Analiza sentimenta u tekstovima i mikroblogovima o izbjegličkoj krizi

Raguzin, Ana

Sveučilište u Rijeci – Odjel za informatiku

Diplomski studij- Informacijski i komunikacijski sustavi

Ana Raguzin

Analiza sentimenta u tekstovima i mikroblogovima o izbjegličkoj krizi

Diplomski rad

Mentor: izv.prof. dr. sc. Sanda Martinčić – Ipšić

Rijeka, rujan 2018.

//ZADATAK

Sažetak

Mišljenje uvelike utječe na način na koji se osobe ponašaju, djeluju i odlučuju. Prije neke važne odluke, ljudi će pitati za mišljenje svoje bližnje, no razvojem Interneta mijenja se način na koji ljudi izražavaju i traže mišljenja. U današnje vrijeme postoje razni tipovi stranica gdje osobe mogu izraziti svoje stavove: društvene mreže, portali, blogovi, forumi i slično. Međutim, zbog prevelike količine dostupnih podataka teško je dobiti jasno izražene stavove i mišljenja kako pojedinaca tako i ciljanih skupina. Jedan od osnovnih problema predstavlja broj različitih izvora koji sadrže mišljenja što može otežati ljudima traženje onih relevantnih. Što dovodi do sve veća potrebe za automatskim otkrivanjem mišljenja, odnosno analizom sentimenta. Ova vrsta analize postaje sve popularnija u mnogim domenama: politici, zdravstvu, proizvodnji potrošačkih proizvoda, raznim uslugama i slično. Jedan od osnovnih zadataka analize sentimenta je klasifikacija prema polaritetu, odnosno odvajanje tekstova koji sadrže sentiment u tri kategorije: pozitivnu, negativnu i neutralnu. Ovaj rad obuhvaća teoretski dio u kojem su opisani i objašnjeni glavni pojmovi analize sentimenta te istraživački dio koji se bazira na analizi komentara s portala Index i Jutarnji list te mikroblogova vezanih uz temu izbjegličke krize. Analiza je podijeljena u dva dijela: u prvom dijelu je izrađena frekvencijska analiza riječi u rečenici te je izrađen program koji uči i testira klasifikator pomoću pripremljenog korpusa. U rezultatima je dobiveno da su najčešće korištene riječi u komentarima: „bravo“, „čovjek“, „izbjeglica“ i „eu“ , te da klasifikator maksimalne entropije ima najveći postotak točnosti za ovaj korpus. Drugi dio analize se sastoji od usporedbe korištenog vokabulara i ukupnog sentimenta dva skupa podataka koji su prikupljeni u dva različita vremenska perioda. Rezultati su prikazani u obliku histograma i tablica, a program je napisan u programskom jeziku Python. Cilj rada je analizirati podatke i zaključiti koji su stav ljudi imaju u vezi izbjeglica te vidjeti postoji li razlika u stavu kroz određeni vremenski period.

Ključne riječi: analiza sentimenta, Index, Jutarnji list, Twitter, klasifikacija, polaritet

Abstract

Sentiment Analysis of Texts and Tweets Related to War and Immigrant Crises

Opinion has a strong influence on how people behave and therefore is a central part of all human activity. Before making an important decision, most people ask for an opinion of their closest friends and family, but big development of the Internet changed the way people express and search opinions. Nowadays, there are a lot of web pages where a person can express their opionon: social networks, portals, blogs, forums, etc. However, due to large amounts of available data, certain issues may arise. In the first place, there are a lot of different sources of data so people may have difficulties finding the ones that are relevant. Therefore, there is an increasing need for automatic detection of opinion, respectively sentiment analysis. This type of analysis is becoming more popular in many domains: politics, healthcare, production of consumer products, and various other services, etc. One of the tasks of sentiment analysis is classfication according to the polarity, respectively separation of texts in three categories: positive, negative and neutral. This paper consists of theoretical part where main concepts are explained and research part which is based on the analysis of the comments collected from portals „Index“ and „Jutarnji list“ and microblogs, that contain opinions about refugees. The analysis is divided into two parts: the first one contains frequency analysis of words in the sentence and results from a program which is using the prepared corpus for learning and testing the classifiers. Results from the analysis show that the most commonly used words in the comments are: „bravo“, „čovjek“, „izbjeglica“ and „eu“, and that the classifier with the highest percentage of accuracy for this corpus is maximum entropy. The second part of the analysis consists of the comparison of the used vocabulary and overall sentiment from two sets of data collected in two different time periods. Results are shown in the form of histograms and tables, and the program is written in programming language Python. The aim of the paper is to analyze the data and figure out what kind of opinion do people have about refugees, and to see if there is a difference in the opinion over a certain period of time.

Key words: sentiment analysis, Index, Jutarnji list, Twitter, classification, polarity

Sadržaj

1. Uvod

Mišljenja predstavljaju bitan faktor u životima ljudi, na način da upravljaju njihovim ponašanjem i pomažu prilikom procesa odlučivanja. Većina ljudi će prije nego što mora donijeti neku važnu odluku pitati za savjet i mišljenje nekog člana obitelji, prijatelja ili poznanika. Ljudima je oduvijek bila važna informacija „što drugi ljudi misle“ prije nego učine neki veći korak u svom životu ili ukoliko žele znati iskustva drugih o nekom proizvodu ili usluzi. Davno prije Interneta i društvenih mreža, ukoliko bi pojedinac htio čuti tuđe mišljenje o nečemu, on bi uglavnom pitao uski krug prijatelja. S druge strane, kada bi neka organizacija htjela saznati mišljenja svojih korisnika o njenim proizvodima i uslugama, ona bi provodila razna ispitivanja, ankete i fokusne skupine. Međutim, nakon razvitka World Wide Weba i socijalnih mreža gdje imamo konstantan protok ogromne količine podataka koji sadrže mišljenje, stvari se mijenjaju. U današnje vrijeme će se pojedinac koji traži mišljenje o nečemu uglavnom služiti Internetom koji je uvelike promijenio način na koji ljudi izražavaju svoje stavove. Organizacija više nema potrebe za trošenjem resursa za provođenje raznih istraživanja na tržištu, kada postoji puno takvih podataka koji su javno dostupni. Postoje mnoge platforme na kojima ljudi mogu izraziti svoje mišljenje, kao što su: društvene mreže (Facebook, Twitter, Instagram, Youtube, LinkedIn..), razni blogovi, portali, forumi i slično. Ljudi su jako dobro prihvatili ove platforme jer im omogućuju da slobodno izraze svoje mišljenje i stajališta o svakoj temi te im omogućuju jednostavno povezivanje s drugim ljudima i međusobno dijeljenje podataka. Brz i jednostavan pristup informacijama mijenja način na koji ljudi „traže“ tuđa mišljenja. Međutim, velik broj informacija donosi i probleme. Na prvom mjestu to je pronalaženje relevantnih izvora i stavova u ogromnoj količini podataka koja je danas dostupna na Internetu. Stoga pojedincu može predstavljati problem kako te podatke razvrstati i organizirati u neki upotrebljiv oblik. Sukladno tome, nastaje sve veća potreba za automatskim otkrivanjem mišljenja, odnosno analizom sentimenta. U posljednjih nekoliko godina, svjedočili smo kako su tekstovi koji sadrže mišljenja imali velikog utjecaja u društvenim i političkim sustavima pogotovo u vrijeme političkih izbora kao što smo imali prilike vidjeti kod zadnjih predsjedničkih izbora u Sjedinjenim Američkim Državama. Upravo zbog toga, puno tvrtki ulaže u prikupljanje i proučavanje mišljenja na Internetu i mnogim drugim izvorima (putem pozivnih centara, knjigi žalbi, rezultata istraživanja, itd.). Analiza sentimenta zbog svoje velike važnosti u praktičnim primjenama postaje sve popularnija u raznim domenama: politici, zdravstvu, proizvodnji potrošačkih proizvoda, raznim uslugama, društvenim događajima i slično. Većina velikih organizacija izgradile su svoje sustave za analizu mišljenja, a neke od njih su: Google, Amazon, Facebook,

Microsoft, SAS i SAP (Internetski izvori). U analizi sentimenta, bitnu ulogu ima proučavanje društvenih mreža, među kojima se ističe Twitter koji je često korišten za predviđanje izbornih rezultata analizom statusa mikroblogova odnosno „tweetova“[1]. Osim toga, koristio se i za predviđanje uspjeha filmova i tržišta dionica na način da su analizirana „raspoloženja“ ljudi na njihovim stranicama (Agarwal, 2011). Tweetovi su kratki jer imaju ograničenje od najviše 140 znakova, što prisiljava njihove autore da budu izravni u izražavanju svojeg mišljenja, zbog čega je onda često lakše postići veću razinu točnosti analize sentimenta. Poznati autor, Bing Liu, je sa svojom grupom stručnjaka koristio posebno razvijen sustav imena „Opinion Parser“ pomoću kojega je analizirao pozitivna i negativna mišljenja o raznim filmovima na Twitteru. Osim na društvenim mrežama, analiza sentimenta je korištena i u opisivanju društvenih odnosa (npr. kako različiti spolovi osjećaju drugačije emocije), no to je samo jedan dio njene široke praktične primjene (Bing, 2012).

U sklopu ovog diplomskog rada analizirala sam komentare i tweetove na temu izbjeglica, a navedenu analizu sam podijelila u dva dijela. U prvom dijelu sam koristila komentare s portala Index i Jutarnji list gdje sam najprije izračunala pojavljivanja pojedinih riječi i to prikazala pomoću histograma i tablica. Zatim sam pripremila korpus na način da sam ručno klasificirala komentare u tri kategorije. Pripremljeni korpus sam potom iskoristila za učenje i testiranje klasifikatora: Naive Bayes, stablo odlučivanja, maksimalne entropije i support vector machine. U drugom dijelu analize koristila sam tweetove sa specifičnim ključnim riječima koji su prikupljeni kroz određeni period. Cilj ovog dijela bio je usporediti korišteni vokabular i ukupni sentiment tweetova kako bi se vidjela razlika u stavu ljudi na tu temu. U radu će najprije biti objašnjen teoretski dio vezan za analizu sentimenta i klasifikaciju prema polaritetu, a nakon njega slijedi pratkični dio. Sveukupni cilj ovog rada je analizirati dobivene podatke te tako zaključiti kakvo su ljudi mišljenje imali u vezi izbjeglica i postoji li razlika u sentimentu prilikom usporedbe podataka iz dva različita perioda.

[1] Naziv za komentar ili poruku napisanu na društvenoj mreži Twitter

2. Analiza senimenta

Analiza mišljenja iz teksta, poznatija kao analiza sentimenta, je studija koja se bavi analiziranjem stavova i osjećaja ljudi o nekom entitetu i njegovim atributima (Bing, 2015). Pod pojmom entitet misli se na bilo kakav proizvod, uslugu, organizaciju, događaj ili temu o kojem se može izraziti određeno mišljenje ili stav. Rečenice koje sadrže mišljenje, su uglavnom subjektivne, suprotno od objektivnih rečenica koje navode činjenice. Općenito, tekstualne informacije (podatke) možemo svrstati u dvije glavne kategorije: činjenice i mišljenja. Činjenice se definiraju kao objektivni izrazi o osobama, događajima i njihovim svojstvima, odnosno kao nešto što možemo neupitno ustanoviti. S druge strane, mišljenja definiramo kao subjektivne izraze koji opisuju osjećaje koje neka osoba ima o entitetima i događajima (Bing, 2010). Međutim, objektivne rečenice također mogu i implicitno iznositi stav njenih autora na način da opisuju poželjne ili nepoželjne činjenice. Na primjer, u rečenicama: „Jučer sam kupila usisavač i danas se pokvario“ i „Nakon jednog pranja majce, izblijedila je boja“, možemo primjetiti opis dviju nepoželjnih činjenica, što nas dovodi do zaključka da autor ima negativno mišljenje o kupljenom usisavaču i majci. Analiza sentimenta se također bavi analiziranjem ovakvih objektivnih rečenica. Nadalje, analiza mišljenja se često smatra potpodručjem računalne obrade prirodnoga jezika (engl. *Natural-language processing- NLP*) s obzirom da su njena istraživanja usmjerena na pisani tekst, to jest na prepoznavanje teme o kojima ljudi govore i njihove osjećaje prema temama. Analiza mišljenja obuhvaća mnoge postupke NLP-a, kao što su: prepoznavanje leksičkoga i rečeničnoga značenja, ekstrakcija informacija, analiza diskursa i slično. No, NLP nije jedino područje koje se bavi analizom sentimenta, 2000. godine analiza sentimenta postaje jedno od najaktivnijih područja u rudarenju podataka (engl. *data mining*), otkrivanju znanja u podacima, dubinskom pretraživanju Interneta, menadžmentu i mnogim drugim područjima koji se bave tekstualnim podacima (Bing, 2015). Fokus analize sentimenta mijenja se ovisno o disciplini u kojoj je proučavana. Na primjer, u menadžmentu je glavni fokus istraživanje utjecaja mišljenja potrošača na posao tvrtke i kako iskoristiti te stavove za poboljšanje poslovne prakse. S druge strane, cilj NLP-a i rudarenja podataka je dizajnirati učinkovite algoritme i modele za izdvajanje mišljenja iz prirodnog jezika.

Generalno, mišljenja dijelimo na četiri osnovne komponente. Prva je nositelj mišljenja, koji predstavlja osobu ili organizaciju koja je izvor mišljenja. Na primjer, u rečenici: „Ivana voli slušati pjesme Olivera Dragojevića.“, nositelj mišljenja je Ivana. Druga je objekt, koji predstavlja cilj mišljenja- proizvod, osoba, događaj ili tema o kojemu se može izraziti mišljenje.

Treća je mišljenje, odnosno pozitivni, negativni ili neutralni stav nositelja mišljenja o određenom entitetu. Posljednja komponenta je vrijeme kad je mišljenje izraženo (Bing, 2015). Sve četiri komponente su jednako bitne. Na primjer, vremenska komponenta je vrlo važna u praksi jer mišljenje koje je izrečeno prije tri godine nije jednako važno kao ono izrečeno danas. Također, ovisno o tome tko je nositelj mišljenja, mijenja se važnost mišljenja. Primjerice, mišljenje vrlo bitne osobe kao što je predsjednik države je puno važnije nego mišljenje običnog građanina. Nadalje, komponenta „objekt“ može sadržavati skup dijelova i atributa odnosno svojstva. Možemo ga zamisliti kao stablo čiji je korijen sam objekt dok svaka grana predstavlja jedan dio objekta. Na primjer, određeni model mobitela je objekt i on sadrži set atributa kao što su kvaliteta slike, vrijeme trajanja baterije, brzina rada, težina i slično. Stoga, osoba može imati mišljenje o cijelom objektu „Ne sviđa mi se taj mobitel“ ili o nekom svojstvu objekta „Ovaj mobitel ima super kvalitetu zvuka“. Također, treba uvest pojam „aspekt“ koji je naziv za atribute i komponente objekta, npr. kvaliteta zvuka i baterija (Bing, 2015). Nadalje, postoje dvije vrste mišljenja: općenito i specifično. Općenito mišljenje je ono o objektu samom, npr. „Ova majca je jako lijepa“. Dok je specifično mišljenje ono koje imamo o pojedinoj osobini objekta, npr. „Boja majce je odlična“ (Bing, 2012).

