Razvoj i značaj analize sentimenta u tehnici

Ivana Zeljković, Nikola Garabandić
Univerzitet u Novom Sadu, Fakultet tehničkih nauka, Softversko inženjerstvo i informacione tehnologije
Novi Sad, Srbija

***Apstrakt* — U ovom radu je dat presek najbitnijih elemenata analize sentimenta, oblasti koja uživa veliku popularnost u poslednje dve decenije. Analiza sentimenta bavi se analiziranjem ljudskih mišljenja, pristupa, sentimenata i emocija, izraženih u pisanoj formi. Kao takva, zastupljena je u skoro svakom domenu ljudske delatnosti, s obzirom na to da je mišljenje centar svih ljudskih aktivnosti i ključ ponašanja pojedinca i zajednice. Sumirajući na jednom mestu bitne koncepte, rad je namenjen široj publici koja želi bolje upoznavanje sa datom temom. Pored objašnjenja ključnih pojmova, opisano je nekoliko izazova u upotrebi ove analize, uz predstavljanje njihove veličine i značaja. Takođe, opisana je podela pristupa analizi sentimenta na tri tehnike: tehniku mašinskog učenja, tehniku zasnovanu na upotrebi leksikona i hibridnu tehniku; uz isticanje prednosti i mana svake od njih. Klasifikaciju prati i dekompozicija tehnika na metodologije, najčešće upotrebljavane kada je u pitanju problem klasifikacije teksta. Osim toga, rad obuhvata i predstavljanje nekoliko primera alata koji se koriste u svrhu analiziranja sentimenta, objašnjavajući način funkcionisanja svakog. Na kraju, sumira se i elaborira značaj pomenute teme u nekoliko različitih oblasti, koje predstavljaju najreprezentativnije primere njene upotrebe: marketing, politika, medicina, obrazovanje…**

Ključne reči — analiza sentimenta, polaritet reči, sentiment reči, pristup mašinskog učenja, pristup zasnovan na upotrebi leksikona

I. Uvod

Društveni odnosi su preneti iz stvarnog u virtuelni svet, čime je postignuto kreiranje onlajn zajednica koje povezuju ljude iz celog sveta, omogućavajući im zabavu, komunikaciju, razmenu znanja. Ponašanje potrošača tokom pretrage informacija o određenom proizvodu ili usluzi je značajno izmenjeno. Glavni uticaj na formiranje i oblikovanje mišljenja pojedinca ima mišljenje zajednice kojoj pripada.

Sa porastom količine javno dostupnih sadržaja na internetu u kojima je iskazano mišljenje, na implicitan ili eksplicitan način, nameće se sve veća potreba za analizom istih kako bi se iskoristili za razumevanje izbora, namera i sentimenta samih konzumenata. Velikom količinom podataka se teško upravlja i njihova pretraga zahteva dosta vremena, te je značajna uloga u potrazi za korisnim i kvalitetnim informacijama kroz adekvatno filtriranje, eksploataciju i analizu pripala metodama i tehnikama *Data mining*-a [1]. Primena ovih metoda i tehnika na sadržaje sajtova društvenih medija se naziva analiza društvenih medija (engl. *Social media analytics* - SMA) koja obezbeđuje prikupljanje podataka sa sajtova i upoznavanje sa bazom konzumenata na najvišem nivou, uključujući analizu njihovih emocija, sentimenta i mišljenja.

Jedan od posebnih aspekata SMA je i analiza sentimenta (engl. *Sentiment analysis*), u literaturi poznata još i kao *opinion mining*. Analiza sentimenta je istraživačka oblast u kojoj se vrši analiziranje javnih mišljenja, stavova, sentimenta i emocija usmerenih ka konkretnom entitetu, koji može biti proizvod, usluga, tema, problem, osoba, organizacija ili događaj.

Osnovni cilj ovog rada je upoznavanje šire publike sa datom temom, gde je primarni akcenat na predstavljanju osnovnog skupa informacija koje ukazuju na značaj i popularnost iste danas. U radu su obrađeni najvažniji aspekti analize sentimenta: definicija pojma, izazovi i problemi, klasifikacija pristupa i metodologija, najčešće korišteni alati i primeri primene u različitim domenima ljudskog života.

U prvom poglavlju, *Razvoj i definicija pojma*, dato je detaljnije objašnjenje predmetne terminologije. U narednom, *Izazovi u analizi sentimenta*, predstavljeno je nekoliko najzanimljivih problema ove analize, poput detekcije sarkazma i ironije, utvrđivanja polariteta reči, itd. Treće poglavlje, *Tehnike analize sentimenta*, obrađuje klasifikaciju pristupa analizi sentimenta, uz jasno definisanje razlika među pristupima i podelu istih na metodologije. U četvrtom poglavlju, *Alati u analizi sentimenta*, navedeni su i bliže objašnjeni najčešće korišteni alati za detekciju sentimenta. U poslednjem poglavlju, *Primena analize sentimenta*, elaborirana je tema primene analize sentimenta u različitim oblastima ljudske delatnosti, počev od marketinga do obrazovnog sistema.

