

Analiza pozitivnog i negativnog polariteta tekstova na Internetu

Raguzin, Ana

Undergraduate thesis / Završni rad

2016

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Rijeka, Faculty of Humanities and Social Sciences / Sveučilište u Rijeci, Filozofski fakultet**

Permanent link / Trajna poveznica: <https://um.nsk.hr/um:nbn:hr:186:767160>

Rights / Prava: [In copyright](#) / [Zaštićeno autorskim pravom.](#)

Download date / Datum preuzimanja: **2024-11-30**



Repository / Repozitorij:

[Repository of the University of Rijeka, Faculty of Humanities and Social Sciences - FHSSRI Repository](#)



Sveučilište u Rijeci – Odjel za informatiku

Dvopredmetni studij informatike i filozofije

Ana Raguzin

Analiza pozitivnog i negativnog polariteta tekstova na Internetu

Završni rad

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

Rijeka, rujan 2016.

//Zadatak

Sažetak

Analiza pozitivnog i negativnog polariteta tekstova na internetu

Mišljenje predstavlja središnji dio svih ljudskih aktivnosti te uvelike utječe na način na koji se osobe ponašaju. Većina ljudi prije neke važne odluke pita za tuđe mišljenje i stav, no razvojem Interneta sve više i više njih poseže za mišljenjima i stavovima nepoznatih osoba koje lako mogu pronaći na brojnim društvenim mrežama, portalima, forumima i slično. Unatoč brzim i jednostavnim pristupom informacija na Internetu, postoje određeni problemi, kao što je preveliki broj različitih izvora koji sadrže mnogo različitih mišljenja. Stoga se javlja sve veća potreba za automatskim otkrivanjem mišljenja, odnosno analizom sentimenta. U današnje vrijeme, analiza sentimenta ima široku primjenu u mnogim domenama kao što je proizvodnja potrošačkih proizvoda, političkim izborima, te raznim drugim uslugama. Veliki broj kompanija ima svoje sustave koji im pomažu u analizi sentimenta njihovih korisnika. Kod analize sentimenta uobičajena je i klasifikacija prema polaritetu, koja svrstava rečenice u tri kategorije: pozitivno, negativno i neutralno. Ovaj rad osim teoretskog dijela sadržava istraživački dio koji se bazira na analizi podataka prikupljenih sa portala Index i Jutarnji list u vrijeme izbjegličkog vala u drugoj polovici 2015. godine. U sklopu ovog rada izrađena je frekvencijska analiza riječi u rečenici te je izračunat njihov ukupni polaritet. Rezultati su prikazani u obliku tablica i histograma. Cilj ovog rada je analizirati dobivene podatke te tako zaključiti koji stav su ljudi imali u vezi izbjeglica i općenito izbjegličkog vala. Analiza je pokazala da najviše ima negativnih komentara, između 50 i 70 posto ovisno o tome radi li se o portalu Index ili portalu Jutarnji list. Najčešće korištene riječi u komentarima, odnosno one riječi koje imaju najveću frekvenciju, su: „eu“, „ljudi“ i „izbjeglice“.

Ključne riječi: analiza sentimenta, izbjeglice, polaritet, Index, Jutarnji list, sarkazam

Abstract

Sentiment analysis of Internet texts

Opinion is a central part of all human activity and therefore has a strong influence on how people behave. Most people before making an important decision ask for someone else's opinion, but nowadays with development of the Internet they usually search for opinions and attitudes of strangers that can easily be found on many social networks, portals, forums etc. Despite fast and easy access to information, Internet has certain problems, such as many different sources that contain lots of different opinions. Therefore, there is an increasing need for automatic detection of opinion, respectively sentiment analysis. Today, sentiment analysis is widely used in many domains such as: production of consumer products, political elections, and various other services. Many companies have their own systems to help them analyze the sentiment of their users. It's common to use classification according to polarity when analyzing a sentiment, which ranks sentences in three categories: positive, negative and neutral. This paper besides theoretical part, contains a research part which is based on an analysis of data collected from portals Index and Jutarnji list at the time of the big refugee wave in the second half of 2015. As a part of this work, I made a frequency analysis of words in the sentences and calculated their total polarity. The results are shown in the form of tables and histograms. The aim of this paper is to analyze the collected data, and then see what kind of attitudes people had about refugees and the refugee wave in general. The analysis showed most of the comments are negative, between 50 and 70 percent depending on whether they are from portal Index or Jutarnji list. The most commonly used words in the comments, which are those that have the highest frequency, are: „eu“, „people“ and „refugees“.

Keywords: sentiment analysis, refugees, polarity, Index, Jutarnji, sarcasam

Sadržaj

1. Uvod	5
2. Analiza sentimenta (mišljenja).....	7
3. Klasifikacija prema polaritetu sentimenta	9
3.1 Analiza sentimenta na razini dokumenta.....	10
3.2 Analiza sentimenta na razini rečenice	10
3.3 Analiza sentimenta na razini značajki	11
4. Priča o izbjeglicama	12
5. Analiza podataka o izbjeglicama (portali: Jutarnji list i Index).....	14
6. Kategorizacija komentara.....	23
7. Sarkazam	28
8. Zaključak.....	29
9. Popis literature.....	30

1. Uvod

Mišljenja i stavovi upravljaju našim ponašanjem i time su središnji dio svih ljudskih aktivnosti. Većina ljudi prije nego što mora donijeti neku važnu odluku pita za tuđe mišljenje, bilo da je to od člana obitelji, prijatelja ili poznanika. Jednoj organizaciji koja se bavi proizvodnjom nekog proizvoda je važno zadovoljstvo kupca, odnosno njegovo pozitivno mišljenje i stav o njihovim proizvodima. U prošlosti kada bi pojedinac želio čuti tuđe mišljenje on bi pitao uski krug poznanika, dok bi s druge strane neka organizacija napravila istraživanja na tržištu. No, u današnjem svijetu, nakon velikog razvitka društvenih medija ljudi će se uglavnom koristiti Internetom. On je uvelike promijenio način na koji ljudi izražavaju svoje stavove i mišljenja. Danas ljudi mogu izraziti mišljenja na mnoge načine: putem Facebooka, Twittera, blogova, foruma, raznih portala i slično. Brzi i jednostavni pristup raznih informacijama mijenja način na koji ljudi „traže“ tuđe mišljenje. Osoba više neće biti ograničena mišljenjima i stavovima uskog kruga prijatelja ili članova obitelji nego će ukoliko želi kupiti neki proizvod pronaći stranice na kojima postoji recenzije i diskusije od ljudi koji već imaju i koriste taj proizvod. Organizaciji više nije potrebno provoditi istraživanja na tržištu na način kako je to prije radila te tako trošiti resurse jer sada postoji puno takvih podataka koji su javno dostupni. Međutim, koliko god je Internet olakšao život svima nama, svejedno se javljaju određeni problemi. Najveći od njih je svakako velika količina podataka, pošto pronalaženje izvora na Internetu nije lagani zadatak jer postoji veliki broj različitih izvora i svaki od njih ima puno tekstova koji pored činjenica sadrže mišljenja i stavove. Stoga pojedincu može biti teško pronaći relevantne izvore i stavove i te iste razvrstati i organizirati u upotrebljivi oblik. Potreba za automatskim otkrivanjem mišljenja, odnosno analizom sentimenta, postaje sve veća (Dobrescu, 2011). Svjedoci smo da su u posljednjih nekoliko godina, tekstovi koji sadrže nečije mišljenje i stav uvelike pomogli u razvitku i preoblikovanju raznih poslova. Oni imaju veliki utjecaj u društvenim i političkim sustavima, pogotovo u vrijeme izbora, kao što je u Hrvatskoj ove godine bilo u vrijeme izbora Vlade. Stoga postaje nužno prikupljati i proučavati mišljenja na Internetu. Tekstovi koji sadrže mišljenja i stavove nisu dostupni samo na Internetu, tako primjerice neke tvrtke imaju svoje podatke koje su prikupili na druge načine kao što je: putem e-maila, rezultate raznih istraživanja i anketa, podaci prikupljeni u pozivnim centrima i tako dalje. Zbog svoje ogromne vrijednosti u praktičnim primjenama, analiza sentimenta postaje sve popularnija u mnogim domenama od proizvodnje potrošačkih proizvoda, zdravlja, raznih usluga, političkim izborima i mnogim drugim

društvenim događajima. Sukladno tome dolazi do povećanog rasta broja istraživanja na mnogim akademijama koje se bave analizom mišljenja. Danas samo na prostoru Amerike možemo pronaći između četrdeset i šezdeset start-up tvrtki koje se bave proučavanjem sentimenta. Primjer jednog takvog sistema je *Opinion Parser*, čiji je osnivač Bing Liu, jedan od najpoznatijih ljudi u ovom području analize (Bing, 2012). Sve veće kompanije, kao što su Google, Microsoft, SAP, SAS i mnoge druge, razvile su svoje sustave koji im pomažu u analizi mišljenja (Internetski izvori). U analizi sentimenta veliku ulogu ima proučavanje društvenih mreža. Primjerice, Twitter je jedna od društvenih mreža koja je korištena za predviđanje izbornih rezultata analizom komentara i statusa ljudi. On je također korišten za predviđanje uspjeha filmova i tržišta dionica tako što su analizirani razni podaci kao što je „raspoloženje“ ljudi na njihovim stranicama i filmske recenzije (Agarwal, 2011). Bing Liu sa svojom grupom stručnjaka koristio je *Opinion Parser* za analizu pozitivnih i negativnih mišljenja i stavova o raznim filmovima na Twitteru, rezultati istraživanja su uglavnom bili točni i precizni. Osim društvenih mreža, provedena su razna istraživanja elektronske pošte te uspoređivanje kako različiti spolovi osjećaju drugačije emocije. (Bing, 2012). Analiza mišljenja je korištena i u opisivanju društvenih odnosa, i to je samo jedan dio njene široke primjene.