Također, mišljenja razlikujemo ovisno o načinu na koje je ono izraženo, stoga postoji regularno i komparativno mišljenje. Regularno mišljenje ima dvije podvrste, a to su izravno i neizravno mišljenje. Izravno mišljenje je ono koje se izražava direktno o objektu, primjerice „Kvaliteta slike je odlična“. Dok se neizravno mišljenje definira kao indirektno izražavanje mišljenja o objektu ili aspektu objekta na temelju pozitivnih i negativnih učinaka na neke druge objekte. Ova se podvrsta regularnog mišljenja često javlja u medicinskoj domeni (Bing, 2015). Na primjer, u rečenici: „Nakon injekcije koja sadrži lijek, zglobovi su me još više boljeli.“ imamo nepoželjni učinak lijeka na zglobove što indirektno daje negativno mišljenje o lijeku. S druge strane, komparativno mišljenje izražava odnose sličnosti ili razlike između dva ili više objekta i/ili dva ili više svojstva objekta. Primjerice, rečenice „Samsung je bolji od Iphone-a.“ i „Samsung je najbolja marka mobitela.“ izražavaju dva komparativna mišljenja. Komparativno mišljenje je uglavnom izraženo pomoću komparativnog ili superlativnog oblika pridjeva ili priloga, no ne uvijek (Dobrescu, 2011). Mišljenja također razlikujemo po jačini. Pozitivno mišljenje izražava osjećaje zadovoljstva, veselja i radosti no ne uvijek u istom intenzitetu. Ljudi koriste dva načina izražavanja svojih osjećaja u tekstu. Prvi način je odabir riječi, na primjer: riječ „dobro“ ima manju vrijednost od riječi „odlično“. Drugi način je pojačavajući ili smanjujući intenzitet izraženog sentimenta korištenjem priloga kao što su „izuzetno“, „jako“,

„užasno“, „malo“ ili „jedva“ i slično (Bing, 2015). Zbog toga se razvila klasifikacija mišljenja na osnovu njihove polarnosti o kojoj nešto više piše u idućem poglavlju.

3. Klasifikacija prema polaritetu sentimenta

Jedan od zadataka analize mišljenja je klasifikacija tekstova koje sadrže sentiment u tri kategorije: pozitivnu, negativnu i neutralnu. Neutralno u ovom smislu pripada objektivnoj kategoriji subjektivne analize, iako će mnogi neutralnost definirati kao mišljenje koje nema jasnu tendenciju prema pozitivnom ili negativnom (Bing, 2015). Za klasifikaciju mišljenja na osnovu polarnosti koristimo leksičke resurse koji sadrže skup podataka s kojim uspoređujemo tekst i tako ocjenjujemo njegovu pozitivnost ili negativnost. Leksički resurs možemo definirati kao listu riječi koje su ocjenjene na temelju izrečene emocije ili na temelju njihove pozitivnosti i negativnosti (Dobrescu, 2011). Riječi se ocjenjuju ocjenama od -5 do 5, gdje ocjenom -5 ocjenjujemo najnegativnije riječi, a ocjenom 5 ocjenjujemo najpozitivnije riječi. Tablica 1 prikazuje nekoliko primjera ocjenjivanja riječi iz leksičkog resursa za hrvatski jezik koji je izradio Marko Modrić u sklopu svog završnog rada (Modrić, 2013). Istraživanje u analizi se usredotočuje na razvoj shema koja bi se koristila za razvoj korpusa za shvaćanje specifičnosti različitih izraza nekog mišljenja u različitim vrstama tekstova u kojima je mišljenje izrečeno (recenzije, blogovi, portali..) (Dobrescu, 2011).

Tablica 1-Primjer leksičkog resursa za hrvatski jezik (Izvor: M. Modrić, *Leksikon za analizu mišljenja iz teksta na hrvatskome jeziku)*

angry	ljut	-2	Izražava negativnu emociju
anger	ljutnja	-2	Izražava negativnu emociju
fury	bijes	-3	Negativnije od riječi ljutnja, ljut
bastard	gad	-5	Najnegativnija ocjena riječi
brilliant	briljantan	4	Izražava pozitivnu emociju
outstanding	izvanredan	5	Najpozitivnija ocjena riječi

Analiza sentimenta klasificira tekstove koje sadrže mišljenja prema polaritetu emocije koja je izražena. Ukoliko je osoba zainteresirana za pronalaženje mišljenja ljudi o nekom određenom filmu, ukupni sentiment će biti dovoljan da ta osoba odluči hoće li ga gledati ili ne. S druge strane, ako osoba želi kupiti određeni proizvod, kao što je mobitel, ukupni sentiment neće biti dovoljan za njenu odluku pošto osoba može biti više zainteresirana za neke od značajki mobitela, primjerice trajanje baterije ili kvaliteta fotoaparata. Stoga možemo zaključiti da analiza mišljenja zahtijeva različite pristupe koje ovise o razini analize, potrebama korisnika i tipu teksta koji se analizira. U idućim potpoglavljima opisane su razine analize sentimenta.

3.1. Analiza sentimenta na razini dokumenta

Ova vrsta analize pretpostavlja da dokument koji sadrži mišljenje (npr. recenzija proizvoda) izražava mišljenja o samo jednom objektu i da je napisan od strane jednog nositelja mišljenja (Bing, 2015). Analiza sentimenta na razini dokumenta je dobila ovaj naziv jer gleda svaki dokument kao cjelinu te ne proučava objekte i aspekte unutar dokumenta. Njen zadatak je klasificirati cijeli dokument ovisno o tome izražava li ukupno pozitivan ili negativan sentiment, odnosno pozitivno ili negativno mišljenje. Različiti autori daju različite pristupe ovoj analizi. Primjerice, Peter Turney analizu koristi za recenzije filmova na način da konačno dobiveni rezultat predstavlja zbroj polariteta pojedinačnih riječi u recenziji (Turney, 2002). Autor Bo Pang daje malo drugačiji pristup klasifikaciji polariteta sentimenta, naime, on se u svom radu koristio Naive Bayes klasifikatorom te je pokazao da korištenje unigrama[2] bolje od korištenje bigrama[3] (Pang, 2002). Godinu dana poslije, autori Pang i Lee klasificiraju recenzije u većim mjerilima vrijednosti, a ne samo pozitivno i negativno. U svome radu koristili su strojno učenje pomoću potpornih vektora (engl. *Support Vector Machines- SVM*). Dobiveni rezultat usporedili su s brojem „zvijezdica“ koje su dane toj recenziji, odnosno ocjenom. Autori Goldberg i Zhu predstavili su grafički pristup klasifikacije sentimenta, gdje je dokument prikazan kao vektor, izračunato na temelju prisutnosti riječi koje sadrže mišljenje. Dokumenti se zatim povežu s najsličnijima sebi te se na kraju klasificiraju na temelju informaciji dobivenih iz grafa i SVM modela (Goldberg, 2006).

[2] frekvencije pojedinih riječi (Izvor: Wikipedia)
[3] frekvencije parova riječi (Izvor: Wikipedia)

3.2. Analiza sentimenta na razini rečenice

Ova razina analize usko je povezana s klasifikacijom subjektivnosti, koja razlikuje rečenice koje sadrže činjenice (objektivne) od subjektivnih rečenica. Međutim, subjektivnost nije ekvivalentna sentimentu ili mišljenju s obzirom da postoji puno objektivnih rečenica koje mogu izražavati mišljenje, kao što je rečenica: „Kupili smo automobil prošli mjesec i pao je brisač za vjetrobran“. S druge strane, mnoge subjektivne rečenice ne izražavaju nikakvo mišljenje, primjerice kao u rečenici: „Mislim da je otišao doma nakon izlaska“. Analiza sentimenta na razini rečenice se uglavnom radi u dva koraka. U prvom koraku se gleda je li rečenica subjektivna ili objektivna. U drugom koraku u slučaju da je rečenica subjektivna, radi se klasifikacija sentimenta prema polaritetu, pod pretpostavkom da svaka rečenica izražava samo jedno mišljenje, pozitivno ili negativno (Dobrescu, 2011). Klasifikacija na ovoj vrsti analize je često teža od one na razini dokumenta budući da su informacije koje sadrži jedna rečenica oskudnije od onih koje sadrži cijeli dokument. Većina klasifikacija na razini dokumenta potpuno izbace neutralnu kategoriju, dok se ona na ovoj razini analize ne može zanemariti pošto se dokument, koji sadrži mišljenje, može sastojati od rečenica koje ne izražavaju nikakvo mišljenje ili sentiment (Bing, 2015). Autori Yu i Hatzivassiloglou koriste ovu vrstu analize s ciljem razdvajanja činjenica od mišljenja (Yu, 2003). S druge strane, Kim i Hovy pokušavaju s obzirom na danu temu naći, pozitivan, negativan i neutralan sentiment te izvor mišljenja, odnosno nositelja mišljenja (Kim, 2004). Autori, nakon što naprave listu sentimenata pomoću leksičkog resursa koji se zove WordNet[4], odabiru rečenice koje sadrže nositelja mišljenja te zatim izračunaju sentiment rečenice.

3.3. Analiza sentimenata na razini značajki

Glavni fokus ove vrste analize nije na jezičnim jedinicama (dokument, paragraf, rečenicu, frazu i slično) već na samom mišljenju, odnosno cilju mišljenja. Shvaćanje važnosti ciljeva mišljenja omogućava nam bolje razumijevanje problema analize sentimenta (Bing, 2015). Unatoč tome što je klasificiranje tekstova na razini dokumenata i rečenice korisno u mnogim slučajevima, ipak postoje određeni primjeri koji traže da se uvede analiza sentimenta na razini značajki. Na primjer, „pozitivan“ tekst o nekom objektu ne znači da autor ima dobro

[4] Izvor: https://wordnet.princeton.edu/

mišljenje o svim značajkama tog objekta. Istu stvar imamo kod „negativnog“ teksta, iako je on ukupno negativan, to ne znači da se autoru ne sviđaju apsolutno sve značajke tog objekta. Dakle, u normalnom tekstu koji izražava nečije mišljenje i stav možemo pronaći pozitivne i negativne aspekte, no generalno mišljenje o objektu može biti pozitivno ili negativno (Dobrescu, 2011). Ova vrsta analize je potpunija i detaljnija od prijašnjih, pošto otkiva mišljenja o svakoj pojedinoj značajki objekta, a ne samo o cijelom objektu. Razlikujemo tri koraka koja se izvode na ovoj razini analize. U prvom koraku je zadatak prepoznati i izdvojiti značajke objekta koje je komentirao nositelj mišljenja, npr. u rečenici „Kvaliteta zvuka na ovom mobitelu je odlična“ gdje je značajka objekta „kvaliteta zvuka“. U drugom koraku utvrđuje se jesu li mišljenja o značajkama pozitivna, negativna ili neutralna, u gore navedenom primjeru bi mišljenje o značajki bilo pozitivno. U posljednjem koraku se grupiraju značajke istog značenja nakon čega se izračuna polaritet i dobije rezultat koji je prikazan u postocima pozitivnog i negativnog mišljenja o svakoj značajki (Dobrescu, 2011).

4. Klasifikatori

Klasifikacija tekstova služi za kategoriziranje dokumenata ili dijelova tekstova. Na temelju riječi koje se upotrebljavaju u tekstu grade se klasifikatori koji tekstu pridružuju određenu oznaku – klasu (Jurafsky, 2000). Postoje dvije metode klasifikacije teksta, a to su ručno napisana pravila i nadzirano strojno učenje. Ručno napisana pravila se temelje na kombinaciji riječi i ostalim značajkama. Točnost ove metode klasifikacije može biti visoka, no pisanje i održavanje pravila je skupo (Arun, 2012). Metoda nadzirano strojno učenje kao ulaz ima dokument d, određeni skup klasa $C = \{c_1, c_2, \ldots c_j\}$ i skup za učenje koji se sastoji od m ručno označenih dokumenata $(d_1, c_1),\ldots,(d_m, c_m)$. Izlaz predstavlja naučeni klasifikator $y:d \rightarrow c$, kao što su primjerice: Naïve Bayes klasifikator, logistička regresija, Support-vector machines, stablo odlučivanja, maksimalna entropija i ostali (Arun, 2012). U ovom diplomskom radu koristila sam četiri klasifikatora čiji ću princip rada ukratko objasniti u idućim potpoglavljima.

4.1. Naivni Bayesov klasifikator

Naivni Bayesov klasifikator temelji se na Bayesovom teoremu i on se smatra statističkim klasifikatorom pošto opisuje vjerojatnost nekog događaja na temelju poznatog uvjeta (Leung Ming, 2007). Bayesov algoritam se oslanja na prikaz dokumenata kojeg nazivamo „bag of words“, odnosno naziv na hrvatskom jeziku glasi „princip vreće riječi“. Ovaj princip osniva se na pretpostavci da se značenja dokumenata mogu dobiti na način da se broje riječi koje se u njima pojave. Naime, ne zanima nas sintaksa ni poredak riječi već se analiza zasniva na neuređenom skupu riječi gdje je tekst reprezentiran kao vreća riječi (Medium, 2018). Bayesov teorem je definiran na sljedeći način: neka je $X = \{x_1, x_2,\ldots, x_n\}$ uzorak čije komponente predstavljaju vrijednosti napravljene na skupu od n atributa i neka je H hipoteza kao što je „uzorak X pripada određenoj klasi C“. Cilj klasifikacije je odrediti P(H|X)-vjerojatnost hipoteze H ukoliko imamo „dokaze“ X, odnosno u ovom slučaju vjerojatnost da uzorak X pripada određenoj klasi C, ukoliko znamo opis X-a (Leung Ming, 2007). Klasifikator će predvidjeti da uzorak X pripada klasi koja ima najveću a posteriori vjerojatnost. Učenje Naivnog Bayesovog modela odnosi se na jednostavno korištenje vjerojatnosti, i u tom smislu se smatra „naivnim“. Na slici 1 možemo vidjeti primjer koji prikazuje jednadžbu koja računa koliko se puta riječ w_i pojavljuje u odnosu na sve riječi u dokumentima s temom c_j:

$$\hat{P}(w_i \mid c_j) = \frac{count\ (w_i, c_j)}{\sum_{w \in V} count\ (w, c_j)}$$

Slika 1- Naive Bayes-jednadžba (Izvor: https://medium.com/@theflyingmantis/text-classification-in-nlp-naive-bayes-a606bf419f8c)

Budući da nam pojavljivanje riječi može biti važnije od frekvencije riječi, potrebno je ukloniti duplikate riječi tipa *w* u svakom dokumentu te se zadržati samo jednu instancu.

4.2. Stablo odlučivanja

Stablo odlučivanja je stablo u kojem svaki čvor predstavlja značajku (atribut), svaka veza (grana) predstavlja odluku/pravilo, a svaki list predstavlja ishod (kategoričku vrijednost) (Medium, 2017). Ideja je podijeliti skup podataka u manje skupove podataka na temelju opisane značajke sve dok se ne dostigne dovoljno mali skup koji sadrži instance sličnih vrijednosti (homogenost). Svaka značajka skupa postaje korijenski čvor (roditelj), a listovi djeca (Saedsayad, 2018). Konačni rezultat je stablo s čvorovima i listovima. Stablo odlučivanja gradi se od vrha prema dnu počevši od korijenskog čvora. Glavni dio algoritma za izgradnju stabla odluka naziva se „ID3“, koji u svom radu koristi entropiju i *information gain.* ID3 algoritam koristi entropiju za izračunavanje homogenosti uzorka. Ukoliko je uzorak homogen, entropija je 0 te ukoliko je uzorak jednakomjerno podijeljen onda mu entropija iznosi 1. Kako bi izgradili stablo odlučivanja potrebno je izračunati dva tipa entropije koristeći frekvenciju. Prva entropija odnosi se na tablicu frekvencije jednog atributa, a druga se odnosi na tablicu frekvencije dva atributa (Saedsayad, 2018). Slika 2 prikazuje jednadžbe kojima računamo njihove vrijednosti. Information gain, odnosno dobivanje informacija temelji se na smanjenju entropije nakon što je skup podataka podijeljen na atribute, a pri tome razlikujemo nekoliko koraka. Prvi korak je računanje entropije cilja. U drugom koraku se skup podataka podijeli na različite atribute i izračuna se entropija za svaku granu. Zatim se dodaje proporcionalno kako bi se dobila ukupna entropija za podjelu. Završna entropija se oduzima od entropije prije razdvajanja, a kao rezultat dobijemo information gain, tj. smanjenje entropije (Saedsayad, 2018). U trećem koraku se odabire atribut koji vraća najveći information gain i on se definira kao čvor za odlučivanje. Nakon toga se skup podataka podijeli na grane i ponavljaja se postupak na svakoj grani. Grana s entropijom 0 je list, a grane koje imaju entropiju veću od 0 se moraju dalje dijeliti. ID3

algoritam se vrši rekurzvino na granama bez listova sve dok se svi podaci ne klasificiraju (Saedsayad, 2018).

$$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i \qquad E(T,X) = \sum_{c \in X} P(c)E(c)$$

Slika 2- Entropija koja se odnosi na tablicu frekvencije jednog atributa (lijevo) /Entropija koja se odnosi na tablicu frekvencije dva atributa (desno)- Izvor: http://www.saedsayad.com/decision_tree.htm

4.3. Maksimalna entropija

Klasifikator maksimalne entropije je probabilistički klasifikator koji za razliku od Naivnog Bayesovog ne pretpostavlja da su značajke uvjetno nezavisne jedna od druge, odnosno ne daje pretpostavke o odnosima između značajki. Ovaj klasifikator se temelji na načelu maskimalne entropije po kojemu se među svim modelima koji pripadaju podacima za učenje, odabire onaj koji imaju najveću entropiju (Datumbox, 2013). Jedna od glavnih formula je prikazana na slici 3.

$$P_{\mathrm{ME}}(c \mid d) := \frac{1}{Z(d)} \exp\left(\sum_i \lambda_{i,c} F_{i,c}(d,c)\right)$$

Slika 3- Maksimalna entropija- formula (Izvor: http://www.aclweb.org/anthology/W02-1011)

Gdje je *Z(d)* normalizacijska funkcija, $F_{i,c}$ funkcija klase/značajke za značajku f_i i klasu *c*, koju definiramo na sljedeći način:

$$F_{i,c}(d,c') := \begin{cases} 1, & n_i(d) > 0 \text{ and } c' = c \\ 0 & \text{otherwise} \end{cases}$$

Slika 4- Funkcija klase/značajke (Izvor: http://www.aclweb.org/anthology/W02-1011)

$\lambda_{i,c}$ su parametri značajki-težina. Definicija „P_{ME}“ nam pokazuje da ukoliko parametar $\lambda_{i,c}$ ima veliku vrijednost to znači da se f_i smatra jakim indikatorom klase *c*. Vrijednosti parametara su postavljene tako da maksimiziraju entropiju inducirane distribucije podložno ograničenju

prema kojem su očekivane vrijednosti funkcija značajki/klase, u odnosu na model, jednake očekivanoj vrijednosti u odnosu na podatke za učenje (Pang, 2002).