II. Razvoj i definicija pojma

Analiza sentimenta kao istraživačka oblast danas uživa veliku pažnju istraživačke zajednice, pri čemu se domenska oblast istraživanja i primene pomenute analize veoma razlikuje. Istraživanja su brojna i obuhvataju oblasti većih i manjih razmera, od marketinga, psiholoških nauka, računarskih nauka, itd. Osim domen-specifičnih istraživanja na temu analize sentimenta, od velikog značaja su i radovi posvećeni ključnim pojmovima: ekspresija subjektivnosti [2], sentiment reči [3], sentiment rečenica [4] i sentiment tema [5,6,7].

Iako je u pitanju veoma popularna oblast današnjice, ona postoji već duži niz godina. Prve korake, u razvoju ove oblasti, predstavljao je rad u domenu mašinske sentiment analize iz 1979. godine [8]. Međutim, prvo pojavljivanje termina u današnjem obliku, dokumentovano je u istraživanju [7] koje je podstaklo naglu ekspanziju oblasti i radova na ovu temu, dok

se značajnije istraživanje na temu sentimenta i mišljenja pojavilo nešto ranije [9,10,11,12,13].

Analiza sentimenta predstavlja kompleksan proces, sačinjen od pet osnovnih koraka analiziranja sentimenta podataka:

1) Prikupljanje podataka – u zavisnosti od oblasti u kojoj se vrši analiziranje. Podaci se najčešće prikupljaju sa blogova, foruma i društvenih mreža. Glavni problem u ovom koraku je neorganizovanost podataka u pogledu pripadnosti različitim rečnicima, kao i razlika u kontekstu napisanih reči.
2) Priprema teksta – podrazumeva pročišćavanje ulaznog skupa podataka. Identifikuju se i odbacuju sadržaji koji su netekstualni i nerelevantni za samu analizu.
3) Detekcija sentimenta – korak najveće odgovornosti u toku analize. Podrazumeva selekciju rečenica koje sadrže subjektivne izraze (mišljenja, verovanja i pogledi) uz odbacivanje rečenica sa objektivnim izrazima (činjenice, činjenične informacije).
4) Klasifikacija sentimenta – klasifikovanje rečenica selektovanih u prethodnom koraku. Klasifikacija se vrši na određene grupe koje se mogu definisati uzimajući u obzir jedan ili više aspekata posmatranja izdvojenih rečenica. Primeri grupa: dobro, loše, pozitivno, negativno.
5) Prezentacija izlaza – najteži korak analize sentimenta. U ovom koraku, neophodno je od nestrukturiranog teksta, izdvojiti informaciju određenog značaja za dati domen u kom se primenjuje analiza. Po završetku analize, ovde će se na adekvatan način (zavisno od domena primene analize) predstaviti grafički prikaz rezultata analize (linijski grafik, stubičasti grafik…).

III. Izazovi u analizi sentimenta

U ovom poglavlju je napravljen pregled nekoliko najzanimljivijih izazova i problema, uz objašnjenje njihove relevantnosti u samoj analizi.

A. *Određivanje polariteta reči i fraza*

Osnovni pristup analizi sentimenta podrazumeva korišćenje rečnika polarizovanih reči i fraza. Pod polaritetom se podrazumeva klasifikovanje reči ili fraza kao pozitivnih ili negativnih. Polaritet konkretne reči može biti promenjen u zavisnosti od konteksta reči, što uzrokuje problem kada je reč o određivanju sentimenta neke fraze. Osim negacije, kao najučestalijeg faktora promene polariteta reči, u druge bitne faktore ubrajamo i kvantifikatore, reči koje gradiraju (uvećavaju ili umanjuju) sentiment reči.

B. *Detekcija negacije, ironije i sarkazma*

Prisustvo negacije dovodi do promene značenja reči, zbog čega se ona ubraja u veoma važne faktore pri određivanju polariteta reči. Promena značenja reči upotrebom negacije podrazumeva promenu orijentacije sentimenta, ali ne nužno i njegovog polariteta.

Sledeća tri primera ilistruju problematiku upotrebe negacije u frazama, čiji je polaritet neophodno utvrditi.

Pr. 1. *Dopada mi se današnje predavanje.*
Pr. 2. *Ne dopada mi se današnje predavanje.*
Pr. 3. *Ne samo da mi se dopada današnje predavanje, nego je i veoma korisno.*

Rečenica u prvom primeru ima pozitivan polaritet, jer predstavlja pozitivan stav. Rečenica iz drugog primera ima negativan polaritet, jer se upotrebom negacije promenilo značenje rečenice iz prvog primera, tako da dobijamo negativan stav. Rečenica iz trećeg primera ima u celini pozitivan polaritet, iako je u prvom delu rečenice upotrebljena negacija. Međutim, u ovom primeru negacija nije uzrokovala promenu polariteta rečenice.