U sklopu ovog završnog rada izrađena je statistička analiza tekstova koja ide u smjeru analize stavova i mišljenja u tekstu. Program najprije izbaci zaustavne riječi koristeći se popisom koji je izradio Jan Šnajder sa suradnicima. Nakon toga, program uspoređuje riječi sa leksičkim resursom za hrvatski jezik, koji je izradio Marko Modrić u sklopu svog završnog rada, te potom zbraja polaritete riječi koje su pronađene na listi. Na kraju prikaže popis svih riječi koje se nalaze u txt datoteci i njihove frekvencije počevši od najveće. Osim toga, program prikaže konačan zbroj svih polariteta riječi u rečenici i sa tim podatkom možemo vidjeti da li je neka rečenica pozitivna ili negativna. Rad će sadržavati teoretski dio o analizi sentimenta i klasifikaciji prema polaritetu te istraživački dio koji se bazira na analizi prikupljenih komentara i članaka sa portala Index i Jutarnji list. Cilj ovog završnog rada je analizirati dobivene podatke te tako zaključiti koji stav su ljudi imali u vezi izbjeglica i općenito izbjegličkog vala.

2. Analiza sentimenta (mišljenja)

Tekstualne podatke možemo svrstati u dvije glavne kategorije: činjenice i mišljenja. Činjenice možemo definirati kao nešto što možemo neupitno ustanoviti. To su objektivni izrazi o osobama, događajima i njihovim svojstvima. Dok s druge strane mišljenja su obično subjektivni izrazi koji opisuju osjećaje koje neka osoba ima o entitetima i događajima (Bing, 2010). Analiza sentimenta, odnosno drugim riječima analiza mišljenja, pokušava identificirati mišljenje/sentiment koji neki entitet ima o nekom objektu. To uključuje dublju analizu teksta koja je onda uspoređena sa našom subjektivnom analizom. Autori kao što su Pang i Lee, smatraju da je analiza mišljenja dio zadatka detekcije subjektivnosti (Pang, 2003). Analiza sentimenta, klasificira tekstove koje sadrže sentiment u tri kategorije: pozitivno, negativno i neutralno. Neutralno u ovom smislu pripada objektivnoj kategoriji subjektivne analize, iako će mnogi neutralnost definirati kao mišljenje koje nema jasnu tendenciju prema pozitivnom ili negativnom.

Mišljenja imaju tri osnovna dijela. Prvi dio je nositelj mišljenja, koji predstavlja izvor mišljenja. Oni su uglavnom autori tekstova u kojima se izražava neko mišljenje. Primjerice u rečenici „Ana je nezadovoljna dobiven ugovorom“, nositelj mišljenja je Ana. Drugi dio je objekt koji predstavlja cilj mišljenja. Posljednji dio je mišljenje, odnosno pozitivni ili negativni stav o objektu koji ima nositelj mišljenja (Bing, 2012). Mišljenja se mogu izraziti o bilo čemu, na primjer mišljenje o proizvodu, usluzi, osobi, organizaciji, nekoj određenoj temi, itd. Oni predstavljaju objekte, odnosno entitete na kojih se odnosi mišljenje. Objekt može sadržavati skup dijelova i atributa (svojstava). Možemo ga zamisliti kao stablo čiji je korijen sam objekt dok svaka grana predstavlja jedan dio objekta. Osoba tako može izraziti mišljenje o cijelom objektu „Ne sviđa mi se ovaj mobitel“ ili o nekom atributu mobitela „Ovaj mobitel ima dobru kvalitetu zvuka“. Međutim, kako bi se olakšala primjena, ovu definiciju je potrebno pojednostaviti. Jedan od razloga je kompleksnost hijerarhijskog prikaza objekta i mišljenja o objektu. Stoga umjesto korištenja pojmova dijelova i svojstava, objekt ćemo definirati kao osobinu, što u biti predstavlja korijen stabla (Bing, 2012). Razlikujemo 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“. Osim vrste mišljenja, razlikujemo i dva načina izražavanja mišljenja: direktno i komparativno mišljenje. Direktno mišljenje je ono koje osoba daje o objektu i njegovim osobinama, tako da bira riječi ili izraze iz

odgovarajućeg skupa sinonima te na taj način izražava pozitivno, negativno ili neutralno mišljenje o objektu. S druge strane, komparativno mišljenje izražava odnose sličnosti ili različitosti između dva ili više objekta i/ili dvije ili više osobine objekta. Takvo mišljenje se obično izražava pomoću komparativnog ili superlativnog oblika pridjeva ili priloga. Ova vrsta mišljenja često je korištena kod pacijenata kada izražavaju svoje mišljenje o lijeku ili opisuju njegove nuspojave (Dobrescu, 2011). Primjerice, rečenica „Leđa me više ne bole nakon što sam popila ovaj lijek“ opisuje pozitivno djelovanje lijeka na leđa, što u biti daje pozitivno mišljenje o lijeku.

Mišljenja se također razlikuju i u jačini. Pozitivno mišljenje može izraziti osjećaje zadovoljstva, veselja, radosti no ne uvijek u istom intenzitetu. Stoga se razvila klasifikacija mišljenja na osnovu njihove polarnosti. Za to 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). Tablica 1 prikazuje primjer načina ocjenjivanja riječi. Rang ocjenjivanja je od -5 do 5 gdje ocjenom -5 ocjenjujemo najnegativnije riječi, a ocjenom 5 ocjenjujemo najpozitivnije riječi. Osim leksičkih resursa, analiza sentimenta se također fokusira na razvoj sheme 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
wrath	srdžba	-3	Izražava negativnu emociju
brilliant	briljantan	4	Izražava pozitivnu emociju
faithful	vjeran	3	Izražava pozitivnu emociju
outstanding	izvanredan	5	Najpozitivnija ocjena riječi
praise	pohvala	3	Izražava pozitivnu emociju
proudly	ponosno	2	Izražava pozitivnu emociju

3. Klasifikacija prema polaritetu sentimenta

Analiza sentimenta klasificira tekstove koje sadrže mišljenja prema polaritetu emocije koja je izražena. Ako osoba želi primjerice saznati nečije mišljenje o nekom filmu, ukupni sentiment je dovoljan da odlučimo hoćemo li ga pogledati ili ne. No, u slučaju da osoba želi rezervirati sobu u hotelu, ukupni sentiment neće biti dovoljan, s obzirom da osoba može više biti zainteresirana za neke značajke hotela za koje druga osoba nije bila (npr. je li hotel blizu centra grada). Analiza mišljenja stoga zahtijeva druge pristupe ovisno o potrebama korisnika, razini analize i tipu teksta koji se analizira. Ideja klasificiranja sentimenta je zamišljena tako da se jednom dijelu teksta dodaje vrijednost pozitivno, negativno ili neutralno. Pozitivno možemo shvatiti kao „svidanje“, a negativno kao „nesvidanje“. Ono što predstavlja problem analizi sentimenta je velika semantička varijabilnost prirodnih jezika.

Nadalje, kod analize sentimenta bitno je razlikovati onu koja je na razini dokumenta. Dokle god imamo činjenice i ljude, odluka o tome da li je rečenica pozitivna ili negativna ne smije utjecati na presudu o sentimentu. Primjerice, rečenica kao „Velike borbe su izvršene od strane Vlade u borbi protiv financijske krize, što je donijelo mnoge tvrtke u bankrot“ se mora smatrati negativnom jer raspravlja o posljedicama financijske krize. No, ona se mora također smatrati pozitivnom kada se analizira sentiment Vlade (Dobrescu, 2011). Ovaj primjer pokazuje kako polarnost sentimenta uvelike ovisi o načinu na koji je napisan dokument. Osim analiziranja sentimenta na razini dokumenta, razlikovati ćemo analizu na razini rečenice te na razini značajki. U idućim poglavljima ću ukratko objasniti njihovu razliku.

3.1 Analiza sentimenta na razini dokumenta

Ova vrsta analize sentimenta se uglavnom koristi za recenzije filmova, knjiga, proizvoda i slično, ali samo pod pretpostavkom da je ta recenzija napisana od strane jednog nositelja mišljenja i da je ona o jednom objektu, temi ili proizvodu. Različiti autori daju različite pristupe ovoj analizi. Peter Turney analizu koristi za recenzije filmova (Turney, 2002). Bo Pang daje drugi pristup klasifikaciji polariteta sentimenta, autor se tamo koristi strojem koji se zove *Naive Bayes*, koji pokazuje da je korištenje unigrama¹ bolje od korištenja bigrama² (Pang, 2002). 2003. godine, Pang i Lee, klasificiraju recenzije u proširenim mjerilima vrijednosti, a ne samo pozitivno i negativno. Pritom su se koristili učenjem modela pomoću strojeva sa potpornim vektorima *Support Vector Machines (SVM)*. Ishod zatim usporede sa brojem „zvijezdica“ koje su dane u recenziji (Pang, 2003). S druge strane, Goldberg i Zhu predstavili su grafički pristup klasifikacije sentimenta, gdje dokument predstavlja vektor, izračunato na temelju prisutnosti riječi koje sadrže mišljenje. Zatim se dokumenti povežu sa najsljelijima sebi te se na kraju dokumenti klasificiraju na temelju informaciji dobivenih iz grafa i SVM modela (Goldberg, 2006).

