije.

Kao metoda odabira značajki (engl. *feature selection*) korištena je metoda χ^2 (engl. *chi squared*). Navedena metoda najčešće se koristi u statistici te se njome određuje nezavisnost dvaju događaja, u ovom slučaju između pojedine značajke i klase sentimenta. Veća vrijednost χ^2 sugerira da su značajka i klasa zavisni te bi takva značajka trebala biti zadržana. U ovom radu eksperimentira se s različitim vrijednostima broja značajki koje se odabiru χ^2 testom. Te vrijednosti su sljedeće: 500, 1000, 2000, 3500, 5000, 7500, 10000, 12500, 15000, 20000, 25000, 30000, 50000 te odabir svih značajki. Ove vrijednosti odabrane su empirijski temeljem dobivenih rezultata nakon nekoliko probnih pokretanja modela.
Nakon odabira najboljih značajki slijedi učenje i odabir modela. Za odabir modela koristi se 10-terostruka unakrsna validacija kako bi se odabrali optimalni parametri, odnosno najbolji model. Za svaki od modela izvršeno je pretraživanje po rešetci (engl. *grid search*) temeljem predefiniranih parametara pretrage za pojedini model. Za model logističke regresije optimira se regularizacijski parametar C. Raspon vrijednosti parametra C je $10^{-2} - 10^{6}$, pri čemu se u svakom koraku parametar uvećava za 10. Za svaku grupu pretražuje se ukupno 80 konfiguracija. Za model stroja s potpornim vektorima korištena je posve druga konfiguracija. Mogući parametri tog modela su sljedeći: jezgra stroja, parametar C te dodatno u slučaju radijalne jezgre parametar γ. Za parametar jezgre moguć je odabir dvije vrste jezgre, linearne te radijalne. Za parametar C, koji određuje način kažnjavanja neispravno klasificiranih primjera, koriste se sljedeće vrijednosti: 1, 10, 50, 100, 200, 500, 1000. U slučaju radijalnih jezgri, optimira se i parametar γ čije vrijednosti mogu poprimiti vrijednosti koje su potencije

Tablica 5.2: Korištene značajke i grupe značajki

Značajka/Grupa	$f1$	$f2$	$f3$	$f4$	$f5$	$f6$	$f7$	$f8$	$f9$	$f10$	$f11$	$f12$	$f13$
(*n*) # hashtagova	+				+	+	+		+		+	+	+
(*t*) # riječi s ponavljajućim slovima	+				+	+	+		+		+	+	+
(*n*) # veznika	+				+	+	+		+		+	+	+
(*n*) # rečenica	+				+	+	+		+		+	+	+
(*n*) # pozitivnih emotikona	+	+			+	+	+		+		+	+	+
(*n*) # negativnih emotikona	+	+			+	+	+		+		+	+	+
(*t*) sadrži li tvit emotikon	+	+			+	+	+		+		+	+	+
(*t*) sadrži li tvit pozitivan i negativan emotikon	+	+			+	+	+		+		+	+	+
(*t*) zadnji znak pozitivan emotikon	+	+			+	+	+		+		+	+	+
(*t*) zadnji znak negativni emotikon	+	+			+	+	+		+		+	+	+
(*t*) # ponavljajući znakova !	+				+	+	+		+		+	+	+
(*t*) # ponavljajući znakova ?	+				+	+	+		+		+	+	+
(*t*) # ponavljajući znakova !?	+				+	+	+		+		+	+	+
(*t*) zadnji znak ?!	+				+	+	+		+		+	+	+
(*t*) # riječi pisanih svim velikim slovima	+				+	+	+		+		+	+	+
(*t*) # tokena	+				+	+	+		+		+	+	+
(*n*) # pozitivnih riječi			+		+	+	+		+		+	+	+
(*n*) # negativnih riječi			+		+	+	+		+		+	+	+
(*n*) sadrži li tvit pozitivne i negativne riječi			+		+	+	+		+		+	+	+
(*n*) sadrži li tvit pozitivne i negativne riječi i veznik			+		+	+	+		+		+	+	+
(*n*) zadnja rijec pozitivna, negativna ili neutralna			+		+	+	+		+		+	+	+
(*n*) 1-grami (tokena)				+	+	+	+	+			+	+	+
(*n*) 2-grami (tokena)				+	+	+			+	+	+		+
(*n*) 3-grami (tokena)				+	+						+	+	+
(*n*) 4-grami (tokena)				+	+								
(*n*) 3-grami (znakova)				+	+								
(*n*) 4-grami (znakova)				+	+								
(*n*) 5-grami (znakova)				+	+								+

broja 10 iz raspona vrijednosti $10^{-2} - 10^{3}$. Za model stroja s potpornim vektorom ukupno je pretraženo 180 konfiguracija.