S druge strane, podjednako veliki problem u analizi sentimenta nastaje i upotrebom ironije ili sarkazma u rečenici. Ovaj slučaj je veoma čest u diskusijama u oblasti politike. Same konstrukcije, koje predstavljaju ironiju ili sarkazam, nemoguće je mašinski detektovati. Čak i ukoliko rečenice ne sadrže pomenute retoričke alate, može se desiti da u sklopu njih postoje reči koje će biti prepoznate kao ključne reči određenog polariteta, iako zapravo rečenica ima suprotan polaritet.

C. *Domenska zavisnost*

Prilikom izražavanja sopstvenog mišljenja i osećanja vezanih za datu temu, ljudi koriste različite vokabulare, stilove pisanja, kao i žargone. Veoma veliki značaj za pravilno analiziranje ima i domen u kom se definiše neki iskaz, čiji se sentiment posmatra. Rečenica „Pročitaj knjigu.“, u domenu kritikovanja knjiga, ima pozitivan sentiment iz razloga što predstavlja preporuku za neku knjigu. S druge strane, ista rečenica, u domenu kritika filma ima sasvim suprotan sentiment, jer njom izražavamo negativnu kritiku. Klasifikator kritika, koji je obučen skupom kritika o jednoj vrsti proizvoda, često ne pokazuje iste performanse kada se primeni nad skupom kritika o drugom proizvodu [14,15].

Iz navedenog se može zaključiti da je izbor vokabulara usko vezan za domen u kom definišemo iskaz, čiji sentiment analiziramo.

D. *Fraze bez konkretnih sentiment reči*

U svakodnevnoj komunikaciji se koristi veliki broj fraza, pri čemu neke od njih predstavljaju problem pri analiziranju sentimenta. Razlog tome je sam sadržaj fraze, koji ne uključuje nijednu reč sa jasno definisanim sentimentom, te je gotovo nemoguće izvršiti određivanje sentimenta posmatrane fraze. Primer takve fraze je i „Kapa dole!“. Ova veoma često korištena fraza oslikava pozitivan stav prema osobi kojoj odajemo priznanje za određeni uspeh. Međutim, odrediti sentiment iste, upotrebom odgovarajućih metodologija je nerešiv problem, iako je čoveku intuitivno jasan.

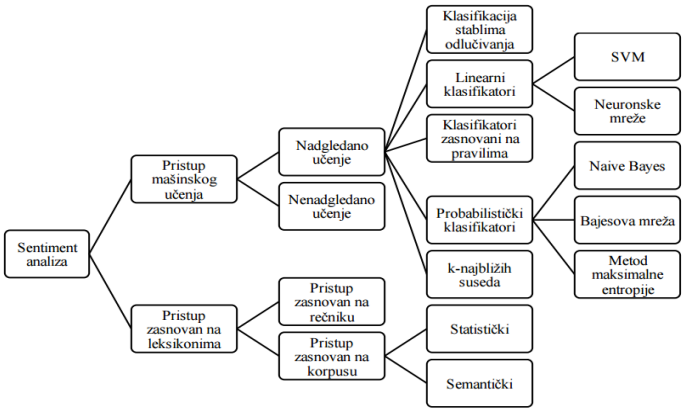
Uprkos postojećim izazovima, analiza sentimenta je pokazala solidnu tačnost, koja se kreće u rasponu od 80% do 95% za sofisticirane sisteme. Glavna prepreka postizanju perfektne tačnosti je nedostatak rešenja za opisane probleme.

IV. TEHNIKE ANALIZE SENTIMENTA

Tehnike klasifikacije sentimenta se, u zavisnosti od načina na koji se sprovodi sama analiza, mogu u osnovi podeliti na:

- tehnike pristupa zasnovanog na mašinskom učenju
- tehnike pristupa zasnovanog na upotrebi leksikona
- tehnike hibridnog pristupa, koji predstavlja kombinaciju prethodna dva.

Na slici 1 je prikazana hijerarhija najčešće korištenih tehnika klasifikacije sentimenta, kao i podela istih na konkretne metodologije, koje su detaljnije objašnjene u nastavku. U tabeli 1 dat je prikaz pristupa klasifikaciji sentimenta uz navođenje dobrih strana, kao i ograničenja svakog.



Slika 1 Tehnike klasifikacije sentimenta [16]

A. *Pristup zasnovan na mašinskom učenju*

Ovaj pristup se zasniva na upotrebi lingvističkih atributa i algoritama mašinskog učenja. Koristi se u situacijama kada se vrši predikcija sentimenta uzoračkog skupa podataka, nakon što je izvršeno obučavanje klasifikatora na odgovarajućem skupu podataka, koji je namenjen treniranju.