3.2 Analiza sentimenta na razini rečenice

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, događa se klasifikacija sentimenta prema polaritetu, pod pretpostavkom da svaka rečenica izražava samo jedno mišljenje, pozitivno ili negativno. Autori Yu i Hatzivassiloglou koriste ovu vrstu analize sa 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 se kreira lista sentimentata pomoću leksičkog resursa koji se zove WordNet³, pronalaze rečenicu koja sadrži nositelja mišljenja te zatim izračunaju sentiment rečenice.

¹ frekvencije pojedinih riječi

² frekvencije parova riječi

³ Izvor: <https://wordnet.princeton.edu/>

3.3 Analiza sentimenta na razini značajki

Unatoč tome što je klasificiranje tekstova na razini dokumenta i rečenice korisno u mnogo slučajeva, postoje ipak neki problemi. Primjerice, „pozitivan“ tekst o nekom objektu ne znači da autor ima dobro mišljenje o svim značajkama tog objekta. Ista stvar je kod „negativnog“ teksta, iako je negativan, to ne znači da se autoru ne sviđa apsolutno sve značajke tog objekta. U normalnom tekstu koji izražava nečije mišljenje i stav možemo pronaći pozitivne i negativne aspekte, iako generalno mišljenje o objektu može biti pozitivno ili negativno (Dobrescu, 2011). Analiza sentimenta na razini dokumenta i rečenice nam ne može pružiti te informacije. Kod analize na razini značajki imamo tri glavna koraka. Prvi korak je odrediti značajku objekta koja je komentirana, npr. u rečenici „Kvaliteta zvuka na ovom mobitelu je odlična.“ značajka objekta je „kvaliteta zvuka“. U drugom koraku određuje se da li su mišljenja o značajkama pozitivna, negativna ili neutralna, u gore navedenom primjeru bi mišljenje o značajki bilo pozitivno. U trećem koraku se grupiraju značajke istog značenja nakon čega se izračuna polaritet i dobije se rezultat koji je prikazan u postocima pozitivnog i negativnog mišljenja o svakoj značajki (Dobrescu, 2011).

4. Priča o izbjeglicama

Rat u Siriji traje od proljeća 2011. godine, no unatoč tome izbjeglice nisu do sada dolazile na područje Europe jer prema europskom zakonu o migraciji, izbjeglice mogu dobiti azil samo u onoj europskoj državi u koju prvu uđu. Pozivom njemačke predsjednice vlade Angele Merkel krajem ljeta 2015. godine, počinje veliki izbjeglički val u Europi i masovna migracija. Pitanje koje se svakome nametnulo na umu je: „Zašto izbjeglice nisu do sada pobjegli u Tursku ili neku drugu obližnju zemlju?“. Turska je do sada uložila približno 2 milijarde američkih dolara za njihov smještaj, no zbog konvencije Ujedinjenih naroda o izbjeglicama iz 1951., Turska je obvezna prihvaćati isključivo izbjeglice iz europskih zemalja (Ženevska konvencija, 1951). Drugim riječima, to znači da se Sirijci tamo ne mogu zaposliti niti dobiti trajno boravište, stoga ne preostaje im ništa drugo nego otići u Europu. Ostale zemlje kao što su Saudijska Arabija, Oman, Katar i slično, ih jednostavno ne žele prihvatiti, unatoč tome što se čini logičnije da otiđu tamo, najviše zbog toga što tamo žive ljudi koji pripadaju istoj vjerskoj zajednici. Također, zemlje Perzijskog zaljeva su im bliže nego zemlje Europe. No, nakon poziva Angele Merkel koja im je obećala posao i smještaj Sirijci kreću prema Njemačkoj, najvećoj europskoj ekonomskoj sili, u nadi da će započeti novi i bolji život, daleko od rata i opasnosti (Marjanović, 2015).

Kada je riječ o izbjeglicama, ljudi su uglavnom podijeljena mišljenja, neki osjećaju strah i prezir dok im drugi žele pružiti pomoć. Ono što je svih zabrinulo jest činjenica da velika većina izbjeglica koje ulaze u Europu čine mladi muškarci sposobni za borbu, a time i obranu svoje zemlje. Stoga se nameće pitanje: zašto nisu ostali u Siriji i borili se za svoju zemlju? Na portalu „dnevno.hr“ to objašnjavaju teškom političkom situacijom u državi. Sukladno tome neki ljudi jednostavno nisu mogli odabrati stranu za koju će se boriti (Dnevno.hr, 2015). No, problem ne leži u tome, pravi problem predstavlja činjenica da je za vrijeme izbjegličkog vala u Europu ušao veliki broj terorista, koji su pripadnici ISIL-a⁴, nepriznate države koja je zauzela veliki dio Iraka i Sirije (Wikipedia). Koliko je to zapravo veliki problem pokazalo se brojnim napadima koji su slijedili nakon dolaska izbjeglica na područje Europe, za čije napade je preuzeo odgovornost upravo ISIL. Najveći od tih napada je onaj u Parizu koji se dogodio 13. studenog u kojem je smrtno stradalo oko 130 ljudi (Wikipedia). Uzevši sve u obzir nije ni čudo da većina ljudi ima negativno mišljenje o izbjeglicama. Na Internet portalima možemo naći brojne negativne

⁴ Islamic State of Iraq and the Levant (Izvor: Wikipedia)

komentare o izbjeglicama i njihovom dolasku u Europu, primjer jednog takvog je komentar jedne osobe na Index portalu, koja smatra: „Islam nikada i nikako ne pripada EU!! Hrvatska Vlada i dalje vodi politiku "šalji dalje" i neodgovorno bez detaljne kontrole propušta ljude među kojima se nalazi preko 90% potencijalnih terorista i islamskih terorista i mirne duše i bez imalo grižnje savjesti ih šalje dalje u EU i time ujedno podržava terorizam islamskih ekstremista“. No, u moru negativnih komentara se može pronaći i koji pozitivan komentar, na primjer: „Podrška dolasku imigranata.“ ili „Ovo je Hrvatska kojom se možemo ponositi!“. Međutim, ako uzmemo sve u obzir ni jedna strana nije u pravu, može se reći da smo prihvaćenjem izbjeglica prihvatili i dobro i loše. Istina je da se među izbjeglicama mogu naći potencijalni teroristi koji mogu učiniti katastrofu, no nemojmo generalizirati, unatoč svemu, veći dio izbjeglica čine ljudi koji bježe od rata, ljudi koji su izgubili sve što su imali na ovom svijetu i koji trebaju našu pomoć i ono što je Europljanima najprivlačnije, oni mogu biti nova jeftina radna snaga. S jedne strane imamo školovane Sirijce koji će poboljšati razvoj u državama diljem Europe, a s druge strane imamo moguće kriminalce, nasilnike, ljude koji neće htjeti raditi i koji mogu biti opasnost za sve nas. Međutim, takvi smo mi ljudi, različiti smo, ima nas dobrih i ima nas loših, da bar postoji svijet u kojem nema rata i u kojem ljudi žive u miru bez da osjećaju mržnju jedni prema drugima, no takav svijet nažalost ne postoji. Stoga, iako postoji veliki rizik od daljnjih katastrofa, mi Europljani smo pokazali veliko srce i solidarnost prema ljudima. Reći „ne“ za pomoć izbjeglicama je praktički ih osuditi na smrt. No, kad je riječ o izbjeglicama se ne smijemo zavaravati, nisu sve izbjeglice loše, ali isto tako nisu niti dobre. Postoji određeni rizik puštanja velikog broja nepoznatih ljudi u našu zemlju, ali isto tako to je rizik koji moramo prihvatiti za dobrobit cijelog čovječanstva.



Slika 1- Izbjeglice u Europi (Izvor: <http://www.express.co.uk/news/world/623651/Refugee-crisis-Germany-planning-move-500000-Syrians-Turkey-EU>, pristupljeno 9.8.2016.)

5. Analiza podataka o izbjeglicama (portali: Jutarnji list i Index)

U ovom završnom radu koristila sam se komentarima i člancima preuzetih sa portala *Jutarnji list* i *Index* u vrijeme kada su izbjeglice useljavale u Europu i time na svom putu prolazile i kroz našu državu, materijale su pripremili kolege Kathrin Maeusl i Edi Čiković. U sklopu prvog dijela svog zadatka, ručno sam odvojila komentare u posebne datoteke ovisno o tome da li je komentar po meni bio pozitivan, negativan ili neutralan. Kao što se moglo pretpostaviti, ljudi su uglavnom pisali negativne komentare, dok je velika većina članaka napisana neutralno. Pomoću programa koju sam izradila u sklopu ovog rada, analizirala sam podatke, frekvencije riječi i njihov polaritet te onda pomoću tablica uspoređivala dobivene rezultate, odnosno datoteke sa zaustavnim riječima sa datotekama koje ne sadrže zaustavne riječi. U tablici 2 možemo vidjeti opće informacije o korištenim datotekama. Prvi red tablice prikazuje ukupan broj txt datoteka (u jednoj txt datoteci se nalazi jedan komentar/članak). Drugi red tablice prikazuje ukupan broj riječi svih txt datoteka uključujući i zaustavne riječi⁵, a treći red prikazuje ukupan broj riječi nakon što se izbace zaustavne riječi. Nadalje, slovo J u prvom stupcu označava članke s Jutarnjeg lista, slovo I u drugom stupcu članke sa Indexa, slovo KJ u trećem stupcu komentare sa Jutarnjeg lista, a KI u posljednjem stupcu označava komentare sa Indexa.