Nakon unakrsne validacije i odabira najboljeg modela, model je vrednovan na primjerima iz skupa za testiranje. Vrednovanje modela napravljeno je korištenjem prethodno opisanih evaluacijskih mjera: P, R i F_1.

Skraćeni prikaz opisanog postupka učenja prikazan je na slici 5.1 u nastavku. Isječak programskog koda sa slike nije u potpunosti identičan stvarnom programskom kodu zbog sažetog zapisa i preglednosti, no prikazuje glavni tok koji se izvodi pri učenju modela.

Dodatno, radi usporedbe implementirana je primitivna metoda za klasifikaciju (engl. *baseline model*). Metoda klasificira tvit ovisno o pronađenom emotikonu ili riječi iz rječnika u tvitu. Ako tvit sadrži pozitivan emotikon ili riječ iz rječnika pozitivnih riječi te negativan emotikon ili riječ iz rječnika negativnih riječi, tvit se klasificira u pozitivnu/negativnu klasu. Ako tvit sadrži samo pozitivan emotikon ili riječ iz rječnika pozitivnih riječi, tvit se klasificira u pozitivnu klasu. Ako tvit sadrži samo negativan emotikon ili riječ iz rječnika negativnih riječi, tvit se klasificira u negativnu klasu. Konačno ako tvit ne sadrži niti jedan tip emotikona te nijednu riječ definiranu u rječnicima pozitivnih i negativnih riječi, tvit se klasificira u onu klasu sentimenta koja je najzastupljenija u primjerima za učenje, odnosno pozitivnu.

```
def main():
  # ucitava konfiguracijsku .json datoteku sa specifikacijom
  # modela (metode i paramatara) te grupama znacajki
  config = load_config()

  data = load_dataset()
  data = normalize(data)

  # f1 - f13
  for featuregroup in config['features']:
    featureset = extract_features(featuregroup)

    # 500, 1000, 2000, ... 50000, sve
    for featuresize in config['featuresizes']:
      if featuresize > len(featureset):
        continue

      # odabire se najbolji |featuresize| znacajki prema Chi2
      featureset = chi2(featureset)

      # podijeli primjere na skup za ucenje i tesitranje
      # u omjeru 0.8, odnosno 0.2
      # primjeri su svaki puta podijeljeni na isti nacin (1)
      train, test = split(featureset, ratio=0.8, state=1)

      # NB, Logit, SVM
      for model in config['models']:
        # 10-erosturka unakrsna validacija sa specificiranim
        # parametrima
        classifier = learn(model, model['params'])

        # klasificiraj primjere
        predicted = classifier.predict(test)

        # racuna i vraca P, R, F1 te konfuzijsku matricu
        results = evaluate(test, predicted)

        # pohrani dobivene rezultate za najbolji model
        # za odredenu grupu znacajki
        store(featuregroup, featuresize, model, results)
```

Slika 5.1: Isječak koda za učenje modela

5.3. Rezultati

U ovom potpoglavlju prikazani su rezultati dobiveni prethodno opisanim postupkom učenja modela. Prvi dio potpoglavlja predstavlja rezultate koji se odnose na odabir najboljih grupa i broja značajki. U drugom djelu detaljnije su prikazani rezultati pojedinog modela za najbolju grupu značajki.

5.3.1. Odabir značajki

Rezultati koji se odnose na odabir najbolje grupe i broja značajki prikazani su tablicama 5.3 i 5.4. Obje tablice prikazuju najbolje dobivene rezultate vrednovanja nad skupom za testiranje koji se sastoji od 594 primjera tvitova. Svaki redak tablice prikazuje najbolje dobivenu konfiguracija u smislu broja značajki za tu grupu značajki. Npr., ako je najbolje dobiveni rezultat za grupu značajki $f6$ bio model SVM koristeći 7500 najboljih značajki dobivenih χ^2 testom, onda se za ostale modele (NB, LG) također prikazuje njihov rezultat dobiven nad istim brojem značajki za tu grupu, iako je moguće da su ti modeli koristeći neki drugu veličinu značajki postigli bolji rezultat. No, promatrajući sve modele, konkretnije u pokaznom slučaju model SVM, takav rezultat nije najbolji.

U tablici 5.3 prikazani su rezultati modela koji nisu koristili zaustavne riječi kao značajke primjera. Rezultati pokazuju da je model SVM daleko bolji od ostalih razmatranih modela te daje najbolje rezultate za sve grupe značajki osim za grupe značajki $f1$ te $f2$ gdje daje isti rezultat kao i NB te LG. Najbolji rezultat koji postiže model SVM jest 0.852 za grupu značajki $f13$ koristeći 25000 najboljih značajki dobivenih korištenjem metode χ^2.