U zavisnosti od dostupnosti skupa podataka koji služi za obučavanje klasifikatora, ovaj pristup se može podeliti na dve osnovne kategorije: nadgledano i nenadgledano mašinsko učenje. Nadgledano mašinsko učenje se koristi kada je dostupan veliki skup označenih podataka, namenjen za obučavanje. S druge strane, nenadgledano mašinsko učenje se koristi u situacijama kada ne postoje pripremljeni podaci za svrhu obuke.

Kod nadgledanog učenja, označeni podaci se koriste u procesu treniranja klasifikatora, koji na osnovu zadatog skupa podataka stiču određeni nivo znanja o datom sadržaju. Nakon što se u potpunosti završi obučavanje, klasifikatoru se daje uzorački skup podataka čiji je cilj da se utvrdi mera obučenosti i performansi rada klasifikatora.

Kod nenadgledanog učenja je relativno teško naći skup označenih podataka koji se mogu koristiti u svrhu treniranja. Zbog izostanka obuke, pomenuta kategorija mašinskog učenja pribegava nešto drugačijoj tehnici. Vrši se određivanje semantičke orijentacije[1] (*Semantic orientation* – SO) fraza samog dokumenta. Osnovno pravilo, koje se koristi u nenadgledanom mašinskom učenju, je da se prosečna semantička orijentacija fraza poredi sa vrednošću koja predstavlja unapred definisani prag iste. Ukoliko je prosečna vrednost veća od zadatog praga, tekst se označava kao pozitivan, dok se u suprotnom smatra negativnim.

U nastavku je dat pregled osnova najčešće upotrebljavanih algoritama u kategoriji nadgledanog mašinskog učenja.

a) *Stabla odlučivanja*

Vrednosti koje predstavljaju rezultat rada klasifikatora su ljudima nerazumljive, ukoliko ne poseduju određeno znanje o mehanizmu rada algoritma. Međutim, ovaj problem je prevaziđen upotrebom simboličkih klasifikatora, od kojih su stabla odlučivanja najpoznatija tehnika. Ova tehnika vrši dekompoziciju polaznog skupa podataka, koji se koristi u svrhe obučavanja klasifikatora, na određene nivoe hijerarhije, predstavljajući podatke čvorovima stabla, koji su povezani u zavisnosti od uslova koje atributi moraju da zadovolje. Krajnji čvorovi u stablu, tzv. listovi stabla predstavljaju klasu kojoj dokument pripada.

Sam proces klasifikacije upotrebom ove tehnologije otpočinje u korenskom čvoru stabla. Vrši se pretraga stabla s leva na desno, krećući se iz svakog čvora na naredni nivo hijerarhije, do novog čvora, onom granom koja zadovoljava definisani uslov [17]. Za primenu stabla odlučivanja potrebno je obezbediti veliku količinu primera za obučavanje koji su opisani diskretizovanim atributima, kao i predefinisane klase kojima primeri pripadaju [18].

b) *Neuronske mreže i Mašine potpornog vektora*

Neuronske mreže i mašine potpornog vektora spadaju u grupu linearnih klasifikatora.

Neuronske mreže se sastoje od određenog broja slojeva neurona. Grupa međusobno povezanih neurona koji primaju ulaze, predstavljaju ulazni sloj neuronske mreže. Neuroni koji primaju signale samo od drugih neurona, čine jedan ili više skrivenih slojeva mreže. Svaki sloj neuronske mreže prima ili impulse u obliku vektora podataka ili izlaze iz prethodnih slojeva mreže i paralelno ih obrađuje. Poslednji sloj neurona koji daje finalni rezultat rada neuronske mreže se naziva izlazni sloj.

Najjednostavniji vid neuronske mreže je mreža sa dva sloja neurona, ulaznim i izlaznim. Iako postoje i kompleksnije neuronske mreže, sa većim brojem skrivenih slojeva, u problemu klasifikovanja dokumenta, najbolje performanse pokazale su najjednostavnije implementacije.

Ključna prednost ove tehnike klasifikacije je što pokazuje odlične rezultate u kompleksnim domenima. S druge strane, najvećim nedostatkom se smatra dug proces obučavanja i činjenica da je stečeno znanje implicitno, tj. duboko u strukturi mreže, pa ga je samim tim veoma teško interpretirati.

[1] Semantička orijentacija (engl. *Semantic orientation* – SO) predstavlja polaritet sadržaja koji se posmatra, određen na osnovu polariteta konkretnih reči tog sadržaja.

Metoda potpornih vektora – SVM (engl. *Support vector machines*) je jedan od efikasnijih metoda klasifikacije koji se svodi na pronalaženje hiperravni[2] separacije. Hiperravan maksimalno razdvaja podatke u prostoru na pozitivne i negativne instance na osnovu klasnog atributa [19]. Glavna prednost ove tehnike se ogleda u mogućnosti klasifikacije podataka velike dimenzionalnosti.

c) Klasifikator zasnovan na pravilima

Metode koji se zasnivaju na pravilima modeluju podatke u prostoru pomoću skupa pravila. Pravila se zadaju u disjunktivnoj normalnoj formi [20]:

$$If\ X_1 = A_1\ AND \ldots AND\ X_n = A_n\ Then\ Y = C\ sa\ CF$$

gde svaka ulazna promenljiva X_i kao vrednost prihvata skup termina ili lingvističkih oznaka A_i, dok izlazna promenljiva (Y) poprima jednu od vrednosti klasa. Uslovi koji se testiraju pravilima se odnose na prisustvo određenog termina u dokumentima koji se koriste za obučavanje. Svako pravilo takođe obuhvata i faktor pouzdanosti (*CF*), vrednost iz intervala [0,1] koja govori o pouzdanosti klasifikacije u klasu predstavljenu skupom pravila.