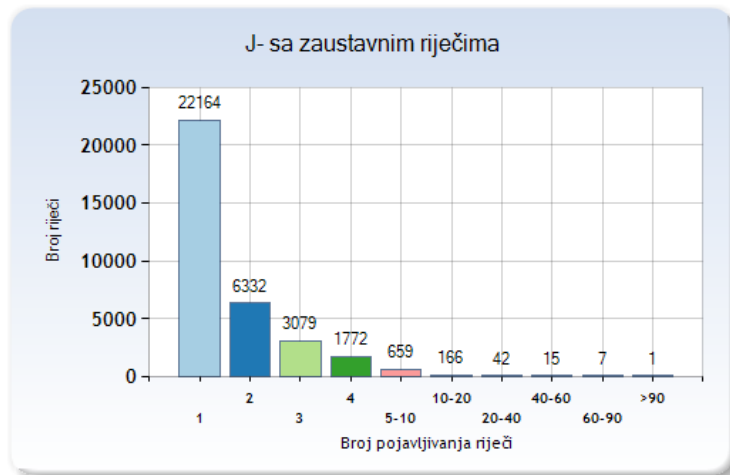
Tablica 2-Opće informacije o korištenim datotekama

	J	I	KJ	KI
Ukupan broj	901	487	1100	1230
Ukupan broj riječi sa zaustavnim riječima	40679	22442	17143	11000
Ukupan broj riječi bez zaustavnih riječi	37385	20400	15645	10381

U nastavku ćemo prikazivati dva različita histograma koji prikazuju frekvencije riječi, odnosno broj pojavljivanja riječi u txt datoteci, a to su: histogrami sa zaustavnim riječima i bez zaustavnih riječi. Na histogramu X-os prikazuje broj pojavljivanja pojedine riječi, a Y-os prikazuje ukupan broj riječi s tom frekvencijom, primjerice ukupan broj riječi sa frekvencijom 3 je 1772. U nastavku ću prvo prikazati tablice i histograme za portal Jutarnji list, a potom za Index.

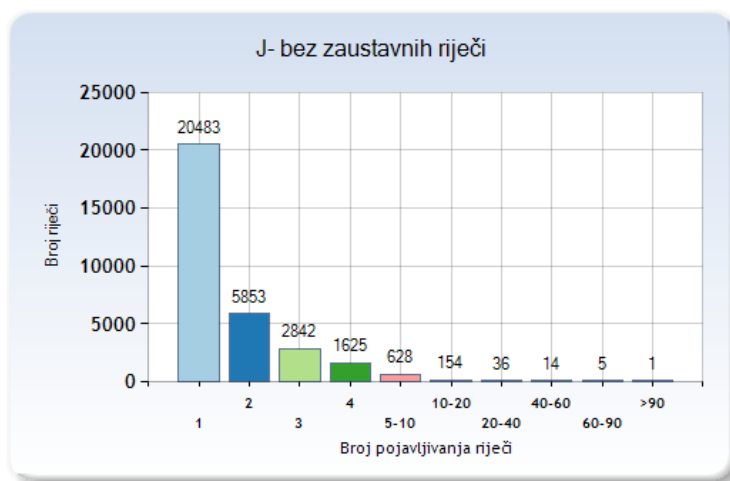
⁵ riječi prirodnog jezika koje imaju sintaksnu ulogu, a nemaju samostalno značenje kao npr. „, a“, „, ali“, „, još“, „itd

Slika 3 prikazuje histogram na kojem možemo vidjeti broj pojavljivanja riječi u člancima sa portala Jutarnji list uključujući i zaustavne riječi. Primjećujemo kako je najviše riječi sa frekvencijom 1, točnije njih 22164, što je 54% od ukupnog broja riječi. Dok s druge strane prosječno postoji samo jedan primjer riječi sa istom frekvencijom većom od 90. Taj broj se dobio tako što se prvo zbroje sve riječi sa istom frekvencijom i onda se dobiveni broj podijeli sa ukupnim brojem različitih frekvencija većih od 90.



Slika 2- Histogram frekvencija riječi- Jutarnji članci sa zaustavnim riječima

Slika 4 prikazuje histogram koji sadrži frekvencije riječi iz članaka Jutarnjeg lista, nakon što se izbace zaustavne riječi. Možemo primjetiti promjene u brojkama, no postotak riječi sa frekvencijom 1 ostaje isti kao i u prethodnom histogramu (54%).



Slika 3- Histogram frekvencija riječi - Jutarnji članci bez zaustavnih riječi

U tablicama 3 i 4 možemo vidjeti popis prvih nekoliko riječi iz članaka Jutarnjeg lista i to onih sa najvećom frekvencijom. Tablica 3 prikazuje listu riječi uključujući i zaustavne riječi. U toj tablici se riječ „izbjeglica“ nalazi tek na 21. mjestu po broju pojavljivanja. S druge strane, u tablici 4, odnosno onoj koja ne sadrži zaustavne riječi, riječ izbjeglica ima najveću frekvenciju, a iza nje slijede riječi „izbjeglice“, „ljudi“, „biti“ i „hrvatska“.

Tablica 3- Lista riječi sa frekvencijom

(sa zaustavnim riječima-J)

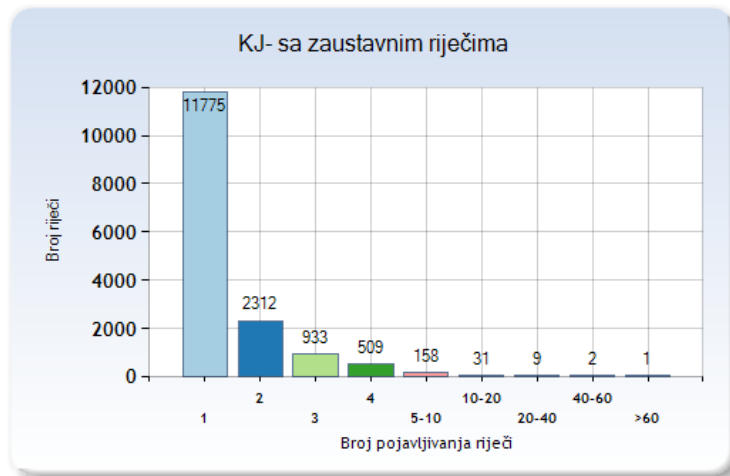
1.	je	11732
2.	u	10735
3.	i	10710
4.	da	6070
5.	se	5639
6.	na	4973
7.	su	4417
8.	za	3041
9.	a	2532
10.	s	2360
11.	od	2301
12.	koji	2293
13.	ne	2178
14.	će	1838
15.	to	1829
16.	iz	1685
17.	što	1658
18.	bi	1620
19.	kako	1509
20.	o	1322
21.	izbjeglica	1138

Tablica 4- Lista riječi sa frekvencijom

(bez zaustavnih riječi-J)

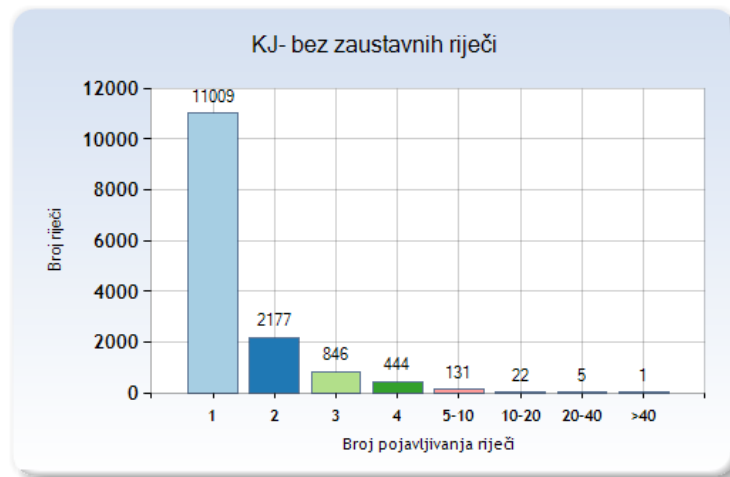
1.	izbjeglica	1151
2.	izbjeglice	961
3.	ljudi	929
4.	biti	668
5.	hrvatska	629
6.	rekao	587
7.	kad	522
8.	eu	503
9.	granice	498
10.	migranata	494
11.	hrvatskoj	437
12.	hrvatske	422
13.	ima	421
14.	dana	408
15.	hrvatsku	394
16.	kaže	387
17.	godine	384
18.	izbjeglicama	383
19.	zemlje	358
20.	nema	331
21.	granicu	323

Slika 5 prikazuje histogram koji sadrži frekvencije riječi izvučenih iz komentara sa portala Jutarnji list, uključujući zaustavne riječi. Ovdje također imamo najveći broj riječi sa frekvencijom 1 i to čak 68%, najmanji broj riječi je onih sa frekvencijom većom od 60.



Slika 4- Histogram frekvencija riječi- KJ sa zaustavnim riječima

Slika 6 prikazuje histogram frekvencija riječi iz komentara s Jutarnjeg lista, bez zaustavnih riječi. Najmanje ima riječi sa frekvencijom većom od 40, a najviše sa frekvencijom 1 kao što je bilo i u prethodnim primjerima.



Slika 5- Histogram frekvencija riječi- KJ bez zaustavnih riječi

Tablica 5 prikazuje popis prvih nekoliko riječi koje su najčešće korištene u komentarima Jutarnjeg lista, uključujući zaustavne riječi. Dok tablica 6 prikazuje popis nakon što se izbace zaustavne riječi. U tablici 5 prvih 40-ak riječi sa najvećom frekvencijom čine zaustavne riječi, a tek na 46. mjestu pronalazimo riječ „izbjeglice“. U tablici 6 vidimo kako su ljudi najčešće koristili riječi „eu“, „ljudi“ i „izbjeglice“. Tablica 7 prikazuje riječi sa frekvencijom 31 i 32 gdje možemo pronaći riječi kao: „dijete“ ili „izbjeglicama“.