Suprotno tablici 5.3, tablica 5.4 prikazuje rezultate modela koji koriste zaustavne riječi kao značajke tvita. Detaljnija usporedba oba provedena vrednovanja nije moguća, pošto su za različite grupe značajki najbolji modeli dobiveni korištenjem ne nužno istog broja značajki. Primjerice, za grupu značajki $f4$ u prvom eksperimentu najbolji model postiže se uporabom 20000 najboljih značajki, dok je drugim eksperimentom pokazano da se najbolji model postiže koristeći 25000 najboljih značajki. No, iz mjerenja je jasno vidljivo da se drugim eksperimentom postižu bolji rezultati, pri čemu je najbolji rezultat također dobiven kao i u prvom, uporabom SVM-a za grupu značajki $f13$ koristeći 25000 najboljih značajki te iznosi 0.877. Također, modelom SVM-a ponovno su postignuti najbolji rezultati, izuzev grupa značajki $f1$, $f5$ te $f7$.

Tablica 5.3: Točnosti modela za različite grupe i veličine značajki (zaustavne riječi odbačene)

Grupa	#	NB	LG	SVM
$f1$	all	**0.724**	**0.724**	0.722
$f2$	all	**0.724**	**0.724**	**0.724**
$f3$	all	0.741	0.749	**0.763**
$f4$	20000	0.800	0.811	**0.832**
$f5$	15000	0.800	0.830	**0.838**
$f6$	5000	0.786	0.803	**0.825**
$f7$	2000	0.805	0.798	**0.810**
$f8$	1000	0.785	0.786	**0.798**
$f9$	3500	0.753	0.774	**0.825**
$f10$	3500	0.741	0.739	**0.828**
$f11$	7500	0.764	0.798	**0.850**
$f12$	5000	0.776	0.803	**0.837**
$f13$	25000	0.815	0.808	**0.852**

Tablica 5.4: Točnosti modela za različite grupe i veličine značajki (sa zaustavnim riječima)

Grupa	#	NB	LG	SVM
$f1$	all	**0.726**	0.722	0.721
$f2$	all	**0.724**	**0.724**	**0.724**
$f3$	all	0.726	0.742	**0.754**
$f4$	25000	0.822	0.822	**0.840**
$f5$	20000	0.818	**0.832**	0.815
$f6$	7500	0.795	0.793	**0.857**
$f7$	1000	**0.818**	0.806	0.806
$f8$	1000	0.785	0.783	**0.806**
$f9$	5000	0.749	0.769	**0.840**
$f10$	5000	0.746	0.751	**0.830**
$f11$	10000	0.766	0.791	**0.867**
$f12$	7500	0.778	0.796	**0.862**
$f13$	25000	0.828	0.816	**0.877**

5.3.2. Evaluacije modela

Rezultati naučenih modela koji koriste najboljih 25000 značajki iz grupe značajki $f13$ prikazani su u nastavku tablicama 5.5, 5.6 i 5.7. U recima tablica prikazani su rezultati za pojedine mjere (P, R, F_1) za pojedinu klasu sentimenta i određeni model. Također, prikazani su rezultati za modele koji su koristili zaustavne riječi te za one koji to nisu. Vrijednosti pojedinih mjera se izračunavaju pomoću vrijednosti iz matrica konfuzije koje ovdje nisu prikazane. Predzadnji redak tablica prikazuje vrijednosti mikro-mjera pojedinog modela. Vrijednosti svih mikro-mjera izračunavaju se koristeći formule navedene u 5.1. Primjerice, ako računamo vrijednost mikro-mjere preciznosti P, potrebno je odrediti omjer ispravno klasificiranih primjera i ukupnog broja svih klasificiranih primjera. Koristeći konfuzijske matrice, lako se dolazi do tog omjera tako da se vrijednosti elemenata na dijagonali matrice zbroje te podijele s ukupnim brojem svih primjera u skupu za testiranje. U zadnjem retku tablica prikazane su vrijednosti makro-mjera pojedinog modela. Vrijednost pojedine makro-mjere izračunava se kao

prosjek dobivenih vrijednosti svih klasa, što se lako izračunava i iz samih tablica.
U tablici 5.5 prikazane su vrijednosti mjere preciznosti postignute korištenim modelima. Iz tablice je vidljivo kako su najprecizniji modeli NB te SVM. Ako se razmatraju modeli koji su koristili zaustavne riječi kao značajke, u vidu mikro-preciznosti daleko najbolji rezultat postiže model SVM s $P_{micro} = 0.87$, dok se najbolja makro-preciznost postiže modelom NB s $P_{macro} = 0.77$. Za modele koji ne koriste zaustavne riječi kao značajke daleko najbolju preciznost u oba slučaja, tj. za mikro-preciznosti i makro-preciznost postiže model SVM s vrijednostima $P_{micro} = 0.84$ odnosno $P_{macro} = 0.72$.