4.4. Support vector machine

Support vector machine (metoda potpornih vektora) algoritam može se koristiti za klasifikaciju i regresiju.. U ovom algoritmu se svaki podatak iscrta kao točka u n-dimenzionalnom prostoru (gdje je n predstavlja broj značajki) na način da vrijednost svake značajke bude vrijednost određene koordinate (Analyticsvidhya, 2017). Kod postupka učenja ideja je pronaći „hyperplane“[5] prezentiran vektorom „w“ koji odvaja dokument vektore od ostalih tako da ih stavlja u odvojenu klasu, na način da veličina margine (udaljenosti) između klasa bude što veća. Ukoliko „$c_j \in \{1,-1\}$“ predstavlja točnu klasu dokumenta d_j onda rješenje može biti zapisano na sljedeći način:

$$\vec{w} := \sum_j \alpha_j c_j \vec{d}_j, \quad \alpha_j \geq 0,$$

Slika 5- Formula za računanje vektora w (Izvor: http://www.aclweb.org/anthology/W02-1011)

Gdje se α_j dobiva rješavanjem dvojnog optimizacijskog problema. Vektore α_j, c_j i d_j nazivamo support vectors jer su jedini dokument vektori koji pridonose vektoru w (Pang, 2002). Na kraju se klasifikacija testnih instanci sastoji se od određivanja na koju stranu hyperplane-a vektora w padaju.

[5] Je potprostor, čija je dimenzija manja od one u ambijentalnom prostoru (Izvor: Wikipedia)

5. Analiza podataka o izbjeglicama

Praktični dio ovoga rada bazirao se na analizi prikupljenih komentara i tweetova na temu izbjeglica. Analizu sam odvojila u dva dijela, budući da sam koristila dva različita izvora podataka, koji se također razlikuju i u jeziku kojim su napisani. Prvi dio analize bazira se na komentarima preuzetih 2015. godine s portala *Index* i *Jutarnji list* u vrijeme velikog izbjegličkog vala u Europi. Korištene materijale su pripremili studenti Kathrin Maeusl i Edi Čiković. U sklopu prvog dijela, ručno sam odvojila navedene komentare u tri kategorije ovisno o tome izražava li komentar po meni pretežno pozitivno, pretežno negativno ili neutralno mišljenje u okviru svojeg završnog rada (Raguzin, 2016). U diplomskom radu sam za ručno pripremljeni korpus izradila program za učenje i testiranje klasifikatora. Drugi dio odnosi se na analizu tweetova koji sadrže ključne riječi: *terrorism*, *isis*, *refugees*, *puppy*, *kitty* i *baby*. Cilj je bio usporediti mišljenja ljudi o navedenim temama kroz određeni period. Kako bi napravila tu usporedbu, koristila sam dvije skupine tweetova koji su prikupljeni u vremenskom razmaku od dvije godine, točnije 2016. i 2018. godine. U idućih nekoliko potpoglavlja prikazati ću dobivene rezultate za obje analize počevši od analize komentara s portala Index i Jutarnji list.

5.1. Analiza komentara s portala Index i Jutarnji list

Tablica 2 prikazuje opis korištenog korpusa u prvom dijelu analize. Korpus se sastoji od komentara prikupljenih 2015. godine s portala Index i Jutarnji list, a spremljeni su na način da je svaki komentar u zasebnom tekstualnom dokumentu. Kratica „KJ“ odnosi se na komentare s Jutarnjeg lista, a „KI“ na komentare s Index portala. Prvi stupac tablice sadrži ukupan broj komentara, odnosno tekstualnih datoteka koje sam naknadno podijelila u tri kategorije. U drugom stupcu se nalazi ukupan broj riječi uključujući i zaustavne riječi[6], a treći stupac prikazuje ukupan broj riječi nakon što se izbace zaustavne riječi. Četvrti stupac sadrži broj različitih pojavnica[7] u tekstu uključujući zaustavne riječi, a peti stupac sadrži broj pojavnica nakon što izbacimo zaustavne riječi.

[6] Zaustavne ili stop riječi su riječi prirodnog jezika koje imaju sintaksnu ulogu, a nemaju samostalno značenje kao npr. „ a“, „ ali“, „ još“,..itd

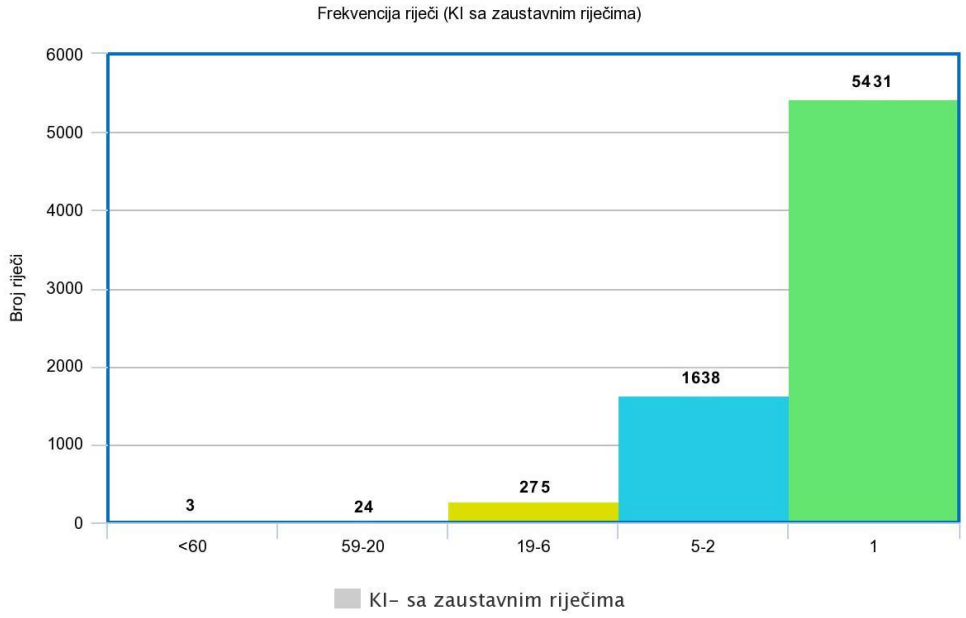
[7] Pojavnice = riječi + interpunkcijski znakovi

Tablica 2-Opis korištenog korpusa (komentari s Index portala i Jutarnjeg lista)

	Ukupan broj komentara	Ukupan broj riječi sa zaustavnim riječima	Ukupan broj riječi bez zaustavnih riječi	Broj različitih pojavnica u tekstu sa zaustavnim riječima	Broj različitih pojavnica u tekstu bez zaustavnih riječi
KJ	1100	27994	26651	11791	11578
KI	1230	19604	18576	8361	8149

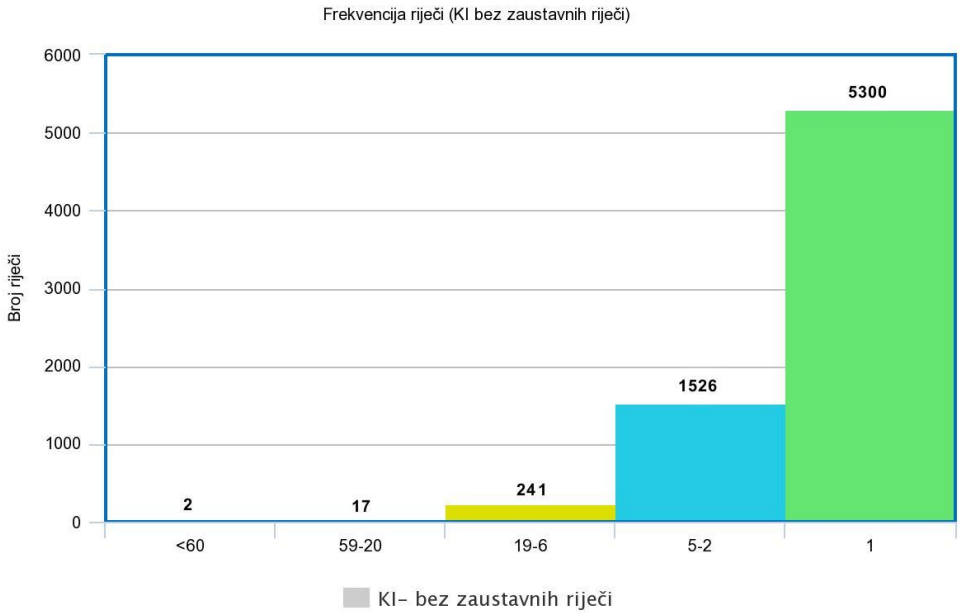
U analizi sam najprije promotrila frekvencije riječi u korpusu, odnosno broj pojavljivanja pojedine riječi. Frekvencije riječi sam dobila korištenjem gotove funkcije „FreqDist“ koja je dio Python biblioteke „nltk“. U nastavku ću prikazati histograme na kojima možemo vidjeti dobivene rezultate te liste riječi s njihovom frekvencijom. Napravila sam dva različita histograma za svaki od portala, u jednom se prikazuju frekvencije riječi uključujući zaustavne riječi, a u drugom bez zaustavnih riječi. Na histogramu os X prikazuje broj pojavljivanja pojedine riječi, a os Y prikazuje ukupan broj riječi s tom frekvencijom. Na primjer, ukupan broj riječi s frekvencijom 1 je 5300. Frekvencije sam odvojila u pet kategorija: više od 60, između 59 i 20, između 19 i 6, između 5 i 2 te 1.

Slika 6 prikazuje histogram koji sadrži podatke o frekvencijama riječi izvučenih iz komentara s portala Index, uključujući i zaustavne riječi. Možemo primjetiti kako najviše riječi ima frekvenciju 1 (5431), što je 73% od ukupnog broja riječi.



Slika 6- Frekvencije riječi (KI sa zaustavnim riječima)

Na slici 7 nalazi se histogram koji prikazuje frekvencije riječi iz komentara s portala Index, nakon što izbacimo zaustavne riječi. Kao i u prethodnom, najviše riječi ima frekvenciju 1 (5300), što je 74% od ukupnog broja riječi.



Slika 7- Frekvencije riječi (KI bez zaustavnih riječi)

U tablicama 3 i 4 možemo vidjeti prvih deset najčešće korištenih riječi u komentarima na portalu Index. Tablica 3 sadrži zaustavne riječi, a tablica 4 ne sadrži. Najčešće korištena riječ je „treba“ s frekvencijom 72, a nakon nje slijedi riječ „eu“ (kratica za Europsku uniju). Riječ „izbjeglice“ ima frekvenciju 43, a riječ „ljudi“ 52.

Tablica 3- Lista riječi s frekvencijom

(KI sa zaustavnim riječima)

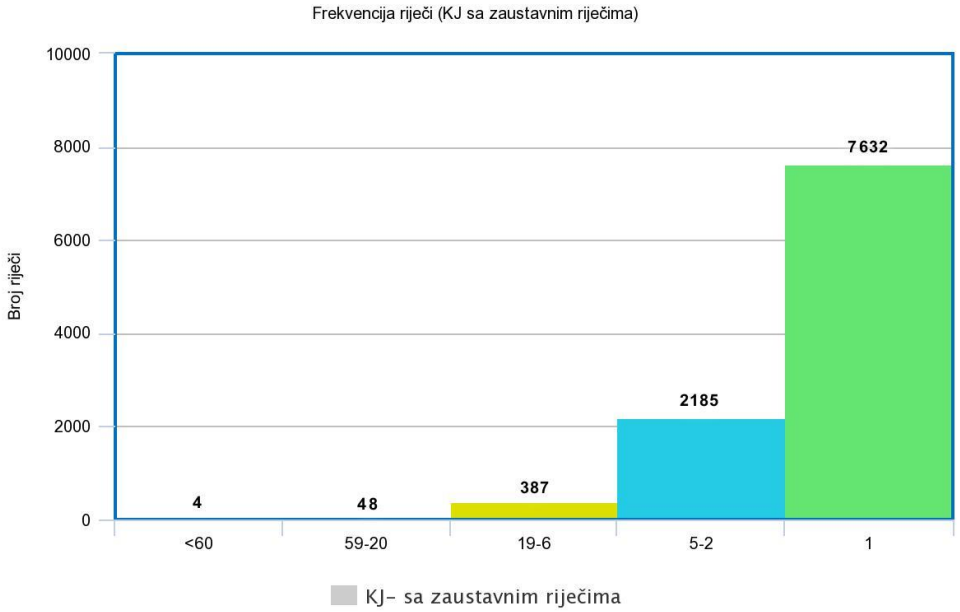
1.	treba	72
2.	eu	70
3.	a	60
4.	i	54
5.	ljudi	52
6.	ce	49
7.	izbjeglice	43
8.	ima	42
9.	ivan	41
10.	u	38

Tablica 4- Lista riječi s frekvencijom

(KI bez zaustavnih riječi)

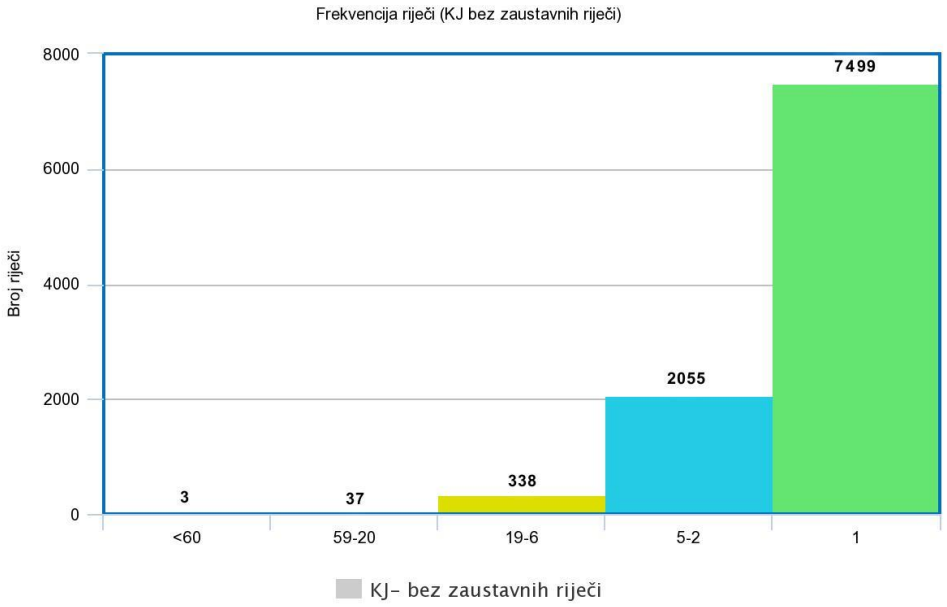
1.	treba	72
2.	eu	70
3.	ljudi	52
4.	izbjeglice	43
5.	ima	42
6.	ivan	41
7.	nema	38
8.	lara	35
9.	wendy	35
10.	ko	33

Na slici 8 nalazi se histogram koji prikazuje frekvencije riječi u komentarima s portala Jutarnji list, uključujući zaustavne riječi. Najviše riječi ima frekvenciju 1 (74%), nakon toga slijede riječi s frekvencijom između 5 i 2 (21%).