Tablica 5- Lista riječi sa frekvencijom (KJ- sa zaustavnim riječima)

1.	i	2483
2.	je	1720
3.	u	1623
4.	da	1392
5.	se	1189
6.	na	764
7.	ne	756
8.	su	746
9.	a	682
10.	to	591
...
40.	ljudi	145
41.	do	142
42.	ce	141
43.	koje	141
44.	biti	139
45.	sad	132
46.	izbjeglice	130

Tablica 6- Lista riječi sa frekvencijom

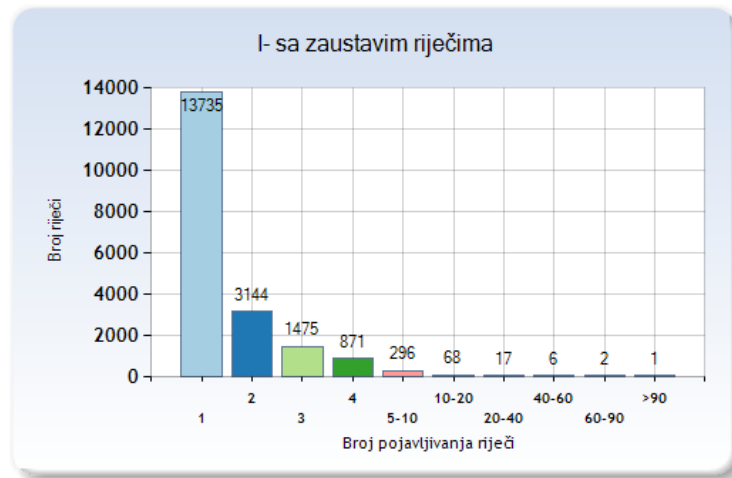
(KJ-bez zaustavnih riječi)

1.	eu	191
2.	ljudi	145
3.	izbjeglice	130
4.	sto	120
5.	izbjeglica	92
6.	nema	85
7.	hrvatska	74
8.	rat	68
9.	zemlje	67
10.	narod	63

Tablica 7- Lista riječi sa frekvencijom 31 i 32

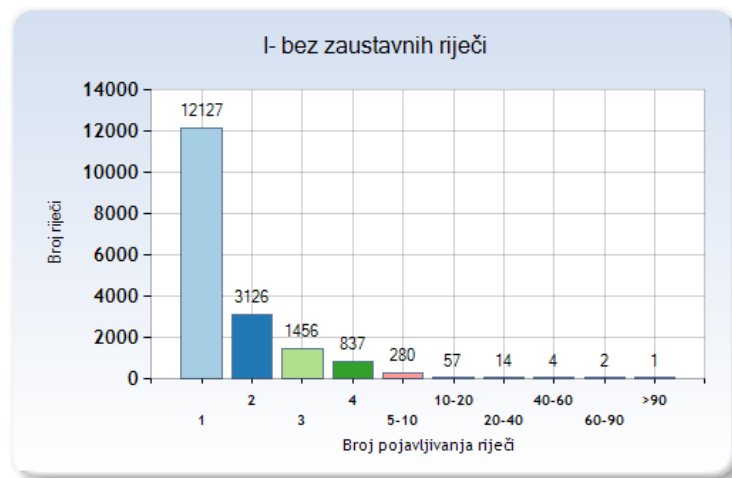
202.	jasno	32
203.	nikad	32
204.	rekao	32
205.	drugi	32
206.	jedna	32
207.	tom	32
208.	vlada	31
209.	nisam	31
210.	dijete	31
211.	izbjeglicama	31

Idućih nekoliko histograma odnose se na podatke prikupljene iz datoteka vezanih za portal Index, uočiti će se neka odstupanja no ne prevelika od Jutarnjeg lista. Slika 7 prikazuje histogram koji sadrži frekvencije vezane za Index članke uključujući i zaustavne riječi. Možemo uočiti kako opet najviše riječi ima frekvenciju 1 (oko 55%). Također, puno je različitih frekvencija većih od 90, stoga u prosjeku jedna riječ ima određenu frekvenciju. Poslije će se u tablici 7 vidjeti koje su riječi najčešće korištene.



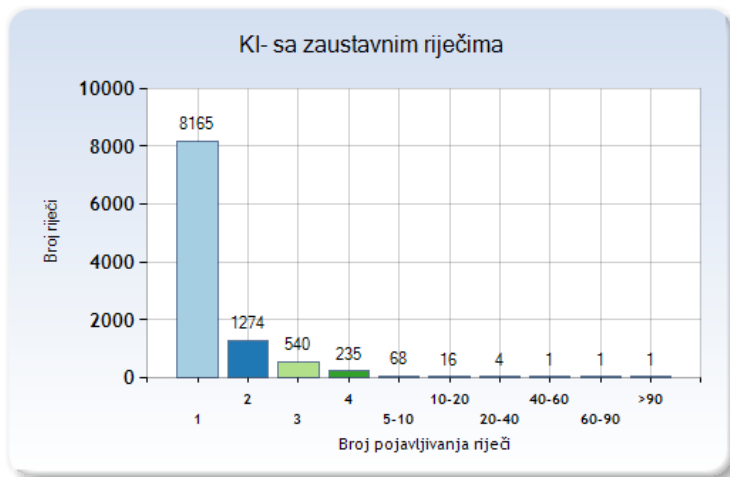
Slika 6- Histogram frekvencija riječi- I sa zaustavnim riječima

Slika 8 prikazuje histogram koji sadrži broj pojavljivanja riječi u člancima nakon što se izbace zaustavne riječi. Najveći broj riječi ima frekvenciju 1, a najmanji broj riječi ima frekvenciju veću od 90.



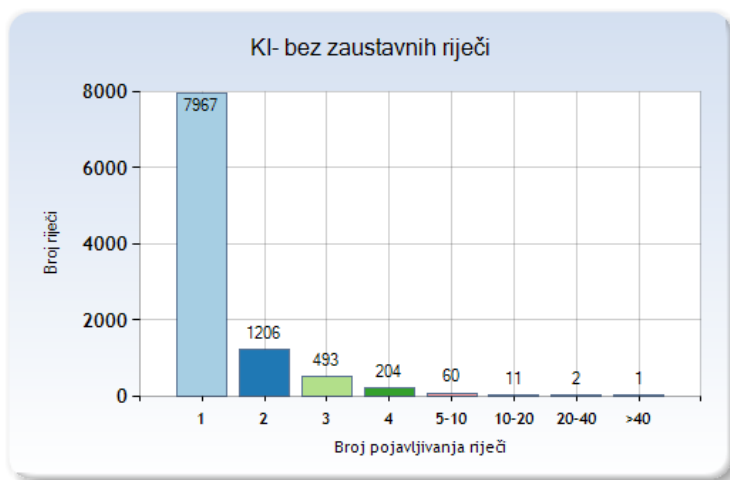
Slika 7- Histogram frekvencija riječi- I bez zaustavnih riječi

Slika 9 prikazuje histogram koji se odnosi na komentare sa Index portala. 74% riječi ima frekvenciju 1, 11% frekvenciju 2, 4% frekvenciju 3, 4% frekvenciju 4, a ostale su frekvencije u jako malom postotku.



Slika 8- Histogram frekvencija riječi- KI sa zaustavnim riječima

Slika 10 također prikazuje histogram koji sadrži frekvencije riječi iz komentara sa Index portala, ali bez zaustavnih riječi. Čak 76% riječi ima frekvenciju 1, a frekvenciju veću od 40 ima samo 11 riječi i one će biti prikazane u tablici 11.



Slika 9- Histogram frekvencija riječi- KI bez zaustavnih riječi

Tablice 8 i 9 prikazuju najčešće korištene riječi u člancima na Index portalu, tablica 8 sadrži zaustavne riječi dok ih tablica 9 ne sadrži. Ako pogledamo tablicu 8, možemo primijetiti kako je riječ „izbjeglica“ korištena 747 puta, a riječ „migranata“ 475 puta i ako ignoriramo zaustavne riječi to su u biti riječi sa najvećom frekvencijom što će i potvrditi tablica 9.

Tablica 8- Lista riječi sa frekvencijom

(I-sa zaustavnim riječima)

1.	je	4878
2.	u	4581
3.	i	3751
4.	da	2380
5.	na	2162
6.	se	2040
7.	za	1510
8.	su	1315
9.	od	988
10.	a	915
11.	s	872
12.	koji	808
13.	će	759
14.	izbjeglica	747
15.	iz	713
16.	bi	690
17.	o	671
18.	ne	662
19.	kako	641
20.	migranata	475

Tablica 9- Lista riječi sa frekvencijom

(I- bez zaustavnih riječi)

1.	izbjeglica	747
2.	migranata	475
3.	rekao	414
4.	izbjeglice	362
5.	eu	313
6.	biti	256
7.	ljudi	250
8.	granice	225
9.	poslova	214
10.	godine	207
11.	hrvatsku	190
12.	granici	182
13.	dana	166
14.	kazao	159
15.	zemlje	157
16.	hrvatske	156
17.	merkel	152
18.	izbjeglicama	151
19.	unutarnjih	148
20.	slovenija	146

U tablicama 10 i 11 možemo vidjeti prvih 20 najčešće korištenih riječi u Index komentarima. Tablica 10 sadrži zaustavne riječi, a tablica 11 ih ne sadrži. U tablici 10 riječ „izbjeglice“ nalazi se na 44. mjestu sa frekvencijom 75, osim te riječi, tu frekvenciju ima i riječ „nisu“. U tablici 11 možemo vidjeti da je najviše korištena riječ u Index komentarima riječ „ljudi“ sa frekvencijom 96.