Tablica 5.5: Preciznosti modela za grupu značajki $f13$

	Model sa zaustavnim riječima			Model bez zaustavnih riječi		
Sentiment klasa	NB	LR	SVM	NB	LR	SVM
Pozitivna (+)	0.84	0.84	**0.93**	0.83	0.83	**0.92**
Pozitvno/negativna (+-)	**0.73**	0.50	0.71	0.20	0.43	**0.78**
Negativna (-)	**0.71**	0.68	0.66	0.59	**0.69**	0.54
Neutralna (O)	**0.82**	0.73	0.76	**0.91**	0.74	0.67
P_{micro}	0.82	0.79	**0.87**	0.79	0.78	**0.84**
P_{macro}	**0.77**	0.68	0.76	0.63	0.67	**0.72**

U tablici 5.6 dani su rezultati odziva pojedinih modela. Najbolje rezultate postiže model SVM za oba slučaja sa mikro-odzivima od $R_{micro} = 0.88$ za prvi, odnosno $R_{micro} = 0.85$ za drugi slučaj, dok su vrijednosti makro-odziva $R_{macro} = 0.68$ za prvi, tj. $R_{macro} = 0.59$ za drugi slučaj. Zanimljivo je primijetiti kako odziv modela NB za oba promatrana slučaja iznosi $R = 1.00$ za pozitivnu klasu sentimenta. Vrijednosti preostalih dvaju modela za tu klasu također su blizu te vrijednosti. No, za razliku od odziva za pozitivnu klasu, odziv modela NB za pozitivno/negativnu klasu sentimenta je samo $R = 0.03$. Moguće tumačenje ovog rezultata jest malen broj primjera za učenje za tu klasu kojih je samo 30. Također, odziv za spomenutu klasu jednako je loš i kod ostalih modela.

Zadnjom tablicom 5.7 prikazane su vrijednosti F_1-mjere. Budući da se F_1-mjera iz-

Tablica 5.6: Odzivi modela za grupu značajki $f13$

	Model sa zaustavnim riječima			Model bez zaustavnih riječi		
Sentiment klasa	NB	LR	SVM	NB	LR	SVM
Pozitivna (+)	**1.00**	0.97	0.98	**1.00**	0.97	0.98
Pozitvno/negativna (+-)	0.27	0.10	**0.40**	0.03	0.10	**0.23**
Negativna (-)	0.55	0.43	**0.70**	0.52	0.45	**0.59**
Neutralna (O)	0.34	0.50	**0.64**	0.34	0.43	**0.58**
R_{micro}	0.83	0.82	**0.88**	0.81	0.81	**0.85**
R_{macro}	0.54	0.50	**0.68**	0.47	0.48	**0.59**

računava kombinacijom mjera P i R, lako je zaključiti da model SVM daje najbolje vrijednosti što je i vidljivo iz tablice. Vrijednosti koje se postižu modelom SVM-a sa zaustavnim riječima kao značajkama su $F_{micro} = 0.87$ te $F_{macro} = 0.71$. Za model bez zaustavnih riječi dobiveni rezultati su $F_{micro} = 0.84$ te $F_{macro} = 0.62$.

Rezultati za primitivnu metode prikazani su u tablici 5.8. Usporedbom rezultata dobivenih korištenjem primitivne metoda klasifikacije vidljivo je da modeli strojnog učenje daju puno bolje rezultate nego primitivne metode.

Tablica 5.7: Vrijednosti F_1-mjera modela za grupu značajki $f13$

	Model sa zaustavnim riječima			Model bez zaustavnih riječi		
Sentiment klasa	NB	LR	SVM	NB	LR	SVM
Pozitivna (+)	0.91	0.90	**0.95**	0.91	0.89	**0.95**
Pozitvno/negativna (+-)	0.39	0.17	**0.51**	0.06	0.16	**0.36**
Negativna (-)	0.62	0.53	**0.68**	0.55	0.55	**0.57**
Neutralna (O)	0.48	0.59	**0.70**	0.50	0.55	**0.62**
F_{micro}	0.80	0.79	**0.87**	0.78	0.78	**0.84**
F_{macro}	0.60	0.54	**0.71**	0.50	0.53	**0.62**

Tablica 5.8: Rezultati primitivnog klasifikatora

	Primitivan klasifikator		
Sentiment klasa	P	R	F_1
Pozitivna (+)	**0.80**	**0.95**	**0.87**
Pozitvno/negativna (+-)	0.23	0.40	0.34
Negativna (-)	0.49	0.45	0.47
Neutralna (O)	0.00	0.00	0.00
$Micro$	0.48	**0.74**	0.68
$Macro$	0.38	**0.45**	0.42

Dobivenim rezultatima ustanovljeno je da je model SVM daleko najbolji model u usporedbi s preostalim dvama modelima i to koristeći 25000 najboljih značajki iz grupe značajki $f13$ uključujući i zaustavne riječi.