Slika 8- Frekvencije riječi (KJ sa zaustavnim riječima)

Slika 9 također se odnosi na podatke o broju pojavljivanja riječi s portala Jutarnji list, no ovaj put bez zaustavnih riječi. Najveći broj riječi ima frekvenciju 1, dok najmanji broj riječi ima frekvenciju veću od 60.



Slika 9- Frekvencije riječi (KJ bez zaustavnih riječi)

Tablice 5 i 6 prikazuju prvih deset najčešće korištenih riječi u portalu Jutarnji list. Možemo primjetiti kako postoji velika sličnost između ove liste i one vezane za komentare s Index portala. Riječ „eu“ ima najveću frekvenciju, a slijedi ju riječ „ljudi“. Riječ „izbjeglice“ se pojavljuje 60 puta u komentarima, a „narod“ 34.

Tablica 5- Lista riječi s frekvencijom

(KJ sa zaustavnim riječima)

1.	eu	90
2.	ljudi	73
3.	i	64
4.	izbjeglice	60
5.	a	57
6.	treba	56
7.	ce	49
8.	ima	48
9.	nema	47
10.	to	42

Tablica 6- Lista riječi s frekvencijom

(KJ bez zaustavnih riječi)

1.	eu	90
2.	ljudi	73
3.	izbjeglice	60
4.	treba	56
5.	ima	48
6.	nema	47
7.	izbjeglica	41
8.	može	36
9.	narod	34
10.	zemlje	34

Tablica 7 prikazuje prosječnu frekvenciju riječi za oba portala. Primjećujemo da je prosjek frekvencija gotovo isti za sve četiri liste. Ukoliko gledamo decimalne vrijednosti, KJ sa zaustavnim riječima ima najveću prosječnu frekvenciju, a KI bez zaustavnih riječi najmanju.

Tablica 7- Prosječne frekvencije

	Prosječna frekvencija
KI sa zaustavnim riječima	1,787
KI bez zaustavnih riječi	1,706
KJ sa zaustavnim riječima	1,825
KJ bez zaustavnih riječi	1,742

5.2. Klasifikacija komentara

U sklopu prve analize koristila sam ručno klasificirane komentare u tri kategorije, ovisno o tome izražava li komentar po meni pozitivno, negativno ili neutralno mišljenje (Raguzin, 2016). Međutim, za potrebe diplomskog rada, odlučila sam izostaviti kategoriju neutralnih komentara te bazirati analizu samo na pozitivnim i negativnim komentarima. Tablica 8 prikazuje rezultat ručnog klasificiranja te sadrži: ukupan broj komenatara (bez neutralne kategorije), broj pozitivnih i broj negativnih komentara. Možemo primjetiti da su komentari na temu izbjeglica većinom bili negativnog mišljenja. Ukoliko izračunamo postotak od ukupnog broja komenatara dobijemo da je njih 54% bilo negativno na portalu Index. Međutim, ako izbacimo iz računice neutralne komentare onda brojka raste na visokih 93%. Kod portala Jutarnji list imamo još veći postotak negativnih komentara, a on iznosi 65% odnosno 97% bez neutralnih komentara.

Tablica 8- Klasifikacija komentara

	Index	**Jutarnji list**
Ukupan broj komentara*	718	739
Broj pozitivnih komentara	47	24
Broj negativnih komentara	671	715

*bez neutralne kategorije

Budući da nas zanima kojim su se vokabularom služili korisnici pri pisanju komentara, odlučila sam izračunati frekvencije riječi za obje kategorije te prikazati nekoliko tablica koje sadrže riječi i broj pojavljivanja tih riječi u tekstu. U nastavku ćemo prikazati dvije liste, jedna lista će sadržavati lematizirane[8] riječi bez zaustavnih riječi, a druga će sadržavati nelematizirane riječi uključujući i zaustavne riječi. Komentare sam lematizirala pomoću alata Nikole Ljubešića koji je dostupan na Internetu (Ljubešić, 2018).

Tablica 9 prikazuje deset najčešće korištenih riječi u pozitivnim komentarima na portalu Index. U ovom slučaju imamo listu lematiziranih riječi, no bez zaustavnih riječi. Riječ koja ima najveću frekvenciju je riječ „bravo“, a nakon nje slijede riječi „čast“ i „podrška“. S druge strane imamo tablicu 10 koja sadrži frekvencije riječi prije lematizacije, a uključuje i zaustavne riječi.

[8] Lematizacija je određivanje osnovnog oblika riječi

U tom slučaju veznik „i“ je najčešće korištena riječ, a riječ „je“ slijedi nakon nje s frekvencijom 18. Primjer jednog pozitivnog komentara s portala Index je: „Bravo Norvezani, ja sam uz vas“.

Tablica 9-Lista riječi s frekvencijom

(KI-pozitivni, lematizirani, bez stop riječi)

1.	bravo	13
2.	čast	7
3.	podrška	6
4.	milan	6
5.	wendy	5
6.	gajski	5
7.	lara	4
8.	trebati	4
9.	pokazati	3
10.	mirko	3

Tablica 10- Lista riječi s frekvencijom

(KI- pozitivni, sa zaustavnim riječima)

1.	i	23
2.	je	18
3.	bravo	13
4.	to	10
5.	na	9
6.	svaka	8
7.	čast	7
8.	u	7
9.	za	6
10.	samo	6

Tablice 11 i 12 prikazuju statistiku za negativne komentare s portala Index. U tablici 11 možemo vidjeti deset najčešće korištenih riječi koje su prethodno lematizirane te su izbačene zaustavne riječi. Najčešće korištena riječ je glagol „htjeti“ s frekvencijom 158. Možemo primjetiti da nam prvih deset riječi ne sugerira da se radi o negativnim komentarima, no ukoliko pogledamo ostatak liste pronaći ćemo mnogo pogrdnih riječi kao što su primjerice: „idiotizam“, „glupi“, „grozno“ i slično. Tablica 12 prikazuje listu bez lematizacije, ali uključuje zaustavne riječi što možemo odmah primjetiti pošto svih deset navedenih riječi spadaju pod zaustavne.

Tablica 11- Lista riječi s frekvencijom

(KI- negativni, lematizirani, bez stop riječi)

1.	htjeti	156
2.	trebati	94
3.	imati	91
4.	čovjek	82
5.	izbjeglica	78
6.	moći	68
7.	nemati	55
8.	ce	52
9.	žica	52
10.	europa	51

Tablica 12- Lista riječi s frekvencijom

(KI- negativni, sa zaustavnim riječima)

1.	i	809
2.	je	466
3.	da	462
4.	u	453
5.	se	356
6.	su	253
7.	ne	249
8.	na	212
9.	a	193
10.	za	169

Primjer negativnog komentara s Index portala je: „primit nesretnike iz ratnog područja i tražit KORIST? gdje je tu ljubav? nije ni čudo što ćeš propasti truli zapade“.

Tablice 13 i 14 prikazuju frekvencije riječi iz pozitivnih komenatara prikupljenih s portala Jutarnji list. Konkretno, tablica 13 prikazuje listu lematiziranih riječi nakon što izbacimo zaustavne riječi, a tablica 14 prikazuje deset najčešće korištenih riječi uključujući i zaustavne. Ukoliko usporedimo s rezultatom kojeg smo dobili s Index portala, možemo primjetiti kako postoje sličnosti među odabirom riječi- u oba slučaja je najčešće korištena riječ „bravo“. Međutim, ako ne izbacimo zaustavne riječi onda je najčešće korištena riječ „da“ s frekvencijom 11 kao što možemo vidjeti u tablici 14, dok je riječ „bravo“ na četvrtom mjestu.

Tablica 13- Lista riječi s frekvencijom

(KJ-pozitivni, lematizirani, bez stop riječi)

1.	bravo	7
2.	dobar	3
3.	hrvatski	3
4.	pravo	3
5.	orban	3
6.	zemlja	3
7.	velik	3
8.	stipe	2
9.	subotić	2
10.	narod	2

Tablica 14- Lista riječi s frekvencijom

(KJ-pozitivni, sa zaustavnim riječima)

1.	da	11
2.	i	11
3.	u	10
4.	bravo	7
5.	je	7
6.	se	6
7.	za	5
8.	to	4
9.	su	4
10.	ne	4

Primjer pozitivnog komentara s portala Jutarnji list je: „Kako je mila ova djevojčica! Ja bi je isto udomio“.

U tablicama 15 i 16 nalazi se lista riječi s frekvencijom prikupljena iz negativnih komenatara na Jutarnjem listu. Tablica 15 prikazuje listu lematiziranih riječi, bez zaustavnih, a tablica 16 nelematiziranu listu sa zaustavnim riječima. U prvom slučaju je riječ s najvećom frekvencijom glagol „htjeti“ kao i kod Index portala, a nakon toga imamo riječ „čovjek“ s frekvencijom 131 i riječ „izbjeglica“ s frekvencijom 127. U drugom slučaju je veznik „i“ najčešće upotrebljena riječ, a nakon nje slijedi riječ „je“ s frekvencijom 771 i samoglasnik „u“ s frekvencijom 716.

Tablica 15- Lista riječi s frekvencijom

(KJ- negativni, lematizirani, bez stop riječi)

1.	htjeti	269
2.	čovjek	131
3.	izbjeglica	127
4.	imati	113
5.	moći	105
6.	zemlja	100
7.	trebati	92
8.	hrvatski	78
9.	nemati	78
10.	europa	75

Tablica 16- Lista riječi s frekvencijom

(KJ- negativni, sa zaustavnim riječima)

1.	i	1207
2.	je	771
3.	u	716
4.	da	653
5.	se	562
6.	na	394
7.	su	389
8.	ne	362
9.	a	314
10.	to	280

Kao i kod Index portala, lista najčešće upotrebljenih riječi ne sugerira da se radi o negativnim komenatarima. No, ukoliko pogledamo ostatak liste, možemo pronaći pogrdne riječi kao što su: „zaglupljivati“, „čmar“, „gnjusan“ i slično. Primjer negativnog komentara s Jutarnjeg lista je: „novinari - izdajnička gamad koja se nada da će poput njihovih gospodara bit udaljena od geta ,premlaćivanja na rasnoj/vjerskoj bazi i svih ostalih prednosti multi-kulti europe. not gonna happen“.

Nakon što smo pogledali sve liste riječi s frekvencijama za oba portala i usporedili ih, možemo zaključiti da je vokabular kojime su se korisnici služili u pozitivnim i negativnim komentarima vrlo sličan. Riječ „bravo“ ima najveću frekvenciju u pozitivnim komentarima u oba slučaja. Kod negativnih komentara nismo mogli u prvih deset najčešće korištenih riječi razaznati da se radi o vokabularu koji je korišten za izražavanje negativnog mišljenja, no uvidom u ostatak liste shvatili smo da puno toga ovisi o samom kontekstu u kojem se riječi nalaze. Također, bitno je naglasiti da su ove tablice rezultat mog ručnog klasificiranja komenatara, što znači da su komentari svrstani isključivo prema mom mišljenju, iako sam ja pokušala biti što objektivnija. Stoga je velika vjerojatnost da bi se dobili drugačiji rezultati ukoliko bi netko drugi sortirao. Nekima će komentar koji sam ja ocjenila pozitivno, biti negativan ili neutralan. Ista rečenica može biti na više načina intepretirana, primjerice: „Održati će se dodatni rok iz xy kolegija“, jednom studentu to može biti pozitivno ukoliko nije uspio proći predmet, dok profesoru koji predaje taj predmet to ne bude pošto mu to znači više posla. Nešto više o mogućim problemima analize sentimenta u poglavlju 6.

5.2.1. Klasifikacija komentara prema polaritetu

Za klasifikaciju komentara prema polaritetu ću iskoristiti pripremljeni korpus za učenje i testiranje klasifikatora, na način da ću upotrijebiti program koji sam napisala u programskom jeziku Python (Privitak 3). Nakon što sam ručno odvojila komentare u dvije kategorije: pretežno pozitivno mišljenje i pretežno negativno mišljenje, napisala sam skriptu koja čita polaritete pojedinih riječi iz datoteka koristeći već spomenuti leksički resurs za hrvatski jezik u kojem se nalazi popis engleskih i hrvatskih riječi te njihov polaritet. Budući da su komentari napisani hrvatskim jezikom, za izgradnju rječnika sam koristila samo hrvatske riječi. Rječnik sadrži riječi kao ključeve, dok su vrijednosti iznosi polariteta. Nakon što sam izgradila dva rječnika- jedan za pozitivne i jedan za negativne riječi, napravila sam rječnik koji sadrži riječi i polaritet

riječi iz oba skupa. Nadalje, bilo je potrebno definirati tri funkcije za izlučivanje značajki iz danog teksta (POS, NEG i ALL). Funkcije kao prostor značajki koriste prethodno definirane prostore, tako da funkcija POS koristi samo pozitivan prostor značajki, a funkcija NEG samo negativan i tako dalje. Ulaz funkcije je tekst, a izlaz je rječnik značajki i njihovih vrijednosti. Izlučavanje se vrši na idući način: najprije se definira prazan izlazni Python rječnik, nakon čega se za svaku riječ iz rječnika značajki zbroji koliko se puta riječ ponovila u ulaznom tekstu, riječ se potom unese u izlazni Python rječnik kao ključ dok se njena vrijednost računa kao umnožak polariteta te riječi iz rječnika značajki i broja ponavljanja te riječi u ulaznom tekstu, na kraju funkcija vrati izlazni Python rječnik koji mora biti jednake duljine kao i rječnik značajki. Glavni dio programa odnosi se na učenje i testiranje klasifikatora. Kako bi to učinili, potrebno je da se pripremljeni korpus podijeli na način da 90% komentara bude za učenje i 10% za testiranje. Učenje klasifikatora sam napravila posebno za svaku funkciju značajki. Za klasifikaciju sam koristila klasifikatore koji su opisani u prethodnom poglavlju, a to su: Naive Bayes, stablo odlučivanja, maksimalna entropija i support vector machine. Rezultati klasifikacije sadrže podatak o točnosti klasifikatora za sve tri funkcije te prikaz petnaest najinformativnijih značajki kod Naive Bayes klasifikatora te klasifikatora maksimalne entropije. Pri svakom pokretanju programa sam dobila različite rezultate, a to je zbog toga što sam prije podijele korpusa na dva dijela (za učenje i testiranje) koristila funkciju „random.shuffle“ koja izmiješa redoslijed komentara.