Iz skupa pravila koja tačno klasifikuju sve primere za obučavanje, najbolja pravila selektuje metod učenja.

d) Probabilistički klasifikatori

U praksi se najčešće sreću tri pripadnika ove grupe klasifikatora, a to su: Naive Bayes, Bayesove mreže i klasifikator maksimalne entropije (engl. *Maximum Entropy*). U kontekstu klasifikacije teksta, ovi klasifikatori su u radu [14] opisani na sledeći način:

Bayesov klasifikator je jednostavan klasifikator koji se u praksi najčešće koristi od sva tri navedena. Njegova osnova je teorija verovatnoće.

Naive Bayes klasifikator određuje verovatnoću klase na osnovu distribucije reči u dokumentu. Dokument se predstavlja kao korpa reči (engl. *Bag of Words*) koja treba da rezultuje listom sentiment reči, pri čemu je neophodno da se ignoriše pozicija reči u dokumentu. Predviđanje verovatnoće da će skup reči pripadati nekoj klasi se sprovodi upotrebom Bayesove teoreme (slika 2):

$$P(klasa|atributi) = \frac{P(klasa) * P(atributi|klasa)}{P(atributi)}$$

Slika 2 Bayesova teorema (uslovna verovatnoća)

P (klasa) je marginalna verovatnoća klase. P (atribut|klasa) je verovatnoća klasifikacije skupa atributa (reči) u datu klasu. P(atributi) je verovatnoća da će se dati skup atributa pojaviti, a u prikazanoj formuli ima ulogu konstante normalizacije.

Ako se pretpostavi da su svi atributi potpuno zavisni, tada se dobija Bayesova mreža, koja reprezentuje usmereni aciklični graf gde je skup čvorova zapravo skup slučajnih promenljivih, a skup grana grafa je skup zavisnosti između promenljivih.

Zbog svoje kompleksnosti, Bayesove mreže se retko upotrebljavaju u praksi za klasifikaciju teksta.

Klasifikator maksimalne entropije (ME), za razliku od Naive Bayes klasifikatora, ne polazi od pretpostavke da su atributi nezavisni. Nasuprot tome, ovaj klasifikator koristi optimizaciju zasnovanu na pretrazi, u cilju pronalaženja težine svakog od atributa, koji se mogu kombinovati kako bi se utvrdila najverovatnija klasa za dati skup atributa.

e) Klasifikator k-najbližih suseda

Klasifikacija teksta koja koristi algoritam k-najbližih suseda (engl. *k-Nearest Neighbor*) je u praksi pokazala veoma kvalitetne rezultate. Klasifikacija upotrebom ovog algoritma se vrši na osnovu sličnosti primera koji se testira i primera koji se nalaze u skupu za obučavanje klasifikatora.

B. Pristup zasnovan na upotrebi leksikona

Ovaj pristup koristi predefinisane liste reči, tzv. leksikone sentimenta, u kojima je svaka reč povezana sa specifičnim sentimentom. Pristup se zasniva na prebrojavanju reči sa pozitivnim i negativnim sentimentom. Međutim, leksikoni u velikoj meri zavise od konteksta u kom su stvoreni. Ovaj pristup klasifikacije sentimenta ne zahteva označavanje podataka, što je njegova dobra strana u odnosu na prethodno opisani pristup, pristup mašinskog učenja. S druge strane, gotovo je nemoguće definisati jedinstven leksički-baziran rečnik koji će se koristiti za različite kontekste, što ovaj pristup dovodi u uravnotežen položaj sa pristupom mašinskog učenja, u pogledu njihovih prednosti i mana.

U zavisnosti od načina kreiranja leksikona, može se izvršiti podela ovog pristupa na tri osnovne kategorije: manuelni pristup, pristup zasnovan na rečnicima i pristup zasnovan na korpusu.

Prvi od nabrojanih, manuelni pristup, gotovo nikada se ne upotrebljava samostalno, već isključivo u kombinaciji sa nekim automatizovanim pristupom u cilju testiranja ispravnosti istog.

Pristup koji se zasniva na upotrebi rečnika podrazumeva da se za svaku reč definiše skup sinonima i antonima, pri čemu su ti skupovi na početku manuelno definisani. Svakom elementu skupa se dodeli određena semantička orijentacija, nakon čega se upotrebom automatizovanih algoritama dati skup iterativno uvećava, dodavanjem novih sinonima i antonima.