5.4. Analiza pogrešaka

U ovom potpoglavlju napravljena je analiza pogrešaka klasifikacije najboljeg modela opisanog u prethodnom poglavlju. Model ispravno klasificira 521 od ukupno 594 pri-

mjera iz skupa za učenje.

Za najbolji model SVM, u tablici 5.9 je prikazana matrica zabune. Reci matrice prikazuju oznake klasa sentimenta koje su primjerima dodijelili označivači, dok se u stupcima matrice nalaze oznake klasa sentimenta koje tvitovima dodjeljuje klasifikator. Na glavnoj dijagonali matrice izraženi su primjeri koji su ispravno klasificirani (TP), odnosno i označivač i klasifikator su takve primjere označili istom klasom sentimenta. U nastavku poglavlja dani su česti primjeri pogrešnih klasifikacija te su opisani mogući razlozi tih pogrešaka.

Tablica 5.9: Matrica zabune SVM modela ($kernel = rbf, C = 10, \gamma = 0.01$)

		Oznaka klasifikatora			
		$+$	O	$-$	$+-$
Oznaka označivača	$+$	420	7	1	2
	O	24	58	8	0
	$-$	4	6	31	3
	$+-$	6	5	7	12

Ukupan broj neispravno klasificiranih primjera u skupu za validaciju je 73, od čega je 10 pogrešnih klasifikacija pozitivnih primjera, 32 neutralna, 13 negativna te 18 primjera klase pozitivno/negativna.

Pogreška u označavanju jedan je od uzroka pogrešne klasifikacije. Mogući uzroci pogreška u označavanju jesu nepažnja ili dekoncentriranost označivača zbog predugog označavanja. Analizom neispravno klasificiranih tvitova primijećeno je nekoliko slučajeva. Sljedeći tvit, označen je kao pozitivan, iako je zapravo trebao biti svrstan u klasu pozitivan/negativan kako je i klasificiran modelom: *@dmgj93: DINO VOLIM TE NISI NEKI PJEVAČ ALI ZGODAN SI A TO JE NAJBITNIJE #TheVoiceHRT*. Nadalje do moguće pogreške u označavanju moglo je doći zbog subjektivnosti shvaćanja sentimenta. Najčešće dvojbe koje se tiču subjektivnosti bile su između pozitivne i neutralne klase sentimenta. Primjeri tvitova koji su mogli biti označeni na oba načina su: *Zabava i napetost ;) #thevoicehrt* ; *Tim Ivan ;) #TheVoiceHRT*; *A nije bas tako lose otpevao #xfactoradria* itd. Takvi primjeri su također najčešće pogrešno klasificirani u pozitivnu klasu premda su neutralni, a glavni razlog tome jest što sadržavaju pozitivan emotikon kao u sljedećim primjerima: *#TheVoiceHRT Idemo Pjerino i Egoon :)*; *Su-*

bota vecer uz #TheVoiceHRT :); *Na Rtl jos dišu :) ..#xfactoradria* itd.
Sarkastični tvitovi također su pogrešno klasificirani u pozitivnu klasu, iako su zapravo svi bili označeni kao negativni. Primjeri sarkastičnih tvitova koji su neispravno klasificirani su: *Obozavam sto je ova obukla sve iz garderobera i ubrus iz kujne #xfactoradria*; *#TheVoiceHRT iva super ti je odjevna kombinacija, jel ti to ostalo od maskara?* itd.
Nadalje, proširenjem rječnika pozitivnih i negativnih riječi, zasigurno bi se poboljšali i rezultati. Primijećeno je da su neki primjeri koji su sadržavali riječi jakog pozitivnog sentimenta neispravno klasificirani, primjerice riječi poput *zanimljivo*, *vrhunski*, *extra*, *bravo*, *highlight* itd. Također, rječnik je sastavljen samo od hrvatskih riječi. No kako se korpus sastoji i od tvitova na srpskom jeziku, dodavanjem srpskih riječi pozitivnog i negativnog sentimenta također bi doprinijelo poboljšanju klasifikacije.

6. Zaključak

Zbog velike dostupnosti informacija te znanja pohranjenog u njima, potrebno je razviti efikasne metode kojima računalo može automatizirati i ubrzati procesiranje podataka te izvoditi zaključke. Jedan od takvih zadataka jest i analiza sentimenta u mikroblogovima. Zbog velike popularnosti i neprestanog generiranja sadržaja društvenih mreža lako je doći do određenog znanja koje može biti upotrebljeno u različite svrhe.