Na kraju analize ću prikazati deset rezultata te uzeti njihov prosjek kako bi dobili što točniju sliku o tome koji klasifikator daje najbolje rezultate za ovaj korpus. S obzirom da sam učenje klasifikatora napravila za sve tri funkcije, rezultate ću prikazati za svaku funkciju posebno. Najprije ću prikazati rezultate vezane za komentare s portala Index (ukupno 47 pozitivnih i 671 negativnih komentara). Na slici 10 možemo vidjeti dobivene rezultate za funkciju POS klasifikatorom Naive Bayes kojim je dobivena točnost 91,66%. Na slici također možemo vidjeti 15 najinformativnijih značajki kao što su riječi „čast“, „super“, „solidarnost“ i tako dalje.

```
Naive Bayes klasifikator, tocnost:  91.66666666666666 %
Most Informative Features
                    cast = 3                 POS : NEG    =     13.7 : 1.0
                   super = 0                 NEG : POS    =      1.0 : 1.0
                  vizija = 0                 NEG : POS    =      1.0 : 1.0
                 ponosan = 0                 NEG : POS    =      1.0 : 1.0
                    cast = 0                 NEG : POS    =      1.0 : 1.0
                    brzo = 0                 NEG : POS    =      1.0 : 1.0
                  istina = 0                 NEG : POS    =      1.0 : 1.0
              nadobudnost = 0                NEG : POS    =      1.0 : 1.0
                 poticati = 0                NEG : POS    =      1.0 : 1.0
                prekrasan = 0                NEG : POS    =      1.0 : 1.0
              solidarnost = 0                NEG : POS    =      1.0 : 1.0
                optimizam = 0                NEG : POS    =      1.0 : 1.0
                  rjesava = 0                NEG : POS    =      1.0 : 1.0
                 interesi = 0                NEG : POS    =      1.0 : 1.0
                    sanse = 0                NEG : POS    =      1.0 : 1.0
```

Slika 10- KI- Naive Bayes (POS)

Slika 11 prikazuje rezultate dobivene klasifikatorom stabla odlučivanja. Točnost ovog klasifikatora je 94,44%.

```
Stablo odlucivanja, tocnost:  94.44444444444444 %
  ==> Training (100 iterations)

     Iteration    Log Likelihood    Accuracy
     ---------------------------------------
             1          -0.69315       0.067
             2          -0.13310       0.933
             3          -0.13271       0.933
             4          -0.13270       0.933
             5          -0.13270       0.933
             6          -0.13269       0.933
             7          -0.13269       0.933
             8          -0.13268       0.933
             9          -0.13268       0.933
            10          -0.13267       0.933
            11          -0.13267       0.933
            12          -0.13266       0.933
            13          -0.13266       0.933
            14          -0.13265       0.933
            15          -0.13265       0.933
            16          -0.13264       0.933
            17          -0.13264       0.933
            18          -0.13263       0.933
```

Slika 11- KI- Stablo odlučivanja (POS)

Na slici 12 možemo vidjeti točnost klasifikatora maksimalne entropije koja iznosi isto kao i kod prethodnog klasifikatora- 94,44%. Također, slika prikazuje i 15 najinformativnijih značajki, kao što su: „vizija“, „cast“, „najveci“ i tako dalje.

```
Maksimalna entropija, tocnost: 94.44444444444444 %
   0.708 ponosan==2 and label is 'POS'
   0.708 vizija==1 and label is 'POS'
   0.708 super==3 and label is 'POS'
   0.514 cast==3 and label is 'POS'
  -0.173 cast==3 and label is 'NEG'
   0.026 najveci==3 and label is 'NEG'
   0.026 rastuci==1 and label is 'NEG'
   0.023 sigurnost==3 and label is 'NEG'
   0.023 pametan==1 and label is 'NEG'
   0.022 stvar==1 and label is 'NEG'
   0.022 odobravanje==1 and label is 'NEG'
   0.022 posten==2 and label is 'NEG'
   0.022 interes==1 and label is 'NEG'
   0.022 interesi==1 and label is 'NEG'
   0.022 sunce==2 and label is 'NEG'
```

Slika 12- KI-Maksimalna entropija (POS)

Posljednji klasifiktor je support vector machine klasifikator s kojim je dobivena točnost od 93,05%.

Nadalje, na slici 13 možemo vidjeti rezultate dobivene za funkciju NEG Naive Bayes klasifikatorom. Točnost klasifikatora za ovu vrstu značajke je 87,5%, a na slici možemo vidjeti i 15 najinformativnijih značajki kao što su: „problemi“, „like“, „mrtav“ i tako dalje.

```
Naive Bayes klasifikator, tocnost:  87.5 %
Most Informative Features
                    like = 2                POS : NEG    =     14.4 : 1.0
                priznati = 0                NEG : POS    =      1.0 : 1.0
              zaboraviti = 0                NEG : POS    =      1.0 : 1.0
                    like = 0                NEG : POS    =      1.0 : 1.0
                 previse = 0                NEG : POS    =      1.0 : 1.0
                problemi = 0                NEG : POS    =      1.0 : 1.0
                    boli = 0                NEG : POS    =      1.0 : 1.0
                slucajno = 0                NEG : POS    =      1.0 : 1.0
                   mrtav = 0                NEG : POS    =      1.0 : 1.0
                      in = 0                NEG : POS    =      1.0 : 1.0
                     dao = 0                NEG : POS    =      1.0 : 1.0
                   staje = 0                NEG : POS    =      1.0 : 1.0
               nametnuti = 0                NEG : POS    =      1.0 : 1.0
                   bljak = 0                NEG : POS    =      1.0 : 1.0
                dogovor = 0                 NEG : POS    =      1.0 : 1.0
```

Slika 13- KI- Naive Bayes (NEG)

Slika 14 prikazuje rezultate dobivene klasifikatorom stabla odlučivanja. Točnost klasifikatora za ovu vrstu značajke je 90,27%.

```
Stablo odlucivanja, tocnost:  90.27777777777779 %
  ==> Training (100 iterations)

      Iteration    Log Likelihood    Accuracy
      ---------------------------------------
              1          -0.69315       0.063
              2          -0.12680       0.937
              3          -0.12658       0.937
              4          -0.12657       0.937
              5          -0.12657       0.937
              6          -0.12657       0.937
              7          -0.12657       0.937
              8          -0.12657       0.937
              9          -0.12656       0.937
             10          -0.12656       0.937
             11          -0.12656       0.937
             12          -0.12656       0.937
             13          -0.12656       0.937
             14          -0.12655       0.937
             15          -0.12655       0.937
             16          -0.12655       0.937
             17          -0.12655       0.937
             18          -0.12655       0.937
```

Slika 14- KI- Stablo odlučivanja (NEG)

Na slici 15 možemo vidjeti rezultate dobivene klasifikatorom maksimalne entropije. Točnost ovog klasifikatora je 91,66%. Prikazane su i najinformativnije značajke, kao što su: „zaboraviti“, „priznati“, „zlocin“ i tako dalje.

```
Maksimalna entropija, tocnost: 91.66666666666666 %
   0.416 priznati==1 and label is 'POS'
   0.416 zaboraviti==1 and label is 'POS'
   0.306 like==2 and label is 'POS'
  -0.098 like==2 and label is 'NEG'
   0.012 davati==2 and label is 'NEG'
   0.012 unistiti==2 and label is 'NEG'
   0.012 tesko==1 and label is 'NEG'
   0.012 dilema==1 and label is 'NEG'
   0.012 kritizira==2 and label is 'NEG'
   0.012 zlocin==3 and label is 'NEG'
   0.012 prevara==2 and label is 'NEG'
   0.012 gubitak==3 and label is 'NEG'
   0.012 platiti==1 and label is 'NEG'
   0.012 spreman==2 and label is 'NEG'
   0.012 vara==3 and label is 'NEG'
```

Slika 15- KI- Maksimalna entropija (NEG)

Točnost support vector machine klasifikatora za ovu vrstu značajke je 90,27%.

Iduća slika nam prikazuje rezultate dobivene za funkciju ALL Naive Bayes klasifikatorom. Točnost klasifikatoraje 88,88%, a neke od najinformativnijih značajki su: „dobar“, „super“, „slucajno“ i tako dalje.

```
Naive Bayes klasifikator, tocnost:  88.88888888888889 %
Most Informative Features
                like_neg = 2                POS : NEG    =     13.1 : 1.0
              dobar_pos = 3                 POS : NEG    =      7.9 : 1.0
              super_pos = 0                 NEG : POS    =      1.0 : 1.0
           priznati_neg = 0                 NEG : POS    =      1.0 : 1.0
             vizija_pos = 0                 NEG : POS    =      1.0 : 1.0
          zaboraviti_neg = 0                NEG : POS    =      1.0 : 1.0
             ponosan_pos = 0                NEG : POS    =      1.0 : 1.0
                like_neg = 0                NEG : POS    =      1.0 : 1.0
              dobar_pos = 0                 NEG : POS    =      1.0 : 1.0
            slucajno_neg = 0                NEG : POS    =      1.0 : 1.0
             previse_neg = 0                NEG : POS    =      1.0 : 1.0
                boli_neg = 0                NEG : POS    =      1.0 : 1.0
                   in_neg = 0               NEG : POS    =      1.0 : 1.0
              mrtav_neg = 0                 NEG : POS    =      1.0 : 1.0
                dao_neg = 0                 NEG : POS    =      1.0 : 1.0
```

Slika 16- KI- Naive Bayes (ALL)

Na slici 17 možemo vidjeti rezultate dobivene stablom odlučivanja, točnost ovog klasifikatora je 97,22%.

```
Stablo odlucivanja, tocnost:  97.22222222222221 %
  ==> Training (100 iterations)

      Iteration    Log Likelihood    Accuracy
      ---------------------------------------
              1          -0.69315       0.070
              2          -0.13900       0.930
              3          -0.13884       0.930
              4          -0.13883       0.930
              5          -0.13883       0.930
              6          -0.13883       0.930
              7          -0.13882       0.930
              8          -0.13882       0.930
              9          -0.13882       0.930
             10          -0.13881       0.930
             11          -0.13881       0.930
             12          -0.13881       0.930
             13          -0.13880       0.930
             14          -0.13880       0.930
             15          -0.13880       0.930
             16          -0.13879       0.930
             17          -0.13879       0.930
             18          -0.13879       0.930
```

Slika 17- KI- Stablo odlučivanja (ALL)

Slika 18 sadrži rezultate dobivene maksimalnom entropijom. Točnost ovog klasifikatora ista je kao kod prethodnog- 97,22%. Također, možemo vidjeti neke od 15 najinformativnijih značajki, kao što su: „super“, „vizija“, „ponosan“ i tako dalje.

```
Maksimalna entropija, tocnost: 97.22222222222221 %
   0.264 super_pos==3 and label is 'POS'
   0.255 priznati_neg==1 and label is 'POS'
   0.255 vizija_pos==1 and label is 'POS'
   0.255 zaboraviti_neg==1 and label is 'POS'
   0.255 ponosan_pos==2 and label is 'POS'
   0.193 like_neg==2 and label is 'POS'
   0.153 dobar_pos==3 and label is 'POS'
  -0.063 like_neg==2 and label is 'NEG'
  -0.033 dobar_pos==3 and label is 'NEG'
   0.009 lud_neg==3 and label is 'NEG'
   0.008 davati_neg==2 and label is 'NEG'
   0.008 veliko_pos==1 and label is 'NEG'
   0.008 udovica_neg==1 and label is 'NEG'
   0.008 krivo_neg==2 and label is 'NEG'
   0.008 prekrasan_pos==3 and label is 'NEG'
```

Slika 18- KI- Maksimalna entropija (ALL)

Točnost support vector machine klasifikatora za ovu vrstu značajke je 97,22% kao prethodna dva klasifikatora.

Kao što je već spomenuto, rezultati se mijenjaju pri svakom pokretanju programa, stoga kako bi dobili što bolji uvid u to koji je klasifikator dao najbolje rezultate za određenu funkciju značajki, uzeti ćemo deset različitih rezultata i napraviti njihov prosjek. U tablici 17 možemo vidjeti prikazane rezultate za funkciju POS, koristeći Index komentare. Najveći prosjek za ovu vrstu značajke, time i najveću točnost imaju klasifikatori stablo odlučivanja i maksimalna entropija s prosječnom točnosti koja iznosi 93,47%.

Tablica 17- KI-usporedba rezultata klasifikatora za POS značajku

	Naive Bayes	**Stablo odlučivanja**	**Maksimalna entropija**	**Support vector machine**
1.	91,66 %	94,44 %	94,44 %	93,05 %
2.	83,33 %	87,5 %	87,5 %	87,5 %
3.	93,05 %	94,44 %	94,44 %	94,44 %
4.	98,61 %	98,61 %	98,61 %	98,61 %
5.	95,83 %	95,83 %	95,83 %	95,83 %
6.	90,27 %	93,05 %	93,05 %	93,05 %
7.	91,66 %	91,66 %	91,66 %	91,66 %
8.	87,5 %	91,66 %	91,66 %	90,27 %
9.	91,66 %	94,44 %	94,44 %	93,05%
10.	90,27 %	93,05%	93,05%	93,05%
PROSJEK	91,38 %	93,47 %	93,47 %	93, 05 %

Tablica 18 također prikazuje usporedbu rezultata, ali ovaj put za funkciju NEG. Najveći prosjek ima maksimalna entropija, a on iznosi 93,47% kao i u prethodnom primjeru.

Tablica 18- KI-usporedba rezultata klasifikatora za NEG značajku

	Naive Bayes	Stablo odlučivanja	Maksimalna entropija	Support vector machine
1.	87,5 %	90,27 %	91,66 %	90,27 %
2.	98,61 %	98,61 %	98,61 %	98,61 %
3.	90,27 %	93,05 %	93,05 %	93,05 %
4.	90,27 %	93,05 %	94,44 %	93,05 %
5.	88,88 %	91,66 %	93,05 %	91,66 %
6.	86,11 %	90,27 %	91,66 %	90,27 %
7.	91,66 %	94,44 %	94,44 %	94,44 %
8.	86,11 %	91,66 %	91,66 %	91,66 %
9.	84,72 %	91,66 %	91,66 %	91,66 %
10.	93,05 %	94,44 %	94,44 %	94,44 %
PROSJEK	89,72 %	92,91 %	93,47 %	92,91 %

U tablici 19 možemo vidjeti prikazane rezultate za funkciju ALL. Kao i za prethodne funkcije, klasifikator maksimalne entropije je opet pokazao najveću točnost, a ona iznosi 93,88%, što je za nijansu veće od one dobivene za stablo odlučivanja i SVM klasifikator.

Tablica 19- KI-usporedba rezultata klasifikatora za ALL značajku

	Naive Bayes	Stablo odlučivanja	Maksimalna entropija	Support vector machine
1.	88,88 %	97,22 %	97,22 %	97,22 %
2.	91,66 %	93,05 %	93,05 %	93,05 %
3.	87,5 %	90,27 %	90,27 %	90,27 %
4.	86,11 %	90,27 %	90,27 %	90,27 %
5.	90,27 %	94,44 %	94,44 %	94,44 %
6.	90,27 %	91,66 %	93,05 %	91,66 %
7.	93,05 %	97,22 %	97,22 %	97,22 %
8.	88,88 %	91,66 %	91,66 %	90,27 %

9.	93,05 %	94,44 %	95,83 %	95,83 %
10.	87,5 %	95,83 %	95,83 %	95,83 %
PROSJEK	89,72 %	93,61 %	93,88 %	93,61 %

Sljedeće što ću prikazati su rezultati s portala Jutarnji list gdje smo imali 715 negativnih i 24 pozitivna komentara. Slika 14 prikazuje dobivene rezultate za funkciju POS klasifikatorom Naive Bayes kojim je dobivena točnost 95,94%. Također, na slici možemo vidjeti 15 najinformativnijih značajki kao što su riječi „nagrada“, „dobar“, „uspjeh“ i tako dalje.

```
Naive Bayes klasifikator, tocnost:  95.94594594594 %
Most Informative Features
                nagrada = 2                POS : NEG    =     25.7 : 1.0
                  dobar = 3                POS : NEG    =     13.6 : 1.0
                    mir = 2                POS : NEG    =     11.5 : 1.0
                  dragi = 3                POS : NEG    =     11.0 : 1.0
                  drago = 3                POS : NEG    =      7.0 : 1.0
                  dobar = 0                NEG : POS    =      1.2 : 1.0
                    mir = 0                NEG : POS    =      1.1 : 1.0
               dijeliti = 0                NEG : POS    =      1.1 : 1.0
                veliko = 0                 NEG : POS    =      1.1 : 1.0
               izvrsno = 0                 NEG : POS    =      1.1 : 1.0
               nagrada = 0                 NEG : POS    =      1.1 : 1.0
                 dragi = 0                 NEG : POS    =      1.1 : 1.0
                 drago = 0                 NEG : POS    =      1.1 : 1.0
                uspjeh = 0                 NEG : POS    =      1.0 : 1.0
               prilika = 0                 NEG : POS    =      1.0 : 1.0
```

Slika 19- KJ- Naive Bayes (POS)

Iduća slika sadrži rezultate dobivene stablom odlučivanja, točnost ovog klasifikatora je bila maksimalnih 100 % što se dogodilo prvi put do sada.

```
Stablo odlucivanja, tocnost:  100.0 %
  ==> Training (100 iterations)

      Iteration    Log Likelihood    Accuracy
      ---------------------------------------|
              1          -0.69315       0.036
              2          -0.07239       0.964
              3          -0.07210       0.964
              4          -0.07209       0.964
              5          -0.07209       0.964
              6          -0.07208       0.964
              7          -0.07207       0.964
              8          -0.07207       0.964
              9          -0.07206       0.964
             10          -0.07205       0.964
             11          -0.07205       0.964
             12          -0.07204       0.964
             13          -0.07203       0.964
             14          -0.07203       0.964
             15          -0.07202       0.964
             16          -0.07201       0.964
             17          -0.07201       0.964
             18          -0.07200       0.964
             19          -0.07199       0.964
             20          -0.07199       0.964
```

Slika 20- KJ- Stablo odlučivanja (POS)

Slika 21 prikazuje rezultate klasifikatora maksimalne entropije kojim je također dobivena stopostotna točnost. Na slici možemo vidjeti 15 najinformativnijih značajki kao što su riječi: „dijeliti“, „veliko“, „izvrsno“ i tako dalje.

```
Maksimalna entropija, tocnost: 100.0 %
   0.867 dijeliti==1 and label is 'POS'
   0.867 veliko==1 and label is 'POS'
   0.867 izvrsno==3 and label is 'POS'
   0.617 nagrada==2 and label is 'POS'
   0.568 dobar==3 and label is 'POS'
   0.502 mir==2 and label is 'POS'
   0.481 dragi==3 and label is 'POS'
   0.365 drago==3 and label is 'POS'
  -0.180 nagrada==2 and label is 'NEG'
  -0.102 dobar==3 and label is 'NEG'
  -0.080 mir==2 and label is 'NEG'
  -0.067 dragi==3 and label is 'NEG'
  -0.041 dobar==0 and label is 'POS'
  -0.037 drago==3 and label is 'NEG'
  -0.028 mir==0 and label is 'POS'
```

Slika 21- KJ- Maksimalna entropija (POS)

SVM klasifikatoru, kao i kod prethodna dva klasifikatora, točnost iznosi 100% za ovu vrstu značajke.