Pristup zasnovan na korpusu je domen-specifičan, jer omogućava identifikaciju sentimenta reči koje su specifične za domen konteksta, odnosno za domen koji predstavlja predmet analize. Međutim, dobra strana ovog pristupa je ujedno i njegova loša strana. Domen-specifični rečnik je veoma često nedovoljan, jer postoje reči koje u zavisnosti od konteksta mogu potpuno da promene orijentaciju, a samim tim i polaritet sentimenta.

[2] U geometrijskom smislu, hiperravan za dati n-dimenzionalni prostor V predstavlja potprostor dimenzije n-1. U zavisnosti od vrste prostora V se daje korespodentna definicija hiperravni.

C. *Hibridni pristup*

Ovaj pristup predstavlja kombinaciju prethodna dva, gde kao takav ima velike predispozicije da unapredi performanse klasifikatora, imajući u vidu dobre i loše strane i jednog i drugog pristupa. Međutim, ovaj pristup je najmanje istražen.

PRISTUPI KLASIFIKACJI SENTIMENTA	**PREDNOSTI I OGRANIČENJA**
PRISTUP MAŠINSKOG UČENJA	**PREDNOSTI** Mogućnost kreiranja modela za obučavanje i njihovog prilagođavanja specifičnim zahtevima i kontekstima **OGRANIČENJA** Nizak nivo primene na nove domene Zahtevanje dostupnosti skupa obeleženih podataka (u svrhu obučavanja) što često može biti veoma skupo
PRISTUP ZASNOVAN NA LEKSIKONIMA	**PREDNOSTI** Pokrivanje širokog skupa pojmova **OGRANIČENJA** Konačan broj reči u leksikonima Fiksno dodeljena orijentacija i polaritet sentimenta rečima koje se nalaze u leksikonima
HIBRIDNI PRISTUP	**PREDNOSTI** Spoj mašinskog učenja i upotrebe leksikona Detekcija i klasifikacija sentimenta na konceptualnom nivou Manja osetljivost na promene u domenu **OGRANIČENJA** Slaba dokumentovanost

Tabela 1 Sumarizacija prednosti i mana različitih pristupa klasifikaciji sentimenta

V. ALATI U ANALIZI SENTIMENTA

Prostoje razne studije koje vrše istraživanje na temu korištenih alata u klasifikaciji sentimenta. Najčešće korišteni alati za određivanje polariteta sentimenta u tekstualnim porukama su zasnovani na detekciji smajlija, koji su bazirani na prikazu lica koje jasno odražava osećanja. Osećanja se na najvišem nivou apstrakcije mogu podeliti u dve kategorije: srećna i tužna. Međutim, sa povećanjem seta smajlija koji se koristi, povećava se i broj kategorija osećanja. Kao takvi, smajliji se veoma često koriste u kombinaciji sa nekom od tehnika nadgledanog mašinskog učenja u cilju definisanja što boljeg i relevantnijeg skupa podataka za obučavanje klasifikatora.

U veoma često upotrebljavane alate spade i *Linguistic Inquiry and Word Count* [21], koji omogućava analiziranje ne samo pozitivnih i negativnih, već i emocionalnih, kognitivnih i strukturalnih komponenti teksta, upotrebom rečnika u kom se za svaku reč jasno definiše skup kategorija kojim pripada. Tako na primer reč „slaganje“ pripada većem broju kategorija: pristanak, pozitivna emocija, pozitivno osećanje i kognitivni proces.

Indeks sreće (engl. *Happiness index*) [22] je skala sentimenta koja koristi popularne *Affective Norms for English Words* (ANEW) [23]. Ovaj indeks kao rezultat daje brojnu vrednost u opsegu 1-9 koja ukazuje na količinu sreće u posmatranom tekstu. Prvo se izračuna frekvencija svake reči iz ANEW koja se pojavljuje u tekstu, a zatim se odredi prosečna brojna vrednost koja ukazuje na stopu sreće u datom skupu reči. Autori veoma često koriste ovaj alat nad skupom reči definisanim sadržajem pesama, a neretko ga primenjuju i na sadržaj blogova.

SentiWordNet je veoma popularan alat današnjice; javno dostupan resurs namenjen klasifikaciji sentimenta i mišljenja. Zasnovan je na upotrebi onlajn engleskog leksičkog rečnika, poznatog kao *WordNet* [24]. Termini iz pomenutog leksikona se na nivou ovog alata organizuju u grupe sinonima, tzv. *synsets*. Svaki od pomenutih skupova je povezan sa tri numeričke vrednosti označene sa: Pos(s), Neg(s) i Obj(s). Ove vrednosti pokazuju koliko su termini, smešteni u posmatranom skupu sinonima (s), pozitivni, negativni i objektivni (neutralni). Svaka od tri ocene može da ima vrednost iz intervala [0,1], pri čemu mora da važi:

Pos(s) + Neg(s) + Obj(s) = 1.