Cilj ovog rada bio je istražiti i proučiti postupke za analizu sentimenta u mikroblogovima na hrvatskome jeziku, s naglaskom na metode temeljene na nadziranom strojnom učenju. U tu svrhu proučeni su postojeći radovi na engleskom jeziku koji se bave analizom mišljenja u mikroblogovima. Za potrebe izrade modela, provedeno je označavanje korpusa koji se u konačnici sastoji od 2967 tvitova označenih jednom od četiri moguće klase sentimenta: pozitivna, negativna, neutralna te pozitivno/negativna. Tvitovi su iz zatvorene domene te se odnose na dva popularna pjevačka natjecanja, The Voice te XFactorAdria. Nadalje, implementirana je programska podrška za ekstrakciju značajki te klasifikaciju tvitova koristeći tri različite metode: Naivan Bayesov klasifikator, logističku regresija te stroj s potpornim vektorima. Konačno, provedeno je eksperimentalno vrednovanje modela, pri čemu su najbolji rezultati dobiveni korištenjem modela SVM, s postignutom F_1-mjerom od 87%.
S obzirom na obećavajuće rezultate, predlaže se izrada web-aplikacije koja će omogućiti praćenje trenda popularnosti pojedinog kandidata u sljedećim natjecanjima s obzirom na sadržaj tvitova koji se odnose na pojedinu osobu. Nadalje, predlaže se i daljnje proširenje korpusa odnosno ujednačavanje broja tvitova različitih klasa sentimenta. Također, predlaže se i izrada modela za analizu sentimenta tvitova iz drugih popularnih i zanimljivih domene poput politike, sporta i sl.

LITERATURA

TEXT MINING SYSTEM 2006 FER. Strojno potpomognuto indeksiranje dokumenata, April 2006. URL `http://textmining.zemris.fer.hr/rektor2006/index.html`.

Inc 2015 Twitter. Twitter statistics, May 2015. URL `https://about.twitter.com/company`.

Johan Bollen, Huina Mao, i Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

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

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

J Richard Landis i Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, stranice 159–174, 1977.

Saif M Mohammad, Svetlana Kiritchenko, i Xiaodan Zhu. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. U *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (SEMSTAR'13)*, 2013.

Jan Šnajder i Bojana Dalbelo Bašić. Naivan bayesov klasifikator. U *Strojno učenje*, stranice 43–52. Sveučilište u Zagrebu, Fakultet elektrotehnike i računarsta, 2012.

Jan Šnajder i Bojana Dalbelo Bašić. Linearni diskriminativni modeli. U *Strojno učenje*, stranice 71–83. Sveučilište u Zagrebu, Fakultet elektrotehnike i računarsta, 2012.

Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, i Theresa Wilson. Semeval-2013 task 2: Sentiment analysis in twitter. 2013.

Alexander Pak i Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. U *LREC*, svezak 10, stranice 1320–1326, 2010.

Bo Pang i Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. U *Proceedings of ACL*, stranice 115–124, 2005.

Bo Pang, Lillian Lee, i Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. U *Proceedings of EMNLP*, stranice 79–86, 2002.

Sara Rosenthal, Preslav Nakov, Alan Ritter, i Veselin Stoyanov. Semeval-2014 task 9: Sentiment analysis in twitter. *Proc. SemEval*, 2014.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M Mohammad, Alan Ritter, i Veselin Stoyanov. Semeval-2015 task 10: Sentiment analysis in twitter. U *Proceedings of the 9th International Workshop on Semantic Evaluation, SemEval*, 2015.

Benjamin Snyder i Regina Barzilay. Multiple aspect ranking using the good grief algorithm. U *HLT-NAACL*, stranice 300–307, 2007.

Peter D Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. U *Proceedings of the 40th annual meeting on association for computational linguistics*, stranice 417–424. Association for Computational Linguistics, 2002.

CJ Van Rijsbergen. Information retrieval. dept. of computer science, university of glasgow. *URL: citeseer. ist. psu. edu/vanrijsbergen79information. html*, 1979.

Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, i Shrikanth Narayanan. A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. U *Proceedings of the ACL 2012 System Demonstrations*, stranice 115–120. Association for Computational Linguistics, 2012.

Princeton University WordNet. Opinion - definition, May 2010. URL `http://wordnet.princeton.edu`.