Tablica 10- Lista riječi sa frekvencijom

(KI-sa zaustavnim riječima)

1.	i	1272
2.	je	798
3.	da	747
4.	u	746
5.	se	614
6.	su	428
7.	ne	422
8.	na	387
9.	za	292
10.	a	274
11.	to	266
12.	bi	203
13.	će	192
14.	od	191
15.	koji	179
16.	ih	177
17.	sve	155
18.	sa	154
19.	nije	147
20.	samo	145
...
44.	izbjeglice	75

Tablica 11- Lista riječi sa frekvencijom

(KI-bez zaustavnih riječi)

1.	ljudi	96
2.	eu	88
3.	kad	88
4.	biti	85
5.	ce	79
6.	izbjeglice	74
7.	sto	65
8.	nema	61
9.	ima	59
10.	sad	57
11.	ko	40
12.	izbjeglica	35
13.	godina	34
14.	hrvatska	33
15.	puno	32
16.	zna	32
17.	žicu	31
18.	jos	31
19.	problem	28
20.	dobro	28

6. Kategorizacija komentara

Kao što je već rečeno, u sklopu prvog djela zadatka za ovaj završni rad ručno sam sortirala komentare ovisno o tome jesu li oni prema mojem mišljenju pozitivni, negativni ili neutralni. Ukupno sam sortirala 1230 komentara sa Index portala i 1100 sa Jutarnjeg lista. Tablica 12 prikazuje koliko ih je prema mom kriteriju bilo ukupno pozitivnih, negativnih i neutralnih. Sa ovom informacijom možemo otprilike vidjeti kakav su stav ljudi imali o izbjeglicama. Čitajući tablicu 12, možemo zaključiti da su ljudi najviše pisali negativne komentare, dok je njih jako mali postotak napisalo nešto pozitivno.

Tablica 12- Opće informacije o komentarima (Jutarnji list i Index)

	Jutarnji list	Index
Ukupan broj	1100	1230
Broj pozitivnih komentara	24	47
Broj negativnih komentara	715	671
Broj neutralnih komentara	361	512

Tablica 13 prikazuje 10 najčešće korištenih riječi u pozitivnim komentarima na portalu Index. Riječ „bravo“ ima najveću frekvenciju, a iza nje slijede riječi „čast“ i „podrška“.

Tablica 13- Lista riječi sa frekvencijom (Pozitivni komentari-KI)

1.	bravo	13
2.	čast	7
3.	podrška	6
4.	hrvatska	3
5.	pravu	3
6.	njemci	2
7.	ima	2
8.	zemlji	2
9.	nema	2
10.	dobro	2

Koliko mržnje se može naći na Internetu pokazuje činjenica da je samo 47 komentara od njih 1230 bilo pozitivno, no nije stvar samo komentara na temu izbjeglica, nego se generalno može reći da se na Internetu uglavnom mogu pronaći negativni komentari i to nažalost u puno većem

broju nego pozitivni. Primjer jednog pozitivnog komentara sa portala Index je: „Svaka čast i hvala policiji, crvenom križu, volonterima i svima ostalima koji s toliko savjesti i humanosti obavljaju taj težak zadatak! To je jedna od rijetkih stvari u ovoj zemlji na koju sam doista ponosan, bez obzira na sve ostale okolnosti i tko što o tome pričao“. S druge strane, negativnih komentara je čak 54% od ukupnog broja komentara. U tablici 14 mogu se vidjeti prvih 10 najčešće korištenih riječi u negativnim komentarima. Iako ova tablica ne sugerira da se radi o negativnim komentarima, ako pogledamo nastavak liste možemo uočiti neke pogredne riječi koje su također bile često korištene u tim komentarima, kao primjerice: „glupi“, „problem“, „terorist“ i slično. Primjer jednog negativnog komentara je: „Index i ekipa koja toliko svesrdno podržava socijalizam i invaziju Europe ljudi iz zemalja trećeg svijeta su ništa nego korisni idioti, koji će, ostvari li se plan koji toliko priželjkuju, biti prvi koji će biti poredani uza zid od strane istih tih koje podržavaju“.

Tablica 14- Lista riječi sa frekvencijom (Negativni komentari-KI)

1.	eu	66
2.	ljudi	55
3.	sad	50
4.	izbjeglice	43
5.	ima	41
6.	biti	41
7.	nema	36
8.	komentara	28
9.	godina	23
10.	žicu	21

Uz negativne i pozitivne komentare postoje i oni koje ne možemo svrstati ni u jednu ni u drugu kategoriju, a nazivamo ih neutralni komentari. Takvih komentara je uz negativne bilo najviše, čak 41% od ukupnog broja komentara. Primjer jednog neutralnog komentara sa Index portala je: „Zašto? Pa kamo će pusti inženjeri, nuklearni fizičari i doktori?“. U tablici 15 mogu se vidjeti 10 najčešće korištenih riječi u neutralnim komentarima.

Tablica 15-Lista riječi sa frekvencijom (Neutralni komentari-KI)

1.	ljudi	48
2.	biti	43
3.	kad	38
4.	izbjeglice	34
5.	nema	29
6.	sad	25
7.	komentara	24
8.	eu	23
9.	izbjeglica	21
10.	ima	17

Tablica 16 prikazuje 10 najčešće korištenih riječi u pozitivnim komentarima na portalu Jutarnji list. Riječ „bravo“, kao i kod Index portala, ima najveću frekvenciju, a nakon nje slijede riječi „dobar“ i „zemlje“ sa frekvencijom 3.

Tablica 16- Lista riječi sa frekvencijom (Pozitivni komentari-KJ)

1.	bravo	7
2.	dobar	3
3.	zemlje	3
4.	pravu	3
5.	ljubavi	2
6.	hrvatske	2
7.	hvala	2
8.	puno	2
9.	jedini	2
10.	vidim	2

Primjer jednog pozitivnog komentara: „Bravo Orban, on jedini javno govori da su Ameri umjetno izazvali ovu krizu. Jedini štiti budućnost svoje zemlje i svog naroda.“ Samo 24 komentara od njih 1100 je pozitivno, što je manje nego na Index portalu. No, negativnih ima više na portalu Jutarnji list, čak 65% od ukupnog broja komentara. Tablica 17 prikazuje 10 najčešće korištenih riječi u negativnim komentarima.

Tablica 17- Lista riječi s frekvencijom (Negativni komentari-KJ)

1.	eu	98
2.	ljudi	79
3.	izbjeglice	63
4.	nema	50
5.	ima	49
6.	izbjeglica	41
7.	zemlje	37
8.	narod	36
9.	godina	32
10.	hrvatski	31

Kao i kod Index portala, 10 najčešće korištenih riječi u negativnim komentarima nije pogrdno niti sugerira na negativne komentare, no ako pogledamo ostatak liste, možemo pronaći riječi kao: „gluposti“, „fuj“, „ubojica“ i slično. Primjer negativnog komentara je: „Budaletine iz Jutarnjeg nemaju pametnijeg posla nego se preseravat u kamionu svojom lažnom sućuti za to lažno izbjegličko smeće koje će okupirat i uništiti Europu...“. Ovo nije jedini primjer u kojem ljudi vrijeđaju novinare, naprotiv, najčešći komentari su bili na temu vrijeđanja novinara i sadržaja napisanog u članku.

Nakon pozitivnih i negativnih, u tablici 18 slijedi 10 najčešće korištenih riječi u neutralnim komentarima kojih je nakon negativnih bilo najviše, oko 32% od ukupnog broja komentara.

Tablica 18- Lista riječi sa frekvencijom (Neutralni komentari-KJ)

1.	eu	41
2.	ljudi	34
3.	izbjeglice	27
4.	ima	23
5.	hrvatska	23
6.	granice	20
7.	rat	18
8.	izbjeglica	18
9.	rata	16
10.	zašto	15

Najčešća korištena riječ u neutralnim komentarima je skraćunica od Europe „eu“, nakon nje slijede riječi „ljudi“ i „izbjeglice“ sa frekvencijama 34 i 27.

Tablica 19 prikazuje 10 nasumično odabranih riječi i njihove frekvencije ovisno o tome da li su napisane u sklopu pozitivnih, negativnih ili neutralnih komentara. Zanimljivo je primjetiti kako se riječ „dobro“ najčešće pojavljuje u negativnim komentarima iako sama po sebi sugerira nešto pozitivno, no ako uzmemo u obzir da je negativnih komentara bilo najviše onda to ne iznenađuje.

Tablica 19- Usporedba riječi i njihovih frekvencija (negativni, pozitivni i neutralni komentari-KI)

Riječi	Negativni kom.	Pozitivni kom.	Neutralni kom.
dobro	21	2	1
podrška	0	6	2
granica	6	1	4
bravo	3	13	1
izbjeglica	19	0	21

Sve ove tablice koje su bile u ovom poglavlju prikazuju isključivo moje mišljenje o tome je li neki komentar negativan, pozitivan ili neutralan. Teško je biti objektivan u procjenjivanju komentara jer svi mi imamo neka svoja razmišljanja o određenoj temi, unatoč tome, ja sam pokušala biti što objektivnija u sortiranju komentara. Međutim, nekima će možda komentar koji sam ja „ocijenila“ negativno biti pozitivan ili neutralan. Jedna rečenica može biti na više načina interpretirana ovisno o tome tko ju čita, kao primjerice rečenica: „Održati će dekanski rok iz xy kolegija“, meni kao studentu predstavlja nešto pozitivno dok profesoru koji drži taj kolegij to ne predstavlja jer mu to znači više posla. Iduće poglavlje govori o još nekom problemu koji sam uočila prilikom pisanja ovog rada.