Nakon POS značajke potrebno je prikazati rezultate NEG funkcije. Na slici 22 možemo vidjeti rezultate Naive Bayes klasifikatora čija točnost iznosi 97,29%. Prikazane su i najinformativnije značajke kao što su riječi: „ludilo“, „nazalost“, „sukob“ i tako dalje.

```
Naive Bayes klasifikator, tocnost:  97.2972972972973 %
Most Informative Features
                  ludilo = 0                NEG : POS    =      1.0 : 1.0
                nazalost = 0                NEG : POS    =      1.0 : 1.0
                   sukob = 0                NEG : POS    =      1.0 : 1.0
                   govno = 0                NEG : POS    =      1.0 : 1.0
                   strah = 0                NEG : POS    =      1.0 : 1.0
                 problem = 0                NEG : POS    =      1.0 : 1.0
                   hitno = 0                NEG : POS    =      1.0 : 1.0
                    kaos = 0                NEG : POS    =      1.0 : 1.0
                     rat = 0                NEG : POS    =      1.0 : 1.0
                 sramota = 0                NEG : POS    =      1.0 : 1.0
                   kriza = 0                NEG : POS    =      1.0 : 1.0
              sprijeciti = 0                NEG : POS    =      1.0 : 1.0
            odugovlacenje = 0               NEG : POS    =      1.0 : 1.0
                  osuditi = 0               NEG : POS    =      1.0 : 1.0
                   zaliti = 0               NEG : POS    =      1.0 : 1.0
```

Slika 22- KJ- Naive Bayes (NEG)

Slika 23 prikazuje rezultate dobivene stablom odlučivanja. Točnost ovog klasifikatora za NEG značajku je 98,64%.

```
Stablo odlucivanja, tocnost:  98.64864864864865 %
  ==> Training (100 iterations)

     Iteration    Log Likelihood    Accuracy
     ---------------------------------------
             1          -0.69315       0.035
             2          -0.06930       0.965
             3          -0.06913       0.965
             4          -0.06913       0.965
             5          -0.06913       0.965
             6          -0.06913       0.965
             7          -0.06913       0.965
             8          -0.06913       0.965
             9          -0.06913       0.965
            10          -0.06913       0.965
            11          -0.06913       0.965
            12          -0.06913       0.965
            13          -0.06913       0.965
            14          -0.06913       0.965
            15          -0.06913       0.965
            16          -0.06913       0.965
            17          -0.06913       0.965
            18          -0.06913       0.965
```

Slika 23- KJ- Stablo odlučivanja (NEG)

Na slici 24 možemo vidjeti rezultat klasifikatora maksimalne entropije čija je točnost za NEG značajku bila 98,64% kao i kod stabla odlučivanja. Također, možemo vidjeti najinformativnije značajke kao što su riječi. „optuziti“, „slijep“, „dosadno“ i tako dalje.

```
Maksimalna entropija, tocnost: 98.64864864864865 %
   0.007 optuziti==2 and label is 'NEG'
   0.007 priznaje==1 and label is 'NEG'
   0.007 sprijeciti==1 and label is 'NEG'
   0.007 slijep==1 and label is 'NEG'
   0.007 bitke==1 and label is 'NEG'
   0.007 beznadno==2 and label is 'NEG'
   0.007 seronja==4 and label is 'NEG'
   0.007 maltretirati==2 and label is 'NEG'
   0.007 dosadno==2 and label is 'NEG'
   0.007 nepovratno==1 and label is 'NEG'
   0.007 prevaranti==4 and label is 'NEG'
   0.007 idiotski==3 and label is 'NEG'
   0.007 rak==3 and label is 'NEG'
   0.007 zaboraviti==1 and label is 'NEG'
   0.007 tuga==2 and label is 'NEG'
```

Slika 24- KJ- Maksimalna entropija (NEG)

SVM klasifikator je dao istu točnost kao i prethodna dva klasifikatora, a to je 98,64%.

Na kraju imamo rezultate ALL značajke. Najprije na slici 25 možemo vidjeti rezultate Naive Bayes klasifikatora čija je točnost 94,59%. Također, možemo vidjeti 15 najinformativnijh značajki kao što su: „nagrada“, „ludilo“, „strah“ i tako dalje.

```
Naive Bayes klasifikator, tocnost:  94.5945945945946 %
Most Informative Features
              nagrada_pos = 2                POS : NEG    =     26.8 : 1.0
                dobar_pos = 3                POS : NEG    =     16.7 : 1.0
                drago_pos = 3                POS : NEG    =      7.3 : 1.0
                  mir_pos = 2                POS : NEG    =      6.1 : 1.0
                dobar_pos = 0                NEG : POS    =      1.2 : 1.0
                  mir_pos = 0                NEG : POS    =      1.1 : 1.0
              izvrsno_pos = 0                NEG : POS    =      1.1 : 1.0
               veliko_pos = 0                NEG : POS    =      1.1 : 1.0
             dijeliti_pos = 0                NEG : POS    =      1.1 : 1.0
              nagrada_pos = 0                NEG : POS    =      1.1 : 1.0
                drago_pos = 0                NEG : POS    =      1.1 : 1.0
               ludilo_neg = 0                NEG : POS    =      1.0 : 1.0
               uspjeh_pos = 0                NEG : POS    =      1.0 : 1.0
                govno_neg = 0                NEG : POS    =      1.0 : 1.0
                strah_neg = 0                NEG : POS    =      1.0 : 1.0
```

Slika 25- KJ- Naive Bayes (ALL)

Iduća slika sadrži rezultate klasifikatora stabla odlučivanja čija je točnost za ovu vrstu značajke 98,64%.

```
Stablo odlucivanja, tocnost:  98.64864864864865 %
  ==> Training (100 iterations)

      Iteration    Log Likelihood    Accuracy
      ---------------------------------------
              1          -0.69315       0.035
              2          -0.06925       0.965
              3          -0.06912       0.965
              4          -0.06912       0.965
              5          -0.06912       0.965
              6          -0.06912       0.965
              7          -0.06912       0.965
              8          -0.06911       0.965
              9          -0.06911       0.965
             10          -0.06911       0.965
             11          -0.06911       0.965
             12          -0.06911       0.965
             13          -0.06910       0.965
             14          -0.06910       0.965
             15          -0.06910       0.965
             16          -0.06910       0.965
             17          -0.06910       0.965
             18          -0.06909       0.965
```

Slika 26- KJ- Stablo odlučivanja (ALL)

Slika 27 prikazuje rezultate klasifikatora maksimalne entropije kojim je dobivena točnost 98,64%. Na slici možemo vidjeti i najinformativnije značajke kao što su riječi: „veliko“, „izvrsno“, „drago“ i tako dalje.

```
Maksimalna entropija, tocnost: 98.64864864864865 %
   0.333 veliko_pos==1 and label is 'POS'
   0.333 izvrsno_pos==3 and label is 'POS'
   0.333 dijeliti_pos==1 and label is 'POS'
   0.256 nagrada_pos==2 and label is 'POS'
   0.233 dobar_pos==3 and label is 'POS'
   0.151 drago_pos==3 and label is 'POS'
   0.136 mir_pos==2 and label is 'POS'
  -0.066 nagrada_pos==2 and label is 'NEG'
  -0.044 dobar_pos==3 and label is 'NEG'
  -0.016 dobar_pos==0 and label is 'POS'
  -0.014 drago_pos==3 and label is 'NEG'
  -0.011 mir_pos==2 and label is 'NEG'
  -0.007 dijeliti_pos==0 and label is 'POS'
  -0.007 izvrsno_pos==0 and label is 'POS'
  -0.007 veliko_pos==0 and label is 'POS'
```

Slika 27- KJ- Maksimalna entropija (ALL)

SVM klasifikator je pokazao točnost od 98,64% za ovu vrstu značajke (istu vrijednost kao i prethodna dva klasifikatora).

Sada ću prikazati rezultate deset različitih pokretanja programa isto kao i kod Index portala. S dobivenim prosjekom točnosti klasifikatora ćemo moći zaključiti koji klasifikator je najbolje radio za određenu značajku za ovaj korpus. Najprije ćemo vidjeti rezultate POS značajke u tablici 20.

Tablica 20-KJ-usporedba rezultata klasifikatora za POS značajku

	Naive Bayes	**Stablo odlučivanja**	**Maksimalna entropija**	**Support vector machine**
1.	95,94 %	100 %	100 %	100 %
2.	98,64 %	100 %	100 %	100 %
3.	90,54 %	93,24 %	93,24 %	93,24 %
4.	95,94 %	97,29 %	97,29 %	97,29 %
5.	95,94 %	98,64 %	98,64 %	98,64 %
6.	97,29 %	97,29 %	97,29 %	97,29 %
7.	97,29 %	97,29 %	97,29 %	97,29 %
8.	94,59 %	95,94 %	95,94 %	94,59 %
9.	93,24 %	98,64 %	98,64 %	98,64 %
10.	93,24 %	94,59 %	94,59 %	93,24 %
PROSJEK	95,27 %	97 %	97 %	97 %

Isti prosjek točnosti klasifikatora, koji je ujedno i najveći, imaju klasifikatori stablo odlučivanja, maksimalna entropija i SVM. Vrijednost tog prosjeka je 97% te je označen crvenom bojom u tablici.

Tablica 21 prikazuje sve rezultate za značajku NEG. Ukoliko pogledamo prosjek svih klasifikatora, možemo primjetiti da, kao i u prethodnoj tablici, klasifikatori stablo odlučivanja, maksimalna entropija i SVM imaju najveći prosjek koji iznosi 97,70%. Također, vidimo da su točnosti tih klasifikatora pri svakom pokretanju programa bile iste.

Tablica 21- KJ-usporedba rezultata klasifikatora za NEG značajku

	Naive Bayes	Stablo odlučivanja	Maksimalna entropija	Support vector machine
1.	97,29 %	98,64 %	98,64 %	98,64 %
2.	91,89 %	95,94 %	95,94 %	95,94 %
3.	97,29 %	98,64 %	98,64 %	98,64 %
4.	94,59 %	98,64 %	98,64 %	98,64 %
5.	98,64 %	98,64 %	98,64 %	98,64 %
6.	93,24 %	97,29 %	97,29 %	97,29 %
7.	97,29 %	98,64 %	98,64 %	98,64 %
8.	95,94 %	97,29 %	97,29 %	97,29 %
9.	97,29 %	98,64 %	98,64 %	98,64 %
10.	91,89 %	94,59 %	94,59 %	94,59 %
PROSJEK	95,54 %	97,70 %	97,70 %	97,70 %

Na kraju imamo tablicu 22 koja prikazuje sve rezultate za značajku ALL. Za ovu značajku najbolju točnost su imali klasifikatori stablo odlučivanja i maksimalna entropija (95,94%). SVM klasifikator je imao nešto manji postotak točnosti- 95,67%.

Tablica 22- KJ-usporedba rezultata klasifikatora za ALL značajku

	Naive Bayes	Stablo odlučivanja	Maksimalna entropija	Support vector machine
1.	94,59 %	98,64 %	98,64 %	98,64 %
2.	93,24 %	95,94 %	95,94 %	95,94 %
3.	90,54 %	95,94 %	95,94 %	95,94 %
4.	95,94 %	97,29 %	97,29 %	97,29 %
5.	97,29 %	97,29 %	97,29 %	97,29 %
6.	94,59 %	97,29 %	97,29 %	97,29 %
7.	89,18 %	90,54 %	90,54 %	90,54 %
8.	95,94 %	97,29 %	97,29 %	95,94 %
9.	90,54 %	93,24 %	93,24 %	93,24 %
10.	93,24 %	95,94 %	95,94 %	94,59 %
PROSJEK	93,51 %	95,94 %	95,94 %	95,67 %

Nakon što smo vidjeli rezultate od oba portala, možemo zaključiti da je generalno najveću točnost za portal Index pokazao klasifikator maksimalne entropije i to za sve tri značajke. S druge strane, kod portala Jutarnji list to su klasifikatori stablo odlučivanja, maksimalne entropije i SVM klasifikator (osim ALL značajke). Naivni Bayesov klasifikator nije ni za jednu značajku pokazao najveću prosječnu točnost, no unatoč tome i on je pokazao velik postotak točnosti koji je uglavnom bio veći od 90%.

5.3. Analiza mikroblogova s društvene mreže Twitter

U ovom poglavlju ću prikazati analizu podataka prikupljenih s društvene mreže Twitter. Podaci su tweetovi koji sadrže ključne riječi: *terrorism, isis, refugees, puppy, kitty* i *baby.* Razlikovati ćemo dva skupa podataka koja su prikupljena u dva perioda, jedni su prikupljeni u prosincu 2016. godine[9], a drugi u srpnju i kolovozu 2018. godine. Tweetove sam podjelila u dvije kategorije: pozitivni tweetovi i negativni tweetovi. Pozitivni tweetovi su oni koji sadrže ključne riječi „puppy“, „kitty“ i „baby“, pod pretpostavkom da su tekstovi koji koriste te riječi pretežno pozitivnog mišljenja. S druge strane, negativni tweetovi su oni koji sadrže ključne riječi „terrorism“, „isis“ i „refugees“, također pod pretpostavkom da tweetovi s tim ključnim riječima su pretežno negativnog mišljenja. Pomoću prikupljenih tweetova napraviti ću analizu na način da ću usporediti korišteni vokabular (frekvenciju riječi u pojedinom skupu) i ukupni sentiment skupa podataka koji ću izračunati pomoću gotovog programa za računanje sentimenta u tekstu. Tablica 23 prikazuje opće informacije o tweetovima. Prvi stupac označava godinu prikupljanja tweetova, drugi označava ukupan broj tweetova, treći označava broj riječi u tekstu, a četvrti broj različitih pojavnica.

Tablica 23- Opće informacije o tweetovima

Godina prikupljanja tweetova	Ukupan broj tweetova	Broj riječi u tekstu	Broj različitih pojavnica
2016.	19 923	285 751	38 864
2018.	19 923	128 197	37 232

[9] Izvor: http://langnet.uniri.hr/resources.html

U tablicama 24 i 25 možemo vidjeti opće informacije o pozitivnim i negativnim tweetovima, no bez zaustavnih riječi.