Dodatak A
Upute za označavanje tvitova

A.1. Opis zadatka

Analiza sentimenta ili analiza mišljenja bavi se proučavanjem ljudskih mišljenja, stavova i emocija koji su izraženi o nekom entitetu (osobi, mjestu, organizaciji), problemu, događaju ili općenito o nekoj temi.
Zadatak ovog diplomskog rada jest izraditi model za automatsko prepoznavanje sentimenta u tvitovima. Prikupljeni tvitovi koji će se koristiti pri označavanju vezani su za dvije vrlo popularne emisije: *The Voice*[1] i *XFactorAdria*[2]. U tvitovima vezanim za te emisije, ljudi najčešće izražavaju svoje mišljenje o nekom od kandidata ili više njih, no mišljenje može biti izraženo i o cjelokupnoj emisiji, članovima žirija, voditeljima itd. U sklopu označavanja, potrebno je označiti kakav je taj stav, odnosno kakvo je mišljenje izraženo u tvitu.

A.2. Korištenje aplikacije

U nastavku je ukratko opisan način korištenja aplikacije (slika 1). Za svaki tvit obavezno je označiti sentiment i jezik tvita. Ako jedno ili oboje od navedenog nije označeno, aplikacija neće dopustiti prelazak na sljedeći tvit (slika 2). Dodatno, ako je u tvitu izražen sarkazam ili ironija, potrebno je označiti polje za sarkazam.
Nakon označavanja tvita, na sljedeći tvit prelazi se pritiskom na gumb *Next* u aplikaciji. U bilo kojem trenutku moguće je vratiti se na prethodno označavane tvitove pritiskom na gumb *Previous* u aplikaciji te ga ponovno označiti. U bilo kojem trenutku, moguće je prekinuti označavanje tako da se ugasi aplikacija pritiskom na gumb *X*. Kod sljedećeg pokretanja aplikacije, za označavanje će biti učitani samo oni tvitovi koji još nisu

[1] `http://voice.hrt.hr/`
[2] `http://xfactoradria.com/`

bili označeni.

A.3. Pravila označavanja

Pri označavanju sentimenta u tvitovima potrebno je označiti jednu od četiri moguće kategorije tvita s obzirom na sentiment u tvitu: pozitivan, negativan, neutralan te pozitivno/negativan sentiment. Pri klasifikaciji tvita potrebno je najprije tražiti korisnikovo mišljenje o nekom entitetu ili njegov stav prema nečemu ako ga ima (npr. *#TheVoiceHRT Dinooo najbolji ikad!*). Nadalje, ako u tvitu nije jasno izraženo mišljenje potrebno je označiti ga prema smislu.
U nastavku su dane upute na koji način je potrebno klasificirati tvit u svaku od spomenutih kategorija:

1. **Neutralan**
   - Ako tvit ne sadrži mišljenje korisnika, odnosno osobe koja je napisala tvit, tvit je potrebno označiti kao neutralan. Primjeri ovakvih tvitova su:
     - *Beng beng Dečaku po glaviii ... OmG ;) @MarijaAmaranth #TheVoiceHRT*
     - *jos 3 minute #TheVoiceHRT ;)*
   - Tvitove koji označavaju neki pozitivan ili negativan događaj, npr., prolazak ili ispadanje kandidata, loše izvedba i sl., potrebno je označiti kao neutralne, ako taj događaj nije poznat iz samog tvita. Primjerice tvitove *melanie :(* ili *Zeznula se :(* potrebno je označiti kao neutralne, iako su najvjerojatnije tvitovi negativni jer su spomenuti kandidati u tvitu ispali iz daljnjeg natjecanja te su korisnici, koristeći emotikone, izrazili tugu za njihovim ispadanjem. U suprotnome, ako je negativan ili pozitivan događaj spomenut u tvitu, tvit se klasificira prema smislu, odnosno prema izraženom stavu o tome događaju.
   - Ostali primjeri:
     - *E bas mi stric pao sa stolicee #TheVoiceHRT*
     - *Pjerino je izgubljeni brat od @_nevits_ :) #TheVoiceHRT*
     - *Pripreme za današnju live emisiju su na svom vrhuncu! Pogledajte kako izgleda naš #backstage :) #TheVoiceHRT http://t.co/vySZ8GvROH*
     - *Pozdrav svim natjecateljima,mentorima i voditeljima :)) #thevoicehrt*