Još jedan zanimljiv alat je PANAS-t [25]. Ovaj alat koristi prilagođenu verziju *Positive Affect Negative Affect Scale* (PANAS) [26], metoda koji se koristi u psihologiji. PANAS-t prati promenu sentimenta (povećanje ili smanjenje) u toku vremena. Zasniva se na velikom skupu reči povezanih sa jedanaest osećanja: veselost, samopouzdanje, spokojnost, iznenađenost, strah, tuga, krivica, odbojnost, stidljivost, umor i pažljivost. Ovaj alat za svaki sentiment računa vrednost koja se kreće u intervalu [-1.0, 1.0] i oslikava nastale promene u posmatranom vremenskom periodu.

Zanimljiv alat je i *Voyant*[3], koji za zadati korpus (ulazni skup sentiment izraza) određuje frekvencije termina, pri čemu reči koje su frekventnije prikazuje većim fontom na grafičkom prikazu i obrnuto. Kako bi alat mogao da prepozna termin, neophodno je da on bude napisan kao jedna reč, što je razlog zbog kog se negacija u ovom alatu piše spojeno sa izrazom koji negira. Obrada ulaznog korpusa ovim alatom rezultuje

[3] Dostupan na stranici: https://voyant-tools.org/

različitim statističkim rezultatima koji se prezentuju pomoću tabela i grafikona. Primeri jednog od oblika grafičkih prikaza ovog alata su dati na slici 3 i slici 4.

Slika 3 Oblak frekventnosti termina sa pozitivnim polaritetom generisan uz pomoć alata Voyant, na osnovu zadatog korpusa

Slika 4 Oblak frekventnosti termina sa negativnim polaritetom generisan uz pomoć alata Voyant, na osnovu zadatog korpusa

VI. Primene analize sentimenta

Analiza sentimenta kao veoma široka istraživačka oblast je našla svoju primenu u različitim domenima ljudskog života. Obzirom na brz razvoj i porast elektronske trgovine, najveći fokus ove analize je na upravljanju ekonomijom i tržištem. Njen glavni značaj u ovom segmentu je ispitivanje uticaja obima i polariteta onlajn recenzija na donošenje odluke o kupovini nekog proizvoda. Povećanje broja onlajn recenzija o nekom proizvođaču/prodavcu ili konkretnom proizvodu rezultuje sticanjem dobre ili loše reputacije, u zavisnosti od polariteta recenzija. Za mnoga poslovanja, mišljenja iskazana na društvenim medijima su se pretvorila u određenu vrstu virtuelne valute koja može da popularizuje ili uništi proizvod na tržištu [27]. Izvori recenzija su obično sajtovi za recenziranje i forumi, ali mogu biti i blogovi, sajtovi društvenih mreža i drugi društveni mediji.

S druge strane, sajtovi društvenih mreža se u velikoj meri koriste kao izvori podataka u analiziranju (ne)saglasnosti javnosti sa određenom političkom temom. U ovom domenu se analiza sentimenta često koristi u cilju praćenja (ne)konzistentnosti između akcija na državnom nivou. Uzimajući u obzir diskusije na forumima i blogovima, kao i njihov sentiment, mogu se uspešno predviđati rezultati političkih izbora. Ovoj temi je posvećen i rad [28] u kom se autori bave analiziranjem i predviđanjem rezultata izbora.

Primena je zastupljena i u oblasti medicine. Upotreba ove analize u medicinskom domenu podrazumeva analiziranje nestrukturiranih komentara pacijenata o zdravstvenoj zaštiti. Komentari se klasifikuju kao pozitivni ili negativni, pri čemu se analiza istih sprovodi sa ciljem predviđanja da će pacijent preporučiti određenu bolnicu, za ocenu čistoće bolnice i ophođenja osoblja ka pacijentima.

Sentiment analiza je pronašla primenu u pojedinim uslužnim sektorima, poput finansijskog i turističkog, ali takođe i u obrazovanju. Razumevanje sentimenta poruka, koje đaci i studenati ostavljaju na društvenim mrežama, pomaže u identifikovanju problema sa obrazovnim sistemom (optimizacija pružanja pomoći i vrednih povratnih informacija, poboljšanje odnosa između djaka/studenata i nastavničkog/profesorskog kadra).

VII. Zaključak

Rad obuhvata osnovni skup informacija za upoznavanje sa temom klasifikacije i analize sentimenta određenog tekstualnog sadržaja. Takođe, izdvaja najbitnije koncepte svakog segmenta pomenute istraživačke oblasti, sa ciljem povećanja zainteresovanosti čitalaca za ovu temu.

Kao takav, rad ne doprinosi ovoj istraživačkoj oblasti sa naučne strane, već je njegova glavna uloga pružanje neophodanog nivoa znanja čitaocima, koji su novi u oblasti analize sentimenta, kako bi lakše razumeli literaturu koja detaljnije razrađuje njene aspekte.