Tablica 24- Opće informacije o pozitivnim tweetovima

Godina prikupljanja tweetova	Ukupan broj tweetova	Broj riječi u tekstu	Broj različitih pojavnica
2016.	9936	83 034	21 335
2018.	9936	100 701	22 718

Tablica 25- Opće informacije o negativnim tweetovima

Godina prikupljanja tweetova	Ukupan broj tweetova	Broj riječi u tekstu	Broj različitih pojavnica
2016.	9987	109 619	20 450
2018.	9987	117 727	17 555

Prvo ću usporediti skupove po frekvenciji riječi u tekstu, na način da ću najprije prikazati liste vezane za pozitivne tweetove, a zatim za negativne. Tablica 26 sadrži dvadeset riječi s najvećom frekvencijom u tweetovima iz 2016. godine, dok s druge strane, tablica 27 sadrži frekvencije riječi iz tweetova prikupljenih 2018. godine. Ukoliko pogledamo tablice možemo vidjeti da riječ „baby“ ima najveću frekvenciju u oba slučaja. U tablici 26 primjećujemo da je riječ „cute“ upotrebljena je 2474 puta, dok je u tablici 27 „samo“ 475 puta. Nadalje, riječ „like“ je u objema tablicama na četvrtom mjestu, no s različitom frekvencijom- u tablici 26 ima frekvenciju 648, a u tablici 27 ima 858. Slično imamo s upotrebom riječi „dog“, u tablici 26 ima frekvenciju 1726 (treće po redu), a u tablici 27 ima 282 (dvanaesto po redu). Postotak isto upotrebljenih riječi, ukoliko gledamo samo ovih prvih 20, je 60% dok se onih 40% razlikuje.

Tablica 26-Lista riječi s frekvencijom (POZ-2016.)

1.	baby	3453
2.	cute	2474
3.	dog	1726
4.	like	648
5.	cat	632
6.	jump	536
7.	teaching	534
8.	love	500
9.	dont	497
10.	im	410
11.	just	399
12.	know	333
13.	puppy	308
14.	cuteemergency	299
15.	animal	288
16.	babyanimalpics	274
17.	day	270
18.	looks	235
19.	good	226
20.	time	222

Tablica 27- Lista riječi s frekvencijom (POZ-2018.)

1.	baby	2543
2.	puppy	2486
3.	kitty	2046
4.	like	858
5.	cute	475
6.	love	470
7.	amp	435
8.	just	399
9.	want	359
10.	little	347
11.	good	292
12.	dog	282
13.	looks	266
14.	cat	261
15.	home	258
16.	new	226
17.	im	219
18.	know	217
19.	got	217
20.	adorable	213

Sljedeće tablice prikazuju dvadeset najčešće korištenih riječi u negativnim tweetovima. Tablica 28 se odnosi na tweetove iz 2016. godine, a tablica 29 na one prikupljene 2018. godine. U prvoj tablici je najčešće korištena riječ „syria“ s frekvencijom 2173, dok u drugoj tablici najveću frekvenciju ima riječ „isis“ (3066), koja je u prvoj tablici na trećem mjestu s frekvencijom 1356. Ukoliko usporedimo riječi koje se nalaze u ovim listama, možemo primjetiti da se samo dvije riječi pojavljuju u obje tablice, a to su: „isis“ i „terrorism“, što je 10% od ukupnog broja riječi u listi. Međutim, postoje neki slični oblici riječi koje možemo također ubrajati pod „isti vokabular“, kao što su riječi: „attacks“ (tablica 28) i „attack“ (tablica 29) te „says“ (tablica 28) i „say“ (tablica 29). Nadalje, primjećujemo da je jedna od najčešće korištenih riječi u tweetovima iz 2018. godine, riječ „toronto“ koja označava najveći grad u Kanadi. Možemo pretpostaviti da su u tom periodu (srpanj i kolovoz 2018.), izbjeglice u Torontu bile aktualna tema. Tu pretpostavku možemo potvrditi pretraživanjem Google tražilice gdje možemo pronaći

članke iz tog perioda, primjerice jedan od naslova je: „I can do great things- Toronto refugees and asylum seekers want voices heard“[10].

Tablica 28- Lista riječi s frekvencijom (NEG-2016.)

1.	syria	2173
2.	muslim	1481
3.	isis	1356
4.	refugees	962
5.	islam	802
6.	terrorist	696
7.	says	582
8.	saudi	452
9.	assad	446
10.	amprt	379
11.	terrorism	344
12.	war	343
13.	iraq	321
14.	man	305
15.	russia	286
16.	syrian	266
17.	helping	257
18.	vows	251
19.	attacks	251
20.	time	249

Tablica 29- Lista riječi s frekvencijom (NEG-2018.)

1.	isis	3066
2.	terrorism	2195
3.	refugees	1582
4.	toronto	1075
5.	responsibility	890
6.	shooting	675
7.	amp	658
8.	claims	595
9.	people	542
10.	world	434
11.	police	413
12.	attack	403
13.	claim	389
14.	state	373
15.	ezralevant	367
16.	year	361
17.	trump	353
18.	syrian	337
19.	just	331
20.	say	299

U drugom dijelu analize ću usporediti ukupni sentiment svih tweetova, zatim posebno sentiment pozitivnih i negativnih tweetova. Pritom ću koristiti gotov softver za računanje sentimenta naziva „AFINN Sentiment Analysis“[11]. Ovaj program pri računanju ukupnog sentimenta koristi leksički resurs za engleski jezik (AFFIN-en-165[12]) u kojem su riječima dodijeljene ocjene od -5 (najnegativnija) do 5 (najpozitivnija) te one predstavljaju polaritet riječi. Najprije se unese tekst kojem želimo izračunati ukupni sentiment. Nakon što to napravimo, u prozoru „Verdict“-što prevedeno na hrvatski jezik znači presuda/sentencija, pojavi se rezultat u kojem se nalazi nekoliko podataka, a to su redom: verdict-je li ukupni sentiment pozitivan, negativan ili neutralan, score-rezultat koji predstavlja zbroj svih polariteta riječi u tekstu, comparative- kolika je vrijednost komparativa (uspoređivanje dva objekta po

[10] Izvor: https://www.ctvnews.ca/canada/i-can-do-great-things-toronto-refugees-and-asylum-seekers-want-voices-heard-1.4021525
[11] Link na stranicu: http://darenr.github.io/afinn/#
[12] Link za leksički resurs: https://github.com/fnielsen/afinn/blob/master/afinn/data/AFINN-en-165.txt

nekom svojstvu), positive- lista pozitivnih riječi te negative- lista negativnih riječi. U nastavku ću prikazati 6 tablica u kojima će se nalaziti dobiveni rezultati.

Tablica 30 prikazuje ukupni sentiment svih tweetova prikupljenih 2016. godine. Analizom je dobiven zbroj polariteta koji iznosi 3767, što znači da je ukupni sentiment svih tweetova pozitivan. Također, u tablici možemo vidjeti iznos komparativa te dio liste pozitivnih i negativnih riječi koje su korištene u tekstu.

Tablica 30- Ukupni sentiment svih tweetova prikupljenih 2016. godine

Ukupni sentiment svih tweetova prikupljenih 2016. godine	
verdict	POSITIVE
score	3767
comparative	0.013182806009427788
positive	justice, likes, loves, good, cute, famous, laugh, loving, friendly, help...
negative	fucking, tortured, hurt, cry, niggas, kill, hell, sloppy, hate, bad...

Iduća tablica prikazuje rezultate koji se odnose na tweetove prikupljene 2018. godine. Dobiven je ukupni sentiment čija vrijednost iznosi 2260, što znači da su tweetovi sveukupno pozitivni ukoliko se zbroje polariteti riječi. Kao i u prethodnoj tablici, možemo vidjeti iznos komparativa i jedan dio pozitivnih i negativnih riječi u tekstu.

Tablica 31- Ukupni sentiment svih tweetova prikupljenih 2018. godine

Ukupni sentiment svih tweetova prikupljenih 2018. godine	
verdict	POSITIVE
score	2260
comparative	0.006611280131055465
positive	xoxo, wish, luck, good, happy, adorable, prepared, cute..
negative	funeral, angers, leave, no, starve, stress, shoot, broken, crying...

Zaključak: Unatoč tome što sveukupno ima više negativnih tweetova (9987) nego pozitivnih (9936), rezultat sentimenta dobiven AFINN programom pokazuje ukupno pozitivan sentiment u oba slučaja. Međutim, treba uzeti u obzir da su tweetovi klasificirani prema tome koje su ključne riječi korištene, a ne ručno, što naravno umanjuje preciznost klasifikacije. Ovime smo pokazali i da ponekad negativne riječi mogu biti korištene u „pozitivnom okruženju“ koje onda umanjuje jačinu negativnosti tih riječi (u ovom slučaju „terorrism“, „isis“ i „refugees“). Primjerice, rečenica: „My cat died yesterday, I loved her very much, I will miss her but I'm glad she had a long and happy life (prevedeno na hrv.: Moja mačka je uginula jučer, jako sam je voljela, nedostajat će mi ali mi je drago da je imala dug i sretan život)“ ima ukupan sentiment 4, što znači da je rečenica pozitivna unatoč tome što govori o nesretnom slučaju. Zbroj je dobiven na sljedeći način, zbrojeni su polariteti riječi „happy“ koji iznosi 3, polaritet riječi „glad“ koji iznosi 3, polaritet riječi „loved“ koji je također 3, zatim polaritet riječi „miss“ koji iznosi -2 i polaritet riječi „died“ koji iznosi -3, što dovodi do zbroja 4. Također, kontekst je bitan faktor pri klasifikaciji tekstova, no ova vrsta analize ne uzima to u obzir pošto se gleda svaka riječ pojedinačno, odnosno gleda se polaritet te riječi koji ne ovisi o kontekstu u kojem je ta riječ upotrebljena.

Tablica 32 prikazuje ukupni sentiment pozitivnih tweetova prikupljenih 2016. godine. No, prije nego pogledamo rezultate, trebamo uzeti u obzir da ovi tweetovi nisu svrstani ručno nego pod pretpostavkom da upotreba određenih ključnih riječi sa sobom vuče pozitivno ili negativno mišljenje.

Tablica 32- Ukupni sentiment pozitivnih tweetova prikupljenjih 2016.godine

Ukupni sentiment pozitivnih tweetova prikupljenih 2016. godine	
verdict	POSITIVE
score	8295
comparative	0.06470510230348604
positive	justice, likes, loves, good, cute, famous, laugh,careful...
negative	fucking, tortured, hurt, cry, niggas, kill, hell, struggle..

Pretpostavka da su tweetovi s ključnim riječima „puppy“, „kitty“ i „baby“ pozitivnog mišljenja pokazala se točnom. Naime, dobiveni zbroj polariteta riječi u tekstu je 8295. Osim toga, možemo vidjeti iznos komparativa te dio liste pozitivnih i negativnih riječi u tekstu.

U tablici 33 možemo vidjeti ukupni sentiment pozitivnih tweetova prikupljenih 2018. godine. Kao i u prethodnoj tablici, ukupni sentiment je pozitivan, ali s nešto manjim zbrojem polariteta riječi (7876). Također, u tablici možemo vidjeti jedan dio riječi pozitivnog i negativnog polariteta.

Tablica 33- Ukupni sentiment pozitivnih tweetova prikupljenih 2018. godine

Ukupni sentiment pozitivnih tweetova prikupljenih 2018. godine	
verdict	POSITIVE
score	7876
comparative	0.05024561403508772
positive	xoxo, luck, good, want, cute, help, agree, like..
negative	no, missed, starve, stress, shoot, broken, suck, hurting, lost...

U oba slučaja se početna pretpostavka pokazala točnom, doista su tweetovi s ključnim riječima „puppy“, „kitty“ i „baby“ većinom upotrebljeni u tekstovima s pretežno pozitivnim mišljenjem. Ukoliko usporedimo zbrojeve polariteta riječi, možemo reći da su tweetovi prikupljeni 2016. godine bili pozitivniji od onih 2018. godine (8295 > 7876).

Iduća pretpostavka se odnosi na tweetove s ključnim riječima „terrorism“, „isis“ i „refugees“, po kojoj su tweetovi koji sadrže takve riječi pretežno negativnog mišljenja. Tablica 34 prikazuje ukupni sentiment negativnih tweetova prikupljenih 2016. godine. Zbroj polariteta riječu u ovim tweetovima je -4528, što znači da je ukupan sentiment negativan. Tablica također prikazuje iznos komparativa te dio pozitivnih i negativnih riječi.

Tablica 34- Ukupni sentiment negativnih tweetova prikupljenih 2016. godine

Ukupni sentiment negativnih tweetova prikupljenih 2016. godine	
verdict	NEGATIVE
score	-4528
comparative	-0.028739170448414838
positive	want, beautiful, peace, love, big, respect, helping, help..
negative	dangerous, terrorist, stop, destroy, escaping, worries, fight..

U tablici 35 možemo vidjeti rezultate negativnih tweetova prikupljenih 2018. godine. U ovom slučaju zbroj polariteta riječi iznosi -5618, što također znači da je ukupan sentiment negativan.

Tablica 35- Ukupni sentiment negativnih tweetova prikupljenih 2018. godine

Ukupni sentiment negativnih tweetova prikupljenih 2018. godine	
verdict	NEGATIVE
score	-5618
comparative	-0.030374299169005022
positive	winning, support, legal, encourage, great, growing, supports, true..
negative	slave, denying, hide, terrorist, charges, fight, loss, attack, saddened..

Zaključak: Početna pretpostavka se opet pokazala točnom, tweetovi s ključnim riječima „terrorism“, „refugees“ i „isis“ su pretežno negativnog mišljenja. Ukoliko usporedimo rezultate, možemo primjetiti da su tweetovi iz 2018. godine negativniji od onih iz 2016. godine. Tablica 36 prikazuje usporedbu ukupnog zbroja polariteta riječi za svaku pojedinačnu kategoriju. U njoj možemo vidjeti promjenu sentimenta u tweetovima kroz period od dvije godine.

Tablica 36- Usporedba ukupnog zbroja polariteta riječi iz tweetova 2016. i 2018. godine

	Ukupni zbroj polariteta riječi		
	Svi tweetovi	Pozitivni tweetovi	Negativni tweetovi
2016.	3767	8295	-4528
2018.	2260	7876	-5628

6. Problem analize sentimenta

Ono što čini analizu sentimenta otežanom jest to što je njen glavni fokus istraživanje mišljenja, koja su za razliku od činjeničnih informacija - subjektivna. Subjektivnost dolazi iz više izvora. Na prvom mjestu moramo istaknuti onaj očiti, a to je da različiti ljudi mogu imati različita iskustva i stavove, čak i o istim stvarima. Na primjer, ukoliko jedna osoba kupi mobitel određene marke i ima jako dobro iskustvo sa njim, ona dakle ima pozitivno mišljenje o njemu. S druge strane imamo osobu koja je također kupila mobitel te marke, no ona je imala nesreću pa je dobila mobitel s greškom te konstantno ima neke probleme sa njim, dakle ona ima negativno mišljenje o njemu. Nadalje, različiti ljudi mogu vidjeti istu stvar na različite načine, primjerice : kada cijena dionica pada, jedna osoba može biti jako tužna zbog toga jer je kupila dionicu dok je njena cijena bila visoka, dok druga osoba može biti sretna jer je to prilika da proda dionicu i profitira. Također, treba uzeti u obzir da ljudi mogu imati različite interese i ideologije, stoga je za analizu sentimenta potrebno analizirati veći broj mišljenja, a ne samo od jedne osobe.