2. **Pozitivan**
   - Ako tvit sadrži pozitivno mišljenje korisnika, odnosno osobe koja je napisala tvit, tvit je potrebno označiti kao pozitivan. Primjeri ovakvih tvitova su:
     - *Nikad nisam vidjela toliko talenata na jednom mjestu. Imamo najbolje pjevace i pjevacice na zemlji a to smo tek sad dokazali (: #TheVoiceHRT*
     - *wow MAKI zvuči svjetski!!! #TheVoiceHRT*
     - *iva savrsenaaa #TheVoiceHRT*
     - *Bravo Jure :') #TheVoiceHRT*
     - *#TheVoiceHRT Elena imas tako predivan i sanzan glas sve pohvale*
     - *Mateo ima super glas #thevoicehrt*
3. **Negativan**
   - Ako tvit sadrži negativno mišljenje korisnika, odnosno osobe koja je napisala tvit, tvit je potrebno označiti kao negativan. Primjeri ovakvih tvitova su:
     - *ovo ide jako jako brzo. Malo su kratke pjesme. #TheVoiceHRT*
     - *žače žače težače, loš odabir pjesme za melanie :(*
     - *zasad su mi svi bezveze, mlako skroz :/*
     - *preduga reklama za karlovačko featuring Dečak -.- #TheVoiceHRT*
     - *Mrzim pp #TheVoiceHRT*
4. **Pozitivan/Negativan**
   - Ako tvit sadrži pozitivno i negativno mišljenje korisnika, odnosno osobe koja je napisala tvit, tvit je potrebno označiti kao pozitivan i negativan. Ovakvi tvitovi najčešće su suprotne rečenice te se u njima najčešće iznosi mišljenje o više kandidata. Primjeri ovakvih tvitova su:
     - *Ova super pjeva ali brate ova 'aljina na njoj... #thevoicehrt*
     - *#TheVoiceHRT Nije me baš očarao kao ostalih dvoje kandidataa,ali stajling mu je superr :)*
     - *Pa zasto Maki nije prosao? I Elena je bila super ali mislim da je on ovaj put bio bolji. Placeem... #TheVoiceHRT*
     - *A svi sto pljuju po Zaku,pjesma koju je Melani pjevala je predivna i ima emociju.Nije kriv odabir pjesme vec Melanin pristup*

#TheVoiceHRT

- *Ajoj... prehladena je.. ma bit ce ona super kao i uvijek! #TheVoiceHRT*

Uz označavanje sentimenta, potrebno je dodatno označiti i jezik kojim je tvit napisan. Moguće je označiti jedan od tri mogućih jezika: hrvatski, srpski, slovenski. Ako je tvit pisan jezikom koji nije naveden, kao jezik je potrebno odabrati opciju ostalo.
Zbog mogućih slučajeva u nastavku su navedena pravila kako označiti iste. Ako je tvit pisan nekim od navedenih jezik, ali sadrži neke strane riječi, tvit je potrebno klasificirati kao da pripada detektiranom jeziku (npr., *Meni je Dora the best #TheVoiceHRT*). U suprotnome, ako je tvit sastavljen od riječi koje ne pripadaju ni jednom od navedenih jezika, tvit je potrebno označiti kao ostalo (npr., *They are soo good wow #TheVoiceHRT*).

Model za analizu sentimenta u tvitovima na hrvatskome jeziku

Sažetak

Analiza sentimenta ili mišljenja je zadatak iz područja obrade prirodnog jezika. Cilj analize sentimenta jest analizirati iznesena mišljenja i stavove korisnika u pisanim tekstovima koji se odnose na neke entitete, događaje ili teme. U okviru ovog diplomskog rada proučeni su postupci za analizu sentimenta u mikroblogovima, s naglaskom na metode temeljene na nadziranom strojnom učenju. Provedeno je označavanje određenog skupa podataka, kojeg čine tvitovi na hrvatskome jeziku, a odnose se na domenu dvaju pjevačkih emisija: The Voice te XFactorAdria. Nadalje, razvijen je model za analizu sentimenta u tvitovima temeljen na postupcima nadziranog strojnog učenja. Provedeno je iscrpno vrednovanje na odgovarajućem skupu podataka, uključujući i analizu značajki. Dobiveni rezultati usporedivi su s rezultatima dobivenim na natjecanju SemEval-2013, uz postignutu F_{micro} od 87%, odnosno F_{macro} od 71%.

Ključne riječi: Obrada prirodnog jezika, analiza sentimenta, analiza mišljenja, mikroblogovi, Twitter, tvit, nadzirano strojno učenje, hrvatski jezik

Sentiment Analysis in Tweets in Croatian Language

Abstract

Sentiment analysis or opinion mining is a task in natural language processing. The goal of the sentiment analysis is to explore users' opinions and their attitudes towards some entities, events or themes. In this research, certain approaches have been studied that consider sentiment analysis in microblogs, focusing on supervised machine learning techniques. A sample of data was annotated, which consisted of tweets in Croatian about two popular music shows: The Voice and XFactorAdria. A system based on the supervised machine learning technique was implemented in order to classify the tweet into one possible sentiment class considering the whole tweet message. The implemented system was then evaluated and yielded results that could be measured with those achieved at SemEval-2013 competition, producing the F_{micro} of 87% and the F_{macro} of 71%.

Keywords: Natural language processing, sentiment analysis, opinion mining, microblogs, Twitter, tweet, supervised machine learning, Croatian