7. Sarkazam

Ono što predstavlja najveći problem i izazov u analizi sentimenta je sarkazam. Bratoljub Klaić u *Rječniku stranih riječi* definira sarkazam kao „zlobnu, ljutu, zajedljivu, pakosnu i oštru porugu“ te ga još naziva pojačanom ironijom (Klaić, 1990). Prepoznati sarkazam nije lako ni ljudima, a kamoli računalu. Analiza sentimenta može lako biti „prevarena“ ako su u rečenici prisutne riječi koje imaju snažan polaritet, a koje su korištene sarkastično, što u principu znači da se mislilo na suprotan polaritet. Taj problem primijetila sam i u sklopu ovog završnog rada. Naime, ako tamo unesem rečenicu: „Naprosto obožavam kada mi zakasni bus“, program će izbaciti zbroj polariteta riječi u rečenici i to će rezultirati sveukupno pozitivnim polaritetom. No svatko može primjetiti da je osoba koja je izrekla ovu rečenicu mislila suprotno od izrečenog, odnosno da ne voli kada joj bus kasni. Međutim, računalo u ovom slučaju to nije moglo otkriti pošto se program bazira na tome da zbroji polaritete svake riječi u rečenici te nakon toga prikaže ukupan zbroj. Zbog toga, neke komentare koje sam ja „ocijenila“ kao negativne, program detektirao kao pozitivne ili neutralne. Primjerice u rečenici: „Dobro došli drage izbjeglice!“ je očito da se radi o sarkazmu jer je autor koristio navodne znakove koji sugeriraju na to. Također, u rečenici: „Baš mi je čudno zašto migranti izbjegavaju Hrvatsku“, možemo lako uočiti sarkazam zbog činjenice da u posljednjih nekoliko godina sve više i više mladih napušta našu zemlju, stoga „čuđenje“ u rečenici nema smisla. Onda postavljamo pitanje: Zašto je računalu tako teško prepoznati sarkazam? Kao prvo, da bi uočili sarkazam mi moramo znati da osoba koja je izrekla neku rečenicu ne misli to za stvarno, što bi značilo da moramo znati njeno pravo mišljenje o nečemu. Upravo to „prepoznavanje laži“ je ono što predstavlja problem računalu. Postoje računalni programi koji mogu prepoznati sarkazam. Jedan od njih radi na principu da svaki put kada uoči naglo mijenjanje polariteta, primjerice u rečenici: „Naprosto obožavam kada mi zakasni bus.“, će prepoznati sarkazam jer uočava kontrast između pozitivne emocije i negativne situacije. Upravo ta struktura rečenice otkriva sarkazam, pošto se u raznim istraživanjima pokazalo kako većina rečenica sa tom strukturom sarkastična (Riloff, 2013). Osim ovog programa, postoje razni drugi koji imaju svoje načine otkrivanja sarkazma.

8. Zaključak

Analiza sentimenta, zbog svoje velike koristi u praktičnim primjenama, postaje sve češće korištena u mnogim kompanijama koji žele istražiti mišljenja svojih klijenata. Tom analizom pokušava se identificirati mišljenje odnosno sentiment koji neki pojedinac ima o nekoj određenoj temi ili proizvodu. Ovaj rad bazirao se na analizi prikupljenih podataka sa portala Index i Jutarnji list na temu izbjeglica i izbjegličkog vala. Analizom se utvrdilo kako je najviše ljudi pisalo negativne komentare, ponajviše zbog straha od terorizma i zbog mnogih neodgovorenih pitanja koje ljudi postavljaju, kao primjerice: „Zašto veliki broj izbjeglica čine mladi muškarci?“ ili „Zašto nisu išli u obližnje zemlje?“. Unatoč velikom postotku negativnih komentara (na Index portalu 54%, a na Jutarnjem listu 65%), postoji mali broj pojedinaca koji su pisali pozitivne komentare na kojima uglavnom podržavaju volontere koji su danonoćno pomagali nemoćnim izbjeglicama u azilima diljem Europe. Najčešće korištene riječi u pozitivnim komentarima su: „bravo“, „podrška“ i „dobar“, dok su u negativnim i neutralnim komentarima najveću frekvenciju imale riječi: „eu“, „ljudi“ i „izbjeglice“. Članci su za razliku od komentara u velikoj većini napisani neutralno.

U svom radu primjetila sam probleme koji su smanjili točnost rada programa koji je izrađen u sklopu ovog rada. Glavni problem je svakako sarkazam, koji računalo nije moglo detektirati. Općenito, analiza sentimenta lako može biti „prevarena“ ako su u rečenici prisutne riječi koje imaju snažan polaritet, a koje su korištene sarkastično. Smatram da bi se u daljnjem radu trebalo u program dodati funkciju pomoću koje bi se prepoznao sarkazam, time bi se poboljšala točnost informacija dobivenih analizom bilo kojih podataka, a ne samo ovih na temu izbjeglica.

9. Popis literature

1. Dobrescu, A. 2011. *Methods and Resources for Sentiment Analysis in Multilingual Documents of Different Text Types*. Doktorski rad. Universitat d'Alacant.
2. Bing, L. 2010. *Sentiment Analysis and Subjectivity*. Izdano u knjizi: N.Indurkha i F.J. Damerau *Handbook of Natural Language Processing*, Second Edition.
3. Bing, L. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.
4. Kathrin Maeusl i Edi Čiković, seminar FJJP1, ak. god. 2015./16, preuzeto ožujak 2016.
5. 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.
6. Jan Šnajder, *Popis hrvatskih stop riječi*, preuzeto: srpanj 2016.
7. Strapparava, C. i Mihalcea, R. 2008. *Learning to Identify Emotions in Text*. University od North Texas.
8. Marjanović, D. 2015. *Razumijevanje tragične izbjegličke krize i odgovori na 2 ključna pitanja: Zašto stižu mahom mladi muškarci i zašto tek sada ako rat u Siriji traje još od 2011.?.*, preuzeto: kolovoz 2016.
9. Raunič, F. 2015. *Postoji jedan vrlo praktičan razlog zašto Njemačka prima izbjeglice*. <http://www.telegram.hr/politika-kriminal/postoji-jedan-vrlo-praktican-razlog-zasto-njemacka-prima-toliko-izbjeglica-2/> (stranica posjećena: 8. kolovoza 2016.).
10. Holjevac, G. 2015. *Istine i zablude o izbjegličkoj krizi! Sve što trebate znati o njihovim namjerama*.<http://www.dnevno.hr/vijesti/komentari/istine-i-zablude-o-izbjeglickoj-krizi-sve-sto-trebate-znati-o-njihovim-kretnjama-i-namjerama-826318> (stranica posjećena: 8. kolovoza 2016.).
11. Riloff, E. i Ashequl Q. 2013. *Sarcasm as Contrast between a Positive Sentiment and Negative Situation*. School Of Computing. Salt Lake City.
12. Sunghwan, K. 2011. *Recognising Emotions and Sentiments in Text*. Doktorski rad. The University of Sidney, preuzeto: ožujak 2016.
13. Klaić, B. 1990. *Rječnik stranih riječi*. Tisak: Tiskara Rijeka 1990.
14. Godbole, N., Srinivasaiah, M., i Skiena, S. *Large-Scale Sentiment Analysis for News and Blogs*. Stony Brook University, preuzeto: ožujak 2016.

15. Vezzosi, P. 2015. *First steps with sentiment analysis in SAP Predictive Analytics 2.4*.
<http://scn.sap.com/community/predictive-analytics/blog/2015/12/01/first-steps-with-sentiment-analysis-in-sap-predictive-analytics-24> (stranica posjećena: 8. kolovoza 2016.)
16. http://www.sas.com/en_ph/software/analytics/sentiment-analysis.html
(stranica posjećena: 8. kolovoza 2016.)
17. https://cloud.google.com/prediction/docs/sentiment_analysis
(stranica posjećena: 8. kolovoza 2016.)
18. https://en.wikipedia.org/wiki/November_2015_Paris_attacks
(stranica posjećena: 5. kolovoza 2016.)
19. https://en.wikipedia.org/wiki/Islamic_State_of_Iraq_and_the_Levant
(stranica posjećena: 5. kolovoza 2016.)
20. <http://www.dnevno.hr/vijesti/komentari/istine-i-zablude-o-izbjeglickoj-krizi-sve-sto-trebate-znati-o-njihovim-kretnjama-i-namjerama-826318>
(stranica posjećena: 8. kolovoza 2016.)
21. http://www.azil.com.hr/download.aspx?f=dokumenti/Razno/Konvencijaostatusuizbjeglica_iz1951.doc. (stranica posjećena: 8. kolovoza 2016.)
22. Pang, B. i Lee, L. 2002. *Thumbs up? Sentiment classification using machine learning techniques*. preuzeto: srpanj 2016.
23. Pang, B. i Lee, L. 2003. *Exploiting class relationship for sentiment categorization with respect to rating scales*.
24. Turney, P. 2002. *Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews*. Institute for Information Technology, Canada.
25. Goldberg A., i Zhu, J. *Seeing stars when there aren't many stars: Graph-based semi-supervised learning for sentiment categorization*.
26. Yu, D. i Hatzivassiloglu. 2003. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
27. Kim, S.-M. i Hovy, E. 2004. *Determining the Sentiment of Opinions*.
28. <http://www.jutarnji.hr/> , stranica Jutarnjeg lista
29. <http://www.index.hr/> , stranica Indexa
30. Agarwal, A., Boyi, X., Vovsha, I., Rambow, O., Passonneau, R. 2011. *Sentiment Analysis of Twitter Data*. Columbia University.