Nadalje, postoje određeni problemi glede polariteta riječi jer riječi mogu imati različite polaritete ovisno o kontekstu u kojem se nalaze ili domeni u kojoj su korištene, primjerice riječ „mekano“ obično označava nešto pozitivno „Ovaj džemper je jako mekan“, ali ona također može označavati nešto negativno: „Ovaj madrac je previše mekan“. Postoje rečenice koje sadrže riječi s polaritetom, ali koje ne izražavaju nikakvo mišljenje, kao što su upitne ili uvjetne rečenice. Na primjer: „Možete li mi reći koji je Samsung mobitel dobar za kupiti?“ i „Ako nađem dobar mobitel u dućanu, onda ću ga kupiti.“, obje rečenice sadrže riječ „dobar“ ali nijedna od njih ne izražava pozitivno ili negativno mišljenje osobe. Međutim, postoje upitne i uvjetne rečenice koje izražavaju mišljenje, kao što su: „Zna li netko kako da popravim ovaj užasan mobitel?“ ili „Ako tražiš dobar mobitel, nemoj kupiti Iphone X“. Nadalje, analizi sentimenta veliki problem mogu zadati sarkastične rečenice, koje je ponekad teško prepoznati ljudima, a kamoli računalu. S obzirom da se većina programa za analizu sentimenta bazira na zbrajanju polariteta pojedinih riječi, onda prisustvo riječi u rečenici koje imaju snažan polaritet može biti problem ukoliko su one korištene sarkastično pošto to onda znači da je polaritet suprotan od uobičajenog. Na primjer: „Naprosto obožavam kad mi zakasni bus“ je sarkastična rečenica, no program će ju klasificirati kao rečenicu koja iskazuje pozitivno mišljenje. Međutim, postoje neki programi koji mogu prepoznati sarkazam na način da svaki put kad uoče naglo mijenjanje polariteta, kao što imamo u navedenom primjeru, će to prepoznati kao

sarkazam jer uoče kontrast između pozitivne emocije i negativne situacije. Istraživanja su pokazala da sarkazam nije uobičajen u recenzijama prozivoda ili usluga, nego je češće korišten, primjerice, u političkim raspravama što otežava analizu sentimenta u toj domeni (Riloff, 2013). Osim ovog programa, postoje još nekolicina koji imaju svoje načine za prepoznavanje sarkazma u rečenici, no nikad toliko dobro koliko to može čovjek.

7. Zaključak

Ljudi sve više koriste Internet kao platformu za izražavanje svojih mišljenja i stavova u vezi bilo čega. Sukladno konstantnom rastu količine tekstova koji sadrže mišljenje na Internetu, raste i popularnost analize sentimenta. U sklopu ovog rada napravljena je analiza tekstova o izbjeglicama koristeći dva izvora podataka. Prvi dio analize odnosio se na komentare prikupljene s portala Index i Jutarnji list u kojoj je utvrđeno da je najviše ljudi pisalo negativne komentare. Postotak negativnih komentara na Index portalu je 93%, odnosno 7% pozitivnih komentara. Jutarnji list ima još veći postotak negativnih komentara, a iznosi 97%, odnosno samo 3% pozitivnih komenatara. Najčešće korištene riječi u pozitivnim komentarima su: „bravo“, „čast“, „podrška“, „dobar“ i tako dalje. S druge strane, najveću frekvenciju kod negativnih komentara imale su riječi: „čovjek“, „izbjeglica“, „europa“, „zemlja“ i tako dalje. Gledajući listu najčešće korištenih riječi u negativnim komentarima možemo zaključiti da kontekst u kojem se riječi nalaze može imati velikog utjecaja na klasifikaciju, odnosno na samo značenje teksta. Nadalje, osim što je izračunata frekvencija riječi, komentari su iskorišteni kao korpus u programu za učenje i testiranje četiri klasifikatora. Svi klasifikatori su dobro radili, s postotkom većim od 80%. Najbolju točnost za sve komentare pokazao je klasifikator maksimalne entropije, no osim njega dobro je radio i klasifikator stabla odlučivanja te support vector machine. Naivni Bayesov klasifikator imao je, ukoliko ga usporedimo sa ostalima, najmanji postotak točnosti, međutim on je kako god pokazao da dobro radi s postotkom koji je uglavnom bio veći od 90%. Drugi dio analize odnosio se na podatke s društvene mreže Twitter, prikupljenih u dva različita perioda. Najprije sam usporedila korišteni vokabular u pozitivnim i negativnim tweetovima. Usporedba se bazirala na dvadeset najčešće korištenih riječi, a njome je utvrđeno: postotak preklapanja riječi u pozitivnim tweetovima je 60%, a u negativnim samo 10%. Nakon toga sam napravila usporedbu ukupnog sentimenta u tweetovima koristeći gotovi alat „AFFIN Sentiment Analysis“. Rezultati su pokazali da su tweetovi s ključnim riječima „puppy“, „kitty“ i „baby“ pozitivni, a tweetovi s ključnim riječima „isis“, „refugees“ i „terrorism“ negativni, kao što smo i na početku pretpostavili. Međutim, ukupni sentiment za tweetove iz oba vremenska perioda je pozitivan unatoč tome što sveukupno ima više negativnih tweetova (9987) nego pozitivnih (9936).

Treba uzeti u obzir da se prva analiza bazirala na ručnom klasificiranim komentarima, stoga postoji velika vjerojatnost da bi se rezultati razlikovali ukoliko bi netko drugi klasificirao komentare. Također, postoje neki slučajevi koji smanjuju točnost programa za učenje i

testiranje klasifikatora. Glavni problem stvaraju, svakako, sarkastične rečenice u kojima su prisutne riječi koje imaju snažan polaritet, a koje su korištene sarkastično. Smatram da je u daljnjem radu potrebno odvojiti sarkastične rečenice i ne koristiti ih u učenju klasifikatora kako bi točnost klasifikacije bila još veća. Također, potrebno je povećati korpus kako bi klasifikacija polariteta radila što bolje i preciznije.

8. Popis literature

1. Bing, L. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

2. Jurafsky, D., Martin, J.H. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing*, Computational Linguistics and Speech recognition

3. Arun, S., Loaiza, F., Rolfe, R. 2012. *Supervised Learning in the Wild: Text Classification for Critical Technologies*. Institute for Defense Analyses.

4. Leung Ming, K. 2007. *Naive Bayesian Classifier.* Polytechnic University.

5. Bing, L. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

6. Dobrescu, A. 2011. *Methods and Resources for Sentiment Analysis in Multilingual Documents of Different Text Types.* Doktorski rad. Universitat d'Alacant.

7. Marko Modrić, *Leksikon za analizu mišljenja iz teksta na hrvatskome jeziku* (leksički resurs za hrvatski jezik), završni rad, ak.god.2012./13, preuzeto: ožujak 2016.

8. Bing, L. 2010. *Sentiment Analysis and Subjectivity*. Izdano u knjizi: N.Indurkhya i F.J. Damerau *Handbook of Natural Language Processing*, Second Edition.

9. Turney, P. 2002. *Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews.* Institute for Information Technology, Canada.

10. Pang, B. i Lee, L. 2002. *Thumbs up? Sentiment classification using machine learning techniques.*

11. Goldberg A., i Zhu, J. 2006. *Seeing stars whem there arenn't many stars: Graph-based semi-supervised learning for sentiment categorization*.

12. Pang, B. i Lee, L. 2003. *Exploting class relathionship for sentiment categorization with respect to rating scales*.

13. Kim, S.-M. i Hovy, E. 2004. *Determing the Sentiment of Opinions*.

14. https://wordnet.princeton.edu/ (stranica posjećena: 9. kolovoza 2018.)

15.https://www.ctvnews.ca/canada/i-can-do-great-things-toronto-refugees-and-asylum-seekers-want-voices-heard-1.4021525 (stranica posjećena: 10. kolovoza 2018.)

16. http://darenr.github.io/afinn/# (stranica posjećena: 10. kolovoza 2018.)

17.https://github.com/fnielsen/afinn/blob/master/afinn/data/AFINN-en-165.txt

(stranica posjećena: 10. kolovoza 2018.)

18. Jan Šnajder, *Popis hrvatskih stop riječi*, preuzeto: kolovoz 2018.

19. Agarwal, A., Boyi, X., Vovsha, I., Rambow, O., Passonneau, R. 2011. *Sentiment Analysis of Twitter Data*. Columbia University.

20. Riloff, E. i Ashequl Q. 2013. *Sarcasm as Contrast between a Positive Sentiment and Negative Situation*. School Of Computing. Salt Lake City.

21.(Medium):https://medium.com/deep-math-machine-learning-ai/chapter-4-decision-trees-algorithms-b93975f7a1f1 (stranica posjećena: 7. kolovoza 2018.)

22.(Saedsayad): http://www.saedsayad.com/decision_tree.htm (stranica posjećena: 7. kolovoza 2018.)

23.(Datumbox):http://blog.datumbox.com/machine-learning-tutorial-the-max-entropy-text-classifier/ (stranica posjećena: 7. kolovoza 2018.)

24.(Analyticsvidhya):https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/ (stranica posjećena: 7. kolovoza 2018.)

25. http://www.jutarnji.hr/, stranica Jutarnjeg lista

26. http://www.index.hr/, stranica Indexa

27. https://en.wikipedia.org/wiki/Hyperplane (stranica posjećena: 7. kolovoza 2018.)

28. (Ljubešić): http://nl.ijs.si/web/ (stranica posjećena: 7. kolovoza 2018.)

29. Yu, D. i Hatzivassiloglu. 2003. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

30.(Medium):https://medium.com/greyatom/an-introduction-to-bag-of-words-in-nlp-ac967d43b428 (stranica posjećena: 30. kolovoza 2018.)

31. Ana Raguzin, *Analiza pozitivnog i negativnog polariteta tekstova na Internetu*, završni rad, ak.god.2015./16., preuzeto: rujan 2016.

32. http://langnet.uniri.hr/resources.html (preuzeto: travanj 2018.)

Privitak 1

U Privitku 1 nalazi se funkcija za čišćenje teksta od interpunkcijskih znakova i računanje frekvencije riječi.

```
import nltk
from string import punctuation

#Funkcija za ciscenje teksta od interpunkcijskih znakova
def ocisti_tekst(tekst):
    for n in punctuation:
        tekst = tekst.replace(n, '')
    return tekst

with open ("2018-TWITTER-POZ_NOVO.txt", "r", encoding='latin-1') as myfile:
    data=myfile.read().replace('\n', ' ').lower()
    ndata= ocisti_tekst(data)

ndata = ndata.split(' ')
fdist1 = nltk.FreqDist(ndata)
#print (sorted(fdist1.most_common(100)))
print (fdist1.most_common())
print('\n')
print('**Opce informacije o tekstu**')
print('\n')
print ('Broj rijeci u tekstu: '+ str(len(ndata)))
print('\n')
print ('Razlicite pojavnice u tekstu: '+ str(sorted(set(ndata))))
print('\n')
print ('Broj razlicitih pojavnica u tekstu: '+ str(len(set(ndata))))
myfile.close()
```

Privitak 2

U Privitku 2 prikazano je izbacivanje zaustavnih riječi iz teksta.

```
def read_words(words_file):
  with open(words_file, 'r', encoding='utf8') as f:
    ret = []
    for line in f:
      ret += line.split()
    return ret

#Izbaci stop/zaustavne rijeci iz teksta
#print (read_words('stop_rijeci.txt'))
write = True
delete_list= read_words('STOP_RIJECI.txt')
with open('2018-TWITTER-NEG.txt', "r", encoding='latin-1') as fin:
    with open('2018-TWITTER-NEG_NOVO.txt', "w+", encoding='latin-1')as
fout:
        for line in fin:
            rijec = line.split(' ')
            for w in rijec:
                for word in delete_list:
                    if w == word:
                        write = False
                        break
                if write:
                    fout.write(w + ' ')
                write = True
            fout.write('')
        fin.close()
        fout.close()
```

Privitak 3

U Privitku 3 nalazi se glavni kod u kojem imamo definiranu funkciju za izlučivanje značajki te učenje i testiranje klasifikatora.

```
from string import punctuation
import random
from os import listdir
from nltk import NaiveBayesClassifier
from nltk import classify
from nltk import MaxentClassifier
from nltk import DecisionTreeClassifier
from sklearn.svm import LinearSVC

#Izrada prostora znacajki za klasifikaciju tekstova
infile = open ('Leksicki_resurs_za_hrvatski_jezik.txt', encoding="utf-8")
lek_resurs= infile.read()
lek_resurs= lek_resurs.split('\n')

pos_dict = {}
neg_dict = {}

#iz svakog retka, dohvati rijec i njen polaritet
for line in lek_resurs:
    word,polaritet   =   line.split()[1],   int(line.split()[-1])   #dohvati
hrvatske rijeci
    if polaritet > 0:
        pos_dict[word] = polaritet
    if polaritet < 0:
        neg_dict[word] = abs(polaritet)

svi_dict = {}

for key, val in pos_dict.items():
    svi_dict[key] = ('pos', val)
for key, val in neg_dict.items():
```

```
    svi_dict[key] = ('neg', val)


#Definiranje funkcije za izlucivanje znacajki

def features_neg(tekst):
    tekst = tekst.split()
    result_neg = {}

    for word, polaritet in neg_dict.items():
        broj_pojavljivanja = tekst.count(word)
        result_neg[word] = polaritet * broj_pojavljivanja

    #Provjeri je li izlazni rjecnik iste duljine kao rjecnik znacajki
    if len(result_neg) == len(neg_dict):
        return result_neg
    else:
        return None

def features_pos(tekst):
    tekst = tekst.split()
    result_pos = {}

    for word, polaritet in pos_dict.items():
        broj_pojavljivanja = tekst.count(word)
        result_pos[word] = polaritet * broj_pojavljivanja

    #Provjeri je li izlazni rjecnik iste duljine kao rjecnik znacajki
    if len(result_pos) == len(pos_dict):
        return result_pos
    else:
        return None

def features_all(tekst):
    result_all = {}
```

```
    neg_feat = features_neg(tekst)
    pos_feat = features_pos(tekst)


    for key, value in neg_feat.items():
        result_all[key + '_neg'] = value
    for key, value in pos_feat.items():
        result_all[key + '_pos'] = value

    return result_all


#Ucitavanje korpusa

podaci = listdir("podaci-Index")
datoteke = [n for n in podaci if n[-4:] == '.txt']

poz_komentari= []
neg_komentari= []

for dat in datoteke:
    if dat[:3] =='poz':
        #dodaj encoding='utf8' za hrv
        with open('podaci-Index/' + dat, 'r') as infile:
            poz_komentari.append(infile.read())
    elif dat[:3] =='neg':
        with open('podaci-Index/'+ dat, 'r') as infile:
            neg_komentari.append(infile.read())

#Micanje interpunkcijskih znakova i prebaci u lower case

def ocisti_tekst(tekst):
    for n in punctuation:
        tekst = tekst.replace(n, '')
    return tekst
```

```
poz_komentari = [ocisti_tekst(p.lower()) for p in poz_komentari]
neg_komentari = [ocisti_tekst(n.lower()) for n in neg_komentari]
svi_komentari = poz_komentari + neg_komentari

print('Broj negativnih komentara: '+ str(len(neg_komentari)))
print('Broj pozitivnih komentara: '+ str(len(poz_komentari)))
print('Broj svih komentara: '+ str(len(svi_komentari)))

#Klasifikacija

dat_poz = [(features_pos(p), 'POS') for p in poz_komentari] +
[(features_pos(n), 'NEG') for n in neg_komentari]
dat_neg = [(features_neg(p), 'POS') for p in poz_komentari] +
[(features_neg(n), 'NEG') for n in neg_komentari]
dat_svi = [(features_all(p), 'POS') for p in poz_komentari] +
[(features_all(n), 'NEG') for n in neg_komentari]
random.shuffle(dat_poz)
random.shuffle(dat_neg)
random.shuffle(dat_svi)

lista = []

for l in [dat_poz, dat_neg, dat_svi]:
    limit = int(len(l) * 0.9)
    train_set, test_set = l[:limit], l[limit:]
    lista.append([train_set, test_set])

lista[0].append("POS")
lista[1].append("NEG")
lista[2].append("ALL")

for l in lista:
    print ('Vrsta znacajke: '+ l[2])
    '''
    Naive Bayes
    '''
```

```
    nb_classifier = NaiveBayesClassifier.train(l[0])
    print("Naive Bayes klasifikator, tocnost: ",
classify.accuracy(nb_classifier, l[1]) * 100, "%")
    print(nb_classifier.show_most_informative_features(15))
    print()

    '''
    Stablo odlucivanja
    '''
    tree_classifier = DecisionTreeClassifier.train(l[0])
    print("Stablo odlucivanja, tocnost: ",
classify.accuracy(tree_classifier, l[1]) * 100, "%")

    '''
    Maksimalna entropija
    '''
    me_classifier = MaxentClassifier.train(l[0], algorithm='gis')
    print("Maksimalna entropija, tocnost:",
classify.accuracy(me_classifier, l[1]) * 100, "%")
    print(me_classifier.show_most_informative_features(15))

    '''
    Support Vector Machine
    '''
    svm_classifier =  classify.SklearnClassifier(LinearSVC())
    svm_classifier.train(l[0])
    print("SVM, tocnost:", classify.accuracy(svm_classifier, l[1]) * 100,
"%")
```