Budući rad u obrađivanoj oblasti bi bio usmeren ka istraživanju naprednih koncepata i rešenja za postojeće izazove ove analize: detekcija ironije i sarkazma, pravilno određivanje sentimenta i njegovog polariteta u prisustvu negacije. Navedeni problemi su trenutno nepremostivi i kao takvi degradiraju performanse analize sentimenta. Njihovim rešavanjem bi se postigla tačnost koja teži perfekciji, što bi dovelo do značajnog širenja opsega primene i značaja ove analize.

Literatura

[1] Liu, Bing. "Sentiment analysis and opinion mining." *Synthesis lectures on human language technologies* 5.1 (2012): 1-167.

[2] Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann. "Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis." *Computational linguistics* 35.3 (2009): 399-433.

[3] Hatzivassiloglou, Vasileios, and Kathleen R. McKeown. "Predicting the semantic orientation of adjectives." *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 1997.

[4] Pang, Bo, and Lillian Lee. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2004.

[5] Yi, Jeonghee, et al. "Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques." *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*. IEEE, 2003.

[6] Hiroshi, Kanayama, Nasukawa Tetsuya, and Watanabe Hideo. "Deeper sentiment analysis using machine translation technology." *Proceedings of the 20th international conference on Computational Linguistics*. Association for Computational Linguistics, 2004.

[7] Nasukawa, Tetsuya, and Jeonghee Yi. "Sentiment analysis: Capturing favorability using natural language processing." *Proceedings of the 2nd international conference on Knowledge capture*. ACM, 2003.

[8] Carbonell, Jaime Guillermo. *Subjective Understanding: Computer Models of Belief Systems*. No. RR-150. YALE UNIV NEW HAVEN CONN DEPT OF COMPUTER SCIENCE, 1979.

[9] Morinaga, Satoshi, et al. "Mining product reputations on the web." *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002.

[10] Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan. "Thumbs up?: sentiment classification using machine learning techniques." *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002.

[11] Tong, Richard M. "An operational system for detecting and tracking opinions in on-line discussion." *Working Notes of the ACM SIGIR 2001 Workshop on Operational Text Classification*. Vol. 1. 2001.

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

[13] Wiebe, Janyce. "Learning subjective adjectives from corpora." *AAAI/IAAI*. 2000.

[14] Dave, Kushal, Steve Lawrence, and David M. Pennock. "Mining the peanut gallery: Opinion extraction and semantic classification of product reviews." *Proceedings of the 12th international conference on World Wide Web*. ACM, 2003.

[15] Reinstein, David A., and Christopher M. Snyder. "The influence of expert reviews on consumer demand for experience goods: A case study of movie critics." *The journal of industrial economics* 53.1 (2005): 27-51.

[16] Medhat, Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment analysis algorithms and applications: A survey." *Ain Shams Engineering Journal* 5.4 (2014): 1093-1113.

[17] Bošnjak, Zita. *Inteligentni sistemi i poslovna primena*. Ekonomski fakultet, 2006.

[18] Xu, Guandong, Yanchun Zhang, and Lin Li. *Web mining and social networking: techniques and applications*. Vol. 6. Springer Science & Business Media, 2010.

[19] Feldman, Ronen, and James Sanger. *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007.

[20] Berlanga, F., et al. "A genetic-programming-based approach for the learning of compact fuzzy rule-based classification systems." *Artificial Intelligence and Soft Computing–ICAISC 2006* (2006): 182-191.

[21] Tausczik, Yla R., and James W. Pennebaker. "The psychological meaning of words: LIWC and computerized text analysis methods." *Journal of language and social psychology* 29.1 (2010): 24-54.

[22] Dodds, Peter Sheridan, and Christopher M. Danforth. "Measuring the happiness of large-scale written expression: Songs, blogs, and presidents." *Journal of happiness studies* 11.4 (2010): 441-456.

[23] Bradley, Margaret M., and Peter J. Lang. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical report C-1, the center for research in psychophysiology, University of Florida, 1999.

[24] Miller, George A. "WordNet: a lexical database for English." *Communications of the ACM* 38.11 (1995): 39-41.

[25] Gonçalves, Pollyanna, Fabrício Benevenuto, and Meeyoung Cha. "Panas-t: A pychometric scale for measuring sentiments on twitter." *arXiv preprint arXiv:1308.1857* (2013).

[26] Watson, David, Lee A. Clark, and Auke Tellegen. "Development and validation of brief measures of positive and negative affect: the PANAS scales." *Journal of personality and social psychology* 54.6 (1988): 1063.

[27] Dobrescu, Alexandra Balahur. *Methods and resources for sentiment analysis in multilingual documents of different text types*. Diss. Universitat d'Alacant-Universidad de Alicante, 2011.

[28] Kim, Soo-Min, and Eduard Hovy. "Automatic identification of pro and con reasons in online reviews." *Proceedings of the COLING/ACL on Main conference poster sessions*. Association for Computational Linguistics, 2006.