Privitak 1

U Privitku 1 nalazi se definicija klase koja se u kodu koristi kao tip elemenata generičke liste koja sadrži podatke iz leksičkog resursa.

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.Threading.Tasks;

namespace AplikacijaZaPolarizaciju
{
    public class LexicalResource
    {
        public string Word { get; set; }
        public int Value { get; set; }
    }
}
```

Privitak 2

U Privitku 2 nalazi se definicija klase koja se u kodu koristi kao tip elementa generičke liste koja sadrži podatke koji se prikazuju na grafičkom sučelju aplikacije kao rezultat obrade.

```
using System;
using System.Collections.Generic;
using System.Linq;
using System.Text;
using System.Threading.Tasks;

namespace AplikacijaZaPolarizaciju
{
    public class PolarizationEntry
    {
        [System.ComponentModel.DisplayName("Riječ")]
        public string Word { get; set; }
        [System.ComponentModel.DisplayName("Broj pojavljivanja")]
        public int WordCount { get; set; }
        [System.ComponentModel.Browsable(false)]
        public int PolarizationValueOfWord { get; set; }
    }
}
```

Privitak 3

U Privitku 3 nalazi se cijela logika aplikacije. U prvom koraku se učitavaju potrebni dokumenti koji uključuju dokument koji sadrži zaustavne riječi, dokument koji sadrži polarizacijske vrijednosti teksta, te dokument koji sadržava tekst za obradu. U drugom koraku se uklanjaju zaustavne riječi iz teksta za obradu. U trećem koraku se broji učestalost ponavljanja riječi u tekstu te ih se prema tome sortira. U četvrtom koraku se izračunava polarizacijska vrijednost riječi i teksta te se rezultat šalje prema grafičkom sučelju.

```
using System;
using System.Collections.Generic;
using System.ComponentModel;
using System.Data;
using System.Data.OleDb;
using System.Drawing;
using System.IO;
using System.Linq;
using System.Text;
using System.Threading.Tasks;
using System.Windows.Forms;

namespace AplikacijaZaPolarizaciju
{
    public partial class MainForm : Form
    {
        List<string> _stopWordList;
        List<string> _TextForProcessList;
        List<LexicalResource> _LexicalResourceList;
        List<PolarizationEntry> _ResultList;
        string _TextForProcessPath;
        string _LexicalResourcePath;
        string _stopWordPath;
        string _sheetName;

        public MainForm()
        {
            InitializeComponent();
            this.Text = "Aplikacija za polarizaciju teksta";
        }

        private void btnStopWordsDocument_Click(object sender,
EventArgs e)
        {
            OpenFileDialog dialog = new OpenFileDialog();
```

```

        dialog.Filter = "Text files | *.txt";
        dialog.Multiselect = false;
        if (dialog.ShowDialog() == DialogResult.OK)
        {
            DisableForm();

            string path = dialog.FileName;
            tbStopWordsDocument.Text = path;
            _stopWordPath = path;

            EnableForm();
        }
    }

    private void btnPolarizationValuesDocument_Click(object sender,
EventArgs e)
    {
        OpenFileDialog dialog = new OpenFileDialog();
        dialog.Filter = "Excel files | *.xls";
        dialog.Multiselect = false;
        if (dialog.ShowDialog() == DialogResult.OK)
        {
            DisableForm();

            string sheetName = tbExcelSheetName.Text;
            string path = dialog.FileName;
            tbPolarizationValuesDocument.Text = path;
            _LexicalResourcePath = path;
            _sheetName = sheetName;

            EnableForm();
        }
    }

    private void btnTextForProcessDocument_Click(object sender,
EventArgs e)
    {
        OpenFileDialog dialog = new OpenFileDialog();
        dialog.Filter = "Text files | *.txt";
        dialog.Multiselect = false;
        if (dialog.ShowDialog() == DialogResult.OK)
        {
            DisableForm();

            string path = dialog.FileName;
            tbTextForProcessDocument.Text = path;
            _TextForProcessPath = path;

```

```

        EnableForm();
    }
}

private List<string> ReadTxtFileSeparatedByRow (string path)
{
    List<string> result = new List<string>();

    try
    {
        using (StreamReader reader = new StreamReader(new
FileStream(path, FileMode.Open), new UTF8Encoding()))
        {
            string line;
            while ((line = reader.ReadLine()) != null)
            {
                result.Add(line.ToLower());
            }

            reader.Close();
        }
    }

    catch (Exception ex)
    {
        MessageBox.Show("Došlo je do pogreške kod učitavanja
dokumenta za stop riječi!", "Aplikacija za polarizaciju teksta",
MessageBoxButtons.OK, MessageBoxIcon.Error);
        return null;
    }

    return result;
}

private List<string> ReadTxtFileSeparatedBySpace(string path)
{
    List<string> result = new List<string>();
    try
    {
        using (StreamReader reader = new StreamReader(new
FileStream(path, FileMode.Open), new UTF8Encoding())) // do anything
you want, e.g. read it
        {
            string text = reader.ReadToEnd();
            string[] lines = text.Split(new Char[] { ' ', '\n'
}, StringSplitOptions.RemoveEmptyEntries);

```

```

        foreach (string s in lines)
        {
            result.Add(s.ToLower());
        }

        reader.Close();
    }
}

catch (Exception ex)
{
    MessageBox.Show("Došlo je do pogreške kod učitavanja
dokumenta teksta za obradu!", "Aplikacija za polarizaciju teksta",
MessageBoxButtons.OK, MessageBoxIcon.Error);
    return null;
}

return result;
}

private List<LexicalResource> ReadPolarizationExcelFile(string
path, string sheetName)
{
    List<LexicalResource> result = new List<LexicalResource>();
    DataTable sheetData = new DataTable();

    if(sheetName == "" || sheetName == null)
    {
        sheetName = "Sheet1";
    }

    try
    {
        using (OleDbConnection conn = new
OleDbConnection("Provider=Microsoft.Jet.OLEDB.4.0;Data Source=" + path
+ "; Jet OLEDB:Engine Type=5;Extended Properties=\"Excel 8.0;\""))
        {
            conn.Open();
            // Retrieve the data using data adapter
            OleDbDataAdapter sheetAdapter = new
OleDbDataAdapter("select * from [" + sheetName + "$]", conn);
            sheetAdapter.Fill(sheetData);
            conn.Close();
        }
    }
}

```

```

        catch(Exception ex)
        {
            MessageBox.Show("Došlo je do pogreške kod učitavanja
dokumenta za polarizacijske vrijednosti!", "Aplikacija za polarizaciju
teksta", MessageBoxButtons.OK, MessageBoxIcon.Error);
            return null;
        }

        result = sheetData.AsEnumerable().Select(x => new
LexicalResource()
        {
            Word = x.ItemArray[0].ToString(),
            Value = Convert.ToInt32(x.ItemArray[1])
        }).ToList();

        return result;
    }

    private void btnStartProcess_Click(object sender, EventArgs e)
    {
        DisableForm();

        InitializeDocuments();

        string[] charsToRemove = new string[] { ",", ".", ";", ":",
"", "!", "?", "(", ")", "-", "%", "'", "\"", "0", "1", "2", "3", "4",
"5", "6", "7", "8", "9", };

        // Removing unwanted characters
        for (int i = 0; i < _TextForProcessList.Count; i++)
        {
            foreach (var c in charsToRemove)
            {
                _TextForProcessList[i] =
_TextForProcessList[i].Replace(c, string.Empty);
            }
        }

        //Remove empty rows
        _TextForProcessList.RemoveAll(x => x == "");

        // Cleaning out stop words
        _TextForProcessList.RemoveAll(item =>
_stopWordList.Contains(item));

        // Sorting and counting elements
        _ResultList = _TextForProcessList.GroupBy(x => x)

```



```

        .Select(g => new PolarizationEntry { Word = g.Key,
WordCount = g.Count() })
        .OrderByDescending(x => x.WordCount).ToList();

        // Setting polarization values of words
        for (int i = 0; i < _ResultList.Count; i++)
        {
            for (int j = 0; j < _LexicalResourceList.Count; j++)
            {
                if (_ResultList[i].Word ==
_LexicalResourceList[j].Word)
                {
                    _ResultList[i].PolarizationValueOfWord =
_LexicalResourceList[j].Value;
                    break;
                }
            }
        }

        // Bind list to data grid view
        dgvResult.DataSource = _ResultList;

        tbPolarizationValue.Text = _ResultList.Sum(item =>
item.PolarizationValueOfWord * item.WordCount).ToString();

        EnableForm();
    }

    private void DisableForm()
    {
        btnStartProcess.Enabled = false;
        btnStopWordsDocument.Enabled = false;
        btnPolarizationValuesDocument.Enabled = false;
        btnTextForProcessDocument.Enabled = false;
        tbExcelSheetName.Enabled = false;
        dgvResult.Enabled = false;
    }

    private void EnableForm()
    {
        btnStartProcess.Enabled = true;
        btnStopWordsDocument.Enabled = true;
        btnPolarizationValuesDocument.Enabled = true;
        btnTextForProcessDocument.Enabled = true;
        tbExcelSheetName.Enabled = true;
        dgvResult.Enabled = true;
    }
}

```

```
private void InitializeDocuments()
{
    _stopWordList = ReadTxtFileSeparatedByRow(_stopWordPath);
    _LexicalResourceList =
ReadPolarizationExcelFile(_LexicalResourcePath, _sheetName);
    _TextForProcessList =
ReadTxtFileSeparatedBySpace(_TextForProcessPath);
}
}
}
```